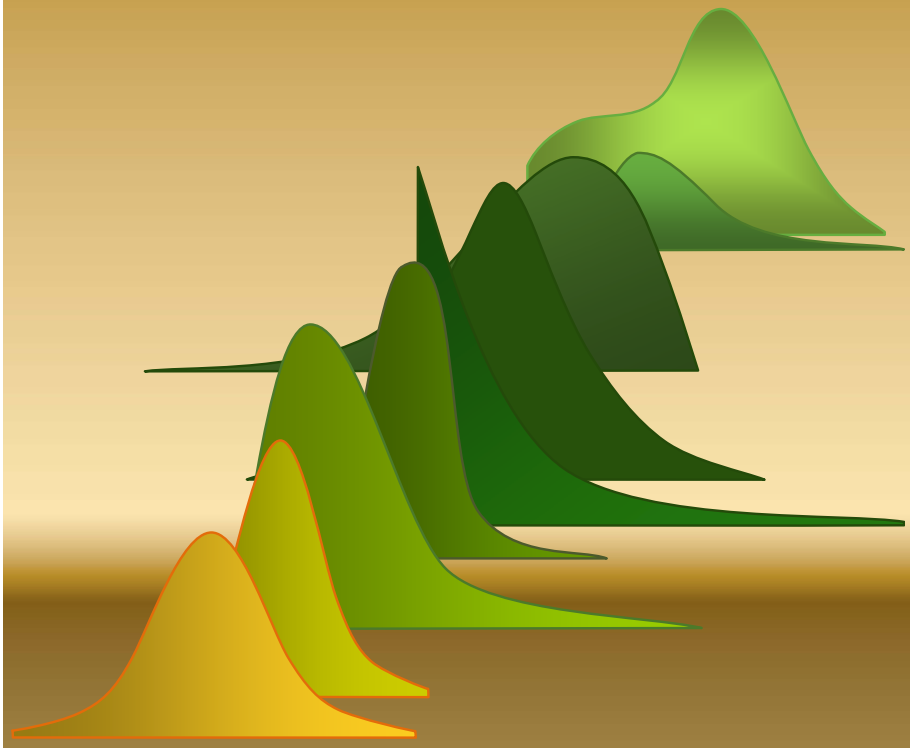
Душан Чакмаков

Веројатност и статистика за инженери



Универзитет „Св. Кирил и Методиј“

Душан Чакмаков

Веројатност и статистика за инженери

Универзитет „Св. Кирил и Методиј“
Скопје, 2015

АВТОР
Проф. д-р Душан Чакмаков, Оддел за математика и информатика, Машински факултет, Универзитет „Св. Кирил и Методиј“, Скопје

РЕЦЕНЗЕНТИ
Проф. д-р Жанета Попеска, Факултет за информатички науки и компјутерско инженерство, Универзитет „Св. Кирил и Методиј“, Скопје
Проф. д-р Никола Тунески, Машински факултет, Универзитет „Св. Кирил и Методиј“, Скопје

Веројатност и статистика за инженери : основен учебник / Душан Чакмаков, Универзитет „Св. Кирил и Методиј“, Скопје, 2015 год., 512 стр., 24 цм., електронско издание.

Издавач: Универзитет „Св. Кирил и Методиј“, Скопје

Техничко уредување и дизајн на корицата: Авторот

Одлука за издавање на учебник бр. 02-2868/ 8 од 27.12.2012 год., Машински факултет - Скопје

Предговор

Целта на оваа книга е да изложи материјал што опфаќа воведни, но и понапредни содржини од теоријата на веројатност и статистиката на еден поприменет начин. За нејзино користење не е потребно никакво предзнаење од веројатност и статистика, но потребни се основни математички предзнаења од околу 2 семестри што вклучуваат елементи од линеарната алгебра, диференцијалното и интегралното сметање. Исто така, потребни се елементарни предзнаења од комбинаторика, софтвер од областа на статистиката за читање вредности на распределби (како на пример Excel) и малку програмирање за додатокот Б.

Книгата содржи значително повеќе материјал од еден почетен курс по веројатност и статистика, прилагоден за инженерите. Изложувањето и содржината на материјалот се такви што книгата има многу поширока примена вклучувајќи ја медицината, економијата, природно-математичките науки итн. Материјалот од наставните програми по предметите веројатност и статистика и одбрани поглавја од веројатност и статистика што повеќе години се предаваат на додипломските и постдипломските студии на Машинскиот факултет во Скопје комплетно е содржан во книгата. Таа исто така делумно или целосно го покрива материјалот по многуте варијации на предметот веројатност и статистика на факултетите низ Универзитетот "Св. Кирил и Методиј", но и на факултетите низ другите универзитети низ државата, како на додипломските така и на постдипломските студии. Такви предмети има не само на "инженерските" факултети како Факултетот за електротехника и информациони технологии, Факултетот за информатика и компјутерско инженерство, Градежниот факултет, туку и на Природно-математичкиот, Економскиот, Медицинскиот факултет, итн.

Текстот е поделен на 15 глави и два додатока.

Во првите три глави се воведуваат елементи од теоријата на веројатност.

Воведната глава ги дава историските корени на теоријата на веројатност како и општите идеи што довеле до толку голем развој и примена на веројатноста и статистиката во другите научни дисциплини. Во втората глава се воведени основните поими од теоријата на веројатност: случаен експеримент и настани, операциите со настани, аксиомите на веројатност, како и класичната и геометриската веројатност. Третата глава е посветена на условната веројатност и соодветните техники за нејзина пресметка. Воведен е важниот поим на независност на настани, како и техниките за пресметка на веројатноста преку: тоталната веројатност, Баесовата формула и сериите независни експерименти.

Следните четири глави, четвртата до седмата, се занимаваат со основниот концепт во теоријата на веројатност - случајните променливи, како и нивните карактеристики.

Во четвртата глава е воведен поимот на случајна променлива и соодветните претставувања преку функцијата и густината на распределба. Разгледани се најчесто користените случајни променливи и нивните распределби, како и случајните вектори. Петтата глава за занимава со бројните карактеристики на случајните променливи и случајните вектори, додека шестата глава е комплетно посветена на функциите од случајните променливи важни за инженерските апликации. Седмата глава ги изложува граничните случаи на низите случајни променливи што е понатаму основа за статистичките модели и оценки. Најголемо внимание е посветено на централната гранична теорема и нејзините импликации.

Од осмата до десеттата глава постапно се преминува од теоријата на веројатност кон статистиката.

Осмата глава ја дискутира логиката и патот кој води од веројатноста кон статистиката воведувајќи го овој "мост" преку емпириската функција на распределба. Во деветата глава се воведуваат статистичките модели што се основа за статистичката анализа на податоци. Десеттата глава е посветена на извлекување на нумерички карактеристики од податоците и нивното графичко претставување (описна статистика).

Главите 11, 12, 13 и 14 ги даваат суштинските концепти на статистичката обработка на податоци и тоа: точкастите оценки на непознати параметри, интервалните оценки и тестирањето хипотези.

Во 11-тата глава во детали е дискутиран проблемот на точкастите оценки и патот по кој може да се дојде до оптимални оценки. 12-тата глава се занимава со проширување на оценките во интервални и техниките за нивно добивање. 13-тата глава е посветена на различните статистички тестови во врска со непознатите параметри, распределби и зависности. Во 14-тата глава се разгледуваат интервални оценки и статистички тестови за повеќе примероци.

Последната 15-та глава е посветена на линеарната регресиона анализа и техниките на оценки, тестови и валидност на ваквите модели.

Додатокот А е посветен на комбинаториката при што е користен елегантен пристап преку формални јазици. Дадени се и алгоритми за генерирање на комбинаторните објекти.

Додатокот Б накусо ја обработува важната тема на веројатносни симулации со генератори на случајни броеви. Додатокот е поткрепен со повеќе репрезентативни примери на симулации дадени со соодветни алгоритми.

Поглавјата означени со * се нешто понапредни и се наменети за читатели со одредени предзнаења од областа. Генерално, тие не се наменети за студентите на додипломските студии и слободно може да се прескокнат и остават за евентуални идни разгледувања.

Книгата содржи околу 200 решени примери, повеќе од 200 задачи и проблеми од кои сите се или решени или за нив има дадено решенија. Вака големиот број решени примери и проблеми овозможуваат книгата да се користи не само како учебник, туку и како збирка задачи со што таа комплетно го покрива материјалот по соодветните предмети, како предавањата така и вежбите.

Од обемната литература од оваа област што постои во светот, најголемо влијание на овој текст го имаа следните 6 одлични книги: [Devore 2012], [Montgomery, Runger 2003], [Spanos 1999], [Mendenhal, Sincich 1992], [Tijms 2007], [Soong 2004].

Некои типографски конвенции користени во оваа книга се дадени во следната табела:

Ако видите	***Тоа значи***
$\mathbb{N}$	Природни броеви
$\mathbb{R}$	Реални броеви
p	Веројатност
$\|$	Условна веројатност
акко	ако и само ако
"…"	Настани
$\|A\|$	Број на елементи во A
$\sim$	Пропорционален
A, B, X, Y, …	Множества или настани
$\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$, $\mathbf{Y}$, …	Матрици
$\mathbf{x}$, $\mathbf{y}$, z, …	Вектори
A, B, X, Y, …	Случајни променливи (СП)
$F()$, $f()$	Функција на распределба и густина на распределба
$\varphi(t)$	Карактеристична функција
Z, z	СП со нормална распределба и нејзина вредност
T, t	СП со студентова распределба и нејзина вредност
F, f	СП со Фишерова распределба и нејзина вредност
$\overline{X}$, $\bar{x}$	Очекување (просек) на СП и негова вредност
S^2, s^2	Дисперзија (варијанса) на СП и нејзина вредност
E, μ	Очекување (просек) на СП
D, σ^2	Дисперзија (варијанса) на СП
$K_{X,Y}$, $\rho_{X,Y}$	Коваријација и коефициент на корелација
θ, $\hat{\theta}$	Параметар на распределба и негова оценка
H_0, H_A	Нулта и алтернативна хипотеза
■	Крај на доказ
■	Крај на пример

Содржина

1

Вовед

До појавата на теоријата на веројатност, објекти на изучување во науката биле појави и случувања коишто практично го определувале исходот еднозначно. Од друга страна, јасно е дека многу од нив (можеби и сите, ако се прифати размислувањето дека сè околу нас подлежи на случајност) имаат неоднозначен и помалку или повеќе непредвидлив, т.е. случаен исход. Теоријата на веројатност е таа што ги изучува математичките модели на ваквите појави и случувања чиишто исходи се случајни, т.е неоднозначни и непредвидливи.

Уште од праисториско време, човештвото било свесно за детерминистичките појави како што се: изгревањето и заоѓањето на сонцето, менувањето на годишните времиња, плимата и осеката на морињата, итн. Се разбира, се забележувале и многу случајни појави како: врнење дожд или снег, пронаоѓање некоја храна, доаѓање на болест итн. Со развојот на цивилизацијата, случајните појави станувале сè позабележливи и сè подоминантно влијаеле на секојдневниот живот. Тука посебно место заземаат појавите и случувањата поврзани со спортските игри и игрите на среќа, каде што случајноста на исходите е очигледна и лесно видлива.

Несомнено е дека основите на теоријата на веројатност произлегле од игрите на среќа. Според ископините на средниот исток и Индија, и уште порано во Египет, познато е дека првите примероци на коцките за играње потекнуваат уште од 3500 година пред нашата ера. Тогашните 4-страни коцки направени од коските на животните (astragalus), се претходници на модерните коцки за играње што се користат уште од средниот век.

Се смета дека првите математички проучувања на коцките за играње (а со тоа и на соодветните веројатности) датираат од 16-тиот век од страна на италијанскиот математичар и физичар Кардано (Gerolamo Cardano, 1501-1576). Тој за првпат ги вовел поимите за настани (outcomes) на експеримент и во случај кога сите настани се еднакво веројатни, тој ја вовел пресметката на веројатноста како однос на бројот на поволни настани врз бројот на сите можни. Овој пристап, заедно со точните пресметки на бројот на настани, му овозможиле на славниот астроном Галилеј (Galileo Galilei, 1564-1642) да објасни зошто при фрлање на 3 коцки веројатноста да се добие збир 10 е поголема од веројатноста да се добие збир 9 (веројатностите се 27/216 и 25/216 соодветно).

Базичните принципи на класичната теорија на веројатност биле воспоставени од Паскал (Blaise Pascal, 1623-1662) и Ферма (Pierre de Fermat, 1601-1665). Сè почнало со нивната кратка писмена кореспонденција во летото 1654 година во којашто тие разгледувале некои специфични проблеми поврзани со игрите на среќа. Еден добро-познат проблем бил поврзан со францускиот писател и математичар-аматер Шевалие (Chevalier de Méré, 1607-1684) кој тврдел дека нашол контрадикција во аритметиката. Имено, тој стекнал богатство кладејќи се на добивање на 6-ка од 4-ри фрлања на коцка. Потоа, тој ја сменил играта кладејќи се на добивање на 2-е 6-ки од 24-ри фрлања на две коцки што се покажало како неповолоно. Според него, контрадикцијата е во тоа што шансите за добивка и во двата случаи треба да се исти поради еднаквоста на односите 4/6 = 24/36. Паскал напишал во коресподенцијата дека Шевалие е голем коцкар и паметен човек, но не е геометар, укажувајќи дека веројатноста не ги следи законите на пропорции. Денеска знаеме дека веројатностите на добивка во првата игра е 0.5177, а во втората 0.4914. Интересно е дека за добивање на приближно рамноправни шанси за 6-ка при фрлање на една коцка се потребни 4-фрлања, а не 3 како што некој би очекувал.

Од ваквите и слични проблеми произлегла класичната теорија на веројатност работена во периодот 1680-1780 година, којашто општо зборувајќи требала да го реши следниот проблем:

> *Како треба да се распореди влогот во една игра на среќа за да се осигура добивка ако доволно долго се игра.*

Особините на ваквите игри се:

1) Теоретски немаат крај (може да се повторуваат произволен број пати);

2) Инструментот со кој се игра останува непроменет (коцка, карти, рулет, итн.);
3) Целиот процес е случаен и се прави според однапред договорено правило.

Некои од термините на теоријата на веројатност денеска наширо-ко се користат во секојдневната комуникација на луѓето. Постојано околу нас ги слушаме фразите: веројатно/неверојатно, возможно/невозможно, сигурно/несигурно, има/нема шанси итн. Изразите како "шансите се 50-50 (фифти-фифти)", "добивката е 7 спрема 4 за победа на ...", "речиси сигурно пропаѓаме", "невозможно е тоа да се случи" итн. се разбирливи и користени од поголемиот број луѓе.

Во денешно време, теоријата на веројатност и нејзината примени реализирани преку статистиката се испреплетуваат во сите дисциплини, тргнувајќи од математиката, информатиката, физиката и инженерството, па преку општествените науки сè до медицината и политологијата. На пример, авторот на книгата имал прилика да помогне во изработката на неколку магистратури и докторски дисертации од областа на медицинските науки. Теоријата на веројатност е интегрален дел на нашиот секојдневен живот. Познатиот математичар Лаплас (Pierre-Simon Laplace, 1749-1827) бил во право кога уште пред повеќе од 200 години напишал

> *Теоријата на веројатност во основа не е ништо друго од секојдневна логика* (*common sense*) *сведена на пресметки.*

Генераторите на случајни броеви што се во основа на веројатносните симулации на реалноста се едно од поголемите откритија во науката. Сепак, едно е сигурно, тој генијалец што го креирал првиот генератор на случајни броеви ќе остане засекогаш анонимен како што е анонимен и пронаоѓачот на можеби најважниот пронајдок во историјата на човештвото, тркалото.

Теоријата на веројатност е математичка дисциплина што е можеби најблиска и најдиректно поврзана со реалноста. Токму од тие причини кај неа подиректно се судруваме со одредени парадокси врзани со поимите за бесконечност, реалните броеви или многу малите големини. Најопшто кажано, тврдењата за веројатноста на случувањата се тврдења за нашите верувања. Различни луѓе имаат различни верувања за веројатноста на исто случување. На пример, такви се спортските обложувања каде што односот на добивката за некој резултат од r спрема s значи дека неговата веројатност p се проценува на $r/(r+s)$. Веројатноста

е нумеричка квантификација на нашите верувања за идните случувања. Еден настан со проценета веројатност p очекуваме да се случи приближно $p \cdot n$ пати во идните n испитувања. Веројатностите ги изразуваме со броеви од интервалот [0, 1] или во % кога вредноста ќе ја помножиме со 100. Така, веројатностите стануваат обични броеви без дополнително значење и тоа е во основата на аксиоматскиот пристап кон теоријата на веројатност. Аксиоматскиот пристап при некои однапред дадени веројатности (комплексно прашање е како) овозможува пресметки на други веројатности на еден математички и логички конзистентен начин. Така, ние можеме да ги пресметуваме веројатностите без филозофски дискусии за нивното значење. За некој што верува дека веројатноста е чисто субјективна категорија, целата теорија ќе биде само некаква шема на пресметки или пак аксиоматскиот пристап ќе му укажува што другите веруваат и колку е тоа во согласност со неговото верување.

За разлика од теоријата на веројатност, статистиката се занимава со собирање и организирање на емпириски и експериментални податоци и ги користи методите од теоријата на веројатност за анализа и изведување заклучоци од собраните податоци. На пример, теоријата на веројатност дава методи за одговор на прашањата од тип: Колкава е веројатноста од 10 фрлања на фер паричка да се добие петка 6 пати?, и го дава одговорот прецизно. Статистиката се обидува да одговори на прашањето: Ако при 10 фрлања на паричка се добила петка 6 пати колку е разумно да се заклучи дека паричката е фер? и го дава одговорот непрецизно, со некоја веројатност. Дефинитивен одговор не е можен бидејќи различни луѓе имаат различна идеја за тоа што е разумно. Во основа, статистичките заклучоци се придружени со ниво на доверба, на пример, со 95% веројатност паричката е фер. Не постои статитичка метода што може да докаже дека паричката е фер, бидејќи тоа е прашање на верување и статистиката може да го даде само степенот на верување преку нивото на доверба, т.е веројатност.

Гледано од инженерски аспект, статистиката се користи како алатка што помага да се опише и разбере *варијабилноста* на разгледуваниот систем. Под варијабилност се подразбира ситуација кога последователни набљудувања на некој систем или феномен не дава точно ист резултат. На пример, да го разгледаме процентот на дефектни производи од една производствена лента. Дали овој процент секој ден е еднаков? Се разбира, не. Може да се очекува дека овој процент значително варира. Оваа варијабилност може да се должи на многу фактори, како на пример: варијабилноста на влезните компоненти, времето од последната калибрација на машините, различни човечки фактори и многу други повеќе или помалку влијателни фактори што може да бидат и непоз-

нати. Статистиката е таа што ни дава методи да се опише ваквата варијабилност и дава одговор на многу прашања за потенцијалните причини на варијабилноста, кој од факторите е со најголемо влијание, дали има корелација меѓу различните фактори итн. Како друг пример да ја разгледаме потрошувачката на гориво на еден автомобил. Дали тој поминува ист број километри со еден полн резервоар? Се разбира, не. Варијабилноста на потрошувачката на гориво зависи од многу фактори: каде се направени километрите (градско возење или отворен пат), брзината на возење, состојбата на гумите, квалитетот на горивото, надворешната температура и временските услови и многу други фактори што може да бидат и непознати. Повторно методите на статистиката се тие што може да ни дадат одговор на многу важни прашања за причините на ваквата варијабилност и со тоа да ни овозможат донесување одлука за евентуално намалување на потрошувачката преку промени во идентификуваните влијателни фактори. И во секојдневниот живот, ние постојано се судираме со варијабилност и тогаш "статистичкото размислување" ни овозможува да ја вклучиме ваквата варијабилност во донесувањето одлуки.

Често пати, физичките закони како Њутновиот (Newton), Омовиот (Ohm), Хуковиот (Hook), итн., се применуваат во развојот на продукти и процеси. Ова е добро познат тип на расудување, од општи закони кон специфични случаи на нивна примена. Од друга страна, исто така е важно расудувањето што оди од конкретни мерења и набљудувања кон поопшти заклучоци корисни за развојот на продуктите и процесите. Расудувањето од земен примерок (неколку производи од фабриката) кон изведување заклучоци за целата популација (производите и процесот на производство) е во основа на статистичката анализа. Историски, термините примерок-популација потекнуваат од расудувањето дека земени податоци од примерок на луѓе може да дадат заклучоци обопштени на целата популација. Јасно е дека расудувањето базирано на примерок од неколку објекти што изведува заклучоци за целата популација е подложно на грешки. Сепак, кога примерокот е избран соодветно, овие грешки може да се квантифицираат и минимизираат со соодветно избрана големина и случајност на примерокот.

2

Случајни настани и веројатност

Аргументите на теоријата на веројатност се случајните настани. Во оваа глава ќе биде прецизиран овој интуитивен поим којшто понатаму овозможува аксиоматско воведување на поимот веројатност.

2.1. Простор на случајни настани

Под поимот *експеримент* подразбираме некоја активност што се одвива зависно (со учество) или независно од набљудувачот и којашто резултира со некакви исходи (резултати) што ќе ги нарекуваме *настани*. Значи под експеримент подразбираме сè, од наједноставна активност како фрлање паричка, преку комплексни лабораториски испитувања, па сè до набљудување на природата и општеството во најширока смисла.

Не сите експерименти се погодни за математичко изучување, туку само оние што задоволуваат одредени услови:

1) (Случајност), т.е. неедноznачност на исходите т.е. настаните;

Тоа значи дека експериментот треба да резултира во повеќе можни настани. На пример, фрлање коцка за играње, победа на тим во натпревар, прогноза на времето, предвидување на победа на избори

се такви експерименти. Од друга страна, паѓање камен на земјата или загревање вода до вриење не се такви експерименти.

2) (Можност за повторување), т.е. можност барем во принцип експериментот да се повтори при исти услови неограничен број пати;

Јасно е дека во која било наука не може да се изучуваат настани што повеќе никогаш нема да се случат. Значи погодни експерименти се оние што можат да се повторат или случат "многу" пати. Ако експериментот се повтори n-пати и притоа некој настан A се појавил k-пати, тогаш бројот $W(A) = k/n$ е релативна честота на A.

3) (Стабилност), што значи при кои било две повторувања (доволен број пати) на експериментот, релативните честоти на појавување на настанот A се приближно еднакви;

Ако n_1 и n_2 се се две повторувања на експериментот доволен број пати, при што настанот A се појавил k_1 и k_2 пати соодветно, тогаш за релативните честоти на појавувањето на настанот A треба да важи $W_1(A) = k_1/n_1 \approx k_2/n_2 = W_2(A)$.

Експериментите со горните особини ќе ги нарекуваме *случајни експерименти*, а нивните настани ќе ги нарекуваме *случајни настани*. Понатаму ќе нè интерсираат само случајните експерименти и случајните настани и нив ќе ги нарекуваме, едноставно, експерименти и настани.

ПРИМЕР 2.1 Да ги разгледаме експериментите со настани:

1) Фрлање паричка со множеството настани $\Omega = \{$"петка", "глава"$\}$;

2) Фрлање 2 коцки со множества настани:

$\Omega_1 = \{(x, y), x=1,2, ...,6; y=1,2, ...,6\}$ - *сите парови*;

$\Omega_2 = \{2, 3, 4, ..., 12\}$ - *збир*;

$\Omega_3 = \{$"парен збир", "непарен збир"$\}$ - *тип на збир*;

$\Omega_4 = \{$"ист број", "различен број", "збир 7"$\}$ - *недисјунктни*;

3) Повици до брза помош во еден месец со множества настани:

$\Omega_1 = \{0, 1, 2, ...\}$;

$\Omega_2 = \{(x, y, z), x, y, z \in \mathbb{N};$ $x =$ "број на дневни повици",
$y =$ "број на ноќни повици",
$z =$ "број на лажни повици"$\}$,

4) Пукање во мета со множество настани

$$\Omega = \{(x, y), x, y \in \mathbb{R}; x^2 + y^2 \le R^2\};$$

5) Брауново движење на честичка во временски интервал $[0, T]$ со множество настани $\Omega = \{(x(t), y(t), z(t)), t \in [0, T]\}$. ■

Од горните примери, множествата настани во 1 и 2 се конечни, а во 4 и 5 бесконечни и непребројиви. Множеството настани во 3 е секогаш конечно иако не сме го ограничиле (тоа би можеле да го направиме знаејќи дека бројот на повици во пракса не може да надмине одреден број). Како што се гледа од примерите 2 и 3, исходите од еден експеримент може да се прикажат со различни множества настани. Кое множество ќе го избереме зависи од проблемот што го решаваме. На пример, во примерот 2, како исход на експериментот имаме дадено 4 различни множества настани. Од овие 4 множества, интуитивно е јасно дека Ω_1 е најинформативното множество настани бидејќи секој настан од другите множества: Ω_2, Ω_3 и Ω_4, може да се претстави како колекција на настани од Ω_1.

2.1.1. Простор на елементарни настани

Множеството настани Ω го нарекуваме простор на *елементарни настани* акко

а) При секоја реализација на експериментот добиениот настан е елемент на Ω;

б) Нема два настана од Ω што може истовремено да настапат (да се случат).

Ако ги разгледаме просторите на настани Ω_1, Ω_2, Ω_3 и Ω_4 од примерот 2, јасно е дека

Ω_1, Ω_2 и Ω_3 се простори на елементарни настани;

Ω_4 не е, бидејќи настаните "различен број" и "збир 7" може истовремено да се случат.

Од Ω_1, Ω_2 и Ω_3, како што веќе нагласивме, Ω_1 е најинформативен бидејќи секој настан од Ω_2 и Ω_3 може да се претстави преку настани од Ω_1, додека обратното не важи. На пример,

$4 \in \Omega_2$ се претставува со $\{(1, 3), (2, 2), (3, 1)\}$ од Ω_1; но

$(5, 4) \in \Omega_1$ не може да претстави со настани од Ω_2 ($9 \in \Omega_2$ опфаќа многу настани од Ω_1).

Јасно е дека просторот на елементарни настани не ги исцрпува сите настани во врска со експериментот. Тој само ги дава сите елементарни

исходи на експериментот. Нека Ω е просторот на елементарните настани во врска со некој експеримент. Тогаш, ако A е настан во врска со тој експеримент, на A секогаш може да му придружиме подмножество од Ω составено од оние елементарни настани што го претставуваат (опишуваат) A. На пример, во врска со експериментот - фрлање на две коцки, даден настан A може да се претстави преку на настани од Ω_1 на следниот начин:

A = "сумата е парен број", $A = \{(x, y), x + y = 2k, x, y = 1, 2, \ldots, 6\}$;

A = "ист број", $A = \{(x, x), x = 1, 2, \ldots, 6\}$;

A = "збир 8", $A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$.

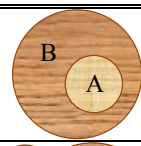
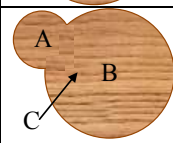
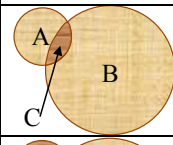
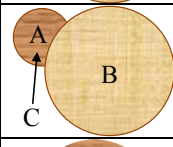
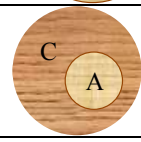
Од дискусијата произлегува дека настан е само едно подмножество на множеството (просторот) на елементарни настани Ω. Понатаму ќе видиме дека овој поглед на настаните останува да важи и во случаите кога Ω е бесконечно и непребројиво.

2.1.2. Операции со настани

Од сите настани во врска со некој експеримент, два настана имаат стандардно значење. Првиот од нив е празниот настан $\varnothing \subseteq \Omega$, т.е. со други зборови *невозможен* настан бидејќи никогаш не се случува. Вториот е целото $\Omega \subseteq \Omega$, коешто се нарекува *сигурен* настан бидејќи секогаш се случува.

Настаните се во основа подмножества и со нив може да се прават истите операции како и со множествата, но со малку различна интерпретација на резултатите. Нека A и B се два настана во врска со некој експеримент. Во табелата дадена подолу, наведени се основните операции со настани.

При операциите со настани важат истите правила како кај операциите со множества. Важи на пример $A - \overline{B} = A \cdot B$ или $A = A - B + A \cdot B$, така што основните операции со настани се сосема аналогни на оние со множествата. Сепак, "читањето" на операциите со настани е малку специфично и не е исто како кај множествата (види ја табелата). Некои автори прават разлика меѓу операциите $\cup$ и + (+ користат само за унија на дисјунктни настани), како и меѓу операциите $\cap$ и $\cdot$ ($\cdot$ користат само за пресек на дисјунктни настани). Во оваа книга, ние ќе ги користиме операциите + и $\cdot$ за унија и пресек со исто значење на $\cup$ и $\cap$. Одбегнувањето на $\cup$ и $\cap$ е со цел да се нагласи дека работиме со настани, а не со множества.

Операција	Интерпретација	Слика	Пример
$A \subseteq B$	А го повлекува В (В следува од А); секогаш кога се случува А се случува и В.		ако А = "ист број", а В = "сумата е парен број", тогаш $A \subseteq B$.
$C = A \cup B$ или $C = A+B$	С е збир на А и В и се случува кога се случил барем еден од А или В.		ако А = "ист број", а В = "различен број", тогаш $A+B = \Omega$.
$C = A \cap B$ или $C = A \cdot B$	С е пресек (производ) на А и В и се случува кога се случил и А и В.		ако А = "ист број", а В = "збир 8", тогаш $A \cdot B = \{(4,4)\}$.
$C = A \setminus B$ или $C = A - B$	С е разлика на А и В и се случува кога се случил А и не се случил В.		ако А = "ист број", а В = "збир помал од 9", тогаш $A-B = \{(5,5), (6,6)\}$.
$C = A^c$ или $C = \bar{A}$	С е спротивен настан (комплемент) на А и се случува кога не се случил А.		важи $\bar{A} = \Omega - A$, $A + \bar{A} = \Omega$, $A \cdot \bar{A} = \varnothing$.

2.2. Аксиоми на веројатност

Нека Ω е множество на елементарни настани. Подмножествата на Ω се настани на кои сакаме да им доделуваме веројатности, или со други зборови, мери за шансите на нивното случување. Во многу ситуации, не сите подмножества се од интерес. Уште повеќе, кога Ω е непребројиво, не постои добар начин за секое подмножество на Ω да се дефинира конзистентна мера, т.е. веројатност, што би задоволувала одредени "логични" правила. Од тие причини, се прави редукција на подмножествата на Ω за кои се дефинира веројатноста. Редуцираната фамилија подмножества се бара да биде затворена во однос на стандардните операции со множества: униите, комплементите и пресеците, и вообичаено се нарекува σ-алгебра.

Дефиниција 2.1 Нека $\mathcal{F}$ биде фамилијата подмножества на Ω коишто ќе ги нарекуваме настани (на кои ќе им доделуваме веројатности). Тогаш $\mathcal{F}$ е σ-алгебра акко важи:

1) $\Omega \in \mathcal{F}$,

2) ако $A \in \mathcal{F}$, тогаш и $\bar{A} \in \mathcal{F}$,

3) ако $A_1, A_2, ..., A_n, ... \in \mathcal{F}$ е пребројива низа настани, тогаш и $A_1 + A_2 + ... + A_n + ... \in \mathcal{F}$.

Аксиомите 1-3 може да сумираат со: σ-алгебрата $\mathcal{F}$ го содржи Ω (1), затворена е во однос на комплементи (2) и пребројивите збирови, т.е. унии (3). Од $\overline{A_1+A_2+ ... +A_n+ ...} \in \mathcal{F}$ следи дека $\bar{A}_1 \cdot \bar{A}_2 \cdot ... \cdot \bar{A}_n \cdot ... \in \mathcal{F}$, т.е. $\mathcal{F}$ е затворена и во однос на пребројивите производи (пресеци).

За конечни или пребројиви Ω, најголемото $\mathcal{F}$ се состои од фамилијата на сите подмножества на Ω бидејќи тие формираат σ-алгебра. Кога $|\Omega| = n$, имаме 2^n подмножества, т.е. настани. Најмали σ-алгебри $\mathcal{F}$ се $\{\Omega, \varnothing\}$ и потоа $\{\Omega, A, \bar{A}, \varnothing\}$. За секој избор на подмножества на Ω, тие, заедно со подмножествата што се добиваат со нивни збирови, производи, спротивности исто така формираат σ-алгебра $\mathcal{F}$.

Најголемата σ-алгебра $\mathcal{F}$ на реалната оска ги содржи сите интервали од секој тип, вклучувајќи ги бесконечните интервали како и сите нивни пребројиви збирови, производи и спротивности. Дали постојат множества на реалната оска што не се во $\mathcal{F}$? Одговорот е да, но нивната конструкција е комплицирана и тие никогаш не се појавуваат во апликациите на веројатноста, па затоа немаат некоја практична вредност. Постоењето на ваквите множества се еден симптом на "нереалност" на реалните броеви.

Откако прецизно е дефиниран просторот на настани $\mathcal{F}$, останува да се дефинира веројатносната мера, т.е. правила по кои на настаните им придружуваме реални броеви - веројатности.

Дефиниција 2.2 Нека $\mathcal{F}$ е σ-алгебра на настани. Веројатност p е пресликување од $\mathcal{F}$ во $\mathbb{R}$ (p: $\mathcal{F} \to \mathbb{R}$) за кое важи:

a1) $p(A) \geq 0, \ \forall A \in \mathcal{F}$,

a2) $p(\Omega) = 1$,

a3) за конечно Ω

$$p(A_1 + A_2) = p(A_1) + p(A_2), \ \forall A_1, A_2 \in \mathcal{F} \text{ и } A_1 \cdot A_2 = \varnothing$$

за бесконечно Ω

$$p(A_1+A_2+ ... +A_n+ ....) = p(A_1) + p(A_2) + ... + p(A_n) + ...$$

$$\forall A_1, A_2, ..., A_n, ... \in \mathcal{F} \text{ и } A_i \cdot A_j = \varnothing.$$

Тројката $(\Omega, \mathcal{F}, p)$ се нарекува простор на веројатност.

Кога множеството елементарни настани Ω е конечно или пребројиво се добива *дискретен простор на веројатност*. Бесконечно и непребројиво Ω води кон *непрекинат простор на веројатност*, за кој ќе стане збор понатаму.

Во пракса, при решавање на веројатносни и статистички проблеми, аксиоматскиот пристап директно не се користи. Тогаш се важни особините и од нив изведените техники на пресметка.

Теорема 2.1 Веројатноста p ги има следните особини:

1) $p(\varnothing) = 0$;

2) $p(\bar{A}) = 1 - p(A)$;

3) $A \subseteq B \Rightarrow p(A) \le p(B)$;

4) $0 \le p(A) \le 1$;

5) $p(A+B) = p(A) + p(B) - p(A \cdot B)$;

6) ако $A_1 \subset A_2 \subset \ldots \subset A_n \subset \ldots$ и $A = \sum_{n=1}^{\infty} A_n$ или

ако $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ и $A = \prod_{n=1}^{\infty} A_n$

тогаш $p(A) = \lim_{n \to \infty} p(A_n)$.

Доказ: Овие особини следуваат директно од аксиомите:

1) од $\Omega \cdot \varnothing = \varnothing$ следува $1 = p(\Omega) = p(\Omega + \varnothing) = p(\Omega) + p(\varnothing) = 1 + p(\varnothing)$;

2) следува од $1 = p(\Omega) = p(A + \bar{A}) = p(A) + p(\bar{A})$;

3) од $B = A + \bar{A} \cdot B$ и $A \cdot (\bar{A} \cdot B) = \varnothing$ следува $p(B) = p(A) + p(\bar{A} \cdot B) \ge p(A)$;

4) од $\varnothing \subseteq A \subseteq \Omega$ следува $p(\varnothing) \le p(A) \le p(\Omega)$ т.е. $0 \le p(A) \le 1$,

5) од $A + B = A + (B - A)$ и $B = A \cdot B + (B - A)$ следува

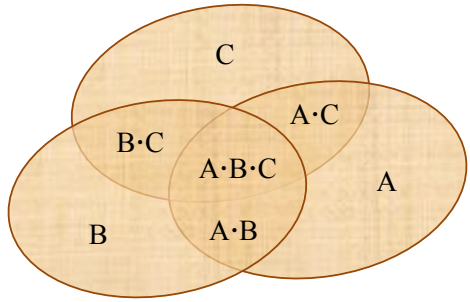
$p(A + B) = p(A) + p(B - A)$ и $p(B) = p(A \cdot B) + p(B - A)$

и со одземање на овие равенства се добива особината 5.

6) следува од особината: секоја монотоно растечка (опаѓачка) низа и ограничена од горе (долу) конвергира (растечката низа веројатности е ограничена со 1, опаѓачката со 0). ■

Особината 5) аналогно се обопштува на случај на повеќе од 2 настана (види слика),

$$p(\mathrm{A}+\mathrm{B}+\mathrm{C}) = p(\mathrm{A}) + p(\mathrm{B}) + p(\mathrm{C}) - p(\mathrm{AB}) - p(\mathrm{AC}) - p(\mathrm{BC}) + p(\mathrm{ABC}),$$



или општо,

$$p(\mathrm{A}_1+\mathrm{A}_2+\ldots+\mathrm{A}_n) = \sum_{i=1}^{n} p(\mathrm{A}_i) - \sum_{1\le i,j\le n} p(\mathrm{A}_i\mathrm{A}_j) + \sum_{1\le i,j,k\le n} p(\mathrm{A}_i\mathrm{A}_j\mathrm{A}_k) - \ldots + (-1)^{n-1} p(\mathrm{A}_1\mathrm{A}_2\cdots\mathrm{A}_n).$$

Оваа формула обично се нарекува просејување (sieve) или принцип на вклучување-исклучување (inclusion-exclusion principle). Таа лесно се докажува со индукција.

Да разгледаме неколку примери во врска со настаните и дефиницијата на веројатност.

ПРИМЕР 2.2 Нека A, B и C се настани во некој простор на веројатност.

а) Ако $p(\mathrm{A}) = 0.3$, $p(\mathrm{B}) = 0.4$ и $p(\bar{\mathrm{A}}\cdot\bar{\mathrm{B}}) = 0.4$, пресметај ја веројатноста $p(\mathrm{A}\cdot\mathrm{B})$.

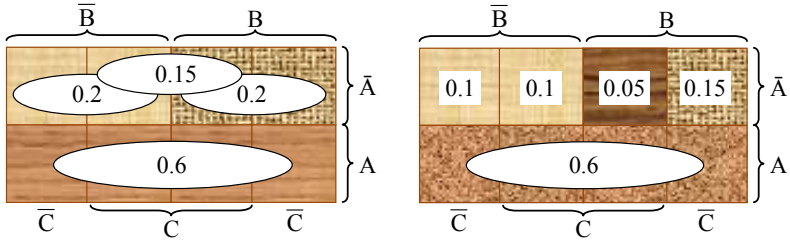
б) Ако $p(\mathrm{A}) = 0.6$, $p(\mathrm{A}+\overline{\mathrm{B}}) = 0.8$, $p(\mathrm{A}+\mathrm{B}+\mathrm{C}) = 0.9$ и $p(\bar{\mathrm{A}}\cdot\mathrm{C}) = 0.15$, пресметај ја веројатноста на $p(\bar{\mathrm{A}}\cdot\mathrm{B}\cdot\mathrm{C})$.

Решение

а) Од $\overline{\mathrm{A}}\cdot\overline{\mathrm{B}} = \overline{\mathrm{A}+\mathrm{B}}$ и $p(\mathrm{A}+\mathrm{B}) = 1 - p(\overline{\mathrm{A}+\mathrm{B}})$, користејќи ја особината 5 од теоремата 2.1 добиваме

$$p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A}) + p(\mathrm{B}) - p(\mathrm{A}+\mathrm{B}) = 0.3 + 0.4 - 1 + 0.5 = 0.2.$$

б) За покомплексни проблеми од овој вид може да се користат таканаречените Карнови мапи (Karnaugh maps). Нивната главна примена е за упростување на Буловите изрази, а во веројатноста поради својата прегледност се од голема помош за решавање на проблеми во врска со операции со настани. Во конкретниот случај имаме 3 променливи A, B и C, што ги ставаме (заедно со обратните настани) на 3 страни од правоаголникот-мапа како што е направено на следната слика



Од условите на проблемот $p(A) = 0.6$, $p(A+\overline{B}) = 0.8$, $p(\bar{A}\cdot C) = 0.15$ и $p(A+B+C) = 0.9$ се конструира мапата (лево). Од неа следува мапата (десно), од којашто веднаш се добива бараната веројатност $p(\bar{A}\cdot B\cdot C) = 0.05$.

Забележи дека Карновите мапи се најпогодни за решавање проблеми со не повеќе од 4 променливи. ■

Во основа, веројатност е секоја функција што ги задоволува горните аксиоми, без разлика на нејзината интерпретација. Во таа смисла, веројатноста е чисто математички концепт исто како што е, на пример реалната оска $\mathbb{R}$. Веројатносниот простор $(\Omega, \mathcal{F}, p)$ дава правила за манипулации со веројатностите на настаните од интерес, но не дава начин како овие веројатности да се добијат. Затоа математичката теорија на веројатност се нарекува и веројатносно сметање (calculus of probabilities). Кога решаваме некој конкретен веројатносен проблем ние одиме многу подалеку од математичката дефиниција, при што основно прашање е како да ги зададеме веројатностите на елементарните настани. Обично за тоа користиме "логични", "очигледни", "разумни" аргументи, како што е физичката симетричност на објектот (пр. фрлање коцка) или претпоставената несиметричност (пр. победа во натпревар). Генерално, може да се разгледуваат 3 интерпретации на веројатноста дадени преку: класичен пристап, релативни фреквенции и степен на верување.

Класичната интерпретација се базира на доделување еднакви веројатности на елементарните настани според претпоставката за физичка симетрија. Интерпретацијата преку релативните фреквенции се базира на статистичка стабилност на настаните при повторувањата на експериментите. Степенот на верување во основа се базира на согледувања, докази и убедувањето во свеста на индивидуата. Во оваа книга е користен "најстандардниот" пристап преку релативните фреквенции. За анализа и моделирање на емпириски податоци во статистиката, овој пристап е секако најсоодветен.

Откако ги имаме аксиомите и основните особини на веројатноста, некој би можел да се запраша што понатаму има да се изучува. Да забе-

лежиме дека во конкретен простор на веројатност познати се веројатностите само на некои настани, најчесто на елементарните. Во теоријата на веројатност развиени се многу техники што овозможуваат пресметка на веројатностите на едни настани преку веројатностите на други настани. Но зошто непознатите веројатности не се проценуваат директно, преку релативните честоти користејќи симулации? Одговорот на тоа не е едноставен. Некои веројатности се премали, или симулацијата е прекомплексна за да би можела да се направи. На пример, надежноста на еден комплексен систем е многу полесно да се пресмета отколку процени со симулации. Како би ја процениле веројатноста на дефект во еден нуклеарен реактор? Уште поважно, многу појави и случувања не може да се симулираат ниту компјутерски ниту реално, на пример, кога експериментот не може да се повторува прозволен број пати (археолошки ископини, болести, несреќни случаи итн.).

За решавање проблеми од теоријата на веројатност најпрво е потребно да се дефинира соодветен простор на веројатност $(\Omega, \mathcal{F}, p)$ во кој ќе се бара решението. Потоа се колектираат познатите веројатности, обично на елементарните настани од Ω. Дури потоа може да се пресметуваат веројатностите на други настани од $\mathcal{F}$. Вообичаен принцип е пресметката на веројатностите на покомплексни настани да оди преку веројатностите на поедноставните настани.

ПРИМЕР 2.3 Една коцка е изработена така што честотата на појавување на еден број е пропорционална со тој број (на пример, со додадени тежини спрот броевите). Да се пресмета веројатноста на добивање парен број.

Решение

$\Omega = \{1, 2, 3, 4, 5, 6\}$, $p(\omega_i) = k \cdot i$

од $\sum_{i=1}^{6} p(\omega_i) = 1$ добиваме $k(1+2+3+4+5+6) = 1 \Rightarrow k = 1/21$

за A = {2, 4, 6} имаме $p(A) = 2/21 + 4/21 + 6/21 = 4/7$. ■

ПРИМЕР 2.4 Да се пресмета веројатноста дека при фрлање 2 парички тие ќе паднат со различни страни.

Решение

Да разгледаме 3 решенија:

а) Ω = {ПП, ПГ, ГП, ГГ}, $p(\omega_i) = 1/4$; за A = {ПГ, ГП}, $p(A) = 1/2$;

б) Ω = {"0 П", "1 П", "2 П"}, $p(\omega_i) = 1/3$; за A = {"1 П"}, $p(A) = 1/3$;

в) Ω = {"две П", "П и Г", "две Г"}, $p(\omega_i) = 1/3$; за A = {"П и Г"}, $p(A) = 1/3$.

Решението (а) е точно, а (б) и (в) се погрешни бидејќи кај нив веројатностите за елементарните настани не се добри, т.е. тие не соодветствуваат на релативните честоти на нивното случување, што може и експериментално да се потврди. Во (б) и (в) би требало да се стави p("2 П") = p("две П") = 1/4, p("1 П") = p("П и Г") = 1/2 и p("0 П") = p("две Г") = 1/4 и тогаш сé би било во ред. Сепак, просторот на елементарни настани во (а) е најинформативен и дава еднаквоверојатни елементарни настани. ■

Кога работиме со елементарни настани во врска со поголем број исти објекти, општо правило е дека самите објекти треба да се третираат како различни и така да се добие најинформативен простор на веројатност (решението а) од примерот 2.4). Постојат многу мал број појави во физиката поврзани со честички на податомско ниво коишто треба да се третираат како физички еднакви и за кои експериментално е потврдено дека треба да се примени простор на веројатност соодветен на решението (б) или (в).

2.3. Класичен простор на веројатност

Класичната веројатност е почетната точка од којашто потекнува денешната модерна теорија и, како што дискутиравме во воведниот дел, таа произлегла во 17 век од игрите на среќа. Интуитивно, класичната веројатност се занимава со конечни множества елементарни настани коишто се еднаквоверојатни. Попрецизно, нека (Ω, $\mathcal{F}$, p) е дискретен простор на веројатност каде што Ω е конечно.

Дефиниција 2.3 (Ω, $\mathcal{F}$, p) е класичен простор на веројатност акко $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ и $p(\omega_i) = 1/n$.

Сега, ако настанот $A = \{\omega_{i1}, \omega_{i2}, ..., \omega_{ik}\}$, веројатноста на А е едноставно дадена со $p(A) = k/n$. Вообичаено k се нарекува *број на поволни случаи*, а n е *број на сите можни случаи*. Токму ова е начинот на кој повеќето луѓе размислуваат за веројатноста.

Кога Ω е пребројиво, $\Omega = \{\omega_1, \omega_2, ..., \omega_n, ...\}$, $p(\omega_i)$ треба да бидат избрани така што редот $\sum_{i=1}^{n} p(\omega_i) = 1$ (да биде конвергентен, со сума 1). Сега, ако настанот А е составен од бесконечно елементарни настани $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}, \ldots\}$, неговата веројатност $p(A)$ се сведува на сума на конвергетен ред, т.е. $p(A) = \sum_{k=1}^{\infty} p(\omega_{i_k})$.

ПРИМЕР 2.5 Колкава е веројатноста при фрлање 2 коцки да се добие: а) збир 7; б) различни броеви; в) збир поголем од 7?

Решение

Едноставно ги броиме поволните случаи и тој број го делиме со бројот на сите можни случаи, т.е. со бројот на елементи во Ω.

$\Omega = \{(x, y) \mid x, y = 1,2,...,6\}, \ n = |\Omega| = 36$

а) $p(\text{"збир 7"}) = \dfrac{6}{36} = 0.1667;$

б) $p(\text{"различни броеви"}) = \dfrac{30}{36} = 0.8333;$

в) $p(\text{"збир поголем од 7"}) = \dfrac{9}{36} = 0.25.$ ■

ПРИМЕР 2.6 (Chevalier de Méré). Колкава е веројатноста да се добие 6-ка од 4 фрлања на коцка? Колкава е веројатноста да се добијат 2-е 6-ки од 24 фрлања на 2 коцки?

Решение

Проблеми со "барем", "најмалку", "најмногу", "не помалку", "не повеќе" често пати позгодно се решаваат преку обратните настани. Во овој случај имаме

$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \mid \omega_i = 1,2,...,6\}, \ n = |\Omega| = 6^4$ и сега ако

A = "барем 1-на 6-ка од 4 фрлања" тогаш

$\bar{A}$ = "нема 6-ка од 4 фрлања"

и користејќи варијации со повторување добиваме

$$p(A) = 1 - p(\bar{A}) = 1 - \frac{5^4}{6^4} = 0.5177$$

За вториот проблем, на сосема идентичен начин добиваме

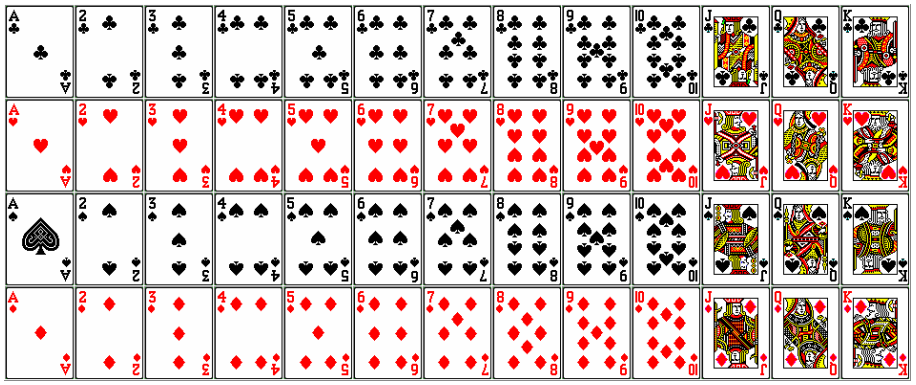
$\Omega = \{((x_1, y_1), (x_2, y_2), ..., (x_{24}, y_{24})) \mid x_i, y_i = 1,2,...,6\}, \ n = |\Omega| = 36^{24}$ и

A = "барем еднаш 2-е 6-ки од 24 фрлања"

$\bar{A}$ = "нема 2-е 6-ки од 24 фрлања"

$$p(A) = 1 - p(\bar{A}) = 1 - \frac{35^{24}}{36^{24}} = 0.4914.$$ ■

ПРИМЕР 2.7 Од шпил се влечат 3 карти. Колкава е веројатноста да се добие: а) еден ас; б) барем една треф карта; в) барем 2 црвени карти?



Решение

Бидејќи редоследот на извлечените карти не е важен туку само содржината, за пресметка на поволните и можните случаи користиме комбинации.

$$\Omega = \{(x, y, z) \mid x, y, z = 1, 2, ..., 52 \text{ и } x \neq y \neq z\},\ n = |\Omega| = \binom{52}{3}$$

$$\text{а) } p(\text{"еден ас"}) = \frac{\binom{4}{1}\binom{48}{2}}{\binom{52}{3}} = 0.2042;$$

$$\text{б) } p(\text{"барем една треф"}) = 1 - p(\text{"ниедна треф"}) = 1 - \frac{\binom{13}{0}\binom{39}{3}}{\binom{52}{3}} = 0.5865;$$

$$\text{в) } p(\text{"барем 2 црвени"}) = p(\text{"2 црвени"}) + p(\text{"3 црвени"}) =$$

$$= \frac{\binom{26}{2}\binom{26}{1} + \binom{26}{3}\binom{26}{0}}{\binom{52}{3}} = 0.4246. \blacksquare$$

ПРИМЕР 2.8 На полица на случаен начин се поставени 40 книги меѓу кои и еден 4-томен роман. Колкава е веројатноста томовите да се во добар редослед, одлево-надесно (не мора соседни)?

Решение

$\Omega = \{\text{сие пермурации од 40 книги}\},\ n = |\Omega| = 40!$

За поволните случаи да забележиме дека 4-те тома може во правилен редослед да се стават меѓу другите 36 книги колку што има избори на 4 еле-

менти од 40 (тоа се комбинации). За секој таков избор другите 36 книги може произволно да се пермутираат. Оттука добиваме

$$p(\text{"томовите се во добар редослед"}) = \frac{\binom{40}{4}\cdot 36!}{40!} = \frac{1}{24} = 0.0417. \blacksquare$$

ПРИМЕР 2.9 Се запишува случаен троцифрен број. Колкава е веројатноста да биде делив барем со еден од броевите 2 и 5.

Решение

$\Omega = \{\text{целите броеви од 100 до 999}\}$, $n = |\Omega| = 900$ и ако

A = "делив со 2", B = "делив со 5" треба да се најде $p(A+B)$

$$p(C) = p(A) + p(B) - p(A\cdot B) = \frac{450}{900} + \frac{180}{900} - \frac{90}{900} = 0.6$$

или со обратниот настан $\overline{A+B}$ = "не е делив со 2 и со 5"

$$p(A+B) = 1 - p(\overline{A+B}) = 1 - (\frac{450}{900} - \frac{90}{900}) = 1 - 0.4 = 0.6. \blacksquare$$

ПРИМЕР 2.10 Изгубен е 4-цифрениот код (pin) на кредитна картичка. Колкава е веројатноста: а) сите цифри да се различни; б) да има 3 различни цифри; в) да има 2 различни цифри; г) сите цифри да се исти?

Решение

Бидејќи редоследот на цифрите е важен, работиме со варијации.

$\Omega = \{(\omega_1, \omega_2, \omega_3, \omega_4) \mid \omega_i = 0,1,2,...,9\}$, $n = |\Omega| = 10^4$

а) $p(\text{"сите цифри се различни"}) = \frac{10\cdot 9\cdot 8\cdot 7}{10^4} = 0.5040;$

б) Имаме $120 = \binom{10}{3}$ различни тројки различни цифри и за секоја тројка имаме 12 пермутации со 2 повторувања на првата, 12 пермутации со 2 повторувања на втората и 12 пермутации со 2 повторувања на третата цифра. Тоа дава

$$p(\text{"има 2 различни цифри"}) = \frac{(12+12+12)120}{10^4} = 0.4320;$$

в) Имаме 45 различни парови различни цифри и за секој пар имаме 4 пермутации со 3 повторувања на првата, 4 пермутации со 3 повторувања на втората и 6 пермутации со 2 повторувања и на првата и на втората цифра. Тоа дава

$$p(\text{"има 3 различни цифри"}) = \frac{(4+4+6)45}{10^4} = 0.0630;$$

г) $p(\text{"сите цифри се исти"}) = \frac{10}{10^4} = 0.0010.$

Се разбира, за контрола проверуваме дека збирот на сите 4 веројатности дава 1 (тоа е сигурен настан). ■

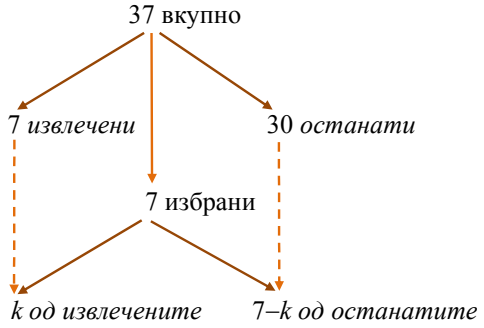
ПРИМЕР 2.11 Се извлекуваат 7 од броевите 1, 2, ..., 37 без враќање (лотарија 7 од 37). Колкава е веројатноста со избрани 7 броја да се добијат k погодоци (k = 0,1, ..., 7)?

Решение

Бидејќи редоследот на броевите не е важен, работиме со комбинации.

$$\Omega = \{(\omega_1, \omega_2, ..., \omega_7) \mid \omega_i = 0, 1, 2, ..., 37\},\ n = |\Omega| = \binom{37}{7} = 10295472$$

Поволните случаи ги пресметуваме со помош на шемата



k	$p(\text{"}k\text{ погодоци"})$
0	0.197737413
1	0.403713885
2	0.290673997
3	0.093164743
4	0.013802184
5	0.000887283
6	0.000020397
7	0.000000097

добиваме $p(\text{"}k\text{ погодоци"}) = \dfrac{\binom{7}{k}\binom{30}{7-k}}{\binom{37}{7}}$. ■

ПРИМЕР 2.12 Колкава е веројатноста во група од k луѓе барем двајца да имаат ист роденден?

Решение

Како и во други случаи, проблем со "барем" позгодно се решава преку обратниот настан. Нека

$\Omega = \{(\omega_1, \omega_2, ..., \omega_k) \mid \omega_i = 1,2,...,356\},\ n = |\Omega| = 356^k$ и сега ако

A = "барем 2-ца имаат ист роденден" тогаш

$\bar{A}$ = "сите имаат различен роденден"

и со помош на варијациите добиваме

$$p(A) = 1 - p(\bar{A}) = 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - k + 1)}{365^k}.$$

Во следната табела се дадени решнијата за некои k:

k	$p(A)$
10	0.1169
20	0.4115
23	0.5073
30	0.7064
40	0.8912
50	0.9704

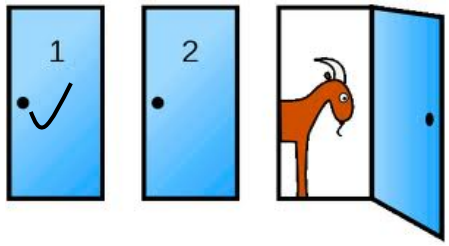
Да забележиме дека ова е еден од проблемите што не е во согласност со нашата интуиција бидејќи доволни се само 23 луѓе за веројатноста да имаат ист роденден биде преку 50%. ■

Интуитивната проверка на резултатот при решавање на веројатносни проблеми може да биде важен чекор за контрола на исправноста на решението. Ако резултатот не се совпаѓа со нашата интуиција, логично е да се направи проверка на решението. Следниот пример е славен по својата репутација на конфликт со интуицијата.

ПРИМЕР 2.13 (Monty Hall) Овој проблем потекнува од американското телевизиско шоу "Let's Make a Deal". Имаме 3 врати зад кои се наоѓаат 2 кози и автомобил (добивката). Играчот одбира врата (не ја отвора), а водителот, кој знае каде е автомобилот, од останатите две ја отвора таа со козата (ако зад двете е коза, одбира која било). Играчот има право да остане на својот избор или да ја избере другата неотворена врата. Што е подобро за него?

Решение

На следната слика е прикажан случајот кога играчот ја избрал првата врата, а водителот ја отворил третата знаејќи дека зад неа има коза. Играчот може да остане на изборот на првата врата или да промени, и да ја избере втората.



Претпоставуваме дека почетниот избор на играчот е случаен, и дека изборот на водителот од двете останати врати кога зад двете има коза е случаен.

Кога играчот на почеток избира врата, веројатноста да е таму автомобилот е 1/3. Ако тој и понатаму остане на тој избор, останува и веројатноста 1/3. Ако тој го промени изборот, се покажува дека веројатноста расте на 2/3 бидеј-

ќи ако автомобилот е во другите 2 врати, со промена на изборот играчот сигурно го добива (се користи знаењето на водителот).

Во следната табела се дадени сите можни случаи зад вратите и резултатите кога играчот останува или го менува изборот:

Врата 1	Врата 2	Врата 3	Резултат при промена на изборот	Резултат при останување на изборот
Автомобил	Коза	Коза	Коза	Автомобил
Коза	Автомобил	Коза	Автомобил	Коза
Коза	Коза	Автомобил	Автомобил	Коза

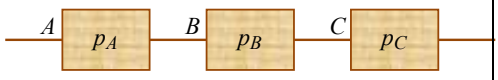
Значи веројатноста на добивка при промена на изборот на вратата е двојно поголема од онаа при останување на почетниот избор.

Наједноставното расудување сугерира дека промената на изборот на вратата губи акко играчот на почетокот ја избрал вратата со автомобил, што се случува со веројатност 1/3. Така, промената на изборот мора да добива со веројатност 2/3. Проблемот е во тоа што по почетниот избор, играчот не прави нов случаен избор туку останува на почетниот или го менува. Ако играчот по отварањето на вратата со козата од страна на водителот, наместо да остане на почетниот избор, на случаен начин би бирал една од двете затворени врати, веројатноста да го добие автомобилот би пораснала на 1/2.

Кога овој проблем за прв пат се покажува на луѓето, големо мнозинство од нив смета дека останатите две врати имаат еднаква шанси на добивка, па промената на одлуката е неважна. Во една студија на 228 испитаници, само 13% од нив се одлучиле на промена на изборот на вратата. Во некои дискусии на когнитивните психолози околу овој проблем може да се најде информација дека дури и физичарите-Нобеловци систематски давале погрешен одговор, инсистирајќи на него. Интересно е да се забележи дека гулабите, кои добро учат од грешките, при симулациите на овој проблем брзо научуваат да го променат изборот (за разлика од луѓето).

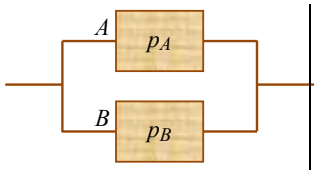
Комплетно математички подржано решение на овој проблем, користејќи ги техниките на условната веројатност (Баесовата формула) е дадено во глава 4 (пример 4.2). ■

ПРИМЕР 2.14 Во инжинерските системи, компонентите (подсистемите) може да се поврзуваат сериски, но за поголема сигурност и паралелно. Секоја компонента има веројатност дека ќе работи исправно. Системот работи исправно кога која било серија од компоненти работи исправно. На пример, подолу е пресметана надежноста на еден сериски и паралелен систем S со дадени надежности на компонентите p_A, p_B и p_C.



Сериски систем работи исправно ако секоја компонента е исправна,

$$p(S) = p_A \cdot p_B \cdot p_C.$$



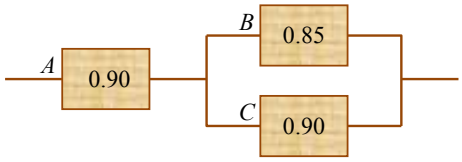
Паралелен систем работи исправно ако барем една од компонентите е исправна,

$$p(S) = p_A + p_B - p_A \cdot p_B.$$

Користејќи обратни настани го имаме истото

$$p(S) = 1 - (1-p_A)(1-p_B).$$

Пресметај ја надежноста на системот



Решение

Ако со A, B и C ги означиме настаните на исправна работа на трите компоненти, системот работи исправно кога се случува настанот

A(B + C) , A е исравно и B или C е исправно. Оттука

$p(\mathrm{A}(\mathrm{B}+\mathrm{C})) = p(\mathrm{AB}+\mathrm{AC}) = p(\mathrm{AB}) + p(\mathrm{AC}) - p(\mathrm{ABAC}) =$

$p(\mathrm{A})p(\mathrm{B}) + p(\mathrm{A})p(\mathrm{C}) - p(\mathrm{A})p(\mathrm{B})p(\mathrm{C}) =$

$0.90 \cdot 0.85 + 0.90 \cdot 0.90 - 0.90 \cdot 0.85 \cdot 0.90 = 0.8865.$

Паралелните компоненти често пати позгодно се работат со обратните настани. Со таков пристап системот работи исправно кога се случува

$\mathrm{A}(\overline{\overline{\mathrm{B}} \cdot \overline{\mathrm{C}}})$, A е исправно и не се неисправни и B и C. Оттука

$p(\mathrm{A}(\overline{\overline{\mathrm{B}} \cdot \overline{\mathrm{C}}})) = p(\mathrm{A})p(\overline{\overline{\mathrm{B}} \cdot \overline{\mathrm{C}}}) = p(\mathrm{A})(1 - p(\overline{\mathrm{B}})p(\overline{\mathrm{C}})) =$

$0.90(1 - (1 - 0.9)(1 - 0.85)) = 0.8865.$ ■

2.4. Геометриска веројатност

Дефинирањето на веројатноста во недискретен случај, кога Ω е бесконечно и непребројиво води кон одредени проблеми. На пример, нека $\Omega = \{x \mid x \in [0, 1]\}$ и нека го повторуваме 10^6 пати експериментот на случаен избор на број од Ω. Сега еден елементарен настан, да речеме појавување на бројот 0.82025342341725, може да се случи еднаш или ниеднаш (поверојатно) во овие 10^6 повторувања на експериментот. Која веројатност да му доделиме на овој настан?

$$p(0.82025342341725) = 10^{-6} \quad \text{или} \quad p(0.82025342341725) = 0.$$

Многу напори се вложени во обидите веројатноста да се дефинира "логично", како граница на релативната честота на случување на настаните

$$p(\mathrm{A}) = \lim_{n\to\infty} \mathrm{W}(\mathrm{A}) = \lim_{n\to\infty} \frac{k}{n}.$$

иако физички секој експеримент се повторува само конечен број пати (n е секогаш конечно). За жал, оваа граница не постои во математичка смисла. Интуитивно е јасно дека за непребројиви множества оваа граница кога би конвергирала, би конвергирала кон 0. Оттука, единствен модел кој би можел да функционира во недискретен случај е на секој елементарен настан да му се додели веројатност 0. Од друга страна, при повторување на вакви експерименти се случуваат елементарни настани па се поставува прашањето, каде е веројатноста?

Решението на оваа ситуација е (ненулта) веројатност да се доделува на области (на пример, на реалната оска на интервали), а не на поединечни настани. Но ако еден интервал има ненулта веројатност, а секој број (елементарен настан) во него има веројатност 0, дали е тоа во спротивност со 3-тата аксиома на веројатност? Тука треба да се има предвид дека аксиомата 3 кажува дека веројатноста на *пребројива* сума од дисјунктни настани е еднаква со сумата на нивните поединечни веројатности. Аксиомата 3 ништо не тврди за непребројива сума на дисјунктни настани (како што се реалните броеви во интервалите). Генерално, во математиката, непребројивоста води кон логика често пати спротивна на интуицијата. Конкретно, во нашиот случај, некој би можел да заклучи дека збир на непребројиво многу 0-ли не дава 0, и тоа не е погрешно. Уште еднаш да напоменеме дека за непребројивите множества елементарни настани, дефинирање на конзистентна мера (како што е веројатноста) на сите подмножества не е возможно. Токму затоа се оди на редукција на подмножествата на кои се дефинира веројатноста, а тоа се σ-алгебрите.

Горенаведените проблеми од теоријата на веројатност имаат свој корен во старите парадокси како што се: Ахил и желката, истрелана стрела и многу други. Имено сите тие сугерираат дека реалните брови се измислен, т.е. нереален концепт. На пример, парадоксот на Ахил и желката е во следното: ако пред почеток на трката Ахил ѝ даде предност на желката, тој веќе никогаш нема да ја стигне. Имено, кога тој ќе стигне до позицијата на желката таа ќе биде малку понапред, па кога тој ќе дојде до следната нејзина позиција таа повторно ќе биде малку

понапред, и така до бескрајност тој ќе ѝ се приближува, но никогаш нема да ја стигне.

Се разбира во реалноста тоа не се случува и Ахил брзо ја престигнува желката. Едно логично објаснување е дека намалување до бескрајност на простор-временското растојание не е можно, туку движењето се одвива по некој минимален интервал (веројатно Планков = $1.616252(81)\times10^{-35}$ метри) во кој Ахил поминува поголем пат од желката и ја престигнува. Тоа значи дека позициите на Ахил и желката треба да се одредат преку поделбата на растојанието на минимални интервали и потоа да се пресмета кој од нив за одредено време колку интервали поминал. Така парадоксот исчезнува.

Ова сугерира дека и многу малите рационални броеви се исто така сомнителен концепт поради кој се јавуваат истите парадокси. Сомнителниот концепт за бескрајност (чија спротивност се бескрајно малите броеви) е основен концепт во модерната математика и затоа таа може да се смета како релативно добра (или не толку добра, но неверојатно корисна) апроксимација на реалноста. Можно е реалноста да се потчинува на логиката на дискретни големини, но така прилагоден (дискретен) математички апарат споредлив со оној на модерната математика базирана на бесконечности едноставно не постои.

Сега ќе ја воведеме геометриската веројатност што е пандан на класичниот простор на веројатност, во непрекинат простор на веројатност. Нека Ω е бесконечно и непребројиво и претставува некој геометриски објект: 1-димензионален ако е на реалната оска, 2-димензионален во рамнината, 3-димензионален во просторот, или општо, n-димензионален во просторот $\mathbb{R}^n$. Нека $S\subseteq\Omega$ означува случаен настан.

Дефиниција 2.4 Геометриска веројатност на S е бројот $p(S)$ даден со

$$p(S)=\frac{m(S)}{m(\Omega)},$$ каде што $m(\cdot)$ е мера[1] на множеството.

[1] Прецизно зборувајќи, мера не може да се дефинира на секое множество. Во математиката се прави рестрикција на множествата за кои мерата се дефинира, со цел таа да задоволи некои "логични" правила.

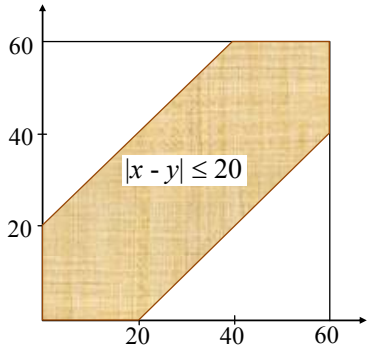
Типична мера на множеството претставува: *должина на интервал* на реалната оска, *плоштина* во рамнината, *волумен* во просторот, или општо, *n-димензионален волумен* во просторот $\mathbb{R}^n$.

ПРИМЕР 2.15 Двајца пријатели закажале средба меѓу 10 и 11 часот со договор секој да го чека другиот најмногу 20 минути. Колкава е веројатноста дека ќе се сретнат ако времето на доаѓање е случајно?

Решение

Ако со x, y ги означуваме времињата на пристигнувањето, за Ω добиваме

$\Omega = \{(x, y) \mid x, y \in [0, 60]\}$, $m(\Omega) = 3600$



$S = \{(x, y) \mid x, y \in [0, 60], |x - y| \le 20\}$

Плоштината на S ја добиваме едноставно од сликата

$m(S) = 3600 - 2 \cdot 40 \cdot 40/2 = 2000.$

Така добиваме

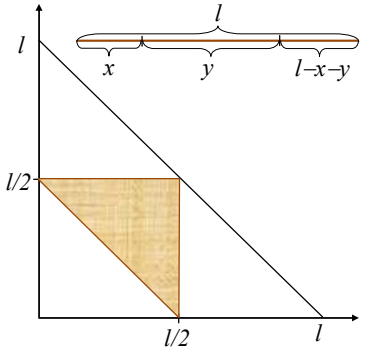
$$p(\text{"ќе се сретнат"}) = \frac{2000}{3600} = 0.5556. \blacksquare$$

ПРИМЕР 2.16 Стап се крши на две случајни места. Колкава е веројатноста дека од деловите може да се направи триаголник?

Решение

Ако со x, y ги означуваме должините на првите два дела, добиваме

$\Omega = \{(x, y) \mid x \ge 0, y \ge 0, x+y \le l\}$, $m(\Omega) = l^2/2$



S ќе го добиеме од условите за постоење на триголник:

$x \le y + (l - x - y) \Rightarrow x \le l/2$
$y \le x + (l - x - y) \Rightarrow y \le l/2$
$(l - x - y) \le x + y \Rightarrow x + y \le l/2$ што дава

$S = \{(x, y) \mid x \le l/2, y \le l/2, x+y \le l/2\}$.

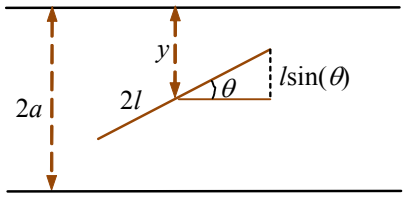
$$p(\text{"може да се формира триаголник"}) = \frac{(l/2)(l/2)/2}{l^2/2} = 0.25. \blacksquare$$

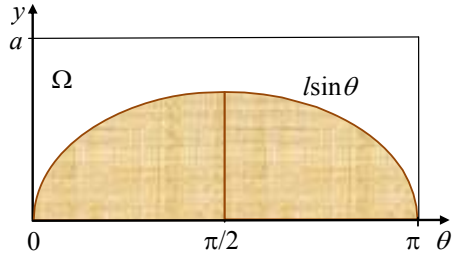
ПРИМЕР 2.17 (Проблем на Буфон (Georges-Louis Leclerc, Comte de Buffon 1707–1788)) Игла со должина $2l$ се фрла на површина со исцртани паралелни прави линии на растојание $2a$ ($l < a$). Да се најде веројатноста иглата да пресече некоја линија.

Решение

Да го означиме со y растојанието од средината на иглата до поблиската линија, а со θ аголот меѓу иглата и паралелните линии (види слика). Тогаш просторот на "елементарни настани" е

$$\Omega = \{(\theta, y) \mid \theta \in [0, \pi], y \in [0, a]\}.$$





Поволните случаи се добиваат кога иглата ја сече линијата, т.е. како што се гледа од сликата

$$S = \{(\theta, y) \mid y \le l\sin\theta\}.$$

$$p(\text{"иглата пресекува линија"}) = \frac{\int_0^{\pi} l\sin\theta d\theta}{a\pi} = \frac{2l}{a\pi}.$$

Користејќи го овој резултат, во периодот 1850-1925 година повеќе истражувачи се обидувале приближно да го пресметаат бројот π. Повеќе информации за овие обиди се дадени во додатокот Б. ■

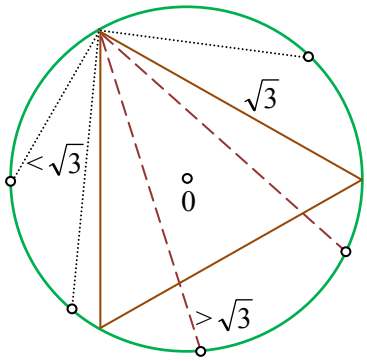
ПРИМЕР 2.18 (Бертранов парадокс (J.L.F. Bertrand, 1822-1900)). Разгледуваме рамностран триаголник впишан во кружница со радиус 1. Случајно избираме тетива на кружницата. Колкава е веројатноста таа да е подолга од страната на триаголникот?

Решение

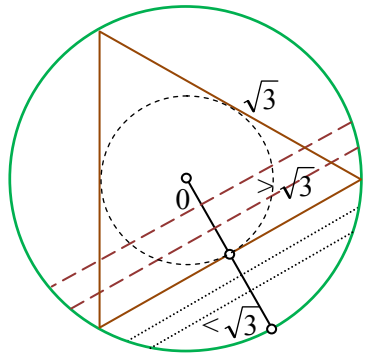
Во литературата често се нагласува дека ова е пример што ја илустрира недоволната прецизност на формулацијата "на случаен начин се бира ...", и дека со нејзино различно толкување може да се добијат различни решенија на проблемот.

Страните на триаголникот имаат должина $\sqrt{3}$. Предложени се три логични решенија.

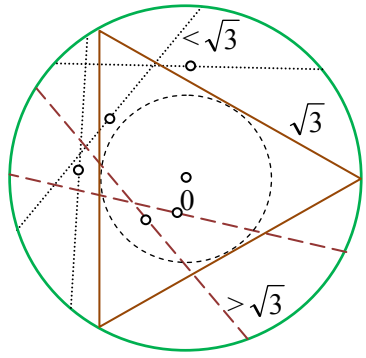
1) Метод со "случајни крајни точки". Случајно избери две точки на кружницата и поврзи ги со тетива. За пресметка на бараната веројатност замисли го впишаниот триаголник чиешто едно теме коинцидира со едната од крајните точки на тетивата (види слика десно). Тетивата ќе биде подолга од страната на триаголникот акко нејзината втора крајна точка лежи на лакот од страната на триаголникот спротивна на првото теме. Оттука следува дека веројатноста случајно избрана тетива да биде подолга од страната на впишаниот рамностран триаголник е 1/3.



2) Метод на "случаен радиус". Случајно избери радиус на кружницата и случајно избери точка на радиусот. Потоа конструирај тетива нормална на радиусот. Тетивата е подолга од страната на впишаниот триаголник акко избраната точка на радиусот е поблиску до центарот на кружницата отколку до точката каде што радиусот ја сече кружницата. Со ротирање на радиусот се добиваат сите тетиви, а бидејќи страната на триаголникот го сече радиусот на пола, веројатноста дека случајно избрана тетива ќе биде подолга од страната на впишаниот рамностран триаголник е 1/2.



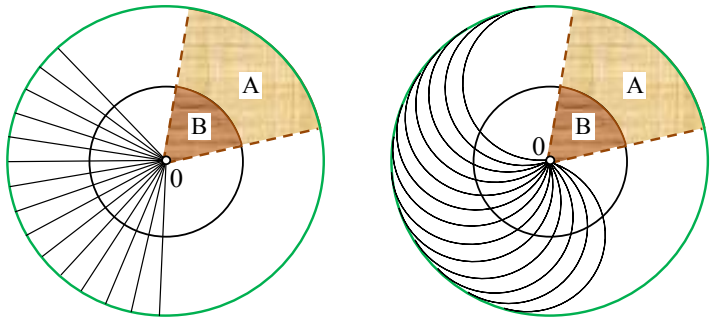
2) Метод на "случајна средна точка". Случајно избери точка во кружницата и конструирај тетива со средина во избраната точка. Тетивата ќе биде подолга од страната на триаголникот акко избраната точка е во внатрешноста на концентричната кружница со радиус 1/2. Плоштината на помалата кружница е 4 пати помала од плоштината на големата. Оттука, веројатноста дека случајно избрана тетива ќе биде подолга од страната на впишаниот рамностран триаголник е 1/4.



По добивањето на 3 различни решенија некој би можел да ја заклучи дискусијата со тврдењето дека тие се различни бидејќи случајниот избор на тетивата се прави на различен начин: во 1) преку две точки на кружницата, во 2) преку точка на радиусот и во 3) преку точка во внатрешноста на кружницата. Но кој е правилниот начин случајно да се избере тетива и дали мерите

на соодветните множества се точно пресметани? Проблемот може да се согледува од повеќе аспекти, од тоа како секоја тетива да има исти шанси да биде избрана (рамномерност) па до "судир" меѓу кардиналноста на множествата (топологија) и мерата на множествата (геометријата).

Проблемот со решението 2) може да се согледа во воспоставената биекција (којашто е коректна) меѓу сите тетиви и точките од сите радиуси. Според топологијата, воспоставената биекција повлекува иста кардиналност (број на елементи) и понатаму геометријата само мери дека половината на радиусот има должина 1/2 колку што е бараната веројатност. Но иста кардиналност (постоење биекција) не значи и иста мера. На пример, интервалот (0, 1) има иста кардиналност со кој било друг интервал (на пример (0, 8)), со целата реална оска $\mathbb{R}$, а дури и со $\mathbb{R}^n$, но геометријата тврди дека мерите на овие множества се различни. Оттука воспоставената биекциија не гарантира и иста мера, т.е. интуитивно, мерата на множеството средини на тетивите не мора да соодвествува (е поголема) на мерата на множествата точки од сите радиуси. Една дводимензионална област (кружница) не може да се покрие со еднодимензионална област (радиус). За тоа се потребни подобласти, како на пример оние меѓу два блиски радиуси (така интегрираме). Додатно, кумулативната мера на половините од радиусите поблиски до кружницата (област A) е поголема од кумулативната мера на половините поблиски до центарот (област B), иако и тука имаме биекција. Оттука, ако се земе предвид дека веројатноста се базира на мера, а не на кардиналност, "бројот на тетиви" со средини во делот на радиусите од областа A има поголема мера. Ова би требало да даде решение на Бертрановиот проблем со веројатност < 1/2 (види слика).



Што ако наместо радиуси, разгледуваме полукружници во чиишто точки се средини на тетивите? Следејќи ја логиката на решението 2) би добиле дека бараната веројатност е 1/3.

Решението 3) се базира на мери на множествата и нема проблеми како решението 2). Најсериозен проблем што тука може да се согледа е дека сите тетиви (бесконечно многу) со средна точка во центарот на кружницата се претставуваат со само една точка. Овие непребројиво многу тетиви низ центарот ја менуваат мерата (ја зголемуваат) на множеството тетиви подолги од $\sqrt{3}$. От-

тука би можеле да заклучиме дека бројот на тетиви со должина поголема од $\sqrt{3}$ би требало да е поголем отколку што ова решение сугерира, т.е. веројатноста би требала да биде >1/4. Генерално, претставувањето на тетивите преку нивните централни точки има незгода во тоа што центарот на кружницата определува многу тетиви за разлика од другите точки за кои таа е единствена.

Останува решението 1), коешто комплетно се базира на геометрија (мера) и кај кое нема "нерегуларни" трансформации ниту пак бесконечно многу тетиви што се "игнорираат" како во решенијата 2) и 3). Бараната веројатност кај ова решение е 1/3, што изгледа логично од аспект дека е меѓу 1/4 и 1/2. Но дали бирањето тетиви преку крајни точки на кружницата е доволно рамномерно? За геометриската веројатност е неопходна еднаковеројатност на "елементарните" настани.

Некој може и праволиниски да размислува дека не се важни позициите на тетивите туку само нивната должина. Бидејќи максималната должина на тетива во кружницата е 2, имаме дека Ω = [0, 2], а S = [0, 2–$\sqrt{3}$] и тогаш $p(\mathrm{S}) = \dfrac{m(\mathrm{S})}{m(\Omega)} = 1 - \dfrac{\sqrt{3}}{2}$. Се разбира, ова решение го игнорира постоењето на сите геометриски елементи на проблемот.

Некои автори, Бертрановиот парадокс го сведуваат на проблем на избор на различни координатни системи.

На решението 1) одговара дефинирање на тетива преку две точки од кружницата (1, α) и (1, β) во поларни координати (во овој случај тие се позгодни од декартовите). Должината L на тетивата во овој случај е

$$L = \sqrt{2 - 2\cos(\alpha + \beta)}\,.$$

На решението 2) одговара дефинирање на тетива преку нејзината средна точка со поларни координати (ρ, φ) на даден радиус на кружницата, па за должината на тетивата веднаш се добива

$$L = 2\sqrt{1 - \rho^2}\,.$$

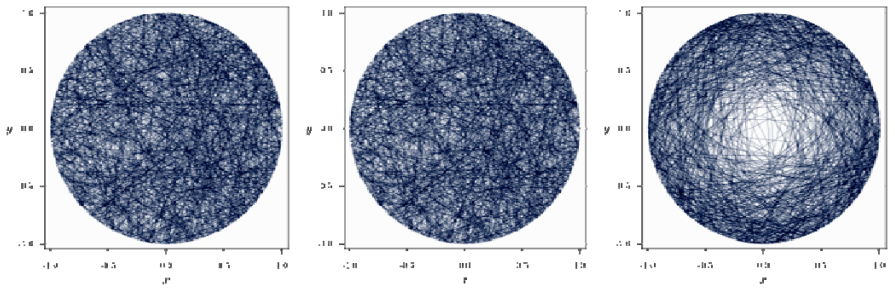
На решението 3) одговара дефинирање на тетивата преку нејзината средна точка со декартови координати (x, y) во внатрешнота на кружницата, и тогаш должината на тетивата е

$$L = 2\sqrt{1 - (x^2 + y^2)}\,.$$

Забележи дека левата страна од последната формула е всушност функција $L(x(\rho, \varphi)\,, y(\rho, \varphi))$, $x = \rho\cos\varphi$, $y = \rho\sin\varphi$, па со директна смена на x и y од десната страна на 3) се добива функција зависна само од аргументите ρ и φ, т.е од облик $L(\rho, \varphi)$ (решението 2)). Значи очигледно е дека двете решенија не се исти. Начинот на случаен избор на тетива во предложените 3 решенија во основа е еквивалентен со 3-те избори на координарните системи.

Ако би сакале изборот на точката и "коректното" задавање на координатите да не зависи од позицијата на кружницата во рамнината или нејзиниот радиус (види задача 23), решението 2 е единствено што ги задоволува овие барања [Jaynes 1973].

На следната слика се прикажани цртежи на случајно избрани тетиви (со симулација) по трите методи.



Оваа слика оди во прилог на решението 2), коешто има најрамномерна распределбата на случајно избраните тетиви. Распределбата очигледно не е рамномерна кај решението 3), но не е и кај решението 1), иако тоа не се гледа јасно на сликата. Уште во 1912 година Поенкаре (Henry Poincare, 1854–1912) укажал дека има само 2 начина на случаен избор на тетива. Првиот е комбинација на фиксна кружница и случајна права и тој води до парадокси. Вториот начин е комбинација на фиксна права и случајна кружница и тој дава единствено решение. Имено, во тој случај, тетивата од пресечните точки на правата и кружницата е подолга од $\sqrt{3}$ акко растојанието меѓу правата и центарот на кружницата е помало од 1/2. Така се добива второто решение.

Од целата претходна дискусија може да се заклучи дека иако во литературата најчесто се преферира решението 2), сепак нема општо прифатен пристап за тоа како да се дојде до вистинското решение на Бертрановиот парадокс. Постојат многу форуми на интернет, на кои овој проблем се дискутира од страна на студенти и професори, и каде што може да се најдат најразлични аргументи за прифаќање или отфрлање на некои од предложените 3 решенија. Ние повторно ќе се навратиме на овој проблем во додатокот Б.2, користејќи веројатносни симулации. ■

ЗАДАЧИ

1. Да се провери кои од следните равенства со настани се точни:

а) $\overline{A}+\overline{B}=\overline{A\cdot B}$; б) $(A+B)-B=A-A\cdot B$; в) $(A+B)-C=A+(B-C)$;

г) $A\cdot\overline{B}\cdot C\subset A+B$; д) $A\cdot B\cdot C=A\cdot B(B+C)$; ѓ) $(\overline{A+B})C=\overline{A}\cdot\overline{B}\cdot\overline{C}$.

2. Нека A, B и C се три произволни настани. Да се најдат изразите со операции со сите три настани такви што:

   а) се случил само A;

   б) се случиле A и B, но не C;

   в) се случиле сите три настани;

   г) се случил најмалку еден од настаните;

   д) се случил еден и само еден од настаните;

   ѓ) не се случил ниеден од настаните;

   е) се случиле не повеќе од два настана.

3. Еден инженерски систем има две компоненти. Ги дефинираме следните настани

   A = "првата компонента е исправна" и
   B = "втората компонента е исправна".

   Опиши ги следните настани преку A, $\overline{A}$, B и $\overline{B}$:

   а) Најмалку една компонента е исправна;
   б) Една компонента е исправна, а една е дефектна.

4. Ако $p(A)=0.3$, $p(B)=0.2$ и $p(AB)=0.1$, најди ги следните веројатности:

   а) $p(A+B)$; б) $p(\overline{A}B)$; в) $p(\overline{A+B})$; г) $p(\overline{A}+B)$.

5. Производител на сијалички за автомобили ги тестира сијаличките на интензитет (на осветлување) и трајност. Во следната табела се дадени резултатите од 130 светилки

		Трајност:	
		задоволува	не задоволува
Интензитет:	задоволува	117	3
	не задоволува	8	2

   а) Оцени ја веројатноста дека една сијаличка ги задоволува двата теста;

   б) Оцени ја веројатноста дека една сијаличка задоволува во интензитет, но не по трајност;

   в) Оцени ја веројатноста дека една сијаличка не ги задоволи и двата теста.

6. Број на кредитна картичка има 16 цифри, но само околу 100 милиони броеви се валидни. Колкава е веројатноста случајно да се погоди валиден број?

7. Треба да се погоди број на регистерска табличка што се состои од 3 цифри следени со 2 букви (A-Z). Колкава е веројатноста таа да се погоди ако:

   а) нема никаква додатна информација,

   б) се знае дека има 8-ца и буква C,

   в) почнува со 4 и завршува со H.

8. Четири машини произведуваат исти делови. Првата машина е најмодерна, додека втората и третата произведуваат само 60% од првата, а најслабата 4-та машина само 30% од првата. Колкава е веројатноста случајно избран дел да е од втората или четвртата машина?

9. Страните на една коцка се нумерирани со броевите 1, 1, 3, 3, 3, 3, а на друга со броевите 2, 2, 2, 2, 4, 4. Да се пресмета веројатноста дека при фрлањето на двете коцки на првата ќе се појави помал број отколку на втората.

10. Во џеп имаме 9 парички и тоа, 4 од 50 денари, 2 од 20 денари и 3 од 10 денари. Вадиме од џепот 5 парички. Пресметај ја веројатноста вредноста на извадените да е поголема од вредноста на оние што останале во џепот.

11. На полици се наоѓаат расфрлани n парови кондури. Ако $2r$ кондури се одберат случајно ($2r < n$), пресметај ја веројатноста дека:

    а) меѓу нив ќе нема комплетен пар;

    б) меѓу нив ќе има точно 2 комплетни пара.

12. Паричка се фрла сé додека два пати последователно не се појави иста страна. Опиши го просторот на елементарни настани и пресметај ги веројатностите дека:

    а) експериментот ќе заврши до 6-тото фрлање;

    б) експериментот ќе заврши по парен број фрлања;

    в) експериментот ќе заврши по конечен број фрлања.

13. Познато е дека во една серија од 100 производи има 10 неисправни. Контролата на квалитетот за проверка случајно избрала 15 производи. Колкава е веројатноста дека меѓу нив има:

    а) точно еден неисправен;

    б) барем 2 неисправни.

14. Од 8 вработени на случаен начин се избираат 3-ца за учество во една комисија. Во наредниот период повторно случајно се избрани 4-ца од нив за друга комисија, а потоа и 5-мина од нив за трета комисија. Пресметај ја веројатноста дека секој од 8-те вработени е член на некоја комисија.

15. Нека A, B и C се три произволни настани. Најди ја веројатноста да се случи точно еден од овие настани.

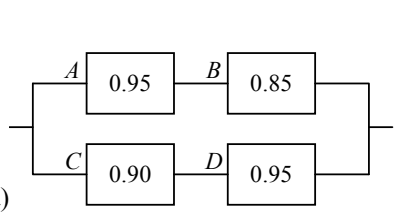
16. Двајца играат руски рулет со пиштол што има буренце со максимум 6 куршуми (по секое повлекување на чкрапалото буренцето што содржи еден куршум се завртува). Пресметај ја веројатноста оној што ја почнува играта да погине.

17. Некој ни кажал дека една непозната фамилија има 2 деца и дека едното е девојче. Се претпоставува дека се еднакви шансите за раѓање на машко и женско дете.

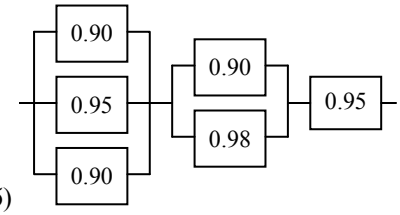
    а) Колкава е веројатноста дека другото дете е девојче?

    б) Се менува ли веројатноста под а) ако им позвониме на врата и ни отвори девојче?

18. Пресметај ја надежноста на следните системи



а)



б)

19. Градовите A, B, C и D се поврзани со следните патишта: $A \Leftrightarrow B$, $A \Leftrightarrow C$, $B \Leftrightarrow C$, $C \Leftrightarrow D$, $B \Leftrightarrow D$. По зимска ноќ снегот може да го блокира секој од патиштата независно, со веројатност p. Колкава е веројатноста дека утредента ќе може да патуваме од градот A до D?

20. Пресметај ја веројатноста во покер (пет поделени карти) да ги има сите 4 типа (каро, херц, треф и пик).

21. Два големи брода треба да пристигнат во едно пристаниште во непознато време, но во следните 24 часа. Двата не може истовремено да се укотват. Пресметај ја веројатноста дека едниот ќе мора да го чека другиот, ако првиот се очекува да стои во пристаништето 1 час, а вториот 2 часа.

22. Во квадрат со страна a се фрла точка. Пресметај ја веројатноста точката да е на растојание поголемо од $a/2$ од темињата на квадратот.

23. Друг начин на случаен избор на тетива од примерот 2.18 е да се избере произволна точка (x_0, y_0) (може и надвор од кружницата) и агол $\phi \in (-\pi/2, \pi/2)$. Тогаш правата што поминува низ (x_0, y_0) со коефициент на правец $k = \tan\phi$ е

$y - y_0 = k(x - x_0)$, со растојание до центарот на кружницата $d = \left| \frac{y_0 - kx_0}{\sqrt{k^2 + 1}} \right|$. Ако $d < 1$ правата ја сече кружницата и добиваме тетива. Колкава е веројатноста таа да е подолга од $\sqrt{3}$? Дали овој метод е независен од положбата на кружницата?

24. На шаховска табла (бесконечна) со квадрати со страна a се фрла паричка со радиус r, $2r < a$. Да се пресмета веројатноста дека:

    а) паричката комплетно ќе падне во еден квадрат,

    б) паричката ќе пресече не повеќе од една страна на квадратот.

25. Да се пресмета веројатноста дека две случајно избрани точки A и B на

    а) една кружница, заедно со центарот O формираат остроаголен триаголник (аголот AOB е остар);

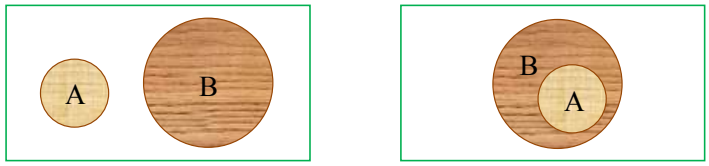
    б) една сфера, заедно со центарот O формираат остроаголен триаголник.

26. Нека V е случајно избрана точка на отсечката меѓу точките (0, 1) и (3, 4) во рамнината. Пресметај ја веројатноста плоштината на триаголникот со темињата (0, 0), (3, 0) и V да е поголема од 2.

3

Условна веројатност

Често се јавува проблем на пресметка на веројатност на случување на некој настан B, кога веќе се знае дека се случил некој друг настан A. Почетнички би било да се смета дека веројатноста на B останува каква што била, т.е. дека случувањето на A не влијае на случувањето на B. На пример, нека се дадени настани A и B како на сликата 3.1.



Слика 3.1 *Илустрација на влијание на еден настан на друг*

На левиот дијаграм од сл. 3.1 важи $A \cdot B = \emptyset$, па очигледно ако се случил A, тогаш B не може да се случи и во таква ситуација неговата веројатност треба да е 0. На десниот дијаграм е $A \subseteq B$, па очигледно ако се случил A, тогаш се случил и B и во таква ситуација неговата веројатност треба да е 1.

Веројатноста на случување на настан B, кога веќе се случил настанот A се нарекува условна веројатност, се означува со $p(B \mid A)$ и се дефинира на следниот начин.

Дефиниција 3.1 Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност. Условна веројатност $p(B \mid A)$ на настан $B \in \mathcal{F}$ при услов $A \in \mathcal{F}$ е дадена со

$$p(\mathrm{B} \mid \mathrm{A}) = \frac{p(\mathrm{A}\cdot \mathrm{B})}{p(\mathrm{A})}, \text{ за } p(\mathrm{A}) > 0.$$

За левиот дијаграм од сл. 3.1 според дефиницијата имаме $p(\mathrm{B} \mid \mathrm{A}) = p(\varnothing) / p(\mathrm{A}) = 0$, а за десниот дијаграм важи $p(\mathrm{B} \mid \mathrm{A}) = p(\mathrm{A}) / p(\mathrm{A}) = 1$.

Мотивацијата зад дефиниција 3.1 е во следното. Нека при n повторувања на некој експеримент, настанот А се случил точно m пати. Случувањата на А може да се сметаат за нов експеримент и нека во тие m случувања на А, настанот В се случил k пати. Тогаш релативната честота на случување на В во m-те експерименти во кои се случил настанот А е

$$\mathrm{W}(\mathrm{B} \mid \mathrm{A}) = \frac{k}{m} = \frac{k/n}{m/n} = \frac{\mathrm{W}(\mathrm{A}\cdot \mathrm{B})}{\mathrm{W}(\mathrm{A})}$$

Дека $p(\mathrm{B} \mid \mathrm{A})$ е добро дефинирано покажува и следната теорема.

Теорема 3.1 Ако $(\Omega, \mathcal{F}, p)$ е простор на веројатност и ако $\mathrm{A} \in \mathcal{F}$ е таков настан што $p(\mathrm{A}) > 0$, тогаш со $p_c(\mathrm{A}) = p(\mathrm{B} \mid \mathrm{A})$, $\mathrm{B} \in \mathcal{F}$ е дефинирана веројатност на Ω таква што $p_c(\mathrm{A}) = 1$.

Доказ: Треба да се покаже дека важат аксиомите на веројатност.

а1) $\forall \mathrm{B} \in \mathcal{F}$ имаме $p_c(\mathrm{B}) = p(\mathrm{A}\cdot\mathrm{B}) / p(\mathrm{A}) \geq 0$

а2) $p_c(\Omega) = p(\mathrm{A}\cdot\Omega) / p(\mathrm{A}) = p(\mathrm{A}) / p(\mathrm{A}) = 1$

а3) $p_c(\mathrm{B}_1+\mathrm{B}_2) = p(\mathrm{A}(\mathrm{B}_1+\mathrm{B}_2)) / p(\mathrm{A}) = p(\mathrm{A}\cdot\mathrm{B}_1+\mathrm{A}\cdot\mathrm{B}_2) / p(\mathrm{A}) = (*)$

Од $\mathrm{B}_1\cdot\mathrm{B}_2 = \varnothing$ следува $(\mathrm{A}\cdot\mathrm{B}_1)(\mathrm{A}\cdot\mathrm{B}_2) = \varnothing$, па имаме

$(*) = (p(\mathrm{A}\cdot\mathrm{B}_1) + p(\mathrm{A}\cdot\mathrm{B}_2)) / p(\mathrm{A}) = p_c(\mathrm{B}_1) + p_c(\mathrm{B}_2)$

На крај, $p_c(\mathrm{A}) = p(\mathrm{A}\cdot\mathrm{A})/p(\mathrm{A}) = p(\mathrm{A}) / p(\mathrm{A}) = 1$. ■

Од условната веројаност следува дека веројатноста може да се дефинира така што покрај сигурниот настан Ω, да има и други настани со веројатност 1. Аналогно, покрај $\varnothing$, може да има и други настани со веројатност 0.

Следните неколку особини на условната веројатност се сосема аналогни со оние на обичната веројатност:

1) $p(\overline{\mathrm{B}} \mid \mathrm{A}) = 1 - p(\mathrm{B} \mid \mathrm{A})$;

2) ако $\mathrm{B} \subseteq \mathrm{C}$, тогаш $p(\mathrm{B} \mid \mathrm{A}) \leq p(\mathrm{C} \mid \mathrm{A})$;

3) ако $\mathrm{B}\cdot\mathrm{C} = \varnothing$, тогаш $p((\mathrm{B} + \mathrm{C}) \mid \mathrm{A}) = p(\mathrm{B} \mid \mathrm{A}) + p(\mathrm{C} \mid \mathrm{A})$ и поопшто

$p((B + C) \mid A) = p(B \mid A) + p(C \mid A) - p(B \cdot C \mid A)$;

4) дури ако се A, B, C и D дисјунктни, во општ случај

$p(B + C \mid A + D) \neq p(B) + p(C \mid A) + p(D)$.

Условната веројатност е основната алатка што се користи во многу области на веројатносната анализа. Таа овозможува користење на секое додатно сознание при анализата на настаните во еден простор на веројатност.

ПРИМЕР 3.1 Веројатноста една светилка да откаже при првото вклучување е 2%. Ако не откаже при првото вклучување, веројатноста да трае една година е 82%. Колкава е веројатноста светилката да трае една година?

Решение

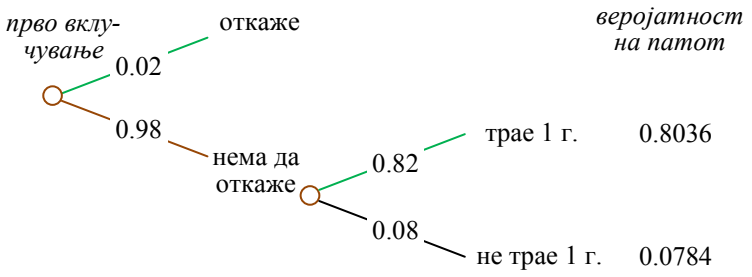
B = "светилката трае една година", P(B) = ?
A = "светилката откажува при прво вклучување", P(A) = 0.02
B | $\bar{A}$ = "светилката трае една година ако не откажала при прво вклучување", P(B | $\bar{A}$) = 0.82

Сега, со директна примена на формулата за условна веројатност добиваме

$p(B \mid \bar{A}) = p(B \cdot \bar{A}) / p(\bar{A})$ т.е. $p(B \cdot \bar{A}) = p(B \mid \bar{A})p(\bar{A}) = 0.82 \cdot 0.98 = 0.8036$

Ако се има предвид дека $B \subset \bar{A}$, имаме $B \cdot \bar{A} = B$, т.е. $p(B) = 0.8036$. ■

Корисна алатака за претставување на веројатностите преку постапно прикажување на несигурноста и ризиците е *дрвото на шанси*. Тоа на многу луѓе им дава појасна слика за условните веројатности преку декомпозиција на настаните на поедноставни случаи. Дрвото на шанси за примерот 3.1 би можело да изгледа вака



Од дрвото на шанси веднаш добиваме дека бараната веројатност се добива како производ на веројатностите по патот "нема да откаже" - "трае 1 г.".

Во случај кога во дрвото на шанси имаме повеќе патеки што одговараат на барањата од проблемот, веројатноста ја добиваме како *веројатност од поволната патека поделена со збир на веројатностите на овие патеки* (види пример 3.6). Дрвото на шанси понатаму ќе го користиме кај некои примери со цел да се добие подобар увид во решението.

ПРИМЕР 3.2 При пописот на население во Англија и Велс 1991 година, покрај другото добиени се и следните податоци:

Татковци со *темни* очи и деца со *темни* очи = 5%,
Татковци со *темни* очи и деца со *светли* очи = 7.9%,
Татковци со *светли* очи и деца со *темни* очи = 8.9%,
Татковци со *светли* очи и деца со *светли* очи = 78.2%.

Ако има статистичка стабилност на врската меѓу боите на очите на татковците и децата да се пресмета веројатноста при случаен избор на едно дете тоа да има *темни* очи ако таткото има *темни* очи; и веројатноста при случаен избор на едно дете тоа да има *темни* очи ако таткото има *светли* очи

Решение

Нека A = "таткото има темни очи" , а B = "детето има темни очи".
Треба да се пресметаат веројатностите $p(\mathrm{B} \mid \mathrm{A})$ и $p(\mathrm{B} \mid \bar{\mathrm{A}})$

Од пописот ги имаме следните податоци

$p(\mathrm{A}\cdot\mathrm{B}) = 0.05,\ p(\mathrm{A}\cdot\bar{\mathrm{B}}) = 0.079,\ p(\bar{\mathrm{A}}\cdot\mathrm{B}) = 0.089,\ p(\bar{\mathrm{A}}\cdot\bar{\mathrm{B}}) = 0.782.$

Сега со помош на овие веројатности добиваме

$$p(\mathrm{B} \mid \mathrm{A}) = \frac{p(\mathrm{A}\cdot\mathrm{B})}{p(\mathrm{A})} = \frac{p(\mathrm{A}\cdot\mathrm{B})}{p(\mathrm{A}\cdot\mathrm{B} + \mathrm{A}\cdot\bar{\mathrm{B}})} = \frac{0.05}{0.05 + 0.079} = 0.3876$$

$$p(\mathrm{B} \mid \bar{\mathrm{A}}) = \frac{p(\bar{\mathrm{A}}\cdot\mathrm{B})}{p(\bar{\mathrm{A}})} = \frac{p(\bar{\mathrm{A}}\cdot\mathrm{B})}{p(\bar{\mathrm{A}}\cdot\mathrm{B} + \bar{\mathrm{A}}\cdot\bar{\mathrm{B}})} = \frac{0.089}{0.089 + 0.782} = 0.1022. \ \blacksquare$$

Понекогаш е корисно да се знае верижното правило (chain rule) за условната веројатност. Имено, од дефиницијата на условната веројатност

$$p(\mathrm{B} \mid \mathrm{A}) = p(\mathrm{A}\cdot\mathrm{B}) / p(\mathrm{A}) \ \text{ следи дека } \ p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{B} \mid \mathrm{A})p(\mathrm{A})$$

и симетрично од

$$p(\mathrm{A} \mid \mathrm{B}) = p(\mathrm{A}\cdot\mathrm{B}) / p(\mathrm{B}) \ \text{ следи дека } \ p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A} \mid \mathrm{B})p(\mathrm{B}).$$

Ова може да се обопшти

$$p(\mathrm{A}\cdot\mathrm{B}\cdot\mathrm{C}\cdot\mathrm{D}\ \ldots) = p(\mathrm{A})p(\mathrm{B} \mid \mathrm{A})p(\mathrm{C} \mid \mathrm{A}\cdot\mathrm{B})p(\mathrm{D}|\mathrm{A}\cdot\mathrm{B}\cdot\mathrm{C})\ \ldots\ .$$

Производот на првите два члена е $p(AB)$, понатаму $p(C \mid AB)p(AB) = p(A \cdot B \cdot C)$, така што производот на првите три члена е $p(A \cdot B \cdot C)$, итн.

Верижното правило може да се примени за пресметка на условните веројатности со условување на сè со даден настан, на пример H, што дава

$$p(A \cdot B \cdot C \cdot D \ldots \mid H) = p(A \mid H)p(B \mid A \cdot H)p(C \mid A \cdot B \cdot H)p(D \mid A \cdot B \cdot C \cdot H) \ldots.$$

ПРИМЕР 3.3 Група од 15 туристи пристигнуваат во град со 4 хотели: A, B, C D, од кои во секој има место за сите 15. Туристите се сместуваат во хотелите случајно на следниот начин: во хотелот A остануваат случаен број од 15-те туристи, во хотелот B се сместуваат случаен број од останатите, во C случаен број од останатите и во D оние што останале. Пресметај ја веројатноста дека во секој од хотелите ќе се смести барем еден турист.

Решение

Со $(n_A, n_B, n_C\, n_D)$ го означуваме настанот во хотелот A да има n_A туристи, во хотелот B да има n_B туристи, во хотелот C да има n_C туристи и во D да има $n_D = 15 - n_A - n_B - n_C$ туристи. Веројатноста на ваквиот настан може да се добие од верижното правило

$$\frac{1}{16} \cdot \frac{1}{16 - n_A} \cdot \frac{1}{16 - n_A - n_B} \cdot 1 \cdot$$

За хотелот A секој број туристи од 0 до 15 има исти шанси да биде избран па за n_A туристи веројатноста е 1/16. За хотелот B секој број туристи од 0 до $15-n_A$ има исти шанси да биде избран па за n_B туристи веројатноста е $1/(16-n_A)$. Слично важи за n_C, додека за n_D веројатноста е 1 бидејќи остатокот од туристите мора да се смести во D.

Сега за веројатноста дека сите 3 хотели ќе примат туристи е

$$\sum_{n_A=1}^{12} \sum_{n_B=1}^{13-n_A} \sum_{n_C=1}^{14-n_A-n_B} \frac{1}{16} \cdot \frac{1}{16 - n_A} \cdot \frac{1}{16 - n_A - n_B} = 0.2856 \cdot$$

Да забележиме дека начинот на "случајното" распоредување на туристите по хотелите е од суштинско значење. На пример, ако би дозволиле секој турист случајно да избере хотел во кој ќе се смести, истата веројатност би била многу поголема (види задача 9). ■

3.1. Тотална веројатност

Формулата за тотална веројатност овозможува пресметка на неусловните веројатности преку условните.

За изведување на формулата тргнуваме од равенството

$B = B\cdot A + B\cdot \bar{A}$ од што следува $p(B) = p(B\cdot A) + p(B\cdot \bar{A})$.

Од друга страна, од дефиницијата за условна веројатност имаме дека

$p(A\cdot B) = p(B \mid A)p(A)$ и $p(\bar{A}\cdot B) = p(B \mid \bar{A})p(\bar{A})$.

Заменувајќи ги овие две равенства во равенството за $p(B)$, ја добиваме формулата за тотална веројатност

$$p(B) = p(B \mid A)p(A) + p(B \mid \bar{A})p(\bar{A}).$$

Се разбира, како и во другите формули за условна веројатност, и тука важи симетричната формула (се добива со замена на местата на A и B). Тоталната веројатност покажува дека веројатноста на некој настан B е тежински просек на веројатностите на B условени со A и $\bar{A}$ (за произволен настан A).

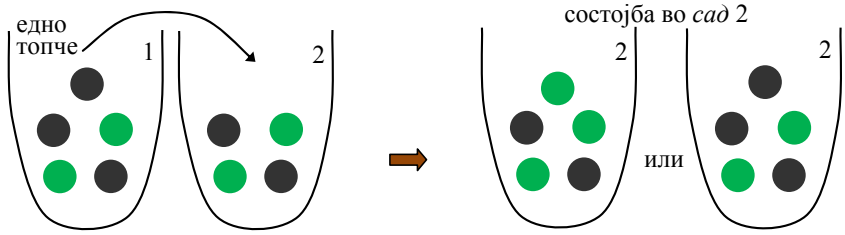
ПРИМЕР 3.4 Во еден сад имаме 2 зелени и 3 црни топчиња, а во друг 2 зелени и 2 црни топчиња. Случајно се одбира топче од првиот сад и без да се види се става во вториот. Потоа се извлекува топче од вториот сад. Колкава е веројатноста тоа да е црно?

Решение

B = "извлечено црно топче од сад 2", $p(B) = ?$

A = "префрленото топче е зелено"

$\bar{A}$ = "префрленото топче е црно"



$$p(B) = p(B \mid A)p(A) + p(B \mid \bar{A})p(\bar{A}) = \frac{2}{5}\frac{2}{5} + \frac{3}{5}\frac{3}{5} = 0.5200. \blacksquare$$

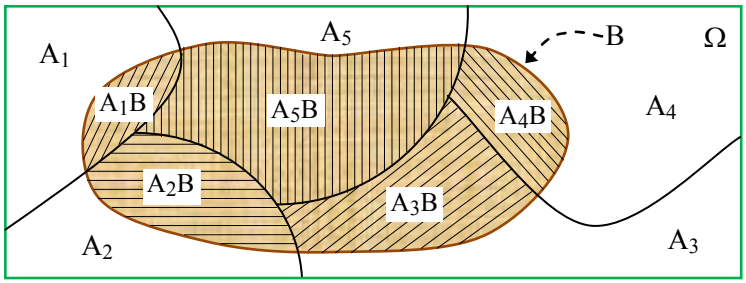
Обопштената формула за тотална веројатост ја даваме во облик на теорема.

Теорема 3.2 Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност и нека $A_1, A_2, \ldots, A_n$ е разбивање на Ω, што значи $A_1 + A_2 + \ldots + A_n = \Omega$ и $A_i\cdot A_j = \varnothing$. Тогаш за произволен настан $B \in \mathcal{F}$ важи

$$p(\mathrm{B}) = p(\mathrm{A}_1)p(\mathrm{B} \mid \mathrm{A}_1) + p(\mathrm{A}_2)p(\mathrm{B} \mid \mathrm{A}_2) + \ldots + p(\mathrm{A}_n)p(\mathrm{B} \mid \mathrm{A}_n).$$

Доказ: Од дисјунктност на A_i следи дисјунктност на $\mathrm{A}_i \cdot \mathrm{B}$. Понатаму, од $\Omega = \mathrm{A}_1 + \mathrm{A}_2 + \ldots + \mathrm{A}_n$ следува $\mathrm{B} = \mathrm{A}_1 \cdot \mathrm{B} + \mathrm{A}_2 \cdot \mathrm{B} + \ldots + \mathrm{A}_n \cdot \mathrm{B}$ што повлекува $p(\mathrm{B}) = p(\mathrm{A}_1 \cdot \mathrm{B}) + p(\mathrm{A}_2 \cdot \mathrm{B}) + \ldots + p(\mathrm{A}_n \cdot \mathrm{B})$. Со примена на условната веројатност $p(\mathrm{A}_k \cdot \mathrm{B}) = p(\mathrm{B} \mid \mathrm{A}_k)p(\mathrm{A}_k)$ на секој собирок се добива тоталната веројатност. ■

На сл. 3.2 е прикажан пример на разбивање на Ω.



Слика 3.2 *Илустрација на разбивање на Ω*

Ако $p(\mathrm{B} \mid \mathrm{A}_j)$ е најмалата веројатност од сите $p(\mathrm{B} \mid \mathrm{A}_k)$, тогаш заменувајќи ги $p(\mathrm{B} \mid \mathrm{A}_k)$ со $p(\mathrm{B} \mid \mathrm{A}_j)$ добиваме

$$p(\mathrm{B}) \le p(\mathrm{B} \mid \mathrm{A}_j)(p(\mathrm{A}_1) + p(\mathrm{A}_2) + \ldots + p(\mathrm{A}_n)) = p(\mathrm{B} \mid \mathrm{A}_j).$$

И аналогно, ако $p(\mathrm{B} \mid \mathrm{A}_i)$ е најголемата веројатност од сите $p(\mathrm{B} \mid \mathrm{A}_k)$, тогаш заменувајќи ги $p(\mathrm{B} \mid \mathrm{A}_k)$ со $p(\mathrm{B} \mid \mathrm{A}_i)$ добиваме

$$p(\mathrm{B}) \ge p(\mathrm{B} \mid \mathrm{A}_i) (p(\mathrm{A}_1) + p(\mathrm{A}_2) + \ldots + p(\mathrm{A}_n)) = p(\mathrm{B} \mid \mathrm{A}_i).$$

Заклучуваме дека

$$\min_j p(\mathrm{B} \mid \mathrm{A}_j) \le p(\mathrm{B}) \le \max_i p(\mathrm{B} \mid \mathrm{A}_i)$$

што е релативно тривијален резултат.

ПРИМЕР 3.5 Од шпил од кој се изгубени 3 карти се влече една карта. Колкава е веројатноста дека е извлечена срце карта?

Решение

B = "извлечена срце карта", $p(\mathrm{B}) = ?$

A_k = "изгубени k срце карти", $k = 0, 1, 2, 3$ е разбивање на множеството елементарни настани $\Omega = \{(x, y, z) \mid x, y, z = 1, 2, \ldots, 52\}$

$$p(\mathrm{A}_k) = \frac{\binom{13}{k}\binom{39}{3-k}}{\binom{52}{3}} \text{ за } k = 0, 1, 2, 3, \text{ т.е.}$$

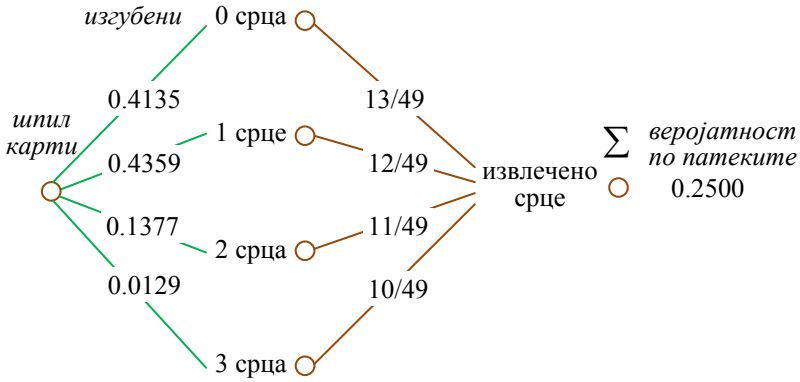
$$p(\mathrm{A}_0) = 0.4135,\ p(\mathrm{A}_1) = 0.4359,\ p(\mathrm{A}_2) = 0.1377,\ p(\mathrm{A}_3) = 0.0129$$

и сега според формулата за тотална веројатност добиваме

$$p(\mathrm{B}) = p(\mathrm{A}_0)p(\mathrm{B} \mid \mathrm{A}_0) + p(\mathrm{A}_1)p(\mathrm{B} \mid \mathrm{A}_1) + p(\mathrm{A}_2)p(\mathrm{B} \mid \mathrm{A}_2) + p(\mathrm{A}_3)p(\mathrm{B} \mid \mathrm{A}_3) =$$

$$= 0.4135\frac{13}{49} + 0.4359\frac{12}{49} + 0.1377\frac{11}{49} + 0.0129\frac{10}{49} = 0.2500. \blacksquare$$

Тоталната веројатност исто така може да се претстави графички, но поради неедноznачноста на одлуките, структурата не е дрво. На пример, решението на примерот 3.5 би можело да се прикаже со следниот дијаграм



Се разбира, ваквиот дијаграм не е ни одблиску корисен како дрвото на шанси и понатаму нема да го користиме.

3.2. Формула на Баес

Нека се дадени настани A и B со $p(\mathrm{A}) > 0$ и $p(\mathrm{B}) > 0$. Тогаш од двете симетрични формули за условна веројатност

$$p(\mathrm{B} \mid \mathrm{A}) = p(\mathrm{A}\cdot\mathrm{B}) / p(\mathrm{A}) \quad \text{и} \quad p(\mathrm{A} \mid \mathrm{B}) = p(\mathrm{A}\cdot\mathrm{B}) / \mathrm{P}(\mathrm{B})$$

ја користиме втората во која броителот $p(\mathrm{A}\cdot\mathrm{B})$ го заменуваме со $p(\mathrm{A}\cdot\mathrm{B})$ од првата формула со $p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{B} \mid \mathrm{A})p(\mathrm{A})$ што дава

$$p(\mathrm{A} \mid \mathrm{B}) = \frac{p(\mathrm{B} \mid \mathrm{A})p(\mathrm{A})}{p(\mathrm{B})}.$$

Добиената формула е наједноставната верзија на Баесовата формула. Понекогаш неа ја нарекуваат Баесова теорема или Баесово правило (што е погрешно). Баесовото правило е поврзано со методологијата на донесување одлуки (decision-making), што е екстремно контраверзна тема во статистиката. Важноста на Баесовата формула е во тоа што таа овозможува пресметка на една условна веројатност $p(\mathrm{A} \mid \mathrm{B})$ преку "обратна" условна веројатност $p(\mathrm{B} \mid \mathrm{A})$.

Следниот пример ја покажува важноста на Баесовата формула во ситуации (чести) кога таа дава "контраинтуитивно" решение што може да нè спаси од кардинални грешки.

ПРИМЕР 3.6 Во тек на рутински преглед, докторот открива тумор (грутка) во градите на една жена што може да биде канцер. Без дополнителни тестови, веројатноста женатa да има канцер е 1%. Мамографијата е тест што со 90% сигурност одредува дали туморот е бениген или малиген. Која е веројатноста дека женатa има канцер на градите ако мамографијата дала позитивен резултат, т.е. дека туморот е малиген ?

Решение

Кога би го имале пред себе овој резултат, многу доктори би помислиле дека веројатноста $p(\text{"канцер"} \mid \text{"позитивен"})$ е само малку пониска од $p(\text{"позитивен"} \mid \text{"канцер"})$, и би ја процениле на околу 80%. Со Баесовата формула добиваме

$$p(\text{"канцер"} \mid \text{"позитивен"}) = \frac{p(\text{"позитивен"} \mid \text{"канцер"}) \cdot p(\text{"канцер"})}{p(\text{"позитивен"})}.$$

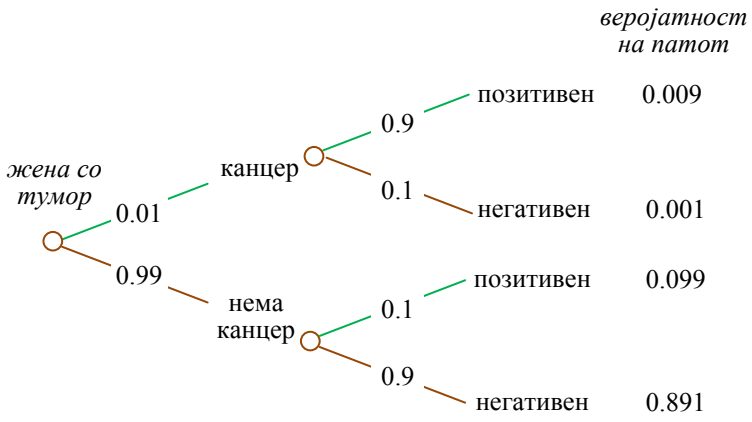
И сега, ако имаме предвид дека:

$p(\text{"позитивен"} \mid \text{"канцер"}) = 0.9$; $p(\text{"канцер"}) = 0.01$;
$p(\text{"позитивен"} \mid \text{"нема канцер"}) = 0.1$; $p(\text{"нема канцер"}) = 0.99$;
$p(\text{"позитивен"}) = p(\text{"позитивен"} \mid \text{"канцер"}) \cdot p(\text{"канцер"}) +$
$+ p(\text{"позитивен"} \mid \text{"нема канцер"})\, p(\text{"нема канцер"}) =$
$= 0.9 \cdot 0.01 + 0.1 \cdot 0.99 = 0.108$, добиваме дека

$p(\text{"канцер"} \mid \text{"позитивен"}) = 0.009/0.108 = 0.0833.$

Оваа веројатност од 8.33% е речиси 10 пати пониска од интуитивната проценка на докторот. Од каде доаѓа ваквиот резултат? Тој логично следува ако се земе предвид (не се заборави) дека почетните шанси за канцер се само 1% и тоа се "искомбинира" со 10%-ната можна грешка на мамографијата.

Дрвото на шанси изгледа вака



и од него веднаш се добива истиот резултат

$$p(\text{"канцер"} \mid \text{"позитивен"}) = \frac{p(\text{"пат 1"})}{p(\text{"пат 1"}) + p(\text{"пат 3"})} = \frac{0.009}{0.009 + 0.099}. \blacksquare$$

Горниот пример покажува колку е важно да се држат на око почетните веројатности меѓу различните категории. Во овој пример тоа се жените со тумор на градите што имаат и немаат канцер. Игнорирањето на почетните веројатности води до чудни заклучоци како "Статистиката покажува дека 10% од сообраќајките се предизвикани од пијани возачи", што значи дека останатите 90% се предизвикани од трезни возачи . . . па оттука може да се заклучи дека е посигурно луѓето да возат пијани.

Поопштата верзија на Баесова формула ја даваме како теорема.

Теорема 3.3 Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност и нека $A_1, A_2, \ldots, A_n$ е разбивање на Ω. Тогаш за произволен настан $B \in \mathcal{F}$ со $p(B) > 0$ важи

$$p(A_k \mid B) = \frac{p(B \mid A_k)p(A_k)}{p(B)} = \frac{p(B \mid A_k)p(A_k)}{p(A_1)p(B \mid A_1) + \ldots + p(A_n)p(B \mid A_n)}.$$

Доказ: Од дефиницијата на условна веројатност имаме дека

$p(A_k \cdot B) = p(A_k)p(B \mid A_k) = p(B)p(A_k \mid B)$ што повлекува

$$p(A_k \mid B) = \frac{p(B \mid A_k)p(A_k)}{p(B)}. \blacksquare$$

Баесовата формула дава начин како веројатностите треба да се ажурираат во светло на ново-добиени информации. Најопштата идеја е во

следното: A_k се хипотези со априорни веројатности $p(A_k)$ коишто не се експериментално потврдени. По извршување на експеримент во врска со избран настан B, проценети се веројатностите $p(B \mid A_k)$. Сега, Баесовата формула овозможува пресметка на постериорните веројатности $p(A_k \mid B)$ откако станал познат исходот на експериментот. Значи веројатностите на хипотезите A_k се проценуваат преку експеримент во врска со погодно избран настан B.

ПРИМЕР 3.7 Три машини изработуваат некој ист производ, и тоа, првата машина покрива 25%, втората 35% и третата 40% од производството. Шкартот на првата машина е 5%, на втората 4% и на третата 2%.

а) Колкава е веројатноста дека случајно избраниот производ е шкартен?

б) Ако случајно избраниот производ е шкартен, колкава е веројатноста дека е произведен на втората машина?

Решение

а) Да ставиме

B = "производот е шкарт", $p(B) = ?$

A_1 = "производот е од првата машина"

A_2 = "производот е од втората машина"

A_3 = "производот е од третата машина"

$p(A_1) = 0.25, \quad p(B \mid A_1) = 0.05$

$p(A_2) = 0.35, \quad p(B \mid A_2) = 0.04$

$p(A_3) = 0.40, \quad p(B \mid A_3) = 0.02$

Со директна примена на формулата за тотална веројатност добиаме

$$p(B) = p(A_1)p(B \mid A_1) + p(A_2)p(B \mid A_2) + p(A_3)p(B \mid A_3) =$$
$$= 0.25 \cdot 0.05 + 0.35 \cdot 0.04 + 0.\ 40 \cdot 0.\ 02 = 0.0345.$$

б) Треба да се пресмета $p(A_2 \mid B)$, а ја имаме $p(B \mid A_2) = 0.04$. Од Баесовата формула веднаш добиваме

$$p(A_2 \mid B) = \frac{p(B \mid A_2)p(A_2)}{p(B)} = \frac{0.04 \cdot 0.35}{0.0345} = 0.4058$$

За другите две машини на истиот начин добиваме

$$p(A_1 \mid B) = \frac{p(B \mid A_1)p(A_1)}{p(B)} = \frac{0.05 \cdot 0.25}{0.0345} = 0.3623$$

$$p(A_3 \mid B) = \frac{p(B \mid A_3)p(A_3)}{p(B)} = \frac{0.02 \cdot 0.4}{0.0345} = 0.2319$$

Се разбира, за контрола проверуваме дека

$$p(A_1 \mid B) + p(A_2 \mid B) + p(A_3 \mid B) = 1,$$

т.е. настанот "шкарт производот да е произведен на некоја од 3-те машини" е сигурен настан. ■

Да ја разгледаме Баесовата формула во светло на прашањето какви се шансите некоја хипотеза H да е точна, или не (точна е $\overline{H}$). На почеток ги имаме априорните, субјективни веројатности $p(H)$ и $p(\overline{H}) = 1 - p(H)$. Овие веројатности ќе ги смениме според некој нов доказ (настан) D. На пример во судница, D може да биде доказ дека крвната група на убиецот е иста со осомниченият, дека некоја физичка карактеристика што очевидец ја забележал е во полза или против осомниченият, итн. Постериорните веројатности $p(H \mid D)$ изразени во терминологија на шанси задоволуваат

$$\frac{p(H \mid D)}{p(\overline{H} \mid D)} = \frac{p(H)}{p(\overline{H})} \frac{p(D \mid H)}{p(D \mid \overline{H})},$$ што може да се парафразира со

постериорните шанси = априорните шанси × шанси од доказот.

Односот $p(H) / p(\overline{H})$ ги дава шансите дека H е точна пред презентација на доказот D. Вториот однос $p(D \mid H) / p(D \mid \overline{H})$ го дава влијанието што го има доказот на верувањето во хипотезата H. Ако имаме повеќе независни докази D_1, D_2, …, тие итеративно може да се применат во Баесовата формула ставајќи ги постериорните шанси како априорни.

ПРИМЕР 3.8 Убиена е жена и осомничен е сопругот за кој се знае дека ја малтретирал. Неговиот адвокат смета дека малтретирањето на жената од страна на сопругот не е битен факт бидејќи само 0.1% од мажите што ги малтретираат жените извршуваат убиство. Дали адвокатот е во право ако се знае дека во една година 4936 жени се убиени од кои 1430 од сопругот и дека генерално, 5% од жените се малтретирани од сопрузите. Додатно, проценето е дека кај околу 50% од убиствата на жените претходело нивно малтретирање.

Решение

Да ставиме

D = "сопругот ја малтретирал жената во минатото",

M = "жената е убиена",

V = "сопругот е убиецот".

Се бара веројатноста $p(V \mid DM)$ со постериорни шанси

$\frac{p(\mathrm{V}\mid \mathrm{DM})}{p(\overline{\mathrm{V}}\mid \mathrm{DM})} = \frac{p(\mathrm{V}\mid \mathrm{M})}{p(\overline{\mathrm{V}}\mid \mathrm{M})}\frac{p(\mathrm{D}\mid \mathrm{VM})}{p(\mathrm{D}\mid \overline{\mathrm{V}}\mathrm{M})}$, каде што од условите на задачата имаме априорни веројатности $p(\mathrm{V} \mid \mathrm{M}) = 1430/4936 = 0.2897$ и $p(\overline{\mathrm{V}} \mid \mathrm{M}) = 0.7103$. Понатаму, дадено е дека генерално $p(\mathrm{D} \mid \overline{\mathrm{V}}\,\mathrm{M}) = 0.05$ и според проценката $p(\mathrm{D} \mid \mathrm{VM}) = 0.5$. Оттука добиваме за шансите

$$\frac{p(\mathrm{V}\mid \mathrm{DM})}{p(\overline{\mathrm{V}}\mid \mathrm{DM})} = \frac{0.2897}{0.7103}\frac{0.5}{0.05} = 4.0786\,.$$

Шансите ги трансформираме во веројатност користејќи додатно дека $p(\overline{\mathrm{V}} \mid \mathrm{DM}) = 1 - p(\mathrm{V} \mid \mathrm{DM})$, што дава веројатност $p(\mathrm{V} \mid \mathrm{DM}) = 80.31\%$. Значи, веројатноста дека сопругот е убиецот во светло на сознанието дека претходно ја малтретирал жената е доста висока, и тоа е сигурно релевантна информација за случајот. ■

3.3. Независност на настани

Во формулата за условна веројатност $p(\mathrm{B} \mid \mathrm{A}) = p(\mathrm{A}\cdot\mathrm{B}) / p(\mathrm{A})$ треба да се пресметува $p(\mathrm{A}\cdot\mathrm{B})$ што почесто е потежок проблем од директната пресметка на $p(\mathrm{B} \mid \mathrm{A})$. Затоа почесто се користат пресметките

$p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{B} \mid \mathrm{A})p(\mathrm{A})$ и
$p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A} \mid \mathrm{B})p(\mathrm{B})$.

Интуитивно, независност на настаните најприродно се дефинира преку условната веројатност.

Дефиниција 3.2 Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност. Настанот $\mathrm{B}\in\mathcal{F}$ е независен од настанот $\mathrm{A}\in\mathcal{F}$ акко $p(\mathrm{B} \mid \mathrm{A}) = p(\mathrm{B})$.

Од дефиницијата 3.2 и формулата за условна веројатност веднаш следува дека ако B е независен од A, тогаш $p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A})p(\mathrm{B})$ и сега ако ова се замени во дуалната формула за условна веројатност се добива

$$p(\mathrm{A})p(\mathrm{B}) = p(\mathrm{A} \mid \mathrm{B})p(\mathrm{B}),$$

што ако се има предвид дека $p(\mathrm{B}) > 0$, дава $p(\mathrm{A} \mid \mathrm{B}) = p(\mathrm{A})$. Значи ако B е независен од A, тогаш и A е независен од B, т.е. независноста е секогаш взаемна. За проверка на независност на настани обично се користи следната тривијална теорема.

Теорема 3.4 Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност. Настаните $\mathrm{A}\in\mathcal{F}$ и $\mathrm{B}\in\mathcal{F}$ се независни акко $p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A})p(\mathrm{B})$. ■

ПРИМЕР 3.9 *Независност*. Од шпил карти случајно се извлекува една карта. Испитај ја независноста на настаните

A = "извлечен ас" и B = "извлечена лист карта".

Решение

а) Имаме дека

A·B = "извлечен е лист ас", од што следува

$$p(\mathrm{A}\cdot\mathrm{B}) = \frac{1}{52} = \frac{4}{52}\frac{13}{52} = p(\mathrm{A})\cdot p(\mathrm{B})$$

што значи дека A и B се независни настани.

Ако сега во шпилот додадеме една карта (да речеме од друг шпил) што не е ниту ас ниту лист карта добиваме

$$p(\mathrm{A}\cdot\mathrm{B}) = \frac{1}{53} \neq \frac{4}{53}\frac{13}{53} = p(\mathrm{A})\cdot p(\mathrm{B}),$$

што значи дека сега A и B се зависни настани. ■

Независноста на настани што тука е воведена понекогаш се нарекува теоретско-веројатносна, стохастичка или статистичка бидејќи таа не секогаш соодветствува на интуитивната проценка (верување) за зависност на настаните. Кога ние априори прифаќаме независност на настани (на пример како во серија независни експерименти) таквата независност се нарекува физичка независност. Физичката независност е во основа особина на самите настани. Ние разумно веруваме дека настаните A и B се физички независни и го изразуваме тоа верување со инсистирањето дека важи $p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A})\cdot p(\mathrm{B})$. Теоретско-веројатносната независност е особина на веројатноста (како мера) и таа не секогаш значи дека настаните се и физички независни. Примерот 3.4 покажува дека со минимална промена во просторот на веројатност, два настана што претходно биле независни, стануваат зависни.

Да нагласиме дека не треба да се меша поимот независност со поимот дисјунктност на настани. Имено, дисјунктни настани A и B со ненулти веројатности се секогаш зависни бидејќи во таков случај имаме дека

$$p(\mathrm{A}\cdot\mathrm{B}) = p(\varnothing) = 0 \neq p(\mathrm{A})\cdot p(\mathrm{B}).$$

Исто така треба да се има предвид дека независноста не е "транзитивна". Имено, ако A, B и C се по пар независни настани тоа не повлекува дека и $p(\mathrm{A}\cdot\mathrm{B}\cdot\mathrm{C}) = p(\mathrm{A})\cdot p(\mathrm{B})\cdot p(\mathrm{C})$. На пример, нека фрламе две коцки и нека A = "на првата коцка паднал парен број", B = "сумата е непар-

на" и C = "на втората коцка паднал парен број ". Тогаш настаните се по пар независни, но $p(\mathrm{A}\cdot\mathrm{B}\cdot\mathrm{C}) = 0 \neq (1/18)(1/18)(1/18) = p(\mathrm{A})\cdot p(\mathrm{B})\cdot p(\mathrm{C})$. Генерално, настаните A_1, A_2, ..., A_n се сметаат за независни ако важи $p(\mathrm{A}_{i_1}\cdot\mathrm{A}_{i_2}\cdots\mathrm{A}_{i_k}) = p(\mathrm{A}_{i_1})\cdot p(\mathrm{A}_{i_2})\cdots p(\mathrm{A}_{i_k})$ за секој $2 \le k \le n$.

Без да ја формулираме како теорема, ќе ја наведеме следната особина. Ако A и B се независни настани, тогаш се независни и настаните $\bar{\mathrm{A}}$ и B, A и $\overline{\mathrm{B}}$, како и $\bar{\mathrm{A}}$ и $\overline{\mathrm{B}}$. На пример, од

$$p(\mathrm{B}) = p(\mathrm{A}\cdot\mathrm{B}+\bar{\mathrm{A}}\cdot\mathrm{B}) = p(\mathrm{A}\cdot\mathrm{B}) + p(\bar{\mathrm{A}}\cdot\mathrm{B}) \text{ имаме } p(\bar{\mathrm{A}}\cdot\mathrm{B}) = p(\mathrm{B}) - p(\mathrm{A}\cdot\mathrm{B})$$

од што понатаму следува

$$p(\bar{\mathrm{A}} \mid \mathrm{B}) = p(\bar{\mathrm{A}}\cdot\mathrm{B}) / p(\mathrm{B}) = 1 - p(\mathrm{A}) = p(\bar{\mathrm{A}})$$

што значи дека се независни $\bar{\mathrm{A}}$ и B. Слично може да се покажат и другите две независности.

3.4. Серии независни експерименти

Ако некој експеримент повторуваме n пати при исти услови и секоја следна реализација не зависи од претходната добиваме серија од n независни експерименти. При секоја реализација на експериментот може да не интересира дали се случил или не, некој настан A. Основниот проблем е да се пресмета веројатноста во n-те повторувања на експериментот, k пати да се случи настанот A.

Теорема 3.5 (Шема на Бернули) Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност и нека реализираме серија од n независни експерименти во кои еден настан A се случува со веројатност $p(\mathrm{A})$. Тогаш, веројатноста A да се случи k пати е

$$p(\text{"A се случил } k \text{ пати од } n\text{"}) = \binom{n}{k} p(\mathrm{A})^k\, p(\overline{\mathrm{A}})^{n-k}\,.$$

Доказ: Веројатноста во една низа од n случувања или неслучувања на A (на пример $\mathrm{A}\cdot\mathrm{A}\cdot\bar{\mathrm{A}}\cdot\mathrm{A}\cdot\bar{\mathrm{A}}\cdot\bar{\mathrm{A}}\cdot\ldots\cdot\bar{\mathrm{A}}$), тој да се случи k пати, поради независноста е $p(\mathrm{A})^k p(\bar{\mathrm{A}})^{n-k}$. Бројот на такви различни низи е еднаков со бројот на избори на k елементи од множество со n елементи, т.е $\binom{n}{k}$. ■

Независноста на серија експерименти попрецизно се нарекува физичка независност и таа не треба да се поистоветува со теоретско-веројатнос-

ната независност воведена во поглавјето 3.3. Физичката независност е верување што ни овозможува веројатноста на производ на настаните да ја пресметаме како производ на веројатностите (на пример $p(\mathrm{A}\cdot\mathrm{A}\cdot\bar{\mathrm{A}}\cdot\mathrm{A}\cdot\bar{\mathrm{A}}\cdot\bar{\mathrm{A}}\cdot\ldots\cdot\bar{\mathrm{A}}) = p(\mathrm{A})p(\mathrm{A})p(\bar{\mathrm{A}})p(\mathrm{A})p(\bar{\mathrm{A}})p(\bar{\mathrm{A}})\ldots p(\bar{\mathrm{A}})$). Физичката независност не може да се докажува, т.е. ние не можеме да докажуваме дека настаните од серијата експерименти се независни, туку само можеме да веруваме дека случувањето на настаните од една реализација на експериментот не влијае на случувањето на настаните од која било друга реализација. Дали е тоа навистина така?

На пример, да ја разгледаме наједноставната серија експерименти, фрлање паричка n-пати. Кога се размислува локално, интуитивно е јасно дека резултатот на секое следно фрлање е независен од резултатот на претходното фрлање (фрлања). Од друга страна, поради стабилноста на експериментот (релативните честоти на 2-та настана одат кон 1/2) јасно е дека ако при претходните реализации имало дебаланс во однос на појавувањето, на пример, на "петка", тогаш при понатамошните реализации ќе има тренд тој дебаланс да се корегира, па би очекувале настанот "петка" да има нешто поголема шанса на случување. Тоа би значело дека верувањето за физичка независност не е реално. Некој би можел да забележи дека ова размислување е погрешно бидејќи стабилноста се достигнува само при доволно големи n (блиску до ∞), така што распоредот на "петка" и "глава" не е стабилен за секое конечно n, и независноста останува. Се чини дека овие проблеми повторно се појавуваат поради воведувањето на поимот ∞ што не може лесно да се усогласи со реалноста и води до парадокси (види поглавје 2.3).

ПРИМЕР 3.10 Што е поверојатно, меч меѓу рамноправни противници да заврши 3:2 или 5:5?

Решение

A = "победил првиот", $\mathrm{P}(\mathrm{A}) = 0.5$,

$$p(\text{"3:2"}) = \binom{3}{2} 0.5^3 (1-0.5)^2 = 0.3125,$$

$$p(\text{"5:5"}) = \binom{10}{5} 0.5^5 (1-0.5)^5 = 0.2461. \blacksquare$$

ПРИМЕР 3.11 Колкава е веројатноста во автобус со 50 луѓе да има повеќе од двајца левораци ако се знае дека леворакоста во популацијата е 13%?

Решение

A = "леворак", $p(\mathrm{A}) = 0.13$

Многу е подобро да се решава со обратниот настан.

p("барем 2-ца левораки") =

= 1 – p("0 левораки") – p("1 леворак") – p("2 левораки") =

$$1-\binom{50}{0}0.13^0 0.87^{50}-\binom{50}{1}0.13^1 0.87^{49}-\binom{50}{2}0.13^2 0.87^{48}=$$

= 1 – 0.0009 – 0.0071 – 0.0259 = 0.9661. ■

Низата веројатности што се добива за $k = 0, 1, ..., n$ во серијата независни експерименти не е монотона. Вредноста на k за која таа достигнува максимум е најверојатниот број на случувања на настанот А. Ако формираме количник од веројатностите за две соседни вредности на k добиваме

$$\frac{\binom{n}{k+1}p(\mathrm{A})^{k+1}\,p(\overline{\mathrm{A}})^{n-k-1}}{\binom{n}{k}p(\mathrm{A})^{k}\,p(\overline{\mathrm{A}})^{n-k}}=\frac{(n-k+1)\,p(\mathrm{A})}{kp(\overline{\mathrm{A}})}=1+\frac{(n+1)\,p(\mathrm{A})-k}{kp(\overline{\mathrm{A}})},$$

и сега овој однос е помал од 1 кога $k > (n + 1)p(\mathrm{A})$ и поголем од 1 кога $k < (n + 1)p(\mathrm{A})$. Возможно е да се случи $k = (n + 1)p(\mathrm{A})$ и тогаш веројатностите за k и $k + 1$ се еднакви. Оттука заклучуваме дека најверојатниот број случувања k_0 на настан А е

$$k_0=\begin{cases}[(n+1)\,p(\mathrm{A})], & \text{ако } (n+1)\,p(\mathrm{A}) \text{ не е цел број}\\ (n+1)\,p(\mathrm{A}) \text{ и } (n+1)\,p(\mathrm{A})-1, & \text{во спротивно}\end{cases}$$

каде што [...] означува цел дел.

ПРИМЕР 3.12 Колкава е веројатноста со пополнети 80 колони (десет ливчиња) на лото 7 од 37 да се добијат барем 5 погодоци. Колкава е веројатноста да се добијат 7 погодоци? Кои се најверојатните броеви случувања за k = 0, 1, 2, 3, 4, 5, 6, 7 погодоци?

Решение

А = "барем 5 погодоци" = "5 погодоци" + "6 погодоци" + "7 погодоци"

Овие веројатности ги пресметавме во примерот 2.10.

p(А) = 0.000887283 + 0.000020397 + 0.000000097 = 0.000807777

Бидејќи бараме А да се случи барем еднаш, многу е подобро да се решава со обратниот настан.

p("од 80 колони барем еднаш А") = 1 – p("од 80 колони ниедаш А") =

$$= 1 - \binom{80}{0} 0.000807777^{0} 0.999192223^{80} = 0.0626 = 6.26\%$$

Сосема аналогно, за барем еднаш да имаме 7 погодоци се добива

p("од 80 колони барем еднаш 7 погодоци") =

= 1 – p("од 80 колони ниедаш 7 погодоци ") =

$$= 1 - \binom{80}{0} 0.000000097^{0} 0.999999903^{80} = 0\ .00000776 = 0.000776\%$$

Во следната табела се дадени веројатностите заедно со најверојатниот број на случувања на погодоци:

k	p("точно k погодоци со 1-а колона")	p("барем еднаш точно k погодоци со 80 колони")	најверојатен број случувања на k погодоци во 80 колони
0	0.197737413	0.999999978	16
1	0.403713885	0.999999999	32
2	0.290673997	0.999999999	23
3	0.093164743	0.999601048	7
4	0.013802184	0.670997002	1
5	0.000887283	0.068551287	0
6	0.000020397	0.001019341	0
7	0.000000097	0.000007760	0

Поради земање цел дел од броевите, изгубен е еден број во најверојатен број случувања на k погодоци во 80 колони (збирот во последната колона е 79 наместо 80). ■

Постои праволиниско обопштување на теоремата 3.5, во коешто наместо да се разгледува случување на само еден настан, се разгледуваат случувања на повеќе настани.

Теорема 3.6 Нека (Ω, $\mathcal{F}$, p) е простор на веројатност и нека реализираме серија од n независни експерименти во кои може да се случат настаните А, В, С, . . . со веројатности p(А), p(В), p(С), ... Тогаш, веројатноста А да се случи k_1 пати, В да се случи k_2 пати, С да се случи k_3 пати, ... е

p("А - k_1 пати, В - k_2 пати, С - k_3 пати, ... од n") =

$$= \frac{n!}{k_1!k_2!k_3!\ldots} p(\mathrm{A})^{k_1} p(\mathrm{B})^{k_2} p(\mathrm{C})^{k_3} \ldots .$$

Доказ: Веројатноста во една низа од n експерименти, A да се случи k_1 пати, B да се случи k_2 пати, C да се случи k_3 пати, ... поради независноста е $p(\mathrm{A})^{k_1}\, p(\mathrm{B})^{k_2}\, p(\mathrm{C})^{k_3} \ldots$. Бројот на вакви различни низи е еднаков на бројот на пермутации со повторување од n елементи каде што имаме $k_1, k_2, k_3, \ldots$ исти, и $k_1+k_2+k_3+ \ldots = n$. ■

Теоремата 3.5 е специјален случај на теоремата 3.6 кога разгледуваме само 2 настани A и $\bar{\mathrm{A}}$ и тогаш $k_2 = n - k_1$, па

$$\frac{n!}{k_1!k_2!}\, p(\mathrm{A})^{k_1}\, p(\overline{\mathrm{A}})^{k_2} = \binom{n}{k_1} p(\mathrm{A})^{k_1}\, p(\overline{\mathrm{A}})^{n-k_1},$$

т.е. се добива формулата од теоремата 3.5 ако се замени k_1 со k.

Формулите за пресметка на веројатност при сериите независни експерименти (теоремите 3.4 и 3.5) не се ништо друго од една додатна шема за пресметка на веројатност во специфична ситуација. Се разбира, при еднаковеројтни елементарни настани, веројатностите кај сериите независни експерименти може да се пресметуваат и на класичен начин (поволни случаи / сите можни) како што е направено во примерот 2.5 за задачата 19, но тоа вообичаено е многу потежок начин.

ЗАДАЧИ

1. Настаните A и B се дисјунктни. Кои од следните равенства се точни:

   а) $p(\mathrm{A} \mid \mathrm{B}) = p(\mathrm{A})$; б) $p(\mathrm{A} + \mathrm{B} \mid \mathrm{C}) = p(\mathrm{A} \mid \mathrm{C}) + p(\mathrm{B} \mid \mathrm{C})$;

   в) $\dfrac{p(\mathrm{A} \mid \mathrm{B})}{p(B)} = \dfrac{p(\mathrm{B} \mid \mathrm{A})}{p(\mathrm{A})}$; г) $p(\mathrm{A}\cdot\mathrm{B}) = p(\mathrm{A})\cdot p(\mathrm{B})$.

   Повтори го сето кога настаните A и B се независни.

2. Се фрлаат две (фер) коцки. Пресметај ја веројатноста:

   а) Сумата да е 6 ако се знае дека паднал различен број;

   б) Сумата да е 7 ако се знае дека паднал различен број;

   в) Првата коцка да е 6 ако се знае дека паднал различен број.

3. Нека играч (во бриџ) има 13 карти меѓу кои има ас.

а) Пресметај ја веројатноста дека играчот има точно еден ас;

б) Што се менува во веројатноста ако се знае дека играчот има срце ас?

4. Секоја вечер, две метеоролошки станици даваат прогноза за времето независно една од друга. Станицата 1 погодува во 90% од случаите, а станицата 2 во 80% без разлика на времето. Една вечер станицата 1 предвидува сончево време, а станицата 2 дожд. Пресметај ја веројатноста дека прогнозата на станицата 1 е точна.

5. Патуваме со едно парче багаж од Скопје за Сиднеј со преседнување во Дубаи и Сингапур. На секое преседнување багажот се префрла во другиот авион. Во Скопје веројатноста багажот да се стави во погрешен авион е 5%. Оваа веројатност во Дубаи е 2%, а во Сингапур 0.5%.

а) Најди ја веројатноста дека багажот нема да стигне во Сиднеј со нас;

б) Ако багажот не стигнал во Сиднеј со нас, пресметај ја веројатноста дека е изгубен во Скопје.

6. На една раскрсница на главна со споредна улица, од секои 100 растојанија меѓу колите на главната улица, 65 од нив се прифатливи, т.е. може да се приклучи кола од споредната улица. Кога пристигнува кола од споредната улица:

а) која е веројатноста дека првото растојание не е прифатливо;

б) која е веројатноста дека првите две растојанија не се прифатливи;

в) првата кола ја поминала раскрсницата. Која е веројатноста дека втората ќе ја помине раскрсницата веднаш на следното растојание.

7. Во еден магацин има 500 контејнери со смрзнат сок од портокал и притоа во просек 1% од нив се расипуваат. Купец купува 2 контејнера. Колкава е веројатноста дека:

а) вториот контејнер е расипан ако се знае дека првиот бил расипан;

б) и двата се расипани;

в) и двата се исправни;

г) ако A="1-от е расипан", B="2-от е расипан", дали A и B се независни ? Дали се независни ако расипаните се враќаат. (замена)

8. Располагаме со 3 карти обоени од двете страни со една боја и тоа: зелена со зелена, црвена со црвена и зелена со црвена. Влечеме една карта (без гледање) и забележуваме дека бојата од едната страна е зелена. Пресметај ја веројатноста дека и другата страна е зелена.

9. Група од 15 туристи пристигнуваат во град со 4 хотели: A, B, C, D, од кои во секој има доволно место за сите 15. Туристите се сместуваат во хотелите

така што секој турист случајно избира хотел во кој ќе се смести. Пресметај ја веројатноста дека во секој од хотелите ќе се смести барем еден турист.

10. Контролата на квалитетот на транзистори ги класифицирала по производител и квалитет во следната табела

Произ-водител	*Квалитет*			
	добар	граничен	лош	вкупно
A	128	10	2	140
B	97	5	3	105
C	110	5	5	120

Процени ја веројатноста дека случајно избран транзитор е:

а) од производителот A со добар квалитет;

б) добар ако се знае дека од производителот C;

в) од производителот B ако се знае дека е со граничен квалитет.

11. Во една елементарна студија за синхронизација на семафорите, разгледуваме систем од 4 последователни семафори со црвено светло во времетраење од 30 секунди во циклус од 50 секунди. Нека

$$p(\mathrm{S}_{j+1} \mid \mathrm{S}_j) = 0.15 \quad \text{и} \quad p(\mathrm{S}_{j+1} \mid \overline{\mathrm{S}}_j) = 0.40, \ \text{за} \ j = 1,2,3$$

каде што S_j означува настан "возачот запрел на j-тиот семафор". Пресметај ја веројатноста на настаните дека возачот ќе го фатат:

а) сите 4 семафори; б) ниеден од семафорите;

в) најмногу еден семафор.

12. Пристигнати повици до еден компјутерски сервис се класифицирани како рекламации (75% од повиците) или барање информација (25% од повиците). Од рекламациите, 40% се хардверски, 57% софтверски, а останатите 3% се проблеми на погрешно следење на упатствата за инсталација од страна на корисникот. Барањата информации се поделени на технички прашања (50%) и нови нарачки (50%). Пресметај ја веројатноста:

а) дека еден повик до сервисот е од корисник кој не го следел упатството за инсталација;

б) дека еден повик е барање за нова нарачка.

13. Воведена е нова медицинска процедура за рано откривање на една болест и спроведен е медицински скрининг. Веројатноста дека тестот коректно идентификува некого со болест е 99%, а веројатноста дека идентификува некого како болен а тој ја нема болеста е 95%. Болеста се јавува во популацијата во 0.01%. Ако некој направил тест и тој бил позитивен, пресметај ја веројатноста дека ја има болеста.

14. Дигиталниот пренос на податоци се состои од пренос на низи од 0-ли и 1-ци. Познато е дека при преносите 0-та се јавува 4 пати почесто од 1-цата. Поради несовршеноста на опремата за пренос веројатноста дека 0-та ќе биде примена како 1-ца е 0.01, а веројатноста дека 1-цата ќе биде примена како 0 е 0.05. Да се пресмета веројатноста дека

    а) е испратена 0, ако е примена 0;

    b) е испратена 1-ца, ако е примена 1-ца.

15. Возачите во една област се тестираат редовно за алкохолизираност од страна на полицијата преку тест со дување. Само после позитивен резултат од дувањето возач се испраќа на крвен тест кој определува дали возачот е под дејство на алкохол. Тестот со дување дава позитивен резултат за 90% од пијаните возачи и 5% позитивен резултат за трезните возачи. Вообичаено, еден возач е подложен на тестот со дување после сомнително возачко однесување, но полицијата во дадената област сметала дека е добро да се тестираат возачите по случаен избор. Тековната статистика покажува дека еден од 20 возачи во областа вози под дејство на алкохол. Пресметај ја веројатноста дека случајно тестиран возач ќе биде непотребно испратен на крвен тест после позитивен резултат од тестот со дување.

16. Се тестира нов метод за детекција на загадувачи во водата. Тој може да детектира (според произведувачот) органски загадувачи со 99.7% сигурност, испарливи загадувачи со 99.95% сигурност и хлорни компоненти со 89.7% сигурност. Ако нема загадување методот тоа исто така го детектира. Припремени се примероци за калибрација од кои 60% се со органски загадувач, 27% со испарливи загадувачи и 13% со хлорни компоненти. Примерокот за тестирање се избира случајно.

    а) Пресметај ја веројатноста дека методот ќе открие загадување;

    б) Ако е откриено загадување, пресметај ја веројатноста дека тоа е од хлорни компоненти.

17. Процент на неисправност во една серија на произведени LED монитори е 5.2%. Колку монитори од серијата треба да се земат за веројатноста меѓу нив да има неисправен монитор биде не помала од 98%?

18. Осум машини независно полнат туби со синтетичко лепило коишто потоа одат на заедничка подвижна лента. Примерок-туба се зема на секои неколку минути. Под претпоставка дека примероците се независни, пресметај ја веројатноста:

    а) дека 5 последователни туби се произведени на првата машина;

    б) дека 5 последователни туби се произведени на иста машина;

    в) дека 4 од 5 последователни туби се произведени на првата машина.

19. (Chevalier de Méré). Користејќи ја шемата за независни експерименти пресметај ја, а) веројатноста да се добие 6-ка од 4 фрлања на коцка и б) веројатноста да се добијат две 6-ки од 24 фрлања на 2 коцки?

20. Стрелец пука во кружна мета со радиус r и секогаш ја погодува. Метата е поделена на 4 области со 3 концетрилни кружници чиишто радуси се зголемуваат за $r/4$ тргнувајќи од центарот на метата. Погодоците во секој од 4-те кружни прстени тргнувајќи од центарот носи 10, 6, 3 и 1 поен последователно. Ако местата на погодување на метата се случајни, пресметај ја веројатноста дека од 10 пукања стрелецот ќе собере најмалку 90 поени.

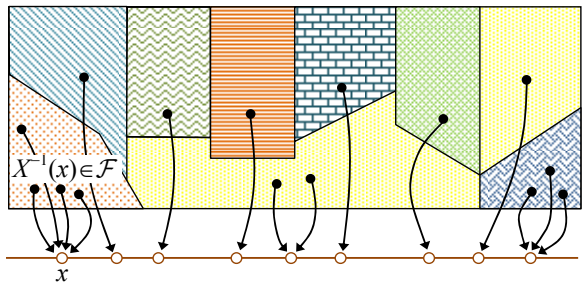
4

Случајни променливи

Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност и нека $X\colon \Omega \to \mathbb{R}$ е пресликување, т.е. реална функција што на секој елементарен настан $\omega \in \Omega$ му придружува реален број $X(\omega) \in \mathbb{R}$. Интуитивно, случајна променлива X е променлива што зема вредности со одредени веројатности. Идејата за нивното воведување е од простори на настани да се префрлиме во просторот на реалните броеви каде што располагаме со добро разработен математички апарат.

Дефиниција 4.1 За реалната фукција $X\colon \Omega \to \mathbb{R}$ велиме дека е случајна променлива акко за секој $x \in \mathbb{R}$ важи $A = \{\omega \mid X(\omega) < x\} \in \mathcal{F}$.

Дефиницијата бара елементарните настаните што се пресликуваат во интервалот $(-\infty, x)$ да формираат настан од $\mathcal{F}$. Ако Ω е конечно или пребројиво овој услов е секогаш исполнет (A е секогаш во $\mathcal{F}$) кога сите подмножества елементарни настани се во $\mathcal{F}$. Кога Ω е непребројиво бараниот услов е битен бидејќи σ-алгебрата $\mathcal{F}$ може да не го содржи A.



Слика 4.1 *Пресликување настан → случајна променлива*

Случајните променливи речиси секогаш се означуваат со испуштање на аргументот: X, Y, A, B, а не $X(\omega)$, $Y(\omega)$, $A(\omega)$, $B(\omega)$. Ова е добро позната практика да се испуштаат аргументите на функциите кога од интерес се функционалните односи, на пример $\mathrm{d}(u\cdot v) = u\cdot \mathrm{d}v + v\cdot \mathrm{d}u$. Понатаму и ние како што е вообичаено ќе го испуштаме ω при работа со настаните определени од случајните променливи, т.е. наместо $X(\omega) < x$, $X(\omega) = x$ или $X(\omega)\in[a, b]$ едноставно ќе пишуваме $X < x$, $X = x$ или $X\in[a, b]$.

Ние и до сега понекогаш имплицитно користевме случајни променливи. На пример, при фрлање коцка наместо се случил настанот "паднала 6-ка" кажуваме се случил бројот 6. Многу случувања во реалноста добро се опишуваат преку нумеричка квантификација, а тоа природно резултира во случајни променливи. На пример:

X = "дневен број на пациенти кај некој стоматолог"; X зема вредности од множеството $\{0, 1, 2, ...\}$;

X = "време меѓу пристигања на автомобили во еден сервис за поправка"; X зема вредности од временскиот интервал $[0, T]$;

X = "локација на пад на метеор на земјата"; X зема вредности од рамнината (x, y) каде што тој пар може да означува географска должина и широчина.

Од дефиницијата за случајна променлива 5.1, веднаш следува

$$(X \geq x) = \overline{(X < x)} \in \mathcal{F},$$

$$(x_1 \leq X < x_2) = (X < x_2) - (X < x_1) \in \mathcal{F},$$

$$(X = x) = \prod_{n=1}^{\infty} (x \leq X < x + \frac{1}{n}) \in \mathcal{F},$$

така што веројатноста на овие настани е дефинирана.

За пресметка на веројатностите на настаните од ваков вид не е потребно да се знае веројатноста на секој елементарен настан, туку е доволно да се знае веројатноста на настаните од облик $(X < x)$. Ова не води до добро познатиот поим на функција на распредеба (distribution).

Дефиниција 4.2 Фукцијата $F(x) = p(X < x)$ е функција на распределба за случајната променлива X.

Интуитивно, функцијата на распределба $F(x)$ ги дава веројатностите на сите "растечки" настани од $\mathcal{F}$ тргнувајќи од $\varnothing$ кога таа има вредност 0, па преку настаните што се состојат од оние елементарни настани што се пресликани во помали реални броеви од x, сè до Ω кога таа добива вредност 1.

Јасно е дека $F(x)$ е дефинирана за секој $x\in\mathbb{R}$. Од равенството меѓу настани

$$(X < x_2) = (x_1 \le X < x_2) + (X < x_1)$$

според аксиомата а3 за збир на дисјунктни настани добиваме

$$p(X < x_2) = p(x_1 \le X < x_2) + p(X < x_1)$$

од што следува

$$p(x_1 \le X < x_2) = F(x_2) - F(x_1).$$

Имајќи предвид дека $X \ge x = \overline{X < x}$, веднаш следува дека

$$p(X \ge x) = 1 - p(X < x) = 1 - F(x).$$

На крај имаме

$$p(X = x) = \lim_{n\to\infty} (F(x + \frac{1}{n}) - F(x)) = F(x + 0) - F(x).$$

Некои основни особини на функцијата на распределба $F(x)$ се дадени во следната теорема.

Теорема 4.1 За секоја фукцијата на распределба $F(x)$ точни се следните тврдења:

1) $0 \le F(x) \le 1$;

2) $F(x)$ е монотоно неопаѓачка функција;

3) $F(x)$ е прекината функција (непрекината е од лево);

4) $\lim_{x\to-\infty} F(x) = F(-\infty) = 0$, $\lim_{x\to\infty} F(x) = F(+\infty) = 1$;

5) $p(a \le X < b) = F(b) - F(a)$.

Доказ: Тврдењата следуваат директно од дефиницијата, а 5) е веќе докажано. ■

Особините изнесени во теоремата 4.1 се карактеристични за функциите на распределба $F(x)$ во таа смисла што секоја функција што ги задоволува овие особини (поточно 2), 3) и 4)) е функција на распределба за некоја случајна променлива. Да ја забележиме особината 5) што овозможува праволиниска пресметка на веројатноста на настаните преку функцијата на распределба.

Секое задавање на X што овозможува да се најде нејзината функција на распределба $F(x)$ вообичено ќе го нарекуваме закон за распределба на X.

4.1. Дискретни случајни променливи

Случајната променлива X е *дискретна* акко зема вредности од конечно $x_1, x_2, ..., x_n$ или пребројиво[2] $x_1, x_2, ..., x_n, ...$ множество броеви и притоа важи

$$\sum_{i=1}^{n} p(X = x_i) = 1, \quad \text{т.е.} \quad \sum_{i=1}^{\infty} p(X = x_i) = 1.$$

Очигледно за дискретна случајна променлива, законот на распределба целосно е определен со задавање на нејзините вредности и соодветните веројатности

x_1	x_2	...	x_n	...
$p(x_1)$	$p(x_2)$	...	$p(x_n)$	...

Функцијата на распределба во овој случај е скалеста со скокови во точките $x_1, x_2, ..., x_n, ...$ еднакви на $p(x_i)$. За случајна променлива со конечен број вредности, функцијата на распределба $F(x)$ го има обликот

$$F(x) = \begin{cases} 0 & \text{за } x < x_1 \\ \sum_{1 \le i \le k-1} p(x_i) & \text{за } x_{k-1} \le x < x_k \\ 1 & \text{за } x \ge x_n \end{cases}$$

Сите веројатносни информации за случајната променлива X се сумирани во нејзиниот закон на распределба и следователно во нејзината функција на распределба.

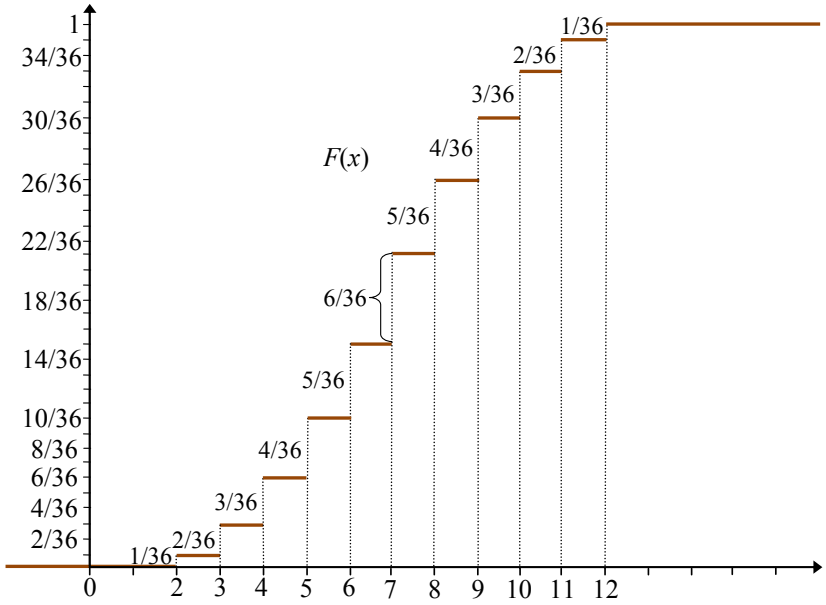
ПРИМЕР 4.1 Да го разгледаме експериментот е фрлање 2 коцки. Дефинираме случајна променлива X = "збир на паднатите броеви". Законот на распределба на X е даден во следната табела

x_i	2	3	4	5	6	7	8	9	10	11	12
$p(X=x_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

На сликата подолу е прикажана функција на распределба за случајната променлива X.

---

[2] Не сè што е пребројиво е дискретно. На пример, рационални броеви од кој било интервал не се дискретни бидејќи меѓу кои било два рационални броја има трет. За дискретност е неопходно да важи $x_i < x_{i+1}$, т.е. броевите да се "дискретно распоредени" што значи да може да се подредат во монотона низа (без точки на натрупување).



За одговор на прашањето, колкава е веројатноста да при фрлање 2 коцки се добие збир поголем од 7, едноставно се пресметува

$$p(X = 8) + p(X = 9) + p(X = 10) + p(X = 11) + p(X = 12) = 15/36 = 0.4167$$

или едноставно

$$p(\text{"збир поголем од 7"}) = F(\infty) - F(7.1) = 1 - 21/36 = 15/36 = 0.4167. \blacksquare$$

ПРИМЕР 4.2 Се враќаме на Monty Hall проблемот (пример 2.13). Имаме 3 врати зад кои се наоѓаат 2 кози и автомобил (добивката). Играчот одбира врата (не ја отвора), а водителот, кој знае каде е автомобилот, од останатите две ја отвора таа со козата. Играчот има право да остане на својот избор или да ја избере другата неотворена врата. Што е подобро за него?

Решение

Слика 4.2 *Monty Hall проблем: играчот ја бира третата врата, а водителот ја отвора првата*

Дефинираме 3 дискретни случајни променливи:

A = "бројот на вратата позади која е автомобилот", $p(A = i) = 1/3$,

I = "бројот на вратата избрана од играчот" и

V = "бројот на вратата отворена од водителот".

$$p(V \mid A,I) = \begin{cases} 0 \text{ ако } V = I & (\textit{водителот не може да ја избере вратата избрана од играчот}) \\ 0 \text{ ако } V = A & (\textit{водителот не може да ја избере вратата зад која е автомобилот}) \\ 1/2 \text{ ако } I = A & (\textit{двете врати зад кои нема автомобил имаат иста шанса да бидат отворени}) \\ 1 \text{ ако } V \neq A \text{ и } I \neq A & (\textit{само една врата може да се отвори}) \end{cases}$$

Играчот сега може да ја искористи Баесовата формула за да ја пресмета веројатноста на наоѓање на козата зад секоја врата, откако се направени почетниот избор и отворање на врата од страна на водителот. Тоа е бараната условна веројатност на A кога се дадени V и I.

$$p(A \mid V,I) = \frac{p(V \mid A,I)p(A \mid I)}{p(V \mid I)} \quad \text{каде што} \quad p(A \mid I) = p(A)$$

за сите вредности на A и I бидејќи положбата на автомобилот не зависи од изборот на играчот. Именителот $p(V \mid I)$ се добива од формулата за тотална веројатност

$$p(V \mid I) = \sum_{A=1}^{3} p(V,A \mid I) = \sum_{A=1}^{3} p(V \mid A,I)p(A \mid I).$$

Така, ако играчот ја избрал вратата 3, а водителот ја отворил вратата 1, (како во примерот 2.13), веројатноста дека автомобилот е зад вратата 2 е условната веројатност

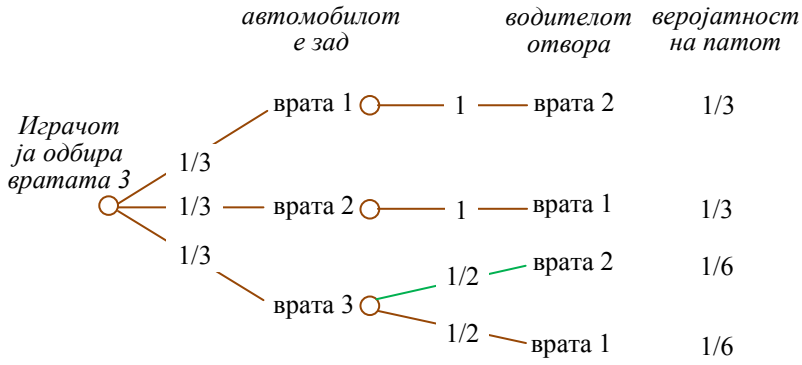
$$p(A = 2 \mid V = 1, I = 3) = \frac{1 \cdot \frac{1}{3}}{0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} = \frac{2}{3}.$$

Едно од многуте обопштувања на овој проблем (природно) е наместо 3 да се работи со n врати при што водителот отвора k врати што губат и потоа му дава можност на играчот за нов избор. Се покажува дека во тој случај промената на изборот добива со веројатност

$$\frac{n-1}{n(n-k-1)}$$

што очигледно за оригиналниот проблем кога $n = 2$, $k = 1$ дава 2/3.

Дрвото на шанси за овој проблем изгледа вака



Промена на вратата е збирот на поволните патеки 1/3+1/2=2/3. ■

Некои дискретни случајни и нивните закони на распределба се јавуваат често во практиката.

БИНОМНА РАСПРЕДЕЛБА. Разгледуваме серија од n независни експерименти при што во секој од нив настанот A се случува со веројатност p. Дефинираме случајна променлива X = "број на случувања на A". Тогаш законот на распределба на X е даден со шемата на Бернули

$$p(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \text{ за } k=0,\ 1,\ 2,\ \ldots,\ n.$$

За $n=1$ се добива Бернулиевата распределба која кажува само дали при реализација на експериментот се случил или не настанот A, $p(X=k)=p^k(1-p)^{1-k}$, за $k=0,\ 1$.

ПРИМЕР 4.3 Раководството на една фирма е загрижено поради високиот одлив на вработените на одредено време. Од претходни испитувања овој број е проценет на 10% годишно. Одбирајќи на случаен начин 8 вака вработени колкава е веројатноста дека повеќе од 2-ца ќе ја напуштат фирмата во следната година?

Решение

Работиме на ист начин како кај сериите независни експерименти.

$p=0.1$, и потребни ни се вредностите на распределбата за

$k=0,\ 1,\ 2$, што дава

$p(X>2)=1-p(X=0)-p(X=1)-p(X=2)=$

$$=1-\binom{8}{0}0.1^0 0.9^8-\binom{8}{1}0.1^1 0.9^7-\binom{8}{2}0.1^2 0.9^6=0.0381.$$ ■

ПУАСОНОВА РАСПРЕДЕЛБА. Разгледуваме број на појавувања на некој настан во некој интервал при што,

а) веројатноста на појавување на настанот е еднаква во сите интервали со иста должина,

б) случувањето на настанот во еден интервал не зависи од неговите случувања во други интервали.

Дефинираме случајна променлива X = "број на случувања на А во еден интервал". Тогаш законот на распределба на X е даден со

$$p(X=k)=\frac{\mu^k}{k!}e^{-\mu}, \text{ за } k=0,\ 1,\ 2,\ ...,\ n,\ ... \text{ каде што}$$

$\mu > 0$ е просечен број случувања во еден интервал. Лесно се проверува дека сумата

$$\sum_{k=0}^{\infty} p(X=x_k)=\sum_{k=0}^{\infty}\frac{\mu^k}{k!}e^{-\mu}=e^{-\mu}\sum_{k=0}^{\infty}\frac{\mu^k}{k!}=e^{-\mu}e^{\mu}=1.$$

ПРИМЕР 4.4 Пациенти пристигаат од брза помош во болницата во време на викенд 6-мина на час. Колкава е веројатноста во викенд 4 пациенти да стигнат во следните 30 минути?

Решение

μ = 6 на час = 3 на 30 минути

$$p(X=4)=\frac{3^4}{4!}e^{-3}=0.1680. \blacksquare$$

Пуасоновата распределба при одредени услови е граничен случај на биномната распределба.

Теорема 4.2 (Пуасон) Ако $n\to\infty$ и $p\to 0$, така што $n\cdot p\to\mu<\infty$, тогаш

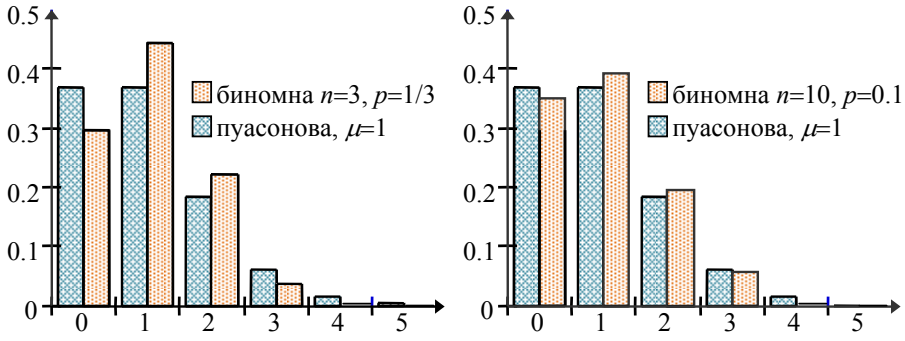
$$p(X=k)=\binom{n}{k}p^k(1-p)^{n-k}\ \to\ \frac{\mu^k}{k!}e^{-\mu}.$$

Доказ: Ставајќи $n\cdot p=\mu_n$ добиваме

$$\binom{n}{k}p^k(1-p)^{n-k}=\frac{n(n-1)\dots(n-k+1)}{n!}\left(\frac{\mu_n}{n}\right)^k\left(1-\frac{\mu_n}{n}\right)^{n-k}=$$

$$\frac{\mu_n^k}{k!}\left(1-\frac{\mu_n}{n}\right)^n\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)\left(1-\frac{\mu_n}{n}\right)^{-\infty} \xrightarrow[n\to\infty]{} \frac{\mu^k}{k!}e^{-\mu}. \blacksquare$$

На овој начин, при големи n и мали p (да речеме $n > 100$ и $n{\cdot}p < 30$) може биномната распределба приближно да се замени со пуасонова.



Слика 4.3 *Разлика меѓу биномна и пуасонова распределба*

Од сл. 4.3, очигледно е кога расте n и се намалува p биномната и пуасоновата распределба стануваат сè поблиски.

ПРИМЕР 4.5 При бомбардирањето на Лондон во втората светска војна, некој се обидувал да ја процени веројатноста дека неговата зграда ќе биде погодена. Тој го поделил неговиот кварт на 10 по 10 блокови при што квартот го погодуваат 400 бомби. Ако паѓањето на бомбите е случајно, колкава е веројатноста дека една зграда ќе биде погодена?

Решение

Решението треба да се бара со биномна распределба каде што $p = 1/100 = 0.01$ и $n = 400$. Нека $X =$ "број на погодоци", тогаш

$$p(X>0) = 1 - p(X=0) = 1 - \binom{400}{0} 0.01^0 0.99^{400} = 0.0180$$

Со пуасоновата распределба добиваме $\mu = n{\cdot}p = 4$, па добиваме

$$p(X=0) = \frac{4^0}{0!}e^{-4} = 0.0183. \blacksquare$$

ХИПЕРГЕОМЕТРИСКА РАСПРЕДЕЛБА. Од N објекти, r се означени како поволни. Се одбираат n објекти и се разгледува случајната променлива $X =$ "број на поволни објекти во n-те избрани". Тогаш, законот на распределба на X е даден со

$$p(X=k)=\frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}, \text{ за } k=0,\ 1,\ 2,\ ...,\ n.$$

ПРИМЕР 4.6 При вадењето на 4 потрошени батерии од еден иструмент, тие по грешка се измешани со новите 4 батерии коишто изгледаат идентично. Колкава е веројатноста при избор на 4 батерии од 8-те барем 3 да се исправни?

Решение

Слични проблеми решаваме кај класичната веројатност, како на пример лотаријата (пример 2.11).

$N=8$, $n=4$, и $r=4$. Потребни се вредностите на распределбата за $k=3$, 4.

$$p(X\geq 3)=p(X=3)+p(X=4)=\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}}+\frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}}=0.2429. \blacksquare$$

ГЕОМЕТРИСКА РАСПРЕДЕЛБА. Се повторува некој експеримент сè до појавување на настан A. Нека p е веројатноста на случување на A, и нека поторувањата на експериментот се независни. Ако ја дефинираме случајната променлива $X=$ "број на повторувања на експериментот додека не се случи A", тогаш законот на распределба на X е даден со

$$p(X=k)=(1-p)^{k-1}p\ , \text{ за } k=1,\ 2,\ ...,\ n,\ ...\ .$$

Речиси тривијално е да се провери дека

$$\sum_{k=1}^{\infty}p(X=x_k)=p\sum_{k=0}^{\infty}(1-p)^k=\frac{p}{1-(1-p)}=1\ .$$

ПРИМЕР 4.7 Веројатноста да во една минута дојде странка на некој шалтер е 26%. Колкава е веројатноста во следните 10 минути да нема странка, а во 11-тата да се појави?

Решение

Ова е типичен случај на геометриска распределба.

$p=0.26$, па бараната веројатност е

$p(X=11)=0.74^{10}\cdot 0.26=0.0128. \blacksquare$

НЕГАТИВНА БИНОМНА РАСПРЕДЕЛБА. Разгледуваме серија независни експерименти при што во секој од нив настанот A се случува со веројатност p. Експериментот се повторува додека точно r пати не се случи A. Дефинираме случајна променлива X = "број на случувања на $\bar{A}$ (не случувања на A) до r-тото случувања на A". Тогаш законот на распределба на X е даден со веројатностите

$$p(X=k)=\binom{k+r-1}{r-1}p^r(1-p)^k \text{, за } k=0,\ 1,\ 2,\ \ldots\ .$$

Името на распределбата доаѓа оттаму што за контраст од биномната распределба, тука бројот на случувања на A е фиксен, а бројот на повторувања на експериментот е произволен. Во специјален случај, кога $r=1$, се добива геометриска распределба.

ПРИМЕР 4.8 Треба да се изберат 5 луѓе-доброволци да учествуваат во испитување на делотворноста на некоја медицинска терапија. Ако веројатноста дека случајно избран човек сака да учествува во испитувањето е 0.2, колкава е веројатноста дека ќе мора да се прашаат 15 луѓе за да се обезбедат 5-тмина за испитувањето? Колкава е веројатноста дека најмногу 10 луѓе ќе одбијат да учествуваат во испитувањето?

Решение

Ова е типичен пример на негативна биномна распределба со:

$p=0.2$, $r=5$ и $k=10$ што дава веројатност

$$p(X=10)=\binom{14}{4}0.2^5 0.8^{10}=0.034\ .$$

Веројатноста дека најмногу 15 луѓе ќе мора да се прашаат е

$$p(X\le 10)=\sum_{k=0}^{10}\binom{k+4}{4}0.2^5 0.8^k=0.164.\ \blacksquare$$

Негативната биномна распределба може праволиниски да се обопшти земајќи r да не биде природен број. Таквата обопштена негативна биномна распределба добро се согласува со податоците од различни апликации.

4.2. Непрекинати случајни променливи

Веќе дискутиравме дека не е можно да се доделуваат веројатности на поединечни реални броеви, туку тоа мора да се прави на интервали.

Непрекинатите случајни променливи земаат вредности на интервал или колекција на интервали.

Формално, случајната променлива X е (апсолутно) непрекината акко постои ненегативна функција $f(x)$, таква што за секор x важи

$$F(x) = p(X < x) = \int_{-\infty}^{x} f(u)du\,.$$

Функцијата $f(x)$ се нарекува *густина* на распределбата на случајната променлива X. Понатаму секогаш ќе претпоставуваме дека $f(x)$ е непрекината скоро секаде, што е потребно за интеграбилност. Веднаш имаме,

$$p(a \le X < b) = \int_{a}^{b} f(x)dx \quad \text{и}$$

$$p(X = a) = \lim_{n\to\infty} p(a \le X < a + \frac{1}{n}) = \lim_{n\to\infty} \int_{a}^{a+\frac{1}{n}} f(x)dx = 0\,.$$

Оттука следува дека за непрекинати случајни променливи важи

$$p(a \le X \le b) = p(\mathrm{a} \le X < \mathrm{b}) = p(a < X \le b) = p(a < X < b)\,.$$

Теорема 4.3 За секоја густина на распределба $f(x)$ точни се следните тврдења:

1) $f(x) \ge 0,\ \forall x \in \mathbb{R}$

2) $\int_{-\infty}^{\infty} f(x)dx = 1\,,$

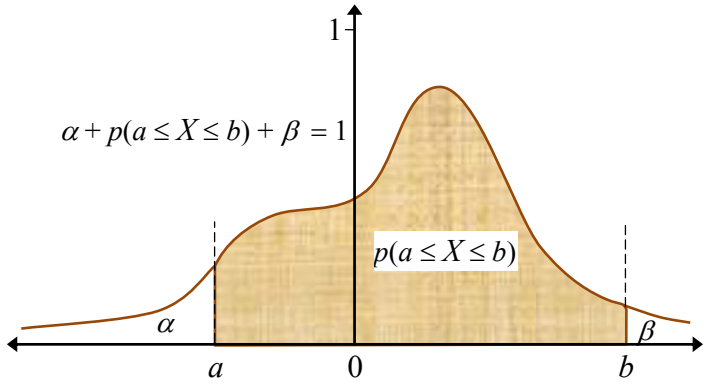
3) $F'(x) = \dfrac{dF}{dx} = f(x)\,.$

Доказ: Следува директно од дефиницијата. ■

Се разбира, важи и тврдењето дека секоја функција $f(x)$ што ги задоволува наведените особини е густина на распределба за некоја случајна променлива X.

Геометриски претставено, веројатноста случајната променлива X да земе вредност во интервалот [a, b] е еднаква на плоштината меѓу нејзината густина на распределба и x-оската во истиот интервал (сл. 4.4).

Интуитивно, а тоа и ќе го видиме понатаму, позгодно е да се работи со густината на распределба отколку со функцијата на распределба. Во многу практични ситуации, густината на распределба за релевантната случајна променлива се обезбедува полесно. Понекогаш тоа се прави приближно, користејќи симулации.



Слика 4.4 *Пресметка на веројатност преку густина на распределба*

Да забележиме дека кога едната од нив, функцијата или густината на распределба е обезбедена, другата може лесно да се добие со диференцирање или интегрирање на другата.

ПРИМЕР 4.9 Доцнењето на воз во минути е дадено со следната густина на распределба

$$f(x) = \begin{cases} \frac{a}{540}(39 - x^2), & \text{за } -3 \le x \le 6 \\ 0, & \text{во спротивно} \end{cases}.$$

Определи го a така што $f(x)$ да биде густина на распределба? Колкава е веројатноста возот да доцни повеќе од 2 минути? Колкава е веројатноста возот да дојде предвреме?

Решение

За $f(x)$ да биде густина на распределба, потребно е и доволно

$\int_{-\infty}^{\infty} f(x)dx = 1$, па оттука добиваме равенка по a,

$$\int_{-3}^{6} \frac{a}{540}(39 - x^2)dx = \frac{39a}{540} x \Big/_{-3}^{6} - \frac{x^3 a}{3 \cdot 540} \Big/_{-3}^{6} = a\left(\frac{351}{540} - \frac{243}{3 \cdot 540}\right) = 1,$$

од што следува $a = 2$.

Веројатноста возот да доцни повеќе од 2 минути е

$$p(x > 2) = \int_2^6 \frac{2}{540}(39 - x^2)dx = \frac{78}{540} x \Big/_2^6 - \frac{2x^3}{3 \cdot 540} \Big/_2^6 = \frac{520}{1620} = 0.3210.$$

Веројатноста возот да дојде предвреме е

$$p(x<0)=\int_{-3}^{0}\frac{2}{540}(39-x^2)dx=\frac{78}{540}x\Big|_{-3}^{0}-\frac{2x^3}{3\cdot 540}\Big|_{-3}^{0}=\frac{216}{540}=0.4000\,.\ \blacksquare$$

Како и во случај на дискретни случајни променливи, некои непрекинати случајни променливи и нивните закони на распределба се исклучително важни во практиката и се темел на статистичките оценки и тестови.

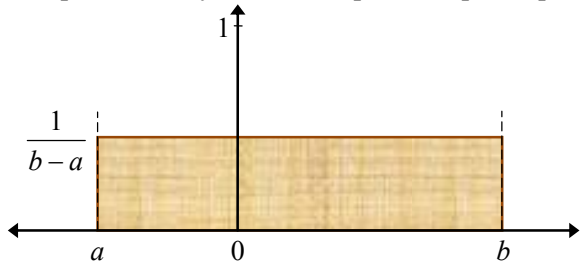
РАМНОМЕРНА РАСПРЕДЕЛБА. Густината на рамномерната распределба, како што името сугерира, е едноставно константа на определен интервал, т.е.

$$f(x)=\begin{cases} c, & \text{за } a\le x\le b \\ 0, & \text{во спротивно}\end{cases}$$, и оттука веднаш добиваме

$$\int_a^b c\,dx=1 \quad\Rightarrow\quad c=\frac{1}{b-a}$$ што ја дава густината

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & \text{за } x\in[a,b] \\ 0, & \text{во спротивно}\end{cases}.$$

На сл. 4.5 е прикажана густината на рамномерната распределба.



Слика 4.5 *Густина на рамномерна распределба*

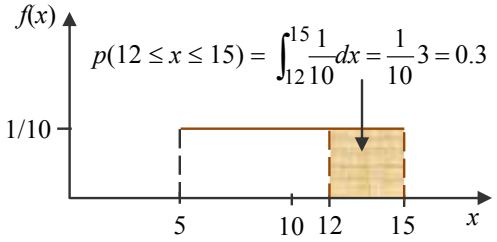
ПРИМЕР 4.10 На гостите на еден ресторан им се наплатува количината на салата што ја земаат. Испитување на примерок сугерира дека количеството земена салата е со приближно рамномерна распределба меѓу 5 и 15 унци

$$f(x)=\begin{cases}\dfrac{1}{10}, & \text{за } x\in[5,15] \\ 0, & \text{во спротивно}\end{cases}$$

каде што x = тежина на салатата во чинија. Колкава е веројатноста гостин да нарача повеќе од 12 унци салата?

Решение

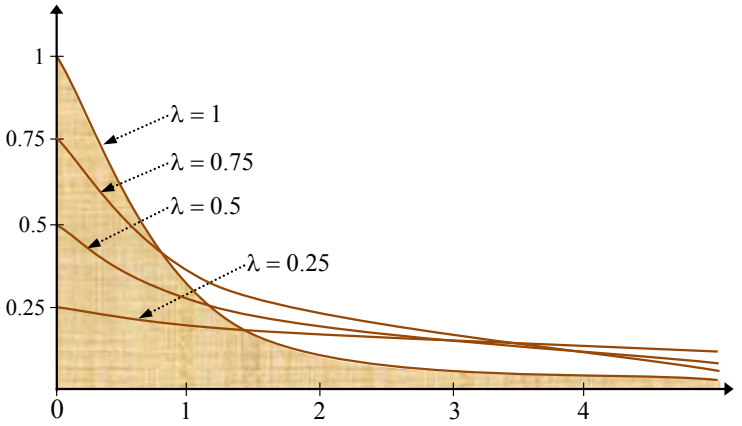
Решението е дадено на следната слика



ЕКСПОНЕНЦИЈАЛНА РАСПРЕДЕЛБА. Експоненцијалната распределба ја користи функцијата $\lambda e^{-\lambda x}$, т.е. дел од неа за кој областа под неа и x оската има плоштина 1. Густината на експоненцијалната распределба е дадена со

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{за } x \geq 0 \\ 0, & \text{во спротивно} \end{cases}$$

На сл. 4.6 се прикажани функции на експоненцијалната распределба за неко вредности на λ.



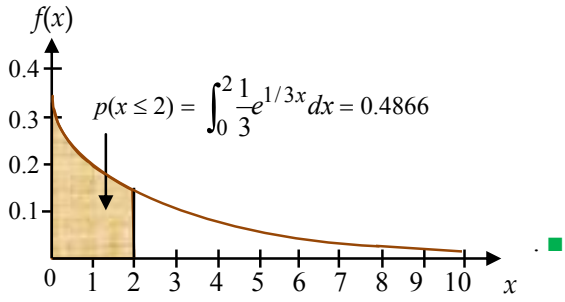
Слика 4.6 *Густина на експоненцијална распределба*

Генерално, експоненцијалната распределба ги опишува процесите во кои настаните се случуваат континуирано и независно во правилни временски растојанија. Се разбира, таа не треба да се меша со фамилијата експоненцијални распределби кои покрај неа, вклучува многу други распределби како нормалната, пуасоновата, гама распределбата итн.

ПРИМЕР 4.11 Пристигнувањето на автомобилите во еден сервис за миење има експоненцијална распределба со просечно време на пристигнување од 3 минути. Колкава е веројатноста времето меѓу две сукцесивни пристигнувања да биде помало од 2 минути?

Решение

Решението е дадено на следната слика.



■

Експоненцијалната распределба игра голема улога во доверливоста на уредите и системите. На пример, животниот век на полупроводниците е експоненцијална случајна променлива со очекување од 40000 часови. Недостатокот на "мемориска особина" на експоненцијалната распределба повлекува дека уредите "не стареат". Тоа значи дека без разлика колку долго уредот работел претходно, веројатноста на откажување во следните, на пример 1000 часа, останува иста. Затоа, уредите што откажуваат со случајни кварови може да се моделираат со експоненцијална распределба. Од друга страна, уредите што покажуваат постепена деградација "стареат" подобро се моделираат со Веибул распределба, којашто ќе ја разгледаме понатаму.

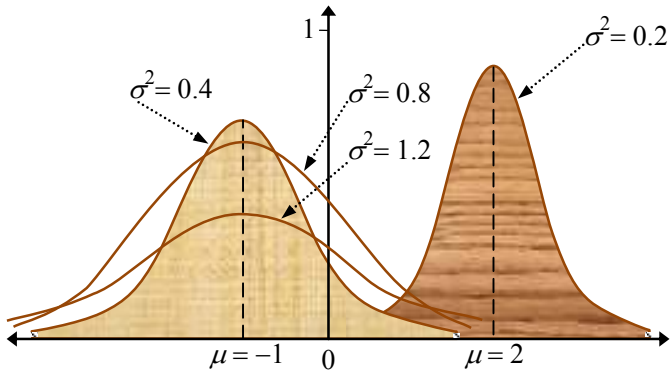
НОРМАЛНА РАСПРЕДЕЛБА. Нормалната (или Гаусовата) распределба е базична и сигурно најкористената распределба во статистиката. Таа ги опишува случајните променливи што имаат тенденција да ги групираат вредностите околу една просечна вредност. Нејзината важност доаѓа од централната гранична теорема што кажува дека при одредени услови, сума на доволен број случајни променливи со произволна распределба има приближно нормална распределба.

Густината на нормалната распределба е дадена со

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

каде што μ е просекот (локацијата на максимумот), а σ^2 е дисперзијата (ширината на распределбата). Просекот и дисперзијата се броеви што ја изразуваат просечната вредност (просекот) и раштрканоста околу просечната вредност (дисперзијата) на една случајна променлива. Тие формално ќе бидат воведени во следната глава. Нормалната распределба вообичаено ќе ја означуваме со $Z(\mu, \sigma^2)$.

На сл. 4.7 се прикажани функции на нормалната распределба за различни вредности на μ и σ.



Слика 4.7 *Густина на нормална распределба*

Густината на нормалната распределба е крива во облик на звоно (bell-like). Поголемо σ води кон посплескано звоно. Во пракса обично се прави смената

$$z = \frac{x - \mu}{\sigma},$$

така што новата случајна променлива има нормална распределба со $\mu = 0$ и $\sigma = 1$, и го добива едноставниот облик

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

кој вообичаено се нарекува стандардна нормална распределба $Z(0,1)$.

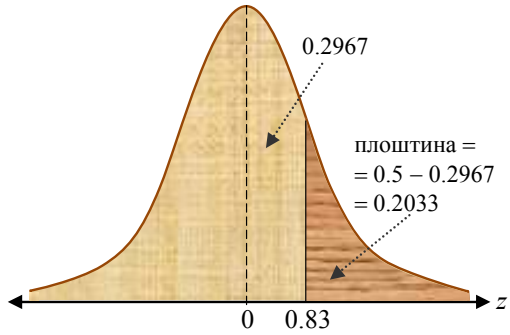
Со оглед на тоа што интегралот на нормалната распределба не е решлив преку примитивна функција, за исчитување на неговите вредности се користат готови табели или подобро, соодветен софтвер.

ПРИМЕР 4.12 Сервис за автомобили продава популарно масло за мотори. Кога резервите ќе паднат на 20 галони се прави нарачка на масло. Менаџерот е

загрижен дека лесно може да се случи недостаток на масло додека не дојде нарачката. Проценето е дека испораките на маслото се со нормална распределба со просек од 15 и дисперзија од 36 галони. Колкава е веројатноста дека ќе се случи недостаток на масло, додека дојде нарачката, т.е. $p(x > 20)$?

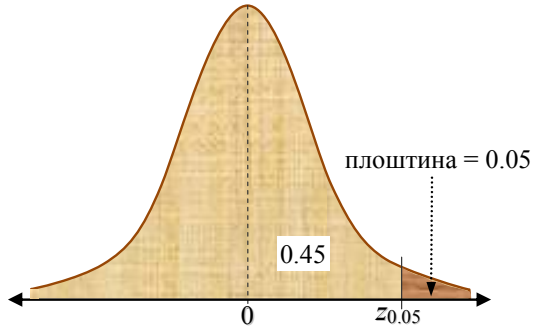
Решение

Преминуваме кон стандардна нормална распределба $z = \dfrac{20 - 15}{6} = 0.83$.



Од таблицата за нормална распределба (додаток B) ја читаме плоштината под кривата од 0 до 0.83 и оттука (види ја сликата) веднаш се добива $p(x > 20) = 0.2033$.

Оваа веројатност му изгледа на менаџерот релативно висока, и тој би сакал веројатноста да се случи недостаток на масло да не биде поголема од 0.05. Тогаш треба да се најде точката од која надесно плоштината межу кривата и x-оската биде 0.05 (види ја сликата).



Од таблицата добиваме $z_{0.05} = 1.645$, па оттука ја добиваме вредноста

$$x = \mu + \sigma \cdot z_{0.05} = 15 + 6 \cdot 1.645 = 24.87$$

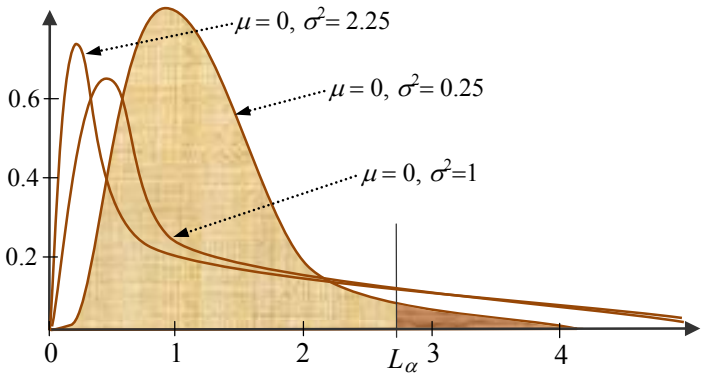
што значи дека со правење нарачки кога резервите ќе паднат на 24.87 галони, веројатноста да се случи недостаток на масло е само 0.05. ■

ЛОГНОРМАЛНА РАСПРЕДЕЛБА. Променливите што се користат во инженерските системи често пати следат експоненцијални зависности. Ако случајната променлива Z има нормална распределбата, тогаш случајната променлива $L = e^Z$ има логнормална распределба. Името доаѓа оттаму што логаритамот на логнормалната распределба L дава нормална распределба, т.е. $Z = \ln L$. Ако μ е просекот, а σ^2 дисперзијата на Z, густината на L е

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad 0 < x < \infty.$$

Забележи дека μ и σ^2 се параметри на нормалната распределба. Соодветните параметри на логнормалната распределба се функции од μ и σ^2 и ќе ги дадеме понатаму.

Животот на производите коишто деградираат со време често се моделира со логнормална распределба. Типичен пример е животот на полупроводнички ласер. За вакви ситуации се користи и Веибул распределбата којашто со погоден избор на параметри може да ја апроксимира логнормалната.



Слика 4.8 *Густина на логнормална распределба*

Сепак, логнормалната распределба е изведена од нормалната со едноставна експоненцијална функција и како таква е лесна за разбирање и пресметки на веројатностите.

ПРИМЕР 4.13 Животниот век на полупроводнички ласер има логнормална распределба со $\mu = 10$ часови и $\sigma = 1.5$ часа. Пресметај ја веројатноста животниот век да надмине 10000 часови. Кој е животниот век што до надминуваат 99% од ласерите?

Решение

Од функцијата на распределба на L добиваме

$$p(L > 10000) = 1 - p(e^Z \le 10000) = 1 - p(Z \le \ln 10000) =$$

$$1 - p\left(\frac{Z-10}{1.5} \le \frac{\ln 10000 - 10}{1.5}\right) = 1 - p(z \le -0.52) = 1 - 0.30 = 0.70.$$

За второто прашање имаме

$$p(L > x) = 1 - p(e^Z \le x) = 1 - p(Z \le \ln x) = 1 - p\left(\frac{Z-10}{1.5} \le \frac{\ln x - 10}{1.5}\right) = 0.99$$

и сега од таблицата за стандардна нормална распределба читаме дека

$$\frac{\ln x - 10}{1.5} = -2.33$$

што повлекува $x = e^{6.505} = 668.48$ часови. ■

ГАМА РАСПРЕДЕЛБА. Гама распределбата е базирана на гама функцијата што се дефинира со

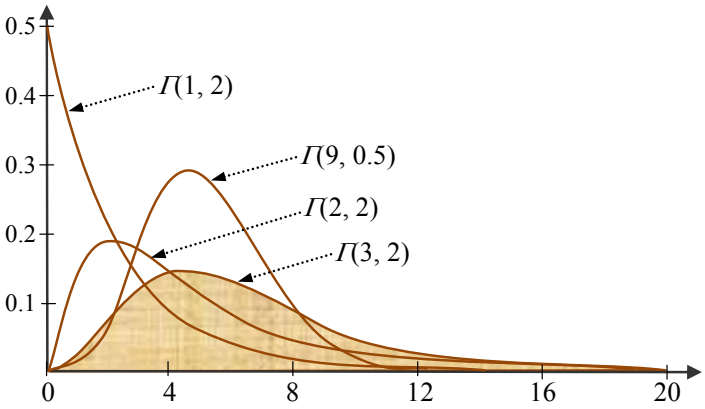
$$\Gamma(\alpha) = \int_0^\alpha x^{\alpha-1} e^{-x} dx,$$

и сега имајќи предвид дека $\int_0^\alpha \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} dx = 1$, веднаш добиваме дека подинтегралната функција е некоја густина на распределба. Со зголемување на флексибилноста преку додавање на уште еден параметар β, се доаѓа до густината на гама распределбата

$$\Gamma(\alpha, \beta) = f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}.$$

Инаку за аргумент цел број m, гама функцијата има вредност $\Gamma(m) = (m - 1)!$ што директно следи од особината $(\alpha - 1)\Gamma(\alpha - 1) = \Gamma(\alpha)$, додека $\Gamma(1/2) = \sqrt{\pi}$. Во многу ситуации гама функцијата се користи за претставување на факториелите.

Да забележиме дека експоненцијалната распределба со $\lambda = 1/\beta$ е специјален случај на гама распределба и се добива од неа ако се стави $\alpha = 1$. Од друга страна, сума на n случајни променливи со експоненцијална распределба со параметар β даваат гама распределба од облик $\Gamma(n, \beta)$.



Слика 4.9 *Густина на гама распределба*

ПРИМЕР 4.14 Должината на престој X во болница поради одредена болест има гама распределба со параметри $\alpha = 2$, $\beta = 1/3$. Цената на третманот е $50X^2 + 500X$. Колкава е веројатноста третмант да чини помалку од 300$.

Решение

Од цената на чинење ќе ја добиеме вредноста за X. Од условот за X имаме $50X^2 + 500X \le 550$.

Позитивното решение на квадратната равенка $50X^2 + 500X - 550 = 0$ е $X = 1$. Забележи дека гама распределбата зема само позитивни вредности.

Густината на гама распределбата за $\alpha = 2$, $\beta = 1/3$ е

$$f(x) = 9xe^{-3x},$$

па се бара веројатноста

$$p(X \le 1) = \int_0^1 9xe^{-3x}dx = -3xe^{-3x}\Big|_0^1 - e^{-3x}\Big|_0^1 = -\frac{3}{e^3} - \frac{1}{e^3} + 1 = 0.8001.$$

Оваа веројатност е малку превисока, па можеби не е добра или распределбата или пресметката на цената на третманот. Да претпоставиме дека имаме гама распределба но со параметри $\alpha = 2$, $\beta = 1/2$. Тогаш добиваме

$$p(X \le 1) = \int_0^1 4xe^{-2x}dx = -2xe^{-2x}\Big|_0^1 - e^{-2x}\Big|_0^1 = -\frac{2}{e^2} - \frac{1}{e^2} + 1 = 0.5940,$$

додека за $\alpha = 2$, $\beta = 3/4$, веројатноста паѓа на

$$p(X \le 1) = \int_0^1 \frac{16}{9}xe^{-(4/3)x}dx = -\frac{4}{3}xe^{-(4/3)x}\Big|_0^1 - e^{-(4/3)x}\Big|_0^1 = -\frac{7}{3e^{(4/3)}} + 1 = 0.3849. \blacksquare$$

ВЕИБУЛ (Weibull) РАСПРЕДЕЛБА. Веибул распределбата комбинира на едноставен начин полиномна и експоненцијална функција за да се добие флексибилна густина на распределба. Еден од начините за нејзино дефинирање е преку рамномерната распределба. Имено, ако случајната променлива X има рамномерна распределба, тогаш случајната променлива

$$W = \alpha(-\ln X)^{1/\beta}$$

има веибул распределба со параметри α и β. Општиот облик на густината на веибул распределбата вообичаено користи 3 параметри α, β и γ и го има следниот облик

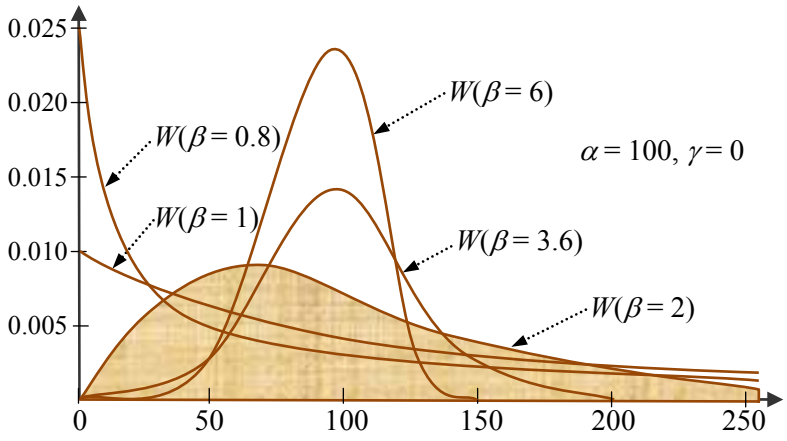
$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}}, x \geq 0, \alpha \geq 0 \text{ и } \beta \geq 0,$$

каде што:

β е параметар на облик, т.е раст и пад на кривата,
α е параметар на скалирање и
γ е параметар на локација.

Често пати γ се зема 0, и тогаш густината тргнува од 0 надесно.

Ако се стави $\beta = 1$, $\gamma = 0$, добиваме експоненцијална распределба со параметар $1/\alpha$. Од друга страна, ако се стави $\beta = 2$, $\gamma = 0$ се добива $\chi 2$ распределба со $n = \beta = 2$ и $\sigma = \alpha$.



Слика 4.10 *Густина на Веибул распределба*

За пресметка на веројатностите корисна е кумулативната функција на распределба, што за Веибул распределбата со $\gamma = 0$ е

$$F(x) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}.$$

Густината на Веибул распределбата има сличности со густината на гама распределбата.

Веибул распределбата се користи во анализа на опстанок на биолошките организми, во испитувања на доверливост на инжинерските системи и анализа на грешките, во индустриското инженерство за претставување на времињата на производство и испорака, во климатските испитувања, во безжичните комуникации, хидрологија итн.

ПРИМЕР 4.15 Во еден членак од индустриско инженерство, авторите сугерираат користење на Веибул распределбата за проценка на траењето на изработката на плочката во производството на полупроводник. Ако со X го означиме ова време за произволно избрана плочка, X има Веибул распределба со $\alpha = 10$, $\beta = 0.3$. Колкава е веројатноста траењето на исработката да биде преку 5 часа? Колкава е веројатноста траењето на исработката да биде од 2 до 7 часа?

Решение

Се бараат веројатности од следната густина на распределба

$$f(x) = \frac{0.3}{10}\left(\frac{x}{10}\right)^{-0.7} e^{-\left(\frac{0.3}{10}\right)^{0.3}}.$$

Оттука имајќи ја предвид кумулативната функција на распределба за Веибул распределбата добиваме

$$p(X \le 5) = 1 - e^{-\left(\frac{5}{10}\right)^{0.3}} = 1 - 0.4439 = 0.5561.$$

Втората веројатност ја наоѓаме на сличен начин

$$\mathrm{p}(2 \le X \le 7) = p(X \le 7) - p(X \le 2) = 1 - e^{-\left(\frac{7}{10}\right)^{0.3}} - 1 + e^{-\left(\frac{2}{10}\right)^{0.3}} =$$

$$= -0.4072 + 0.5395 = 0.1323. \blacksquare$$

ХИ-КВАДРАТ РАСПРЕДЕЛБА. Сума на квадрати на n случајни променливи со стандардна нормална распределба ($\mu = 0$, $\sigma = 1$) дава случајна променлива со $\chi 2$ (хи-квадрат) распределба со n степени на слобода.

Формално, за заедничката густина на распределба на n независни случајни променливи со стандардна нормална распределба може да ставиме

$$p(Q)dQ = \int_D \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} dx_i = \int_D \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} dx_1 dx_2 \ldots dx_n$$

каде што D е површина на инфенитезималната сфера определена со $Q = \sum_{i=1}^{n} x_i^2$ и радиус $R = \sqrt{Q}$, т.е. $dR = (1/2\sqrt{Q})dQ$. Подинтегралниот експонент едноставно се изразува со Q и е константа во интегралот што дава

$$p(Q)dQ = \frac{e^{-Q/2}}{(2\pi)^{n/2}} \int_D dx_1 dx_2 \ldots dx_n \ .$$

Плоштината на n-димензионална сфера е $\frac{nR^{n-1}\pi^{n/2}}{\frac{n}{2}\Gamma(\frac{n}{2})}$, што дава

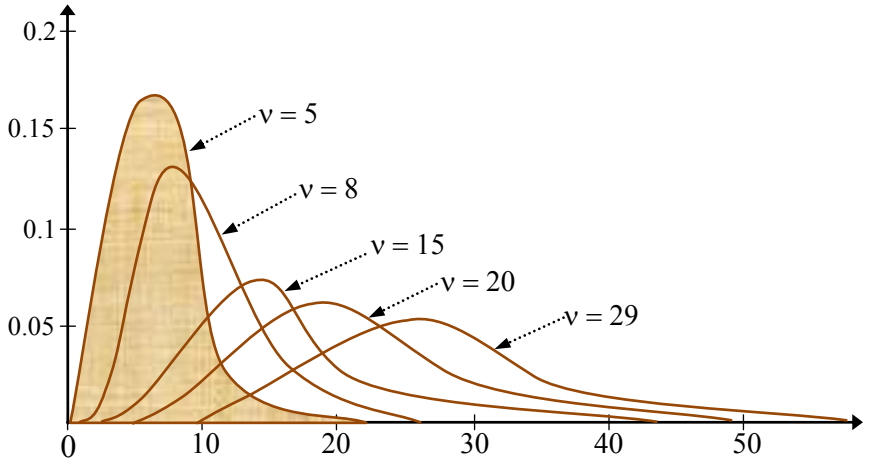
$$p(Q)dQ = \frac{e^{-Q/2}}{(2\pi)^{n/2}} \frac{nR^{n-1}\pi^{n/2}}{\frac{n}{2}\Gamma(\frac{n}{2})} dR \ = \ \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} Q^{n/2-1} e^{-Q/2} dQ \ .$$

Значи густината на $\chi 2$ распределбата е $f(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$.

Степените на слобода опишуваат колку од параметрите се слободни да варираат. На пример, ако разгледуваме 3 коцки и вкупниот збир 13, тогаш со случаен избор на два броја третиот е веќе одреден до збир 13. Значи од 3-те коцки, две може да варираат, а третата е секогаш определена од другите 2, па имаме 2 степени на слобода. Во многу ситуации, бројот на степени на слобода е еднаков на бројот на параметри намален за 1. Ова не е случај кај $\chi 2$ распределбата бидејќи сите n случајни променливи што се во сумата на квадрати може да варираат, па бројот на степени на слобода е n.

Хи-квадрат распределбата е специјален случај на гама распределбите и се добива за $\alpha = n/2$ и $\beta = 2$.

На сл. 4.11 се прикажани густини на $\chi 2$ распределба за различни степени на слобода (ν).



Слика 4.11 *Густина на хи-квадрат распределба*

$\chi 2$ распределбата се користи за тестирање на случајните променливи на независност, хомогеност, квалитет на апроксимација (најмали квадрати), итн.

ПРИМЕР 4.16 Компанија за производство на батерии за мобилни телефони има развиено нов тип литиумска батерија што во просек трае 250 часа со стандардна девијација од 14 часа. Траењето на батериите има нормална распределба. При контрола на квалитетот случајно се избрани и испитани 7 батерии и добиена е стандардна девијација од 22 часа. Истиот тест е повторен на други случајно избрани 7 батерии. Колкава е веројатноста дека стандардната девијација на повториот тест да биде поголема од 22?

Решение

Ако случајната променлива Y_i = "траење на батерија", тогаш сумата на квадратите на $\frac{Y_i - \mu}{\sigma}$ каде што μ е просекот, а σ стандардната девијација (квадратен корен од дисперзијата), е сума на квадрати на случајни променливи со стандардна нормална распределба, т.е. случајна променлива X со $\chi 2$ распределба. Случајната променлива X може да се претстави со

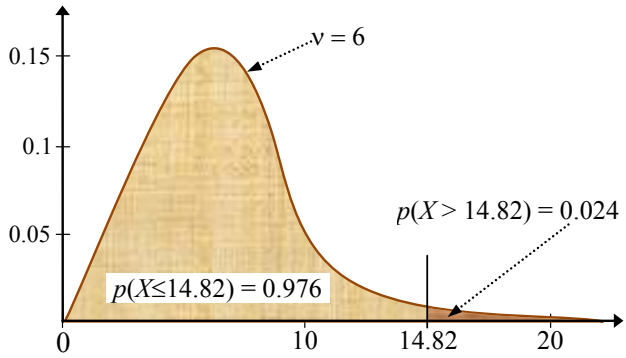
$$X = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2 = (n-1) \frac{1}{\sigma^2} \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \mu)^2 = (n-1) \frac{S^2}{\sigma^2},$$

каде што S^2 е случајна променлива $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \mu)^2$ што ја дава дисперзијата на траењето на батериите. Нејзиниот корен е стандардната девијација.

Во нашиот случај имаме дека бројот на степени на слобода е 7 – 1 = 6, $\sigma =$ 14 и $S = 22$ и оттука добиваме

$$X = 6\frac{22^2}{14^2} = 14.8163.$$

Значи X има $\chi 2$ распределба со 6 степени на слобода и ние ја бараме веројатноста $p(X > 14.8163)$, како што тоа е прикажано на следната слика.



Значи веројатноста стандардната девијација при повториот тест да биде поголема од 22 е само 2.4%. ■

СТУДЕНТОВА РАСПРЕДЕЛБА. Ако случајната променлива X има стандардна нормалната распределба, а Y има $\chi 2$ распределба со n степени на слобода и ако X и Y се независни, тогаш случајната променлива T дефинирана со

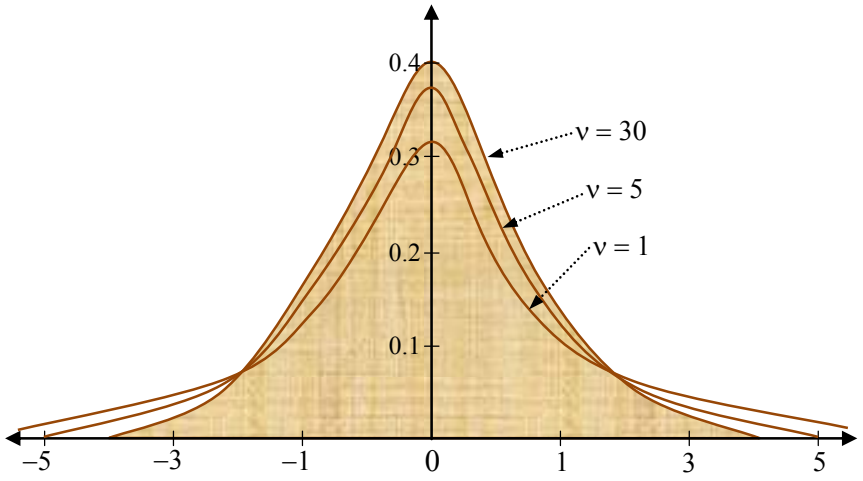
$$T = \frac{X\sqrt{n}}{\sqrt{Y}}$$

има студентова распределба со n степени на слобода. Густината на студентовата распределба е

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$$

што бара користење на таблица за исчитување на нејзините интеграли (веројатности).

Генералниот облик на густината на студентовата распределба е сличен на густината на стандардната нормална распределба ($\mu = 0$ и σ = 1), со разлика дека е малку поспуштена и поширока.



Слика 4.12 *Густина на студентова распределба*

Со растење на бројот на степени на слобода, *T*-рапределбата се приближува кон стандардната нормална распределба.

ПРИМЕР 4.17 Компанија произведува сијалици. Таа тврди дека просечниот век на нивните сијалици е 300 денови. Контролата на квалитет избрала случајно 15 сијалици за тестирање. Овие сијалици просечно траеле 290 денови со стандардна девијација од 50 денови. Ако тврдењето на компанијата е точно, колкава е веројатноста дека 15 случајно избрани сијалици ќе имаат просечен век на траење не поголем од 290 денови?

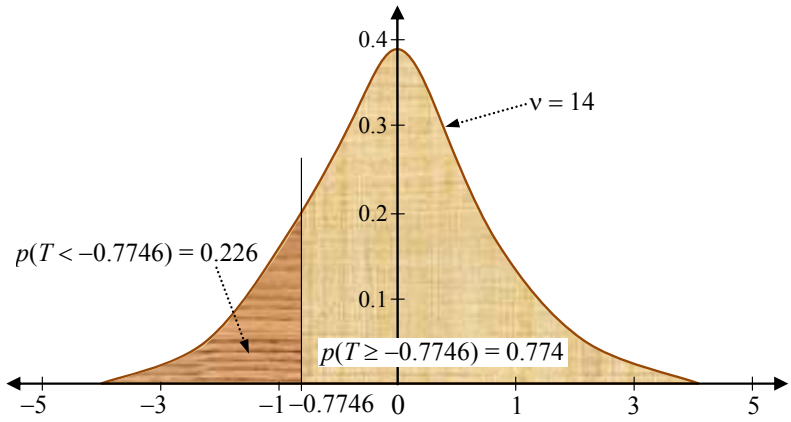
Решение

Ако случајната променлива X = "просечен век на траење на сијалица", таа има нормална распределба со просек μ и стандардна девијација $\sigma/\sqrt{n}$ додека $(n-1)S/\sigma$ има $\chi 2$ распределба со $n-1$ степени на слобода. Тогаш, случајната променлива T дадена со

$$T = \frac{\dfrac{X-\mu}{\sigma/\sqrt{n}}\sqrt{n-1}}{\sqrt{n-1}\,\dfrac{S}{\sigma}} = \frac{X-\mu}{S}\sqrt{n}$$

има студентова распределба со n–1 степени на слобода. Во нашиот пример имаме

$$T = \frac{290-300}{50}\sqrt{15} = -0.7746 \quad \text{со 14 степени на слобода.}$$



Така добиваме дека веројатноста просечниот век на траење на сијалиците да биде не поголем од 290 денови е плоштината под густината на студентовата распределба налево од –0.7746 и изнесува 22.6%. ■

ФИШЕРОВА РАСПРЕДЕЛБА. Ако случајната променлива X има $\chi 2$ распределба со n_1 степени на слобода, а Y има $\chi 2$ распределба со n_2 степени на слобода и ако X и Y се независни, тогаш случајната променлива F дефинирана со

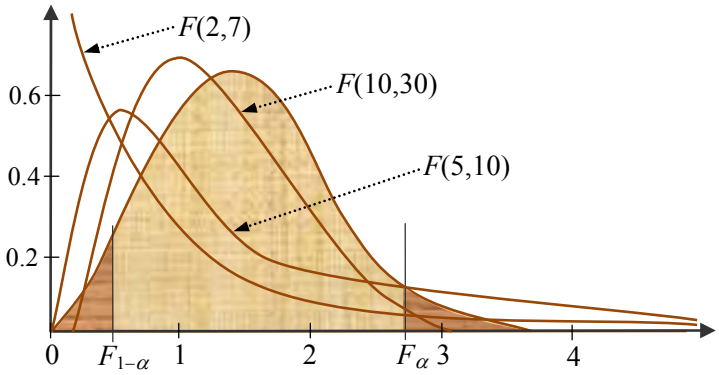
$$F = \frac{X/n_1}{Y/n_2} = \frac{X \cdot n_2}{Y \cdot n_1}$$

има фишерова распределба со (n_1, n_2) степени на слобода. Густината на фишеровата распределба е релативно комплицирана

$$f(x) = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} n_1^{\frac{n_2}{2}} n_2^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} (n_2+n_1x)^{-\frac{n_1+n_2}{2}}, \quad x \geq 0.$$

Јасно е дека ако случајната променлива F има фишерова распределба, тогаш и $1/F$ има исто така фишерова распределба. Ако пак T има студентова распределба со n степени на слобода, тогаш T^2 има фишерова распределба $F(1, n)$.

На сл. 4.13 е прикажана типична густина на фишеровата распределба за некои вредности на n_1 и n_2. Очигледно, нејзиниот облик за определени вредности на параметрите n_1 и n_2 наликува на $\chi 2$ распределбата иако, се разбира, таа се добива како нивен количник и како таква има 2 параметра – степени на слобода.



Слика 4.13 *Густина на фишерова распределба*

F распределбата се користи за споредба на случајни променливи со $\chi 2$ распределба. Имено, разликата меѓу 2 случајни променливи со $\chi 2$ распределба не дава $\chi 2$ распределба, па оттука наместо разлика се разгледува однос. Така, F распределбата може да се користи за проверка која од 2 случајни променливи има поголема или помала дисперзија.

ПРИМЕР 4.18 Од група на жени и мажи пливачи, на случаен начин се избрани 7 жени и 12 мажи со цел да се испитаат варијациите во нивните висини. Од порано познатите стандардни девијации заедно со стандардните девијации за избраните мажи и жени се дадени во следната табела:

Популација	*стандардна девијација*	*стандардна девијација на избраните*
жени	30	35
мажи	50	45

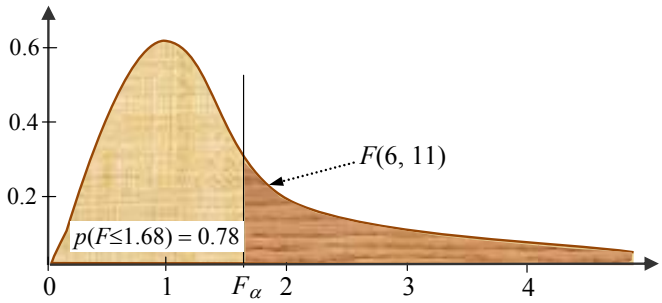
Колкава е веројатноста дека стандардната девијација кај жените е поголема од онаа кај мажите?

Решение

Ако случајната променлива X = "стандардна девијација во висината на жените пливачи", а Y = "стандардна девијација во висината на мажите пливачи" тогаш случајната променлива

$F = \dfrac{X^2}{Y^2}$ има фишерова распределба со (6, 11) степени на слобода.

Нејзината вредност ја добиваме со $F = \dfrac{(X/\sigma_1)^2}{(Y/\sigma_2)^2} = \dfrac{(35/30)^2}{(45/50)^2} = 1.68$ и оттука добиваме веројатност од 0.78.



Ако пак пресметуваме реципрочно, добиваме фишерова распределба со (11, 6) степени на слобода и вредност

$$F = \frac{(Y/\sigma_2)^2}{(X/\sigma_1)^2} = \frac{(45/50)^2}{(35/30)^2} = 0.595\,,$$

што исчитувајќи од таблица дава веројатност 0.22. ■

БЕТА РАСПРЕДЕЛБА. Ако случајната променлива X има има гама распределба $\Gamma(\alpha, \theta)$, а случајната променлива Y има гама распределба $\Gamma(\beta, \theta)$, и ако X и Y се независни, тогаш случајната променлива B дефинирана со

$$B(\alpha, \beta) = \frac{X}{X+Y}$$

има бета распределба.

Густината на бета распределбата е дадена со

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,,\ x \geq 0.$$

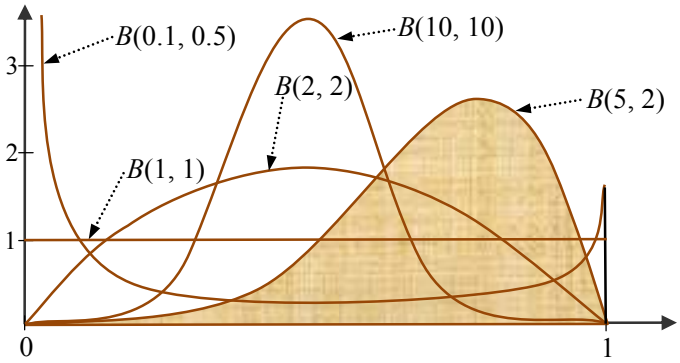
од што веднаш се гледа дека

$$B_{\beta,\,\alpha}(x) = B_{\alpha,\,\beta}(1-x).$$

Врската меѓу фишеровата и бета распределба е дадена со

$$\frac{\frac{n_1}{n_2}F}{1+\frac{n_1}{n_2}F} = B\left(\frac{n_1}{2}, \frac{n_2}{2}\right).$$

На сл. 4.14 е прикажана неколку густини на бета распределбата за некои вредности на α и β.



Слика 4.14 *Густина на бета распределба*

Бета распределбата се користи за моделирање на настани што се случуваат во определен (временски) интервал со минимална и максимална вредност. Такви се проектите поврзани со системите на планирање - контрола каде што бета распределбата се користи за опис на потребното време за завршување на работата.

ПРИМЕР 4.19 Процентот на нечистотија во еден производ е случајна променлива X со бета распределба со параметри $\alpha = 2$, $\beta = 3$. Производ со процент на нечистотија од преку 40% не може да се продаде. Колкава е веројатноста случајно избран производ поради нечистотија да не може да се продаде?

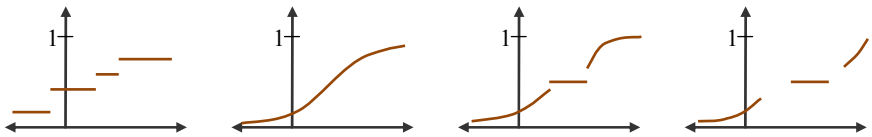
Решение

Бета распределбата на X, за $\alpha = 2$, $\beta = 3$, го има следниот облик

$$f(x) = \begin{cases} 12x^2(1-x), & \text{за } 0 \le x \le 1 \\ 0, & \text{во спротивно} \end{cases}$$

$$p(X > 0.4) = \int_{0.4}^{1} 12x^2(1-x)dx = 4x^3 \Big/_{0.4}^{1} - 3x^4 \Big/_{0.4}^{1} = 0.8208. \blacksquare$$

На крај, важно е да се напомне дека не секоја случајна променлива е од дискретен или непрекинат тип. Такви се на пример случајните променливи што во една област се дискретни, а во друга област непрекинати или пак во некои области се недефинирани.



Дискретна *Непрекината* *Мешана* *Сингуларна*

Сликата прикажува примери на функци на распределба за дискретни, непрекинати, мешани и сингуларни случајни променливи.

Дискретните случајни променливи имаат скалеста функција на распределба додека непрекинатите имаат непрекината функција на распределба што е диференцијабилна скоро секаде[3]. Мешаните случајни променливи имаат диференцијабилна по области функција на распределба со скокови на прекин. Сингуларните случајни променливи имаат диференцијабилна по области функција на распределба, којашто не е диференцијабилна скоро секаде.

Како што се гледа, случајните променливи се карактеризираат преку нивната функција на распределба. Од друга страна, функцијата на распределба е многу понезгодна за користење од законот на распределба во дискретен случај или густината на распределба во непрекинат случај. Некој би поставил прашање зошто тогаш таа воопшто се дефинира и изучува. Концептот на функција на распределба е многу општ и се применува идентично на секоја случајна променлива, дискретна, непрекината или мешана. Секоја случајна променлива X има функција на распределба дефинирана на идентичен начин $F(x) = p(X < x)$, за $-\infty < x < \infty$. Дискретните случајни променливи пак имаат закон на распределба, непрекинатите имаат густина на распределба, мешаните имаат комбинација на двете. Значи заедничките особини на сите типови случајни променливи може да се сумираат во форма на функција на распределба наместо да се користат различни формули. Сепак, изучувањето на особините на случајните променливи преку функцијата на распределба е често пати заплеткана работа и затоа се користат законите и густините на распределба. Функцијата на распределба вообичаено се користи за добивање на општи тврдења за случајните променливи.

4.3. Случајни вектори

Во многу случаи од интерес се повеќе од една случајна променлива и тоа разгледувани заеднички а не индивидуално. На пример, при попис на население, некој вообичаено е заинтересиран за повеќе случајни променливи како заработувачката, возраста, полот итн. Секоја од овие променливи е интересна, но кога 2 или повеќе се изучуваат симултано, се добиваат нови информации за населението. На пример, изучу-

[3] Терминот "скоро секаде" значи дека множеството точки каде што некоја особина не е исполнета има мера 0. (На пример, мера 0 имаат конечните и пребројивите множества, а дури и некои непребројиви).

вањето на овие променливи симултано може да даде увид во еманципацијата на жените. Во екологијата се изучуваат случајните променливи на растителните и животинските видови. На пример, симултаното изучување на случајните променливи на предаторите и пленовите даваат увид во еколошката рамнотежа. Заедничката распределба на просторните компоненти, на пример компонентите на брзината на ветровите се користи во студиите за атмосферските турболенции.

Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност и нека $X_1, X_2, ..., X_n$ се случајни променливи со домен во Ω. Тогаш, подредената n-торка $(X_1, X_2, ..., X_n)$ се нарекува случаен вектор. Како и во други случаи, заради едностаност се ограничуваме на разгледување на дводимензионалните случајни вектори, а дискусијата праволиниски се обопштува на n-димензионалните случајни вектори.

Веројатностите поврзани со случаен вектор (X, Y) вообичаено се означуваат со $p(X = x, Y = y) = p((X = x) \text{ и } (Y = y))$.

За случајниот вектор (X, Y) може да се дефинира следната функција на распределба (од 2 променливи)

$$F(x, y) = p(X < x, Y < y)$$

определена на целата $\mathbb{R}^2$, со особини аналогни на секоја функција на распределба. Имено, важи:

1) $0 \le F(x, y) \le 1$,

2) $F(x, y)$ е монотоно неопаѓачка по секој од аргументите,

3) $F(x, y)$ е непрекината од лево по секој аргумент,

4) $F(x, -\infty) = F(-\infty, y) = 0$, $F(+\infty, +\infty) = 1$,

5) ако $F_1(x)$ и $F_2(x)$ се функции на распределба на X и Y соодветно, тогаш $F(x, +\infty) = F_1(x)$ и $F(+\infty, y) = F_2(y)$.

6) $p(X < x, c \le Y < d) = F(x, d) - F(x, c)$ за $c < d$ и
$p(a \le X < b, c \le Y < d) = F(b, d) - F(b, c) - F(a, d) + F(a, c)$
за $a < b$ и $c < d$.

Случајниот вектор (X, Y) е дискретен, ако се дискретни случајните променливи X и Y. Во дискретен случај, законот на распределба на случајниот вектор (X, Y) е даден со матрица

$$p_{ij} = p(X = x_i, Y = y_j), \quad \sum p_{ij} = 1 .$$

Сега, која било веројатност се добива со сумирање на соодветните елементи од матрицата на веројатности

$$p(a \le X \le b,\ c \le Y \le d) = \sum_{\substack{i,j:a\le x_i\le b,\\ c\le y_j\le d}} p_{ij} \ .$$

Засебните закони на распределба на X и Y праволиниски се добиваат од заедничкиот закон $p(X = x_i,\ Y = y_j)$ со

$$p(X = x_i) = \sum\nolimits_j p_{ij} \quad \text{и} \quad p(Y = y_j) = \sum\nolimits_i p_{ij} \ .$$

ПРИМЕР 4.20 Двајца експерти оценувале еден производ при што првиот на примероците на призводите им давал оценка од 1 до 5 (случајна променлива X), а вториот од 1 до 3 (случајна променлива Y). Заедничкиот закон на распределба на дадените оценки на производите е даден во следната табела:

		X					
		1	2	3	4	5	
Y	1	0.02	0.04	0.05	0.04	0.01	0.16
	2	0.05	0.10	0.13	0.12	0.07	0.47
	3	0.01	0.05	0.08	0.10	0.13	0.37
		0.08	0.19	0.26	0.26	0.21	

На пример, веројатноста првиот експерт да го оценил производот со 4, а вториот со 2 е $p(X = 4,\ Y = 2) = 0.12$.

Да се најдат законите на распределба на X и Y.

Решение

Индивидуалните закони на распределба на X и Y се добиваат со сумирање на веројатностите по колони за X и по редици за Y (тоа е веќе направено во горната табела).

X	1	2	3	4	5
$p(X = x_i)$	0.08	0.19	0.26	0.26	0.21

Y	1	2	3
$p(Y = x_i)$	0.16	0.47	0.37

. ■

Случајниот вектор $(X,\ Y)$ се смета за непрекинат, ако постои ненегативна функција $f(x, y)$, таква што за секој x, y важи

$$F(x,y) = p(X < x, Y < y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$$

што повлекува $f(x,y) = \dfrac{\partial^2 F(x,y)}{\partial x \partial y}$.

Функцијата $f(x,\ y)$ е заедничка густина на распределбата на случајните променливи X и Y, т.е. густина на распределба на случајниот вектор $(X,$

Y). Секогаш претпоставуваме дека $f(x, y)$ е интеграбилна функција. Јасно е дека

$$p((x,y) \in D) = \iint_D f(x,y)dxdy .$$

Индивидуалните функции на распределба $F_1(x)$ и $F_2(x)$, како и густините на распределба $f_1(x)$ и $f_2(x)$ се добиваат од заедничката со

$$F_1(x) = p(X < x) = F(x,+\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)dudv \quad \text{и}$$

$$F_2(x) = p(Y < x) = F(+\infty,x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)dvdu \text{ , т.е.}$$

$$f_1(x) = \int_{-\infty}^{\infty} f(x,v)dv \quad \text{и} \quad f_2(x) = \int_{-\infty}^{\infty} f(u,x)du .$$

Кога распределбата на случајниот вектор (X, Y) е позната, распределбата на секоја од компонентните случајни променливи може да се добие. Овие индивидуални распределби на случајните променливи што се елементи на случајниот вектор се нарекуваат *маргинални* распределби. Во општ случај, обратното не важи. Но ако компонентните случајни променливи се независни, тогаш маргиналните распределби ја определуваат распределбата на случајниот вектор.

ПРИМЕР 4.21 Во некој регион познато е дека има 4 видови карпи. При испитувањата, земани се примероци (карпести структури) со одреден волумен. Нека X_1, X_2 и X_3 бидат пропорциите на волумените на карпите од вид 1, 2 и 3 во случајно избран примерок (пропорцијата за видот 4 е редундантна и се добива од $X_4 = 1 - X_1 - X_2 - X_3$). Заедничката густина на распределбата на волумените на карпите е

$f(x, y, z) = kxy(1 - z)$, $x, y, z \in [0, 1]$, $x + y + z \le 1$;

$f(x, y, z) = 0$, во спротивно.

Определи го k? Најди ја веројатноста во случајно избрана карпеста структура, видот 3 да учествува со најмаку 50%.

Решение

Од условот интегралот на густината на распределба да биде 1 добиваме

$$\int_0^1\int_0^1\int_0^{1-y-z} kxy(1-z)dxdydz = \frac{k}{2}\int_0^1\int_0^1 (1-y-z)^2 y(1-z)dydz =$$

$$\frac{k}{2}\int_0^1 (1-z)\left(\int_0^1 (y + y^3 + yz^2 - 2y^2 - 2yz + 2y^2 z)dy\right)dz =$$

$$\frac{k}{2}\int_0^1 (1-z)(\frac{1}{12}-\frac{1}{3}z+\frac{1}{2}z^2)dz = \frac{k}{2}(\frac{z}{12}-\frac{5z^2}{24}+\frac{5z^3}{18}-\frac{z^4}{8})\Big|_0^1 = \frac{k}{72}.$$

Оттука следува $k = 72$.

Веројатноста видот 3 да учествува со повеќе од 50% е еднаква на интегралот

$$72\int_{0.5}^1\int_0^1\int_0^{1-y-z} xy(1-z)dxdydz = 36\left(1-(\frac{1}{24}-\frac{5}{96}+\frac{5}{144}-\frac{1}{128})\right) = 0.4060. \blacksquare$$

Распределбите на случајните вектори, дури и ако се составени од случајни променливи со релативно едноставни распределби, често се доста комплицирани. Во поглавјето 5.5.3 посебно ќе ги разгледуваме случајните вектори со нормална распределба.

Досега разгледуваните случајни вектори беа дискретни (со компоненти дискретни случајни променливи) или непрекинати (со компоненти непрекинати случајни променливи). Во некои случаи може да се разгледуваат мешани случајни вектори (X, Y), каде што на пример, X е дискретна случајна променлива, а Y е непрекината. Во таков случај веројатностите би се наоѓале со сумирање на $f(x, y)$ по едната променлива и интегрирање по втората.

Постојат случаи на комплицирани случајни вектори што произлегуваат од праксата [DeGroot 1989]. На пример, нека X и Y се времиња на откажување на две специфични компоненти од некој систем. Може да постои веројатност p ($0 < p < 1$) дека двете компоненти ќе откажат во исто време и соодветна густина на распределба $f(x)$. Од друга страна, времињата на нивното откажување е случаен вектор со густина на распределба $g(x, y)$. Заедничката распределба на X и Y не е непрекината бидејќи за која било непрекината распределба, веројатноста дека (X, Y) лежи на линијата $x = y$ мора да биде 0, додека во овој пример таа веројатност е $p > 0$. Сепак, оваа дискусија наликува на парадоксите со реалните броеви дискутирани во Глава 3, бидејќи веројатноста два реални броја да се еднакви ($x = y$) треба да биде 0, или со други зборови, невозможна е истовременост на настани ако времето се мери со реални броеви.

4.3.1. Независност на случајни променливи

Нека (Ω, $\mathcal{F}$, p) е простор на веројатност и нека X и Y се случајни променливи со домен во Ω.

Дефиниција 4.3 Случајните променливи X и Y се независни акко за кој било $x, y \in \mathbb{R}$ важи $p(X < x, Y < y) = p(X < x)p(Y < y)$.

Значи случајните променливи X и Y се независни ако се независни случајните настани $(X < x)$ и $(Y < y)$. Сега ако $F(x, y)$ е заедничка функција на распределба на (X, Y), а $F_X(x)$ и $F_Y(y)$ се функции на распределба на X и Y, веднаш следува дека

$$F(x, y) = p(X < x,\ Y < y) = p(X < x)p(Y < y) = F_X(x)F_Y(y)$$

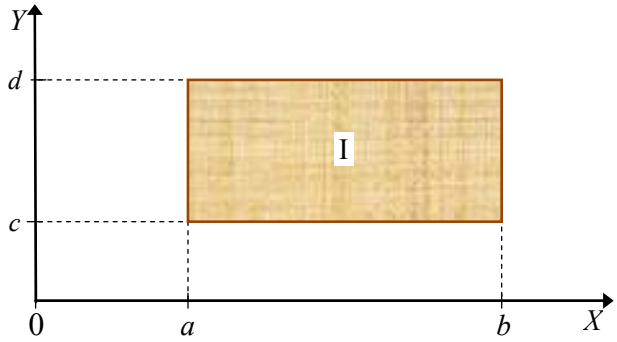
што е еквивалентно тврдење на дефиницијата 4.2. Често од полза е и следното еквивалентно тврдење.

Теорема 4.4 Случајните големни X и Y се независни акко за кои било две множества I_1 и I_2 од σ-алгебра на реалната оска важи

$$p(X \in I_1,\ Y \in I_2) = p(X \in I_1)p(Y \in I_2).$$

Доказ: Ќе докажеме дека ако случајните променливи X и Y се независни на бесконечните полуинтервали $(-\infty, x)$ според дефиницијата 4.3, тие се независни и на произволни полуинтервали. Теоремата следува од фактот што секоја σ-алгебра на реалната оска е генерирана од полуинтервали.

Нека $I = (a \le X < b,\ c \le Y < d)$, како што е прикажано на слика 4.15.



Слика 4.15 З*аедничка настан на X и Y*

Тогаш имаме дека

$$(X < b,\ Y < d) = I + (X < a,\ Y < d) + (X < b,\ Y < c)$$

и според аксиомата а3 на веројатноста важи

$$p((X < a,\ Y < d) + (X < b,\ Y < c)) =$$
$$= p(X < a,\ Y < d) + p(X < b,\ Y < c) - p(X < a,\ Y < c)$$

од што следува

$$p(I) = p(X < b,\ Y < d) + p(X < a,\ Y < c) + p(X < a,\ Y<d) - p(X < b,\ Y < c) =$$
$$= p(X < b)p(Y < d) + p(X < a)p(Y < c) +$$

$$+ p(X < a)p(Y < d) - p(X < b)p(Y < c) =$$

$$= (p(X < b) - p(X < a))\,(p(Y < d) - p(Y < c)) = p(a < X < b)\,p(c < Y < d).\ \blacksquare$$

Во дискретен случај, заедничката распределба на случајните променливи X и Y е дадено со матрица

$$p_{ij} = p(X = x_i,\ Y = y_j), \quad \sum p_{ij} = 1.$$

Тогаш X и Y се независни акко за секој i и j важи

$$p_{ij} = p_{i.} \cdot p_{.j}$$

каде што $p_{i.} = p(X = x_i) = \sum_j p_{ij}$, $p_{.j} = p(Y = y_j) = \sum_i p_{ij}$.

ПРИМЕР 4.22 Да се најдат законите на распределба на случајните променливи $S = X + Y$ и $T = X \cdot Y$ ако X и Y се независни дискретни случајни променливи со закони на распределба

X	10	11	12	16
$p(X = x_i)$	0.4	0.2	0.1	0.3

Y	2	3
$p(Y = x_i)$	0.4	0.6

Решение

Условот за независност $p(X = x_i,\ Y = x_j) = p(X = x_i)p(Y = x_j)$ овозможува лесно да се најдат бараните закони на распределба.

S	12	13	14	15	18	19
$p(S = x_i)$	0.16	0.32	0.16	0.06	0.12	0.18

T	20	22	24	30	32	33	36	48
$p(T = x_i)$	0.16	0.08	0.04	0.24	0.12	0.12	0.06	0.18

. ■

Во непрекинат случај, нека заедничката распределба на случајните променливи X и Y е дадена со заедничката густината $f_{X,Y}(x, y)$. Тогаш X и Y се независни акко во секоја точка на непрекинатост на функциите $f_{X,Y}(x, y), f_X(x)$ и $f_Y(y)$ важи $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$.

ПРИМЕР 4.23 Случајниот вектор (X, Y) е рамномерно распределен на кругот со центар во координатниот почеток и радиус r. Да се најде густината на распределба $f_{X,Y}(x, y)$ на (X, Y) и да се провери дали X и Y се независни.

Решение

Рамномерната распределба во овој случај има густина

$$f_{X,Y}(x,y)=\begin{cases} c \text{ за } x^2+y^2\le r^2 \\ 0 \text{ во спротивно}\end{cases}$$ каде што c се наоѓа од барањето

$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy=1$ од што се добива $c=\frac{1}{r^2\pi}$. Откако ја имаме заедничката густина, ги наоѓаме подинечните густини $f_X(x)$ и $f_Y(x)$. Поради тоа што $f_{X,Y}(x, y)$ е константно, двете поединечни густини се еднакви, па имаме

$$f_X(x)=\int_{-\infty}^{\infty}\frac{1}{r^2\pi}dy=\int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}}\frac{1}{r^2\pi}dy=\frac{2\sqrt{r^2-x^2}}{r^2\pi}$$

од што веднаш следува дека

$$f_X(x)=f_Y(x)=\begin{cases}\frac{2\sqrt{r^2-x^2}}{r^2\pi} \text{ за } x^2+y^2\le r^2 \\ 0 \text{ во спротивно}\end{cases}.$$

На крај, имајќи предвид дека

$$f_{X,Y}(x,y)=\frac{1}{r^2\pi}\neq\frac{4\sqrt{r^2-x^2}\sqrt{r^2-y^2}}{r^4\pi^2}=f_X(x)f_Y(y)$$

на пример за (0, 0), следува дека X и Y не се независни. ■

Дефиницијата за независност на случајни променливи и во дискретен и во непрекинат случај праволиниски се проширува и на n случајни променливи. Така, случајните променливи X_1, X_2, ..., X_n се независни акко

$$p(X_1=x_{k1}, X_2=x_{k2}, ..., X_n=x_{kn})=p(X_1=x_{k1})p(X_2=x_{k2})...p(X_n=x_{kn})$$

за секои x_{k1}, x_{k2}, ..., x_{kn}, во дискретен случај и

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)=f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$$

за секоја точка (x_1, x_2, ..., x_n), во непрекинат случај.

4.3.2. Условни случајни променливи*

Како што кај веројатноста разгледувавме условни веројатности, истиот пристап може да го примениме тука и да разгледуваме условни случајни променливи и условни распределби. Самиот поим на независност, било на настани или на случајни променливи, може згодно да се дефинира преку условни веројатности и распределби.

Ако X е случајна променлива со функција на распределба $F(x)$, а B е случаен настан со $p(\mathrm{B}) > 0$, тогаш под условна функција на распределба на X при услов да се случил настанот B е функцијата

$$F(x \mid \mathrm{B}) = p(X < x \mid \mathrm{B}) = \frac{p((X < x) \cdot \mathrm{B})}{p(\mathrm{B})}.$$

Сите особини што важат за стандардната функција на распределба важат и за условната функција на распределба (и таа е функција на распределба). Нас тука не интересираат случаите кога B е настан поврзан со друга случајна променлива.

Во дискретен случај, кога случајниот вектор (X, Y) како и X и Y се дадени со законите на распределба, едноставно имаме

$$p(X = x_i \mid Y = y_j) = \frac{p(X = x_i, Y = y_j)}{p(Y = y_j)}.$$

Јасно е дека за ваквата условна распределба важи

$$\sum_i p(X = x_i \mid Y = y_j) = \sum_i \frac{p(X = x_i, Y = y_j)}{p(Y = y_j)} = \frac{p(Y = y_j)}{p(Y = y_j)} = 1.$$

ПРИМЕР 4.24 Во следната табела е дадена распредбата на случајниот вектор (X, Y) за пропорциите на гласачите од дадена религиозна припадност што гласале за некоја политичка партија:

Y \ X	0	1	2	3	4	5	
0	0.017	0.034	0.020	0.030	0.039	0.002	0.142
1	0.013	0.121	0.042	0.065	0.082	0.005	0.328
2	0.037	0.116	0.053	0.061	0.090	0.016	0.373
3	0.020	0.062	0.021	0.014	0.037	0.003	0.157
	0.087	0.333	0.136	0.170	0.248	0.026	

X = "религија"
0 - други
1 - протестанти$_1$
2 - протестанти$_2$
3 - протестанти$_3$
4 - католици
5 - евреи

Y = "партија"
0 - друга, 1 - републиканска, 2 - демократска, 3 - негласачи

Да се најде условната распределба на религијата ако се знае дека е гласано за републиканци. Да се најде условната распределба на гласовите ако се знае дека гласале протестанти[4].

[4] Има повеќе групи протестански религии, а во оваа статистичка анализа разгледани се 3 групи.

Решение

Најпрво ја наоѓаме веројатноста на гласање за републиканците

$p(Y=1)=0.328$

Веројатностите $p(X=x_i \mid Y=1)=\dfrac{p(X=x_i, Y=1)}{p(Y=1)}$ се дадени во табелата

x_i	0	1	2	3	4	5
$p(X=x_i \mid Y=1)$	0.040	0.369	0.128	0.198	0.250	0.015

За вториот проблем, најпрво ќе ја најдеме веројатноста да гласачот е протестант

$$p(X=1)+p(X=2)+p(X=3)=0.333+0.136+0.170=0.639$$

и понатаму

$$p((X=1)+(X=2)+(X=3),\ Y=0)=0.034+0.020+0.030=0.084$$
$$p((X=1)+(X=2)+(X=3),\ Y=1)=0.121+0.042+0.065=0.228$$
$$p((X=1)+(X=2)+(X=3),\ Y=2)=0.116+0.053+0.061=0.230$$
$$p((X=1)+(X=2)+(X=3),\ Y=2)=0.062+0.021+0.014=0.097.$$

Бараната распределба ги има вредностите: 0.084/0.639, 0.228/0.639, 0.230/0.639 и 0.097/0.639. Таа е дадена во следната табела

y_j	0	1	2	3
$p((X=1)+(X=2)+(X=3) \mid Y=y_j)$	0.131	0.357	0.360	0.152

. ■

Во непрекинат случај, кога случајниот вектор (X, Y) е даден со густината на распределба $f(x, y)$, а X и Y со густините $f_1(x)$ и $f_2(y)$, условната густина на распределба е дефинирана со

$$f(x \mid y)=\frac{f(x,y)}{f_2(y)}.$$

И тука, сосема аналогно на дискретниот случај важи

$$\int_{-\infty}^{\infty} f(x \mid y)dx=\int_{-\infty}^{\infty}\frac{f(x,y)}{f_2(y)}dx=\frac{f_2(y)}{f_2(y)}=1.$$

Следната теорема ги сумира особините на условната густина на распределба и ја нагласува аналогијата меѓу веројатностите и густините на распределба.

Теорема 4.5 Условните густини на распределба ги имаат особините

1) $f(x \mid y)=\dfrac{f(x,y)}{f_2(y)}$, $f(y \mid x)=\dfrac{f(x,y)}{f_1(x)}$,

2) $\int_{-\infty}^{\infty} f(x \mid y)dx = \int_{-\infty}^{\infty} f(y \mid x)dy = 1$,

3) ако X и Y се независни, тогаш $f(x \mid y) = f_1(x)$ и $f(y \mid x) = f_2(y)$. ■

На пример, за случаен вектор (X, Y) со нормална распределба, условната густина на распределба е дадена со

$$f_{X|Y}(x,y) = Z(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x-\mu_X),\ (1-\rho^2)\sigma_Y^2),$$

каде што ρ е коефициент на корелација за X и Y (ќе бите дискутиран во следната глава).

ПРИМЕР 4.25 Распределбата на коефициентот на интелигенција IQ за жените (X) и нивните ќерки (Y) е случаен вектор (X, Y) со нормална распределба. И двете распределби имаат просек 100 ($\mu_X = \mu_Y = 100$) и стандардна девијација 15 ($\sigma_X = \sigma_Y = 15$), а коефициентот на корелација е 0.60 ($\rho = 0.60$). Која е густината на распределба на IQ за жените чиишто мајки имаат IQ од 145? Колкав е процентот на ќерките што имаат IQ не помал од своите мајки со IQ од 145?

Решение

Според условната нормална распределба имаме

$$f_{X|Y=145}(x,y) = Z(100 + \frac{15}{15}0.6(145-100),\ (1-0.6^2)15^2) = Z(127{,}144),$$

што значи дека густината на распределба на IQ за жените чиишто мајки имаат IQ од 145 е нормална распределба со $\mu = 127$ и $\sigma = 12$.

За вториот проблем, ние едноставно ја бараме веројатноста

$$p(Y \geq 145) = p(z \geq \frac{145-127}{12}) = p(z \geq 1.5) \overset{\text{од таблица}}{=} 0.0668 = 6.68\%.$$

Кои се социјалните последици од овој резултат? ■

ЗАДАЧИ

1. Нека X е случајна променлива што означува број на грешни битови при дигитална трансмисија на 8 бита. Нејзиниот закон на распределба е

x_i	0	1	2	3	4	5	6	7	8
$p(X=x_i)$	0.6245	0.3132	0.0562	0.0054	0.0004	0.0002	0.00008	0.00002	0.00000

Пресметај ја веројатноста при една трансмисија на 8 бита:

а) да има не повеќе од 2 грешни бита;

б) да има повеќе од 3 грешни бита;

в) да има непарен број грешни бита.

2. Една машинска компонента се состои од 3 механички компоненти. Веројатностите секој од 3-те компоненти да ги задоволува спецификациите се 0.95, 0.98 и 0.99. Под претпоставка дека компонентите се независни, најди го законот на распределба на случајната променлива X = "број на компоненти што ги задоволуваат спецификациите".

3. Дебелината на даските (во инчи) што купувачите ги нарачуваат е случајна променлива X со функција на распределба

$$F(x)=\begin{cases}0, & \text{за } x<1/8\\ 0.2, & \text{за } 1/8\le x<1/4\\ 0.9, & \text{за } 1/4\le x<3/8\\ 1, & \text{за } x\ge 3/8\end{cases}$$

Определи ги веројатностите

а) $p(X\le 1/4)$; б) $p(3/16\le X\le 5/16)$;

в) $p(X>1/4)$; г) $p(1/5\le X\le 1/2)$.

4. Се извршува серија независни експерименти сé додека по втор пат не се случи настанот А. Ако случувањето на А е со веројатнос p, најди го законот на распределба на случајната променлива X = "број на извршени експерименти".

5. Економично е да се лимитираат меѓудржавните телефонските разговори на помалку од 3 минути. Нека счучајната променлива X = "траење на меѓудржавен телефонски разговор". Функцијата на распределба на X може да биде

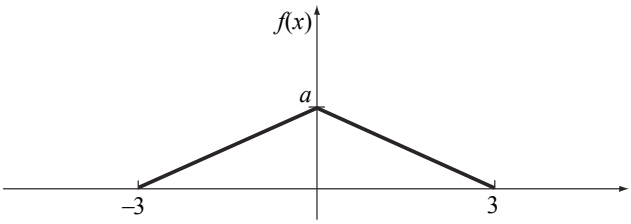
$$F(x)=\begin{cases}0, & \text{за } x<0\\ 1-e^{-x/3}, & \text{за } 0\le x<3\\ 1-e^{-x/3}/2, & \text{за } x\ge 3\end{cases}$$

а) Најди ја веројатноста телефонскиот повик да е подолг од 2 минути;

б) Најди ја веројатноста телефонскиот повик да е меѓу 2 и 6 минути;

в) Нацртај ја густината на распределба. Каква е оваа случајна променлива?

6. Густината на распределба на случајна променлива X е како на сликата



а) Определи го a;

б) Скицирај ја $F(x)$;

в) Пресметај ги веројатностите $p(X > 2)$ и $p(X > 2 \mid X > 1)$.

7. Нека случајната променлива X означува дијаметар на отворот што дупчалка го прави во еден метален дел. Дијаметарот треба да биде 12.5 милиметри, а секој направен отвор со дијаметар поголем од 12.6 милиметри го прави делот неупотреблив. Вообичаено, грешките настануваат од вибрации што резултира во зголемен дијаметар. Од поранешни податоци познато е дека густината на распределба на X е приближно $f(x) = 20e^{-20(x-12.5)}$. Пресметај ја веројатностите дека делот ќе биде добар и дека делот ќе биде неупотреблив.

8. Животниот век X (во часови), на една електронска компонента има густина на распределба

$$f(x) = \begin{cases} 0, & \text{за } x < 100 \\ \dfrac{200a}{x^2}, & \text{за } x \geq 100 \end{cases}$$

Определи го a и веројатноста компонентата да преживее 150 часови работа.

9. Дебелината на обвивката на фоторезистентност при производството на полупроводниците е рамномерно распределена од 0.2050 до 0.2150 микрометри.

а) Определи го процентот на обвивките подебели од 0.2125 микрометри;

б) Која дебелина ја надминуваат 10% од обвивките.

10. Концетрацијата на некој загадувач во една средина предизвикан од изворот на загадувањето може да се моделира со следната густина на распределба ($a > 0$)

$$f(r) = \begin{cases} 0, & \text{за } r < 0 \\ ae^{-ar}, & \text{за } r \geq 0 \end{cases}$$

каде што r е растојанието од изворот на загадувањето. Определи го радиусот во кој е содржано 95% од загадувањето.

11. Времето на откажување на вентилаторот за ладење на еден тип процесор може да се моделира со експоненцијална распределба со $\lambda = 0.0003$.

    а) Која е пропорцијата на вентилаторите што траат најмалку 10000 часа?

    б) Која е пропорцијата на вентилаторите што траат најмногу 7000 часа?

12. Времето меѓу пристигнувањата на таксијата на базната станица е експоненцијално распределена случајна променлива со очекување од 10 минути.

    а) Пресметај ја веројатноста дека таксито ќе се чека подолго од 1 час;

    б) Под претпоставка дека некој веќе чека еден час за такси, пресметај ја веројатноста дека такси ќе дојде во следните 10 минути.

13. Радиусот на држач во еден оптички мемориски уред (како DVD, Bluray) е со нормално распределена димензија со очекување 0.2508 инчи и стандардна девијација 0.00005 инчи. Спецификациите на држачот се 0.2500 ± 0.0015 инчи. Која пропорција на држачите е во согласност со спецификациите?

14. Дијаметарот на точката што ја печати еден печатач е со нормална распределба со очекување 0.002 инчи и стандардна девијација од 0.0004 инчи.

    а) Пресметај ја веројатноста дијаметарот на точката да надмине 0.0026 инчи;

    б) Колкава е вројатноста дијаметарот да е меѓу 0.0014 и 0.0026 инчи?

    в) Која стандардна девијација на дијаметарот е потребна за веројатноста под б) да биде 0.995?

15. Дебелината на ламинатната покривка за дрвени површини е нормално распределена со просек од 5 милиметри и стандардна девијација од 0.2 милиметри.

    а) Колкава е веројатноста дебелината на покривката да надмине 5.5 милиметри?

    б) Ако спецификациите бараат дебелината да биде меѓу 4.5 и 5.5 милиметри, колкава пропорција од покривките е во спецификацијата?

    в) Под која дебелина се 95% од покривките?

16. Бројот на азбестни честици на квадратен метар прашина е со Пуасонова распределба со очекување 1000. Пресметај ја веројатноста дека при анализа на квадратен метар прашина ќе бидат најдени помалку од 950 азбестни честици.

17. Печатач за масовно печатење прави минорни грешки во квалитетот на печатењето на тестен примерок од 1000 страни по Пуасонова распределба со очекување од 0.4 по страна.

а) Зошто бројот на грешки на страните се независни случајни променливи?

б) Кој е просечниот број на страни со грешки (една или повеќе)?

в) Пресметај ја веројатноста дека повеќе од 350 страни содржат грешки.

18. Нека животниот век X (во месеци) на еден електронски уред има логнормална распределба со $\mu = 2$ и $\sigma^2 = 4$. Определи ги

а) Веројатноста $p(X < 500)$;

б) Условната веројатност дека $X < 1500$ ако се знае дека $X > 1000$.

19. Нека скорот на тест на интелегенција (IQ) е нормално распределена случајна променлива X со очекување 100. Тестирани се 20 случајно избрани луѓе при што е добиена стандардна девијација 15. Колкава е веројатноста дека просечниот скор на групата ќе биде најмногу 110?

20. Животниот век на спакуван магнетен диск изложен на корозивен гас има Веибул распределба со $\beta = 2$ и просечен животен век од 600 часови.

а) Пресметај ја веројатноста дека дискот ќе трае најмалку 500 часови;

б) Пресметај ја веројатноста дека дискот ќе откаже пред 400 часови.

21. Заедничката распределба на дискретните случајни променливи X, Y и Z е дадена во табелата:

X	Y	Z	$p(x, y, z)$
1	1	1	0.05
1	1	2	0.10
1	2	1	0.15
1	2	2	0.20
2	1	1	0.20
2	1	2	0.15
2	2	1	0.10
2	2	2	0.05

Определи ги веројатностите:

а) $p(X = 2)$; б) $p(X = 1,\ Y = 2)$;
в) $p(Z < 1.5)$; г) $p(X = 1$ или $Z = 2)$;
д) $p(X = 1 \mid Y = 1)$;
ѓ) $p(X = 1,\ Y = 2 \mid Z = 2)$;
е) $p(X = 1 \mid Y = 1,\ Z = 2)$;
ж) Определи ја распределбата на X под услов дека е $Y = 1$ и $Z = 2$.

22. Во една нарачка од 15 печатачи: 4 се со проширени графички можности, 5 со проширена меморија и 6 со двете карактеристики. Случајно се бираат 4 печатачи. Нека X, Y и Z се бројот на печатачите во примерокот со проширени графички можности, проширена меморија и двете карактеристики соодветно.

а) Дефинирај ја областа на распределбата на случајниот вектор (X, Y, Z);

б) Определи ја условната распределба на X ако $Y = 2$;

в) Пресметај ги веројатностите: $p(X = 1,\ Y = 2,\ Z = 1)$, $p(X = 1,\ Y = 1)$, $p(X = 1,\ Y = 2 \mid Z = 1)$ и $p(X = 2 \mid Y = 2)$.

23. Една квалитетна WEB страна за мал бизнис содржи 100 страни при што 60%, 30% и 10% од страните содржат мала, средна и висока графичка содржина соодветно. Земен е примерок од 4 страни и нека X и Y означуваат број на страни со средна и висока графичка содржина соодветно. Определи ги

    а) $f_{XY}(x, y)$; б) $f_X(x)$; в) $f_{Y|3}(y)$; г) Дали се X и Y независни?

24. Нека случајната променлива X означува потребно време (во милисекунди) за конекција на компјутер со компјутерски сервер, а Y потребно време (во милисекунди) до авторизацијата на корисникот на компјутерот на серверот ($Y = X +$ ауторизација). Распределбата на счучајниот вектор (X, Y) е

    $f_{XY}(x, y) = 6 \cdot 10^{-6} e^{(-0.001x - 0.002y)}$, за $x < y$.

    а) Покажи дека $f_{XY}(x, y)$ е густина на распределба;

    б) Најди ја веројатноста дека конекцијата е покуса од 1000, а авторизацијата од 2000 милисекунди;

    в) Најди ја веројатноста авторизацијата да е покуса од 2000 милисекунди;

    г) Пресметај ја веројатноста дека авторизацијата трае подолго од 2000 милисекунди ако конекцијата траела 1500 милисекунди;

    д) Дали X и Y се независни?

25. Популарен производител на автомобилски делови прима Интернет нарачки преку 2 независни WEB адреси. Времето меѓу нарачките за секоја адреса во типичен ден е со експоненцијална распределба со очекување од 3.2 минути.

    а) Пресметај ја веројатноста дека нема да има нарачка во 5-минутен период. Колку е истото во 10-минутен период.

    б) Колкава е веројатноста дека на двете адреси ќе примат по 2 нарачки меѓу 10 и 15 минути откако сајтовите официјално се отворени за нарачување?

26. Тежината на циглите што се користат во одредени градби е со нормална распределба со очекување од 3 фунти и стандардна девијација од 0.25 фунти. Под претпоставка дека тежините на циглите се независни земен примерок од 20 цигли.

    а) Колкава е веројатноста дека сите цигли од примерокот се потешки од 2.75 фунти?

    б) Колкава е веројатноста дека најтешката цигла во примерокот ќе надмине 3.75 фунти?

5

Бројни карактеристики на случајни променливи

Една случајна променлива X е наполно определена со својот закон на распределба, т.е. густината на распределба ако X е непрекината или со функцијата на распределба. Во многу практични ситуации законот, густината или функцијата на распределба не може да се определат или пак од интерес е некоја "сумарна" карактеристика на X што се обезбедува многу поедноставно отколку распределбата.

5.1. Очекување

Нека е дадена случајна променлива X со закон на распределба

x_1	x_2	...	x_n
p_1	p_2	...	p_n

или густина $f(x)$.

Дефиниција 5.1 Очекување или просек (expectation) на X е бројот $E(X)$ даден со

$$E(X) = x_1p_1 + x_2p_2 + \ldots + x_np_n = \sum_{i=1}^{n} x_i p_i \quad \text{или} \quad \int_{-\infty}^{\infty} xf(x)dx\,.$$

Очекувањето е мера за локација во смисла што тоа дава идеја за тоа каде се наоѓа распределбата на X.

Често пати наместо терминот очекување и ознака $E(X)$, се користи терминот просек и ознака μ. Сепак, μ понекогаш се користи како пара-

метар во некои распределби, а не е очекување на распределбата, па е потребно малку внимание при користење на ознаките.

Во пракса, во дискретен случај, често се случува вредностите на X да се еднаквоверојатни ($p_i = 1/n$) и тогаш очекувањето се сведува на стандарден просек $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$. На пример, просечен успех на студирање, просечна старост на жител на една држава, просечна тежина на пастрмка во Охридското езеро итн.

Најважните особини на очекувањето се дадени во следната теорема.

Теорема 5.1 Очекување ги има следните особини:

1) $E(C) = 0$, кога C е константа,

2) $E(CX) = CE(X)$,

3) $E(X + Y) = E(X) + E(Y)$,

4) $E(X \cdot Y) = E(X)E(Y) + K_{X,Y}$, каде што $K_{X,Y} = E(X–E(X))(Y–E(Y))$ и се нарекува момент на корелација или коваријација. Ако X и Y се независни, $K_{X,Y} = 0$, т.е. $E(X \cdot Y) = E(X)E(Y)$.

Доказ: Следува директно од дефиницијата. ■

Понатаму, вообичаено ќе ги испуштаме заградите при работа со бројните карактеристики на случајните променливи и наместо, на пример $E(X)$, ќе пишуваме едноставно EX или во случаи кога е тоа соодветно ќе користиме едноставно μ.

5.2. Дисперзија

Нека е дадена случајна променлива X со закон на распределба

x_1	x_2	...	x_n
p_1	p_2	...	p_n

или густина $f(x)$.

Дефиниција 5.2 Дисперзија (variance) на X е бројот DX даден со

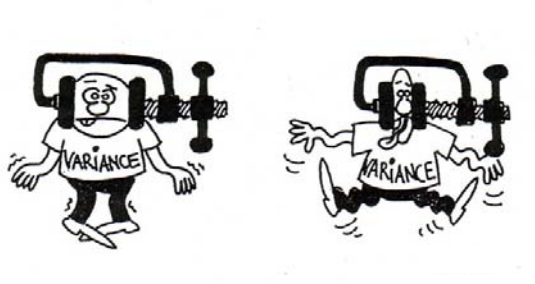
$$DX = E(X - EX)^2 = (x_1 - EX)^2 p_1 + (x_2 - EX)^2 p_2 + \ldots + (x_n - EX)^2 p_n =$$

$$= \sum_{i=1}^{n} (x_i - EX)^2 p_i$$

или во непрекинат случај $\int_{-\infty}^{\infty} (x - EX)^2 f(x)dx$.

Диспезијата на X е мера за раштрканоста на вредностите на X околу нејзиното очекување.

Во пракса, во дискретен случај, често се случува вредностите на X да се еднаквоверојатни ($p_i = 1/n$), и тогаш дисперзијата се сведува на $DX = \frac{1}{n}\sum_{i=1}^{n}(x_i - EX)^2$. На пример, при пописот на население во една држава, од голема важност е просечната старост на населението, но речиси од слична важност е и дисперзијата на староста на населението. Имено, поголема дисперзија значи дека во државата има поголем број на старо и младо население.



Најважните особини на дисперзијата се дадени во следната теорема.

Теорема 5.2 Дисперзијата ги има следните особини:

1) $DC = 0$, кога C е константа,

2) $D(CX) = C^2D(X)$,

3) $D(X + Y) = DX + DY + 2K_{XY}$; Ако X и Y се независни, $K_{X,Y} = 0$, т.е. $D(X + Y) = DX + DY$,

4) $DX = EX^2 - (EX)^2$.

Доказ: Следува директно од дефиницијата. На пример, за особината 4) имаме

$$DX = E(X-EX)^2 = E(X^2-2XEX + (EX)^2) =$$
$$= EX^2 - 2\,(EX)(EX) + (EX)^2 = EX^2 - \,(EX)^2. \blacksquare$$

Како што за очекувањето EX се користи μ, за дисперзијата често пати се користи ознаката σ^2. И двете ознаки потекнуваат од нормалната распределба. Квадратниот корен на дисперзијата $\sigma = \sqrt{DX}$ се нарекува стандардна девијација и има голема примена во статистичките оценки.

ПРИМЕР 5.1 Се разгледува дизајн на три нови производи од аспект на можна повратна добивка. Одделот за маркетинг има проценки дека дизајнот А ќе да-

де добивка од 3 милиони евра. Дизајнот B е понесигурен од аспект на добивката и грубо е проценето дека со 30% шанси ќе донесе 7 милиони, а со 70% шанси само 2 милиони евра. Проценката за дизајнот C е дека со 30% шанси ќе донесе 6 милиони, со 50% шанси 3 милиони и со 20% шанси само 1 милион евра. Кој дизајн би го преферирале?

Решение

Нека X, Y и Z се случајни променливи дефинирани со:

X = "добивка од дизајнот A",

Y = "добивка од дизајнот B",

Z = "добивка од дизајнот C".

Очекувањата на секоја од трите случајни променливи се:

$EX = 3\cdot 1 = 3$ милиони €,

$EY = 7\cdot 0.3 + 2\cdot 0.7 = 3.5$ милиони €,

$EZ = 6\cdot 0.3 + 3\cdot 0.5 + 1\cdot 0.2 = 3.5$ милиони €,

Според очекувањето, дизајните B и C се во предност во однос на A. Од друга страна, за дисперзиите добиваме:

$DX = 0 \Rightarrow \sigma_X = \sqrt{DX} = 0$ милиони €,

$DY = (7-3.5)^2\cdot 0.3 + (2-3.5)^2\cdot 0.7 = 5.25 \Rightarrow \sigma_Y = \sqrt{DY} = 2.29$ милиони €,

$DZ = 2.5^2\cdot 0.3 + 0.5^2\cdot 0.5 + 2.5^2\cdot 0.2 = 3.25 \Rightarrow \sigma_Z = 1.80$ милиони €.

Дизајните B и C имаат исто очекување, но поради помалата стандардна девијација, дизајнот C секако е посигурен дека ќе ја донесе добивка според неговото очекување. Како да се одлучиме меѓу дизајните A и C? Тука би било логично да дадеме лесна предност за дизајнот A, бидејќи разликата во очекуваната добивка во корист на дизајнот C не е доволно голема да ја компензира стандардната девијација од 1.80 милиони €. Но некој што е наклонет на ризици може да го форсира дизајнот C и со добри шанси да оствари добивка од 3.25 ± 1.80 милиони €. ■

Една од предностите на користење на σ наместо σ^2 како мера за дисперзираност е во тоа што таа е изразена во исти единици како очекувањето и може директно да се споредува со него. Бездимензионалниот број $v = \sigma/\mu$ се нарекува *коефициент на варијација* и понекогаш се користи за олеснување на споредбите на случајните променливи што се со различни димензии или се изразуваат во различни единици. Односот меѓу σ и μ е сличен како односот меѓу апсолутна и релативна грешка. Самата σ не е доволно информативна бидејќи таа не кажува на колкаво очекување е "измерена" раштрканоста на вредностите на случајната

променлива околу нејзиното очекување. На пример, недоволно е да кажеме дека една вага за мерење тежина има стандардна девијација од 0.05 килограми ако не додадеме на "очекувана тежина од 100 килограми" или на "очекувана тежина од 3 килограми" (вага во кујна). Иако стандардната девијација е иста, коефициентот на варијација во првиот случај (100 килограми) е очигледно многу помал. Како овие девијации би ги споредиле со стандардна девијација на потрошувачка на гориво од 0.2 литри на просечна потрошувачка од 8 литри кај едно возило? Коефициентот на варијација кај вагата во кујна е 0.05/3 = 0.0167, додека кај потрошувачката на возилото е 0.2/8 = 0.025, па заклучуваме дека варијациите на случајната променлива "потрошувачка на гориво на едно возило" се поголеми.

ПРИМЕР 5.2 Парадокс на чекање: Зошто кога чекаме, на пример автобус што поминува на 10 минути, обично го чекаме поголго од очекуваното време од 5 минути? Зошто тоа е така може да се види од следниот исконструиран пример. Нека автобус поминува во 20% случаи на 30 минути и во 80% случаи секоја секунда. Колкаво е времето на чекање X на случајно пристигнат патник ако се знае дека случајната променлива T = "време меѓу пристигнување на два автобуса" има распределба

$\frac{1}{2}(1+v_T)ET$, каде што v_T е коефициентот на варијација на T.

Решение

$ET = 30\cdot 0.2 + (1/60)\cdot 0.8 = 6.0133$ минути,

$DT = (30-6.0133)^2\cdot 0.2 + (1/60-6.0133)^2\cdot 0.8 = 143.8398 \Rightarrow \sigma_T = 11.9933$

За коефициентот на варијација имаме $v_T = 11.9933/6.0133 = 1.9945$.

На крај добиваме $EX = 0.5(1+1.9945)6.0133 = 14.9667$ минути.

Значи иако просечното време на пристигнување на автобусите е околу 6 минути, просечното време на чекање е околу 15 минути. ■

5.3. Бројни карактеристики на некои случајни променливи

Очекувањето и дисперзијата се основни бројни карактеристики на случајните променливи. Во следната табела е прикажана аналогијата во пресметката на очекувањето и дисперзијата во случај на случајна променлива зададена со законот на распределба (дискретен случај) и густината на распределба (непрекинат случај).

Дискретна случајна променлива **Закон на распределба**		*Непрекината случајна променлива* **Густина на распределба**
$f(x):\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix},\ \sum p_i = 1$	$\Leftrightarrow$	$f(x)\colon \mathbb{R} \to \mathbb{R},\quad \int_{-\infty}^{\infty} f(x)dx = 1$
$p(a \le X \le b) = \sum_{i:x_i \in [a,b]} p_i$	$\Leftrightarrow$	$p(a \le X \le b) = \int_a^b f(x)dx$
$\mu = EX = \sum_1^n x_i p_i$	$\Leftrightarrow$	$\mu = EX = \int_{-\infty}^{\infty} x f(x)dx$
$DX = \sum_1^n (x_i - \mu)^2 p_i = \sum_1^n x_i^2 p_i - \mu^2$	$\Leftrightarrow$	$DX = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

Да ги разгледаме бројните карактеристики на случајните променливи со некои од познатите дискретни и непрекинати распределби.

БИНОМНА РАСПРЕДЕЛБА. Случајната променлива X = "број на случувања на настан A во серија од n независни експерименти", со закон на распределба

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ за } k = 0, 1, 2, \dots, n.$$

Очекувањето може да се најде со помош на биномен развој

$$EX = \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} =$$
$$= np(p + (1-p))^n = np.$$

Секое повторување на експериментот резултира со случајна променлива X_k со распределба

0	1
$1-p$	p

, при што $X = X_1 + X_2 + \dots + X_n$.

Сега, дисперзијата лесно се добива од условот за независност

$$DX = DX_1 + DX_2 + \dots + DX_n, \text{ и од } DX_k = EX_k^2 - (EX)^2 = p - p^2$$

добиваме

$$DX = np(1-p).$$

ПУАСОНОВА РАСПРЕДЕЛБА. Законот на распределба на X е

$p(X=k)=\frac{\mu^k}{k!}e^{-\mu}$, за $k = 0, 1, 2, ..., n, ...$ каде што $\mu > 0$ е просечен број случувања во еден интервал. Очекувањето е

$$EX=\sum_{k=0}^{\infty}k\frac{\mu^k}{k!}e^{-\mu}=\mu e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^{k-1}}{(k-1)!}=\mu e^{-\mu}e^{\mu}=\mu$$

Дисперзијата исто така лесно се добива од

$$EX(X-1)=\sum_{k=0}^{\infty}k(k-1)\frac{\mu^k}{k!}e^{-\mu}=\mu^2e^{-\mu}\sum_{k=2}^{\infty}\frac{\mu^{k-2}}{(k-2)!}=\mu^2,$$

и сега имајќи предвид дека $EX(X-1) = EX^2 - EX = \mu^2$ добиваме $EX^2 = \mu^2 + \mu$, па дисперзијата е

$$DX = EX^2 - (EX)^2 = \mu^2 + \mu - \mu^2 = \mu.$$

ХИПЕРГЕОМЕТРИСКА РАСПРЕДЕЛБА. Од N објекти, r се означени како поволни. Се одбираат n објекти и се разгледува случајната променлива X = "број на поволни објекти во n-те избрани". Тогаш, законот на распределба на X е даден со

$$p(X=k)=\frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}, \text{ за } k = 0, 1, 2, ..., n.$$

За да го најдеме очекувањето, наместо да пресметуваме комплицирана сума, ќе забележиме дека веројатноста на поволен избор на еден објект е r/N. И сега бидејќи избираме n објекти, очекувањето е просто

$$EX=\frac{nr}{N}.$$

На пример, во лотарија 7 од 37, просечниот број погодоци во една пополнета колона е $7\cdot7/37 = 1.3243$.

Дисперзијата ќе ја добиеме на сличен начин. Имено, дисперзијата на еден поволно избран објект е биномна променлива $p(1-p)$ каде што $p = r/N$. Оттука за дисперзијата добиваме

$$DX=n\frac{r}{N}\left(1-\frac{r}{N}\right)=\frac{nr(N-r)}{N^2}.$$

За лотарија 7 од 37, дисперзијата за една пополнета колона е $7 \cdot 7 \cdot 30/37^2 = 1.0738$.

ГЕОМЕТРИСКА РАСПРЕДЕЛБА. Законот на распределба на X е

$$p(X = k) = (1-p)^{k-1} p \text{ , за } k = 1, 2, ..., n, ... \ .$$

каде што p е веројатноста на случување на некој настан А. Очекувањето се добива лесно од редот

$$EX = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1}$$

и сега ако ставиме $f(p) = \sum_{k=1}^{\infty} k(1-p)^{k-1}$ имаме

$$\int f(p)dp = -\sum_{k=1}^{\infty}(1-p)^k = -(1-p)\sum_{k=0}^{\infty}(1-p)^k = -\frac{1-p}{1-(1-p)} = 1-\frac{1}{p}$$

од каде следува $f(p) = \frac{1}{p^2}$, што дава $EX = \frac{1}{p}$.

Со сличен, но малку пообемен пристап може да се добие дека

$EX^2 = \frac{2-p}{p^2}$ од каде веднаш следува

$$DX = EX^2 - (EX)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

НЕГАТИВНА ГЕОМЕТРИСКА РАСПРЕДЕЛБА. Законот на распределба на X е

$$p(X = k) = \binom{k+r-1}{r-1} p^r (1-p)^k \text{ , за } k = 0, 1, 2, ... ,$$

каде што k е бројот на не случувања на настан А пред тој да се случи r пати.

Очекуваниот број повторувања до првото случување на А е $1/p$ (геометриска распределба), така што очекуваниот број на не случувања на А до неговото прво случување е само за 1 помалку, т.е. $1/p - 1 = (1 - p)/p$. Оттука се очекува $EX = r(1 - p)/p$ не случувања на А до негови r случувања.

За дисперзијата се добива $DX = r(1 - p)/p^2$.

За непрекинатите распределби, очекувањето и дисперзијата се пресметуваат со интеграли наместо со суми.

РАМНОМЕРНА РАСПРЕДЕЛБА. Густината на случајна променлива X со рамномерна распределба е

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & \text{за } x\in[a,b]\\ 0, & \text{во спротивно}\end{cases},$$

и оттука лесно го добиваме очекувањето

$$EX=\int_a^b x\frac{1}{b-a}dx=\frac{a+b}{2},$$

додека за дисперзијата добиваме

$$DX=EX^2-(EX)^2=\int_a^b x^2\frac{1}{b-a}dx-\frac{(a+b)^2}{4}=\frac{(b-a)^2}{12}.$$

ЕКСПОНЕНЦИЈАЛНА РАСПРЕДЕЛБА. Густината на експоненцијалната распределба е дадена со

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & \text{за } x\geq 0\\ 0, & \text{во спротивно}\end{cases}$$

и оттука очекувањето и дисперзијата се добиваат од интегралите

$$EX=\lambda\int_0^\infty xe^{-\lambda x}dx=\frac{1}{\lambda}\quad\text{и}\quad DX=\lambda\int_0^\infty x^2e^{-\lambda x}dx-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}.$$

НОРМАЛНА РАСПРЕДЕЛБА. За нормалната распределба веќе знаеме дека очекувањето μ и дисперзијата σ^2 се параметри во нејзината густина на распределба

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

и се разбира може да се добијат со решавање на соодветните интеграли.

СТУДЕНТОВА РАСПРЕДЕЛБА. Густината на студентовата распределба е

$f(x)=\dfrac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}(1+\dfrac{x^2}{n})^{-\frac{n+1}{2}}$, при што очекувањето е 0 бидејќи $f(x)$ е парна функција (тогаш $xf(x)$ е непарна), а дисперзијата е $n/(n-2)$ кога $n>2$ и се добива со покомплицирани пресметки.

Очекувањето и дисперзијата за некои непрекинати распределби се:

Распределба	Густина $f(x)$	EX	DX
Бета $x\geq 0$, $\alpha, \beta\in\mathbb{R}^+$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$
Хи-квадрат $x\geq 0$, $n\in\mathbb{Z}^+$	$\dfrac{1}{2^{n/2}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$	n	$2n$
Експоненцијална x, $\lambda\geq 0$	$\lambda e^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Фишерова $x\geq 0$, $m,n\in\mathbb{Z}^+$	$\dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\dfrac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}(1+\dfrac{m}{n}x)^{-\frac{m+n}{2}}$	$\dfrac{n}{n-2}$	$\dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$
Гама $x\geq 0$, $\alpha, \beta\in\mathbb{R}^+$	$\dfrac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$	$\alpha\beta$	$\alpha\beta^2$
Нормална $\sigma\in\mathbb{R}^+$	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Логнормална $x>0$, $\sigma\in\mathbb{R}^+$	$f(x)=\dfrac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	$e^{\mu+\sigma^2/2}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
Рамномерна $[a, b]$	$\dfrac{1}{b-a}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Студентова $n\in\mathbb{Z}^+$	$f(x)=\dfrac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}(1+\dfrac{x^2}{n})^{-\frac{n+1}{2}}$	0 за $n>1$	$\dfrac{n}{n-2}$ за $n>2$
Веибул $x\geq 0$, $\alpha, \beta\in\mathbb{R}^+$ $\gamma=0$	$f(x)=\dfrac{\beta}{\alpha}\left(\dfrac{x-\gamma}{\alpha}\right)^{\beta-1}e^{-\left(\frac{x-\gamma}{\alpha}\right)^\beta}$	$\alpha\Gamma(1+\dfrac{1}{\beta})$	$\alpha^2\Gamma(1+\dfrac{2}{\beta})-(M\xi)^2$

5.4. Моменти

Нека е дадена случајна променлива X со закон на распределба

x_1	x_2	...	x_n
p_1	p_2	...	p_n

или густина $f(x)$.

Дефиниција 5.3 Момент од n-ти ред на X е бројот $E(X^n)$ (очекување на X^n) даден со

$$EX^n = \sum_1^n x_i^n p_i \quad \text{или} \quad EX^n = \int_{-\infty}^{\infty} x^n f(x)dx .$$

Ние веќе го разгледавме моментот од прв ред EX како очекување или просек на случајната променлива X. Тој момент е сигурно најважен и може да се смета како центар на масата (тежиште) на распределбата.

Покрај очекувањето, постојат уште две други мери за центрираност на случајните променливи што често се користат во праксата, *медијаната* и *модот*.

Медијана на случајната променлива X е која било точка што ја дели распределбата на две по веројатност еднакви области. Попрецизно, медијана на X е точката x_0 за којашто важи

$$p(X < x_0) = \frac{1}{2} .$$

Додека очекувањето на X може и да не постои (на пример е $\pm\infty$), X секогаш има медијана. Во споредба со просекот, медијаната понекогаш се преферира како мера на централната тенденција кога распределбата е "искривена", особено ако има мал број екстремни вредности. На пример, медијаната е згодна централна мера за заработувачката во популацијата, бидејќи таа е помалку осетлива на малиот број екстремно високи (или екстремно ниски заработки).

ПРИМЕР 5.3 Нека T е време меѓу емисии на честички од радиоактивен атом. Добро е познато дека T е случајна променлива со експоненцијална распределба

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & \text{за } t \geq 0 \\ 0, & \text{во спротивно} \end{cases}$$

T вообичаено се нарекува животен век на атомот, а мерата на просекот на животниот век се нарекува време на полураспаѓање (половина животен век) што се пресметува како медијана на T. Пресметај ја медијаната и спореди ја со очекувањето.

Решение

Времето на полураспаѓање τ (медијаната) се добива од равенството

$$\int_0^{\tau} f(t)dt = \lambda\int_0^{\tau} e^{-\lambda t}dt = \frac{1}{2} \quad \text{што дава} \quad \tau = \ln\frac{2}{\lambda}.$$

Како што веќе знаеме, очекувањето за експоненцијална распределба е

$$ET = \lambda\int_0^{\infty} e^{-\lambda t}dt = \frac{1}{\lambda}. \blacksquare$$

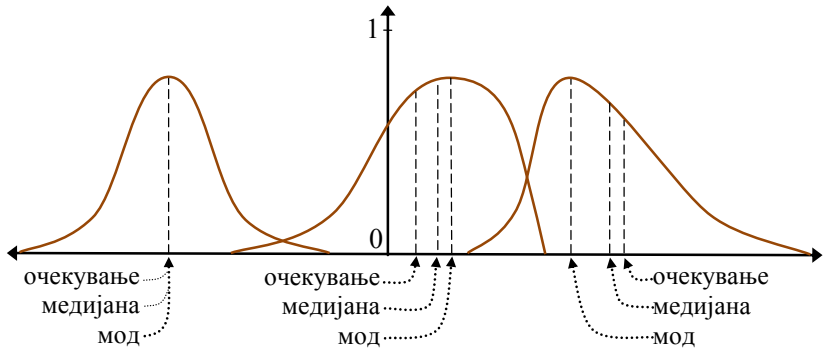
Мод на случајната променлива X е точката x_k за којашто важи

$p(x_{k-1}) < p(x_k) > p(x_{k+1})$, за дискретен случај или

$f(x_k - \varepsilon) < f(x_k) > f(x_k + \varepsilon)$, $\varepsilon > 0$, за непрекинат случај.

Модот кореспондира со точките на максимум во законот или густината на распределбата. Терминот *унимодална* распределба се користи за распределба што има само еден мод.

За да ги нагласиме разликите меѓу очекувањето, медијаната и модот како мери на централна тенденција на една случајна променлива, ги даваме нивните положби за некои облици на густини на распределба (слика 5.1).



Слика 5.1 *Очекување медијана и мод за различни густини на распределба*

Едно обопштување на моментите се централните моменти. Тие се моменти во однос на очекувањето.

Дефиниција 5.4 Централен момент од n-ти ред на X е бројот

$$E(X-\mu)^n = \sum_1^n (x_i - \mu)^n p_i \quad \text{или} \quad E(X-\mu)^n = \int_{-\infty}^{\infty} (x-\mu)^n f(x)dx,$$

каде што $\mu = EX$.

Од сите централни моменти, од најголема важност е централниот момент од 2 ред, т.е. дисперзијата на случајната променлива X. Централните моменти од повисок ред даваат одредени додатни карактеристики на распределбата. На пример, коефециентот на "искривеност" (*skewness*) дефиниран со

$$\gamma_1 = \frac{E(X-\mu)^3}{\sigma^3}$$

е мера за симетричност на распределбата. Тој е позитивен кога една унимодална распределба има доминатна десна страна. Доминантна лева страна на распределбата дава негативен γ_1. Кога $\gamma_1 = 0$, унимодалната распределба е симетрична околу очекувањето. Всушност симетрична распределба околу очекувањето повлекува дека сите централни моменти од непарен ред се 0.

Степенот на рамност на една унимодална распределба околу нејзиниот максимум може да се мери со коефициентот на *ексцес* дефиниран со

$$\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4} - 3\,.$$

Позитивно γ_2 повлекува остар врв околу модот на унимодалната распределба, додека негативно γ_2 по правило значи тап врв. Значењето на 3-ката во изразот е поврзано со нормалната распределба за која γ_2 се зема да биде 0. За централните моменти на нормалната распределба се добива

$$E(X-\mu)^n = \begin{cases} 1\cdot 3\cdot \ldots \cdot (n-1)\sigma^n, & \text{за парни } n \\ 0, & \text{за непарни } n \end{cases}$$

па оттука веднаш следува дека за неа $E(X-\mu)^4 = 3\sigma^4$ што и дава смисла на 3-ката како "гарант" дека за нормалната распределба $\gamma_2 = 0$.

5.5. Бројни карактеристики на случајни вектори

Очекувањето и дисперзијата на праволиниски начин може да се обопштат на случајните вектори. Имено, нека е даден случајниот вектор (X, Y) со закон на распределба

		X			
		x_1	x_2	. . .	x_n
Y	y_1	p_{12}	p_{13}	. . .	p_{1n}
	y_2	p_{21}	p_{22}	. . .	p_{2n}
	⋮	⋮	⋮	. . .	⋮
	y_m	p_{m1}	p_{m2}	. . .	p_{mn}

или густина на распределба $f(x, y)$.

Тогаш, очекувањето и дисперзијата на случајниот вектор (X, Y) се дадени едноставно со (EX, EY) и (DX, DY), каде што $E(\cdot)$ и $D(\cdot)$ се очекување и дисперзија на соодветните случајни променливи добиени од соодветните индивидуални (маргинални) закони или густини на распределба.

Многу поважни и поинтерсни се случите на наоѓање на бројни карактеристики на функции од случајни променливи. Имено, нека $h(X, Y)$ е функција од случајните променливи X и Y. Тогаш, очекувањето на случајната променлива $h(X, Y)$ е дадено со

$$Eh(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}h(x_i,y_j)p_{ij} \text{ или } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)f(x,y)dxdy .$$

Ако е пак $h(X, Y) = g(X)$, тогаш

$$Eg(X)=\sum_{i=1}^{n}g(x_i)\sum_{j=1}^{m}p_{ij}=\sum_{i=1}^{n}g(x_i)p_i \text{ или}$$

$$Eg(X)=\int_{-\infty}^{\infty}g(x)dx\int_{-\infty}^{\infty}f(x,y)dy=\int_{-\infty}^{\infty}g(x)f_X(x)dx .$$

Сосема аналогно се може да се добијат и заедничките моменти од повисок ред на X и Y. Така, момент од ред (k, l) на случајниот вектор (X, Y) е бројот

$$E(X^k\cdot Y^l)=\sum_{i=1}^{n}\sum_{j=1}^{m}x_i^k y_j^l p_{ij} \text{ или } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^k y^l f(x,y)dxdy .$$

Централните моменти од ред (k, l) се добиваат на идентичен начин како $E(X-\mu_X)^k(Y-\mu_Y)^l$. Од овие централни моменти, ние веќе го воведовме моментот од ред (1, 1) за кој се користи терминот коваријација $K_{X,Y}$ на X и Y. Интуитивно, таа ја дава зависноста меѓу случајните променливи X и Y, и добива вредност 0 само кога X и Y се независни. Оваа особина е таа поради која коваријацијата е основа на коефициентот на корелација – основната мера за оценка на степенот на линеарна зависност меѓу случајните променливи.

5.5.1. Коефициент на корелација

Во особините на очекувањето и дисперзијата кога се работи со две зависни случајни променливи X и Y се појавува ненулта коваријација дадена со $K_{X,Y} = E(X–EX)(Y–EY)$. Значи коваријацијата може да се користи како мера за зависноста на случајните променливи. Да дојдеме до таква мера ќе тргнеме од линеарната комбинација на n случајни променливи

$$Y = c_1X_1 + c_2X_2 + \ldots + c_nX_n.$$

Лесно се проверува дека

$$Y - EY = \sum_{i=1}^{n} c_i(X_i - EX_i) \quad \text{и сега}$$

$$(Y - EY)^2 = \sum_{i,j=1}^{n} c_i c_j (X_i - EX_i)(X_j - EX_j)\text{, т.е.}$$

$$E(Y - EY)^2 = DY = \sum_{i,j=1}^{n} c_i c_j K_{X_i,X_j} .$$

Десната страна на $E(Y - EY)^2$ е квадратна форма по c_1, c_2, ..., c_n којашто е секогаш позитивна. Познато е дека тоа е можно тогаш и само тогаш, кога се ненегативни сите главни минори (детерминанти составени од квадратни подматрици на матрица) на матрицата $[K_{X_i,X_j}]_{n\times n}$ составена од коефициентите K_{X_i,X_j} на квадратната форма (матрица на коваријација). Со други зборови, детерминантите

$$\begin{vmatrix} K_{X_1,X_1} & K_{X_1,X_2} & \cdots & K_{X_1,X_k} \\ K_{X_2,X_1} & K_{X_2,X_2} & \cdots & K_{X_2,X_k} \\ \cdots & \cdots & \cdots & \cdots \\ K_{X_k,X_1} & K_{X_k,X_2} & \cdots & K_{X_k,X_k} \end{vmatrix} \geq 0\text{, за } k = 1, 2, \ldots .$$

Во спеијален случај, за $k = 2$, имаме

$$\begin{vmatrix} DX_1 & K_{X_1,X_2} \\ K_{X_2,X_1} & DX_2 \end{vmatrix} = DX_1DX_2 - K^2_{X_1,X_2} \geq 0$$

и оттука следува дека

$$\left|K_{X_1,X_2}\right| \leq \sqrt{DX_1DX_2} .$$

Најпозната едноставна мера на зависноста меѓу две случајни променливи изведена од коваријацијата е Пирсоновиот (Pearson) коефици-

ентот на корелација. Тој се добива со делење на коваријацијата $K_{X,Y}$ на X и Y со производот на нивните стандардни девијации.

Дефиниција 5.5 Коефициент на корелација на случајните променливи X и Y е бројот $\rho_{X,Y}$ дефиниран со

$$\rho_{X,Y} = \frac{K_{X,Y}}{\sqrt{DX}\sqrt{DY}} = \frac{E(X-EX)(Y-EY)}{\sqrt{DX}\sqrt{DY}} = \frac{E(XY)-EX\cdot EY}{\sqrt{D\xi}\sqrt{D\eta}}.$$

Коефициентот на корелација е дефиниран само ако дисперзиите на двете случајни променливи се ненулти. Неговите основни особини се дадени во следната теорема.

Теорема 5.3 За коефициентот на корелација $\rho_{X,Y}$ важи:

1) $\rho_{X,Y} = \rho_{Y,X}$ (симетричност),

2) $|\rho_{X,Y}| \le 1$,

3) ако X и Y се независни, $\rho_{X,Y} = 0$,

4) ако $Y = aX + b$, a *и* b се константи, тогаш $|\rho_{X,Y}| = 1$.

Доказ: Првите 3 особини се веќе покажани и следуваат директно од особините на коваријацијата. Ќе ја докажеме 4-тата особина. Нека, $Y = aX + b$ и $EX = \mu$, $DX = \sigma^2$, тогаш

$$EY = a\mu + b,\ DY = a^2\sigma^2 \text{ и}$$

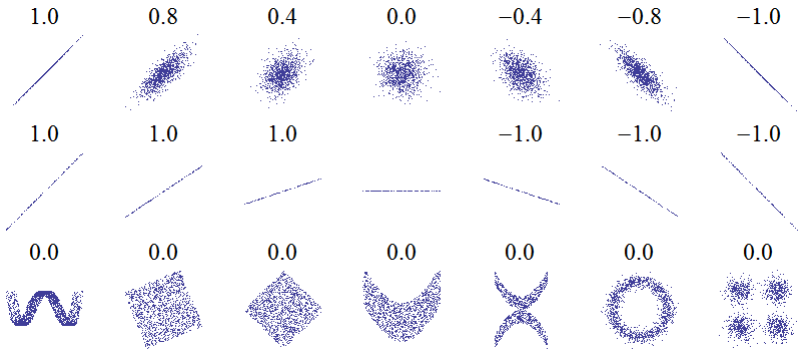
$$K_{X,Y} = E(X-\mu)(Y-a\mu-b) = E(X-\mu)(aX-a\mu) =$$

$$= aE(X-\mu)(X-\mu) = a\sigma^2.$$

Оттука веднаш се добива

$$\rho_{X,Y} = \frac{a\sigma^2}{\sqrt{a^2\sigma^2\sigma^2}} = \frac{a}{|a|} = \pm 1. \blacksquare$$

Коефициентот на корелација има вредност +1 во случај на идеална позитивна (растечка) или −1 во случај на идеална негативна (опаѓачка) линеарна зависност. Вредностите меѓу −1 и +1 го одразуваат степенот на линеарната зависност меѓу случајните променливи. Кога неговата вредност се приближува кон 0, случајните променливи се сè поблиску до некорелираност во линеарна смисла, и обратно, колку коефициентот е поблиску до −1 или +1, толку е посилна корелацијата меѓу нив. На сл. 5.2 се прикажани вредностите на некои парови случајни променливи и соодветниот коефициент на корелација.



Слика 5.2 *Вредности на парови случајни променливи и соодветниот коефициент на корелација*

Значи ако променливите се независни, коефициентот на корелација е 0, но обратното не важи бидејќи коефициентот ги детектира само линеарните зависности меѓу случајните променливи. На пример, ако случајната променлива X е симетрично распределена околу 0 (на пример, со нормална или студентова распределба) и ако $Y = X^2$, тогаш Y е комплетно одредена со X, т.е. Y и X се комплетно зависни, но нивната корелација е 0 (тие се некорелирани). Сепак, во специјален случај кога X и Y се со заедничка нормална распределба, некорелираноста е еквивалентна со независноста (види поглавје 5.5.3).

ПРИМЕР 5.4 Еден камион за достава секој ден оди до супермаркет и назад, враќајќи се по друг пат. При одење до супермаркетот поминува 3 семафори, а при враќање 2 семафори. Нека случајната променлива X = "број на црвени светла при одење" и Y = "број на црвени светла при враќање". Сообраќајниот инженер со испитување на релативните честоти на црвени светла ја добил следната заедничка распределба

Y \ X	0	1	2	3	
0	0.02	0.05	0.10	0.03	0.20
1	0.04	0.09	0.13	0.08	0.34
2	0.05	0.15	0.17	0.09	0.46
	0.11	0.29	0.40	0.20	

На пример, веројатноста да се случат 2 црвени светла при одење и едно црвено светло при враќање е 0.13.

Најди го коефициентот на корелација за бројот на црвени светла при одење и враќање и интерпретирај го добиениот резултат.

Решение

$EX = 1(0.05+0.09+.15) + 2(0.10+0.13+0.17) + 3(0.03+0.08+0.09) = 1.69$

$$EY = 1(0.04+0.09+0.13+0.08) + 2(0.05+0.15+0.17+0.09) = 1.27$$

$$DX = 0.29\cdot 1^2 + 0.40\cdot 2^2 + 0.20\cdot 3^2 - 1.69^2 = 0.83$$

$$DY = 0.34\cdot 1^2 + 0.46\cdot 2^2 - 1.27^2 = 0.58$$

$$E(XY) = 1\cdot 0.09 + 2\cdot 0.13 + 3\cdot 0.08 + 2\cdot 0.15 + 4\cdot 0.17 + 6\cdot 0.09 = 2.11$$

$$\rho_{X,Y} = \frac{E(XY) - EX\cdot EY}{\sqrt{D\xi}\sqrt{D\eta}} = \frac{2.11 - 1.69\cdot 1.27}{\sqrt{0.83\cdot 0.58}} = -0.06$$

Бидејќи коефициентот на корелација е многу близок до 0, заклучуваме дека нема линеарна зависност меѓу бројот на црвени светла при одењето и враќањето на камионот. ■

Генерално, може да сметаме дека ако коефициентот на корелација $|\rho| \geq 0.9$, корелацијата е многу силна. За $0.7 \leq |\rho| \leq 0.9$ имаме силна корелација, додека за $0.5 \leq |\rho| \leq 0.7$ може да сметаме на умерена корелација. Вредностите $|\rho| \leq 0.5$ значат дека немаме корелација меѓу случајните променливи.

Во практиката, често пати заедничката распределба на случајните променливи X и Y (со ист број вредности) се задава просто како низа на соодветни парови вредности

X	x_1	x_2	...	x_n
Y	y_1	y_2	...	y_n

каде што веројатностите на заедничката распределба се имплицитно дадени со

$$p(X = x_i, Y = y_j) = \begin{cases} 1/n & \text{за } i = j \\ 0 & \text{за } i \neq j \end{cases}.$$

ПРИМЕР 5.5 На пример, едно европско истражување за сигурност на патиштата било за тоа како староста на автомобилите влијае на способноста на кочење. За таа цел биле тестирани 10 автомобили со различна старост (X) и мерени должините на запирање во метри (Y) при брзина од 40 км/час. Податоците се дадени во следната табела:

X	9	15	24	30	38	46	53	60	64	76
Y	28.4	29.3	37.6	36.2	36.5	35.3	36.2	44.1	44.8	47.2

Најди го коефициентот на корелација и дискутирај го добиениот резултат.

Решение

Низата парови вредности ја означува следната заедничка распределба

$$p(X = x_i, Y = y_j) = \begin{cases} 1/10 = 0.1 & \text{за } i = j \\ 0 & \text{за } i \neq j \end{cases}$$

од каде добиваме

$$EX = (9+15+24+30+38+46+53+60+64+76)\cdot 0.1 = 41.5$$

$$EY = (28.4+29.3+37.6+36.2+36.5+35.3+36.2+44.1+44.8+47.2)\cdot 0.1 = 37.56$$

$$DX = (9^2+15^2+24^2+30^2+38^2 + 46^2+53^2+60^2+64^2+76^2)\cdot 0.1 - 41.5^2 = 440.05$$

$$DY = (28.4^2+29.3^2+37.6^2+36.2^2+36.5^2 + 35.3^2+36.2^2+44.1^2+44.8^2+47.2^2)\cdot 0.1 - 37.56^2 = 35.02$$

$$E(XY) = (255.6+439.3+902.4+1086+1387+ 1623.8+1918.6+2646+2867.2+3587.2)\cdot 0.1 = 1671.31$$

$$\rho_{X,Y} = \frac{R(XY) - EX \cdot EY}{\sqrt{DX}\sqrt{DY}} = \frac{1671.31 - 41.5 \cdot 37.56}{\sqrt{440.05 \cdot 35.02}} = 0.915$$

Коефициентот на корелација е близок до 1, па заклучуваме дека има силна линеарна зависност меѓу X и Y. ■

Коефициентот на корелација има важно значење при анализата на две случајни променливи. Тој ја мери линеарната зависност на две случајни променливи, т.е. гледано на друг начин, тој ја дава прецизноста со која една случајна променлива може да биде апроксимирана со линеарна функција од другата. Нека случајната променлива Y ја апроксимираме со линеарна функција од случајната променлива X, т.е. со $aX + b$, каде што a и b се избираат така што просечната квадратна грешка

$$E(Y - (aX + b))^2,$$

да биде минимална. Изедначувајќи ги парцијалните изводи по a и b на 0, со директна пресметка добиваме дека минимумот се достигнува за

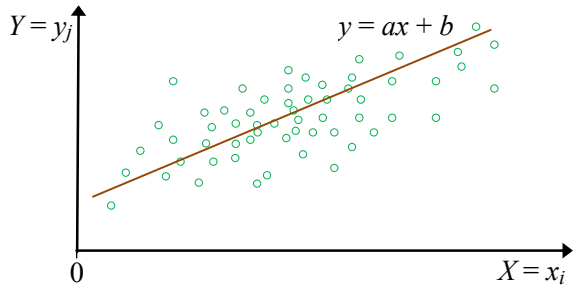
$$a = \frac{K_{X,Y}}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X} \quad \text{и} \quad b = \mu_Y - a\mu_X .$$

Заменувајќи ги овие вредности во просечната квадратна грешка се добива минималната квадратна грешка од

$$\sigma_X^2(1 - \rho^2) .$$

Оттука е јасно дека комплетно совпаѓање во смисла на најмали квадрати се добива за $|\rho| = 1$, а линеарната апроксимација е најлоша кога $\rho = 0$. Значи кога $\rho = \pm 1$, случајните променливи X и Y се идеално корели-

рани во смисла да нивните вредности лежат на права линија со позитивен или негативен пад. Правата линија $y = ax + b$ се нарекува и најдобар линеарен предвидувач за X и Y, бидејќи таа може да се користи за приближно одредување на вредностите за едната случајна променлива ако се дадени вредностите на другата.



Слика 5.3 *Најдобар линеарен предвидувач за X и Y*

Геометриски, најдобриот линеарен предвидувач изгледа како повлекување линија централно низ точките – вредности на случајните променливи X и Y.

ПРИМЕР 5.6 (Продолжение на примерот 5.5) Најди го најдобриот линеарен предвидувач за староста на автомобилот (X) и должините на запирање во метри (Y) при брзина од 40 км/час. Податоците се

X	9	15	24	30	38	46	53	60	64	76
Y	28.4	29.3	37.6	36.2	36.5	35.3	36.2	44.1	44.8	47.2

Решение

Директно решение: Ја формираме $z(a,b) = \sum_{i=1}^{n}(y_i - b - ax_i)^2$. Функцијата $z(\cdot)$ е диференцијабилна, па минимумот по a и b го бараме со изедначување на парцијалните изводи на 0

$$\frac{\partial z}{\partial b} = -2\sum_{i=1}^{n}(y_i - b - ax_i) = 0\,, \quad \frac{\partial z}{\partial a} = -2\sum_{i=1}^{n}(y_i - b - ax_i)x_i = 0\,.$$

Решението на овој систем равенки е

$$a = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,,\ b = \bar{y} - a\bar{x}\,,\ \text{каде што}\ \bar{x} = \frac{1}{n}\sum_{1}^{n}x_i\,,\ \bar{y} = \frac{1}{n}\sum_{1}^{n}y_i$$

За нашите податоци имаме дека $\bar{x} = EX = 41.5$, $\bar{y} = EY = 37.56$ и понатаму, $a = 0.2559$, $b = 26.9402$.

Ако пак a го пресметаме од $a = K_{X,Y} / \sigma_X^2$ = 112.57/440.05 = 0.2558, се разбира добиваме ист резултат (поради заокружувања имаме разлика во 4-тата децимала).

Квадратната грешка на апроксимацијата е

$$\sigma_X^2 (1 - \rho^2) = 440.05 \cdot (1 - 0.915^2) = 71.6291$$

каде што ρ е пресметано во примерот 5.5. ■

Коефициентот на корелација, како сумарна карактеристика, не може да ги замени индивидуалните карактеристика на случајните променливи. Тој само ја изразува силината на линеарната врска меѓу две случајни променливи, но неговата вредност се разбира, не ја карактеризира комплетно нивната врска. На пример, ако условното очекување на Y за дадено X, означено со $E(Y \mid X)$, не е линеарно по X, кофициентот на корелација не го одредува обликот на $E(Y \mid X)$. Додатно, секогаш треба да се има предвид дека корелацијата не повлекува (не изразува) некаква причинско-последична врска меѓу случајните променливи. На пример, познато е дека има висока корелација меѓу тежината на новороденче и неговите достигнувања во образованието на 25 годишна возраст. Од ова никако не би можело да заклучиме дека причината што некој има повисоко образование е што тој бил потешко новороденче. Корелацијата само укажува дека потешко новороденче има добри шанси да има повисоко образование.

Ако парот случајни променливи (X, Y) има заедничка нормална распределба, условното очекување $E(X \mid Y)$ е линеарна функција од Y, а $E(Y \mid X)$ е линеарна функција од X. Коефициентот на кореација ρ заедно со индивидуалните очекувања на X и Y ја дефинира оваа линеарна зависност со

$$E(Y \mid X) = EY + \rho \sigma_Y \frac{X - EX}{\sigma_X}$$

каде што σ_X и σ_Y се стандардните девијации на X и Y.

Покрај Пирсоновиот коефициент на корелација, постојат и други коефициенти на корелација. Во непараметарската статистика се користи Спирмановиот (Spearman) коефициент на корелација што се пресметува преку ранговите на податоците, а не преку нивните вредности. Тој оценува колку добро односот меѓу две случајни променливи може да се опише со монотона функција. Неговата пресметка може да се направи со

$\rho = 1 - \frac{6\sum_{j=1}^{n}(R_{x_j} - R_{y_j})^2}{n(n^2-1)}$, каде што R_{x_j} и R_{y_j} се ранговите на податоците x_j и y_j (нивните позиции кога ќе се сортираат во опаѓачки редослед).

Понатаму, во регресионата анализа (глава 15), ќе разгледаме додатни техники за оценки и тестирања на можната корелираност на податоците.

5.5.2. Бројни карактеристики на условни случајни променливи*

Веќе видовме (поглавје 4.3.2) дека условната распределба на X при дадени вредности на Y во дискретен случај е дадена со

$$p(X = x_i \mid Y = y_j) = \frac{p(X = x_i, Y = y_j)}{p(Y = y_j)}.$$

каде што $p(X = x_i,\ Y = y_j)$, $i = 1, 2,..., m$, $j = 1, 2,..., n$ е распределба на случајниот вектор (X, Y). Условното очекување на X при услов $Y = y_j$ едноставно бројот

$$E(X \mid Y = y_j) = \sum_{i=1}^{m} x_i \frac{p(X = x_i, Y = y_j)}{p(Y = y_j)}$$

што е сосема аналогно на формулата за пресметка на безусловните очекувања. Ако со $E(X \mid Y)$ ја означиме низата вредности $E(X \mid Y = y_j)$, тогаш $E(X \mid Y)$ може да се смета за случајна променлива што прима вредности за различните y_j со веројатности $p(Y = y_j)$. Оттука има смисла да се бара нејзиното очекување за кое важи следната формула

$$EX = E(E(X \mid Y)) = \sum_{i=1}^{m} E(X \mid Y = y_j)\, p(Y = y_j),$$

којашто логично може да се нарече формула на тотално очекување по аналогија на формулата на тотална веројатност.

ПРИМЕР 5.7 Преживувањето на еден човек соочен со снежна бура зависи од тоа по кој од три пата ќе одлучи да продолжи да оди. Првиот пат води до сигурност за 1 час одење, додека вториот пат води до сигурност по 3 часа одење. Третиот пат е кружен и ако тргне по него ќе се врати на почетната локација по 2 часа. Одреди го просечното време на стигнување до сигурност ако изборот на патот е случаен.

Решение

Нека X биде времето на стигнување до сигурност во часови, а нека $Y = 1, 2, 3$ биде избраниот пат. Тогаш $p(Y = i) = 1/3$ за $i = 1, 2, 3$, па просечното време на стигање до сигурност е

$$EX = \sum_{i=1}^{3} E(X \mid Y = i)\, p(Y = i) = \frac{1}{3} \sum_{i=1}^{3} E(X \mid Y = i)\,.$$

Условните очекувања во сумата се:

$p(X \mid Y = 1) = 1,$

$p(X \mid Y = 2) = 3,$

$p(X \mid Y = 3) = 2 + EX$, (проблемот по 2 часа се враќа на почеток).

Оттука добиваме

$EX = \frac{1}{3}(1 + 3 + 2 + EX)$, што дава речение $EX = 3$ часа. ■

Во непрекинат случај, условната густина на распределба е

$$f(x \mid Y = y) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)\,dx}$$

каде што $f(x, y)$ е густина на распределба на случајниот вектор (X, Y). Во овој случај, условното очекување на X при услов $Y = y$ е едноставно бројот

$$E(X \mid Y = y) = \int_{-\infty}^{\infty} x f(x \mid Y = y)\,dx$$

што повторно соодвествува на формулата за пресметка на безусловното очекување. Ако сега повторно го разгледуваме $E(X \mid Y) = E(X \mid Y = y)$ како случајна променлива, имаме дека

$$p(a \le X \le b) = E\left(\int_a^b f(x \mid Y)\,dx \right).$$

Основните особини на условното очекување се сумирани во следните точки:

1) $E(h(Y) \mid Y) = h(Y)$, за функција $h(\cdot)$,

2) $E(h(Y)X \mid Y) = h(Y)E(X \mid Y)$,

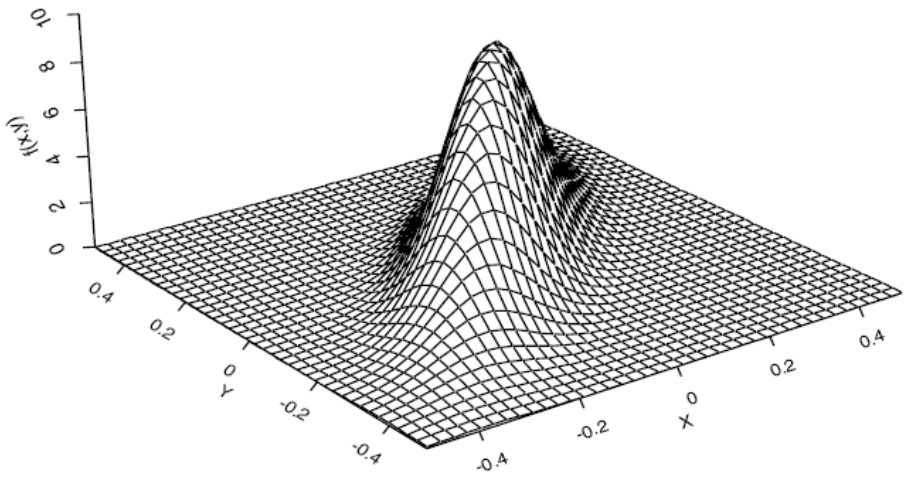
3) $E(X_1 + X_2 \mid Y) = E(X_1 \mid Y) + E(X_2 \mid Y)$,

4) Ако X и Y се независни, $E(X \mid Y) = EX$.

5.5.3. Случајни вектори со нормална распределба

Како што веќе знаеме, еднодимензионалната нормална распределба е наполно определена со очекувањето и дисперзијата. Нејзиното проширување, дводимензионалната нормална распределба е определена со очекувањата и дисперзиите на маргиналните густини на распределба, како и нивниот коефициент на корелација. На сосема сличен начин може да се дојде до повеќедимензионална нормална распределба што се јавува кога разгледуваме сума на независни случајни вектори. Во тој случај, распределбата е наполно определена со очекувањата и матрицата на коваријантност.

На сл. 5.4 е прикажана густина на распределба на случаен вектор со нормална распределба.



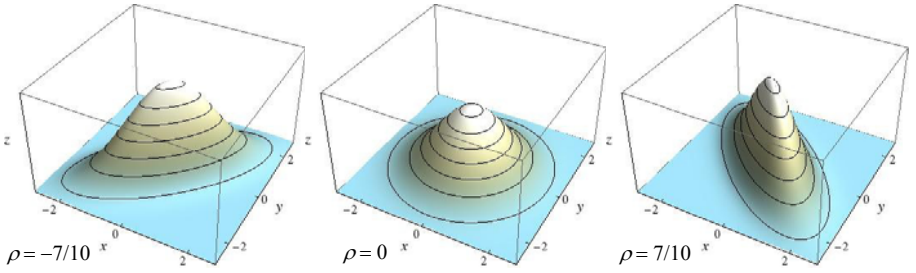
Слика 5.4 *Густина на распределба на случаен вектор (X, Y) со нормална распределба*

Заедничката распределба $f_{X,Y}(x, y)$ на две случајни променливи X и Y со нормална распределба го има следниот облик

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} e^{-Q(x,y)/2}, \text{ каде што}$$

$$Q(x,y) = \frac{1}{(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

Тука ρ е кофициентот на корелација на случајните променливи X и Y. На следната слика се дадени густините на дводимензионални нормални распределби за некои вредности на ρ.



Маргиналните распределби може да се добијат со интегрирање заедничката распределба $f_{X,Y}(x, y)$. Резултатот, се разбира се нормалните распределби $Z(\mu_X, \sigma_X^2)$ и $Z(\mu_Y, \sigma_Y^2)$ за X и Y соодветно.

Кога параметарот ρ е 0 добиваме

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-((x-\mu_X)/\sigma_X)^2 - ((y-\mu_Y)/\sigma_Y)^2} =$$

$$\frac{1}{\sigma_X\sqrt{2\pi}} e^{-((x-\mu_X)/\sigma_X)^2} \frac{1}{\sigma_Y\sqrt{2\pi}} e^{-((y-\mu_Y)/\sigma_Y)^2} = f_X(x) f_Y(y),$$

што значи дека X и Y се независни. Обратното не важи. За случајниот вектор (X, Y) да има дводимензионална нормална распределба не е доволно X и Y да имаат нормална распределба. Контра пример е: Нека X има стандардна нормална распределба и нека $Y = X$ со веројатност 0.5 и $Y = –X$ со веројатност 0.5. Тогаш Y има исто така стандардна нормална распределба. Веднаш се гледа дека $K_{X,Y} = 0$ ($\rho_{X,Y} = 0$) што би повлекувало дека X и Y се независни ако (X, Y) има дводимензионална нормална распределба. Но, X и Y се очигледно зависни што повлекува дека распределбата на (X, Y) не може да биде дводимензионална нормална.

Тука фундаментален резултат е дека случајниот вектор (X, Y) има дводимензионална нормална распределба акко $aX + bY$ (за a и b произволни константи) има еднодимензионална нормална распределба. Доказот (во едната насока) бара користење додатни математички алатки и тука не го даваме.

ПРИМЕР 5.8 *Пијано шетање* (Drunkard's walk). Во рамнина тргнувајќи од (0, 0) се придвижува честичка за единечно растојание во случајно избран правец $\theta \in [0, 2\pi]$. Во секој чекор движењето се повторува од новата локација. Која е

густината на распределба на случајниот вектор (x, y) = "координати на честичката по n чекори"? (Ова е еден од најкорисните веројатносни модели во физиката.)

Решение

Нека θ е правецот на движење во секој чекор. Тогаш положбата на честичката се менува за векторот ($\cos\theta$, $\sin\theta$) каде што случајната променлива θ има рамномерна распределба во [0, 2π]. Ако со X_k и Y_k ја означиме промената на положбата на честичката во k–тиот чекор, нејзината положба по n чекори е дадена со случајниот вектор (U, V) каде што

$$U = X_1 + X_2 + \ldots + X_n \quad \text{и} \quad V = Y_1 + Y_2 + \ldots + Y_n .$$

Случајните вектори (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) се независни и со иста распределба, па според централната гранична теорема нивната сума за доволно големо n (на пример ≥ 30) приближно има нормална распределба со

$$\mu_U = EX_j,\ \ \mu_V = EY_j,\ \ \sigma_U^2 = DX_j\,,\ \sigma_V^2 = DY_j\,,$$

$$\rho = \rho(X_j,\, Y_j).$$

Во конкретниот случај

$$\mu_U = E\cos\theta = \int_0^{2\pi} f(\theta)\cos\theta\, d\theta = \frac{1}{2\pi}\int_0^{2\pi}\cos\theta\, d\theta = 0\,,\ \text{слично и } \mu_V = 0,$$

$$\sigma_U^2 = E\cos^2\theta = \int_0^{2\pi} f(\theta)\cos^2\theta\, d\theta = \frac{1}{2\pi}\int_0^{2\pi}\cos^2\theta\, d\theta = \frac{1}{2}\,,\ \text{исто и } \sigma_V^2 = \frac{1}{2}\,,$$

а $K_{X_j,Y_j} = E(\cos\theta\sin\theta) = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi}\cos\theta\sin\theta\, d\theta = 0$.

Според централната гранична теорема (поглавје 7.1)

$$p(U < x, V < y) \approx \frac{1}{\pi\cdot n}\int_{-\infty}^{x}\int_{-\infty}^{y} e^{-(u^2+v^2)/n}\,dudv\,,\ \text{па оттука}$$

густината на распределба на случајниот вектор (U, V) е приближно

$$f_{U,V}(x,y) = \frac{1}{\pi\cdot n}e^{-(x^2+y^2)/n}\,,\ \text{за доволно големо } n.$$

Веројатноста честичката по n чекори да се најде во мал правоаголник со страни Δa и Δb е $f_{U,V}(a, b)\Delta a\Delta b$. Од разложувањето

$$f_{U,V}(x,y) = \frac{1}{\sqrt{2\pi}\cdot\sqrt{n/2}}e^{-\frac{1}{2}x^2/\frac{1}{2}n}\,\frac{1}{\sqrt{2\pi}\cdot\sqrt{n/2}}e^{-\frac{1}{2}y^2/\frac{1}{2}n}$$

следува дека секоја од координатите на честичката по n чекори има приближно $Z(0,\, n/2)$ распределба, кога n е доволно големо. Ако случајната променлива

D = "растокание од координатниот почеток до позицијата на честичката по n чекори", од $D = \sqrt{U^2 + V^2}$ имаме дека

$$ED \approx \frac{1}{\pi \cdot n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{x^2 + y^2} e^{-(x^2+y^2)/n} dxdy \;=\; (1/2)\sqrt{\pi \cdot n}\,. \;\blacksquare$$

Повеќедимензионалната нормална распределба најчесто се користи да опише (најмалку приближно) кое било множество од (можно) корелирани случајни променливи, од кои секоја групира вредности околу својата очекувана вредност.

Дефиниција 5.6 n-димензионалниот случаен вектор (X_1, X_2, ..., X_n) има повеќедимензионална нормална распределба ако секоја линеарна комбинација

$$a_1X_1 + a_2X_2 + \ldots + a_nX_n, \qquad a_i \in \mathbb{R}$$

има (еднодимензионална) нормална распределба.

Од дефиницијата следува дека и секоја линеарна комбинација на повеќедимензионални нормални случајни променливи има еднодимензионална нормална распределба, што има големо значење во практиката.

Во n-димензионален случај, нормалната распределба на случајниот вектор (X_1, X_2, ..., X_n) е определена со просекот

$$\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^n$$

и матрицата на коваријантност

$$\Sigma = [\, K_{X_i, X_j} \,]_{n \times n} \in \mathbb{R}^{n \times n}$$

којашто е симетрична и позитивно дефинитна. Оттука, нормалната распределба е $Z(\vec{\mu}, \Sigma)$, дадена со

$$f_{X_1, X_2, \ldots, X_n}(\vec{x} = (x_1, x_2, \ldots, x_n)^T) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}\,.$$

Во дводимензионален случај, кога X и Y се независни, едноставно имаме

$$\vec{x} = [x \quad y]^T, \;\; \vec{\mu} = [\mu_X \quad \mu_Y]^T, \;\; \Sigma = \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{bmatrix} \;\Rightarrow\; \Sigma^{-1} = \begin{bmatrix} 1/\sigma_X^2 & 0 \\ 0 & 1/\sigma_Y^2 \end{bmatrix}$$

и оттука добиваме

$$(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu}) = \begin{bmatrix} \frac{x-\mu_X}{\sigma_X^2} & \frac{y-\mu_Y}{\sigma_Y^2} \end{bmatrix} \begin{bmatrix} x-\mu_X \\ y-\mu_Y \end{bmatrix} = \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}$$

што претходно веќе го имавме воведено.

Кога X и Y не се независни, имаме дека

$$\vec{\mu} = [\mu_X \quad \mu_Y]^T, \quad \Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

и оттука се добива нормалната густина на распределба за случајниот вектор (X, Y), која исто така претходно ја воведовме.

Кога земаме случаен примерок X_1, X_2, ..., X_n во врска со некој експеримент за статистичка анализа, секоја реализација на експериментот е случајна променлива X_i. За случајниот примерок вообичаено се прават следните претпоставки: распределба: *произволна*, зависност: *независни случајни променливи* и хетерогеност: *идентично распределени случајни променливи*. Во таков случај, имаме дека $EX_i = \mu$ и $DX_i = \sigma^2$, $i = 1, 2,\ldots, n$, па за очекувањето и дисперзијата на просекот на примерокот $\overline{X} = (X_1 + X_2 + ... + X_n)/n$ имаме

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}(n\mu) = \mu,$$

$$D\overline{X} = E(\overline{X}-\mu)^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right)^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Кога се работи за распределбата на $\overline{X}$, според централната гранична теорема таа се приближува кон нормална кога $n \to \infty$. Попрецизно, распределбата на случајната променлива $(\overline{X}-\mu)(\sqrt{n}/\sigma)$ конвергира кон $Z(0,1)$ кога $n \to \infty$.

ПРИМЕР 5.9 Безалхохолен пијалок се полни во шишиња од страна на автоматизиран полнач. Просекот на волуменот на полнењето е 12.1 унци течност со стандардна девијација од 0.1. Нека волумените на полнење се независни, нормално распределени случајни променливи. Пресметај ја веројатноста дека просечниот волумен на 10 избрани шишиња е помал од 12 унци течност.

Решение

Нека X_1, X_2, ..., X_{10} се волумените на полнење на 10-те шишиња. Просечниот волумен $\overline{X}$ на полнењето е случајна променлива со нормална распределба и

$E\overline{X} = 12.1$, $D\overline{X} = \frac{0.1^2}{10} = 0.001$, па оттука

$$p(\overline{X} < 12) = p\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < \frac{12 - 12.1}{\sqrt{0.001}}\right) = p(z < -3.16) = 0.00079 .$$ ■

ЗАДАЧИ

1. Трговска фирма има 8 компјутери со кои договара размена на стоки и услуги. Веројатноста дека компјутер ќе откаже во еден ден е 0.005 (откажувањето е независно). Компјутерите се поправаат навечер секој ден, што е независен настан.

   а) Пресметај ја веројатноста дека сите 8 компјутери ќе откажат во ист ден;

   б) Кој е просечниот број денови до откажување на еден компјутер?

   в) Кој е просечниот број денови до откажување на сите 8 компјутери во ист ден?

2. Еден автоматизарана лента за полнење конзерви застанува кога се детектираат 3 конзерви со помала тежина. Веројатноста конзерва да има помала тежина е 0.001, а секое полнење е независно.

   а) Кој е просечниот број полнења пред линијата да запре?

   б) Која е стандардната девијација на број на полнења пред линијата да запре?

3. Грешките во пластичните панели што се користат за ентериери во автомобилите се со Пуасонова распределба со просек од 0.02 грешки по панел.

   а) Колкава е веројатноста дека во 50 случајно избрани панели нема да има грешка;

   б) Кој е просечниот број панели што треба да се испитаат за да се појави грешка?

4. (*Продолжение на задача 7 од глава 4.*) Случајната променлива X означува дијаметар на отворот што дупчалка го прави во еден метален дел. Дијаметарот треба да биде 12.5 милиметри, а секој отвор со дијаметар поголем од 12.6 го прави делот неупотреблив. Вообичаено, грешките се од вибрации што резултираат во зголемен дијаметар. Од поранешни податоци познато е

дека густината на распределба на X е приближно $f(x) = 20e^{-20(x-12.5)}$. Пресметај го очекувањето и дисперзијата на X.

5. Времето на чекање X (во минути) пред шалтер има густина на распределба

$$f(x) = \begin{cases} 2e^{-2x}, & \text{за } x \geq 0 \\ 0, & \text{во спротивно} \end{cases}$$

Пресметај го просечното време на чекање и дисперзијата.

6. Дебелината X на кондуктивната обвивка во микрометри има густина на распределба $600x^{-2}$ за $100 < x < 120$ микрометри.

а) Определи го просекот и дисперзијата на дебелината на обвивката;

б) Ако обвивката чини 0.50 евра по микрометар колкава е просечната цена на обвивката по дел?

7. Нака времето X потребно еден оператор да пополни електронска форма со податоци е со рамномерна распределба меѓу 1.5 и 2.5 минути. Определи го просекот и дисперзијата на времето потребно на операторот да пополни форма.

8. Времето на чекање T на патник за авионски лет се карактеризира со мешана функција на распределба

$$F(t) = \begin{cases} p + (1-p)(1-e^{-\lambda t}), & \text{за } t \geq 0 \\ 0, & \text{за } t < 0 \end{cases}$$

Опредечи го просечното време на чекање ET, како и просечното време на чекање ако чекањето е неопходно, $E(T \mid T > 0)$.

9. Тежината на тркачките патики е нормално распределена со просек од 12 унци и стандардна девијација од 0.5 унци.

а) Пресметај ја веројатноста патиките да тежат повеќе од 13 унци?

б) Колкава треба да биде стандардната девијација на тежината на патиките за компанијата да може да тврди дека 99.9% од патиките се полесни од 13 унци?

в) Ако стандардната девијација е 0.5 унци, колкав треба да биде просекот за компанијата да може да тврди дека 99.9% од патиките се полесни од 13 унци?

10. Грешки на некој тип оптички дискови се јавуваат со рата од една грешка на на секој 10^5 бита. Под претпоставка дека грешките се со Пуасонова распределба определи го

а) Просечниот број на битови додека не се појават 5 грешки;

б) Стандардната девијација на бројот на битови додека не се појават 5 грешки.

11. Животот (во часови) на еден уред за магнетна резонанца (MRI) се моделира со Веибул распределба со параметри $\beta = 2$ и $\delta = 500$ часа.

    а) Определи го просечниот животен век на уредот;

    б) Определи ја дисперзијата на животен век на уредот;

    в) Пресметај ја веројатноста уредот да откаже пред 250 часови работа.

12. (*Продолжение на задача 22 од глава 4.*) Во една нарачка од 15 печатачи: 4 се со проширени графички можности, 5 со проширена меморија и 6 со двете карактеристики. Случајно се бираат 4 печатачи. Нека X, Y и Z се бројот на печатачите во примерокот со проширени графички можности, проширена меморија и двете карактеристики соодветно. Определи ги EX и DX.

13. (*Продолжение на задача 23 од глава 4.*) Една квалитетна WEB страна за мал бизнис содржи 100 страни при што 60%, 30% и 10% од страните содржат мала, средна и висока графичка содржина соодветно. Земен е примерок од 4 страни и нека X и Y означуваат број на страни со средна и висока графичка содржина соодветно. Определи ги а) EX, б) $E(Y \mid X = 3)$.

14. (*Продолжение на задача 23 од глава 4.*) Нека случајната променлива X означува потребно време (во милисекунди) за конекција на компјутер со компјутерски сервер, а Y потребно време (во милисекунди) до авторизацијата на корисникот на компјутерот на серверот ($Y = X +$ ауторизација). Распределбата на счучајниот вектор (X, Y) е $f_{XY}(x, y) = 6 \cdot 10^{-6} e^{-0.001x - 0.002y}$, $x < y$. Пресметај го условното очекување на Y за дадено $X = 1500$.

15. Определи го c така што $f_{XYZ}(x, y, z) = c$ да биде густина на распределба за случајниот вектор (X, Y, Z) во областа $x > 0$, $y > 0$, $z > 0$ и $x + y + z < 1$. Потоа пресметај го EX.

16. Инженер сака да ја тестира применливоста на законот за идеален гас на затворен резервоар со фиксна количина гас. Тој смета дека во такви услови, законот предвидува линеарна зависност меѓу температурата и притисокот на гасот. Мерењата што ги направил се дадени во следната табела:

Притисок (atm)	1.2	1.3	1.5	1.7	1.8	2	2.1	2.3	2.8
Температура (K)	298	301	302	310	312	323	337	341	349

    Дали мерењата покажуваат линеарна зависност температура – притисок?

17. Определи го коефициентот на корелација за X и Y со дадена заедничка густина на распределба $f_{XY}(x, y) = e^{-x-y}$ во областа $0 < x$ и $0 < y$.

18. Членак опишува студија за врската меѓу изложувањето на бучава и високиот крвен притисок. Во членакот се дадени следните податоци каде што X = "покачување на крвниот притисок", Y = "бучава во децибели".

X (mm/hg)	1	0	1	2	5	1	4	6	2	3	5	4	6	8	4	5	7	9	7	6
Y (atm)	60	63	65	70	70	70	80	90	80	80	85	89	90	90	90	90	94	100	100	100

а) Пресметај го коефициентот на корелација;

б) Креирај го најдобриот линеарен предвидувач по најмали квадрати и најди ја минималната квадратна грешка;

ц) Колкав пораст на крвниот притисок предвидува овој модел за бучава од 85 децибели.

19. Нека (X, Y) има дводимензионална нормална распределба со $DX = DY$. Покажи дека случајните променливи $X+Y$ и $X-Y$ се независни и нормално распределени.

20. Да претпоставиме дека тежините на луѓето се независни и нормално распределени со просек од 160 фунти и стандардна девијација од 30 фунти. Нека 25 луѓе влезат во индустриски лифт со носивост од 4300 фунти.

а) Колкава е веројатноста дека тие ќе ја надминат носивоста на лифтот?

б) Која носивост се надминува од 25 луѓе со веројатност 0.0001?

6

Функции од случајни променливи

Има многу начини да се направат нови случајни променливи од веќе постоечките. Се разбира тоа не е цел сама за себе. Обично новите случајни променливи настануваат природно при решавање на практични проблеми. Јасно е дека секоја функција применета на постоечка случајна променлива може да даде нова случајна променлива. Природни прашања што тука се наметнуваат се: Што се случува кога се прави трансформација на случајни променливи? Што се случува со нивните функции на распределба? Дали се сочувува независноста?

Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност и нека $X = X(\omega)$ е некоја случајна променлива. Композицијата на реалната функција $X(\omega)$ со домен во Ω и реалната функција $h(x)$ со домен во реалните броеви дава реална функција $Y = h(X(\omega)) = Y(\omega)$ со домен во Ω. За дискретен простор на веројатност, Y е секогаш случајна променлива, па тогаш нема потреба од никакви ограничувања на $Y(\omega)$, т.е. на $h(x)$. За произволен простор на веројатност, се разбира треба за секој x да важи

$$(Y < x) \in \mathcal{F}.$$

За произволна функција $h(x)$ ова не е секогаш исполнето иако таквите "егзотични" случаи не се јавуваат во пракса. Покажано е дека $(Y < x) \in \mathcal{F}$ е исполнето ако за произволно множество B од σ-алгебрата на реалната оска, множеството $\{u: h(u) \in B\}$ е повторно од σ-алгебрата на реал-

ната оска. Со други зборови, инверзната слика $h^{-1}(B)$ е множество од σ-алгебрата на реалната оска. Функции $h(x)$ за кои ова е исполнето се сите непрекинати функции, функции со конечен или пребројив број на прекини или најопшто, функции коишто се непрекинати скоро секаде, т.е. множеството прекини има мера 0. За ваквите функции се користи терминот мерливи функции. Ова значи дека секоја "нормална" функција е мерлива функција, и практично секогаш може да сметаме дека функција од случајна променлива е повторно некоја случајна променлива. Целата оваа дискусија е повторно поврзана со одредената "незграпност" на реалните броеви за опишување на реалноста.

6.1. Функции од дискретни случајни променливи

Да го разгледаме најпрво дискретниот случај кој е доста поедноставен. Нека е дадена дискретна случајна променлива X

x_1	x_2	...	x_n	...
p_1	p_2	...	p_n	...

и нека $h(x)$ е функција дефинирана во сите точки x_i. Ја разгледуваме случајната променлива $Y = h(X)$ што ги прима вредностите $y_i = h(x_i)$, $i = 1, 2, \ldots$.

Ако сега сите вредности се различни, $y_i \neq y_j$ за $i \neq j$ (на пример $h(x)$ е стриктно монотона) тогаш Y има закон на распределба

y_1	y_2	...	y_n	...
p_1	p_2	...	p_n	...

што има исти веројатности како распределбата на X.

Ако пак меѓу вредностите на y има еднакви, тогаш на нив едноставно им се доделуваат веројатности што се сума на сите p_k такви што соодветните x_k се пресликани во иста ведност. Генерално имаме

$$p(Y = y) = p(h(X) = y) = \sum_{x_k : h(x_k) = y} p(X = x_k) .$$

Имено, ако за $x_{i_1}, x_{i_2}, \ldots, x_{i_s}$ се добива исто $y_i = h(x_{i_k})$, $k = 1, 2, \ldots, s$ тогаш земаме дека $p(Y = y_i) = \sum_{k=1}^{s} h(x_{i_k})$ бидејќи настанот $Y = y_i$ може да се претстави како сума на дисјунктните настани ($X = x_{i_k}$), $k = 1, 2, \ldots, s$.

На пример, ако X е дадена со законот на распределба

X	–2	–1	0	1	2	3
$p(X = x_i)$	0.1	0.2	0.2	0.3	0.1	0.1

и $Y = X^2 - 1$, тогаш Y е определена со законот на распределба

Y	–1	0	3	8
$p(Y = y_i)$	0.2	0.5	0.2	0.1

.

6.2. Функции од непрекинати случајни променливи

Сега нека е дадена непрекината случајна променлива Y со густина на распределба $f(x)$ и нека $h(x)$ е функција дефинирана во сите точки $f(x)$ Ја разгледуваме случајната променлива $Y = h(X)$ што има непозната густина на распределба $g(y)$. Функцијата на распределба $G(y)$ на случајната променлива Y може да се определи со

$$G(y) = p(Y < y) = p(h(X) < y) = \int_{x:h(x)<y} f(x)dx .$$

Ако $G(y)$ е функција на распределба на непрекината случајна променлива, нејзината густина се добива едноставно со диференцирање

$$g(y) = \frac{\partial G(y)}{\partial y},$$

во точките во кои $G(y)$ е диференцијабилна.

ПРИМЕР 6.1 Најди ја густината на распределба на линеарна функција од случајна променлива, а потоа примени го резултатот на нормалната распределба.

Решение

Нека случајната променлива X има густина на распределба $f(x)$. Ја разгледуваме случајната променлива $Y = aX + b$, каде што a и b се константи. Најпрво го разгледуваме случајот кога $a > 0$. Тогаш за функцијата на распределба $G(y)$ на Y имаме

$$G(y) = p(Y < y) = p(aX+b < y) = p(X < \frac{y-b}{a}) = \int_{-\infty}^{\frac{y-b}{a}} f(x)dx .$$

Оттука, имајќи ја предвид диференцијабилноста на $G(y)$ имаме

$$g(y) = G'(y) = f(\frac{y-b}{a})(\frac{y-b}{a})' = \frac{1}{a} f(\frac{y-b}{a}) .$$

Во случај кога $a < 0$ имаме

$$G(y) = p(Y < y) = p(aX+b < y) = \ p(X > \frac{y-b}{a}) = 1 - \int_{-\infty}^{\frac{y-b}{a}} f(x)dx$$

што дава $g(y) = -\frac{1}{a} f(\frac{y-b}{a})$.

Обединето, ова дава густина на распределба на Y

$$g(y) = \frac{1}{|a|} f(\frac{y-b}{a}).$$

Во случај X да има нормална распределба, за густината на Y добиваме

$$g(x) = \frac{1}{|a|} f(\frac{x-b}{a}) = \frac{1}{|a|} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\frac{x-b}{a}-\mu)^2}{2\sigma^2}} = \frac{1}{|a|} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-b-a\mu)^2}{2a^2\sigma^2}},$$

што е повторно нормална распределба $Z(b + a\mu,\ a^2\sigma^2)$. ■

За пресметка на густините на распределба на функциите од случајните променливи, корисен е следниот резултат.

Теорема 6.1 Нека X е непрекината случајна променлива во некој интервал (a, b) (a или/и b може да се ∞) со функција на распределба $F(x)$ и густина на распределба $f(x)$. Ако $h(x)$ е стриктно монотона функција во (a, b) со непрекинат извод во тој интервал, тогаш густината на распределба $g(y)$ на $Y = h(X)$ е

$$g(y) = f(h^{-1}(y)) \cdot |(h^{-1}(y))'|.$$

Доказ: Нека $h(x)$ стриктно монотоно расте во (a, b). Тогаш и $h^{-1}(y)$ е еднозначна и монотоно расте во интервалот $(h(a), h(b))$ и е диференцијабилна. За функцијата на распределба $G(y)$ на Y имаме

$$G(y) = p(Y < y) = p(h(X) < y) = p(X < h^{-1}(y)) = F(h^{-1}(y)).$$

Со диференцирање на $G(y)$ добиваме

$$g(y) = G'(y) = F'(h^{-1}(y)) \cdot (h^{-1}(y))',$$

и сега бидејќи $h^{-1}(y)$ монотоно расте, имаме дека $(h^{-1}(y))' > 0$, па може да се стави

$$g(y) = f(h^{-1}(y)) \cdot |(h^{-1}(y))'|.$$

Ако $h(x)$ стриктно монотоно опаѓа во (a, b) тогаш и $h^{-1}(y)$ монотоно опаѓа во $(h(a), h(b))$ и затоа $(h^{-1}(y))' < 0$. Работејќи како претходно добиваме

$$G(y) = p(h(X) < y) = p(X > h^{-1}(y)) = 1 - p(X < h^{-1}(y)) = 1 - F(h^{-1}(y)),$$

и сега повторно имаме

$$g(y) = -f(h^{-1}(y))\cdot(h^{-1}(y))' = f(h^{-1}(y))\cdot|(h^{-1}(y))'|. \blacksquare$$

ПРИМЕР 6.2 Дадена е густината на распределба $f(x)$ на случајната променлива X. Најди ја густината на распределба на случајната променлива Y во следните случаи:

а) $Y = e^{-X}$, б) $Y = \ln X$,
в) $Y = \sqrt{X}$, г) $Y = 1/X^2$.

Решение

а) Функцијата $h(x) = e^{-x}$ е монотоно опаѓачка па според теоремата веднаш добиваме

$h^{-1}(x) = \ln\frac{1}{x}$, што дава $g(x) = \left|-\frac{1}{x}\right| f(\ln\frac{1}{x})$ што е позитивна и дефинирана во (0, 1). Оттука бараната густина е

$$g(x) = \begin{cases} \frac{1}{x} f(\ln\frac{1}{x}) & \text{за } x \in (0,1) \\ 0 & \text{во спротивно} \end{cases}.$$

б) Функцијата $h(x) = \ln x$ е монотоно растечка во (0, ∞) па инверзната $h^{-1}(x) = e^x$ е растечка во (ln0, ln∞) = (–∞, ∞) и така

$$g(x) = e^x f(e^x) \text{ за } x \in (-\infty, \infty).$$

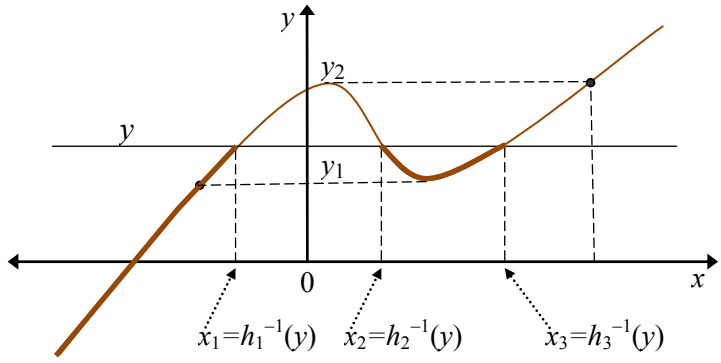
в) Функцијата $h(x) = \sqrt{x}$ е монотоно растечка во (0, ∞) па инверзната $h^{-1}(x) = x^2$ е растечка во (0, ∞) и така

$$g(x) = \begin{cases} 2xf(x^2) & \text{за } x \in (0,\infty) \\ 0 & \text{во спротивно} \end{cases}.$$

г) Функцијата $h(x) = 1/x^2$ е монотоно растечка во (–∞, 0) и монотоно опаѓачка во (0, ∞), а инверзната функција $h^{-1}(x) = \pm 1/\sqrt{x}$ е дефинирана во (0, ∞) и така

$$g(x) = \begin{cases} \frac{1}{2x\sqrt{x}} f(\frac{1}{\sqrt{x}}) & \text{за } x \in (0,\infty) \\ 0 & \text{во спротивно} \end{cases}. \blacksquare$$

Кога функцијата $h(x)$ не е стриктно монотона во некоја област, неа можеме да ја разгледуваме посебно во подобласти во кои таа е монотона. На пример, да ја разгледаме немонотоната функција од сл. 6.1.



Слика 6.1 *Илустрација на немонотона функција на случајна променлива*

Функцијата го задоволува условот за монотоност во секој интервал од областите $y < y_1$ и $y > y_2$. Во овие области, за добивање на густината на распределба на Y, може да се искористи теоремата 6.1. Од друга страна, во областа $y_1 < y < y_2$ мораме одново да ја разгледуваме функцијата на распределба $G(y) = p(h(X) < y)$ на Y. За y во таа област, $Y = h(X) < y$ е означено со здебелена функција $h(x)$ на сл. 6.1. Оттука имаме

$$\begin{aligned} G(y) &= p(h(X) < y) = p(X < h_1^{-1}(y)) + p(h_2^{-1}(y) < X < h_3^{-1}(y)) = \\ &= p(X < h_1^{-1}(y)) - p(X < h_2^{-1}(y)) + p(X < h_3^{-1}(y)) = \\ &= F(h_1^{-1}(y)) - F(h_2^{-1}(y)) + F(h_3^{-1}(y)), \quad \text{за } y_1 < y < y_2, \end{aligned}$$

каде што $x_1 = h_1^{-1}(y)$, $x_2 = h_2^{-1}(y)$ и $x_3 = h_3^{-1}(y)$ се корени по x за функцијата $y = h(x)$ за дадено y. На сликата тоа се x–овците, во кои вредноста y го сече графикот на функцијата $h(x)$. Како и претходно, за добивање на густината на распределба диференцираме и добиваме

$$g(y) = f(h_1^{-1}(y))\cdot(h_1^{-1}(y))' - f(h_2^{-1}(y))\cdot(h_2^{-1}(y))' + f(h_3^{-1}(y))\cdot(h_3^{-1}(y))'$$

и сега ако земеме предвид дека $(h_2^{-1}(y))' < 0$ бидејќи тука $h^{-1}(y)$ опаѓа (бидејќи опаѓа и функцијата $h(y)$), функцијата $g(y)$ може да се напише во облик

$$g(y) = \sum_{i=1}^{3} f(h_i^{-1}(y)) \cdot |(h_i^{-1}(y))'| \quad \text{за } y_1 < y < y_2.$$

Да забележиме дека сумата е од 3 елемента што соодвествува на бројот на корени на равенката $y = h(x)$ за дадено y.

Оваа дискусија може да се обопшти за општа немонотона функција $h(x)$ каде што $h^{-1}(y)$ е нееднозначна од r-ти ред, т.е. кога $y = h(x)$ има r корени за дадено y.

Теорема 6.2 Нека X е непрекината случајна променлива во некој интервал (a, b) (a или/и b може да се ∞) со функција на распределба $F(x)$ и густина на распределба $f(x)$. Ако $h(x)$ е функција во (a, b) со непрекинат извод во тој интервал и $h^{-1}(y)$ има најмногу пребројиво многу корени $x_1 = h_1^{-1}(y)$, $x_2 = h_2^{-1}(y)$, ..., $x_r = h_r^{-1}(y)$, тогаш густината на распределба $g(y)$ на $Y = h(X)$ е

$$g(y) = \sum_{i=1}^{r} f(h_i^{-1}(y)) \cdot |(h_i^{-1}(y))'| .$$

Доказ: Директно следи од претходната дискусија во врска со сл. 6.1. ■

Да забележиме дека теоремата 6.1 е специјален случај на теоремата 6.2, кога $r = 1$, т.е. функцијата $h(x)$ е стриктно монотона па $h^{-1}(y)$ секогаш има само еден корен.

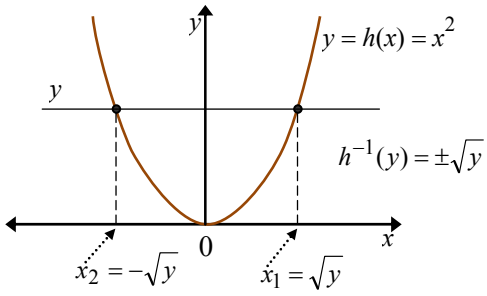
ПРИМЕР 6.3 Случајната променлива X има нормална распределба $Z(0,1)$ со густина

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} .$$

Да се пресмета густината на распределба на случајната променлива $Y = X^2$.

Решение

Најпрво е јасно дека $g(y) = 0$ за $y < 0$ бидејќи Y е секогаш позитивна, т.е. $h^{-1}(y)$ нема корени во таа област ($y < 0$). За $y \geq 0$ ја имаме ситуацијата како на сликата подолу.



Сега директно користејќи ја теоремата 6.1, добиваме

$$g(y)=\sum_{i=1}^{2} f(h_i^{-1}(y))\cdot|(h_i^{-1}(y))'|=\frac{f(\sqrt{y})}{2\sqrt{y}}+\frac{f(-\sqrt{y})}{2\sqrt{y}}=\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}.$$

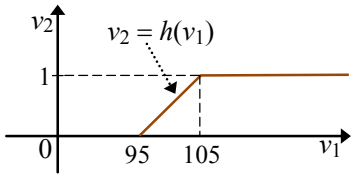
Така, бараната густина на распределба на Y е

$$g(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}} & \text{за } x\geq 0\\ 0 & \text{во спротивно}\end{cases}$$

и тоа е χ^2 густина на распределба (за $n=1$) со 1 степен на слобода. ■

ПРИМЕР 6.4 Случаен напон v_1 со рамномерна распределба во интервалот [90, 110] волти се става на нелинеарен лимитер што прави негова трансформација според функцијата

$$h(v)=\begin{cases}0 & \text{за } v<95\\ \dfrac{v-95}{10} & \text{за } 95\leq v\leq 105\\ 1 & \text{за } v>105\end{cases},$$ дадена на следната слика



Да се пресмета густината на распределба на излезниот напон.

Решение

Тука теоремата 6.1 не е применлива бидејќи во областите $v_1 < 95$ и $v_1 > 105$, постојат непребројиво многу корени за v_1. Но затоа, директно од сликата гледаме дека

$$p(v_2=0)=p(v_1<95)=\int_{90}^{95}\frac{1}{20}dv=\frac{1}{4}\ \text{и}$$

$$p(v_2=1)=p(v_1>105)=\int_{105}^{110}\frac{1}{20}dv=\frac{1}{4}.$$

Имајќи предвид дека $G(y)=F(h^{-1}(y))$, за средниот дел имаме

$$G(v_2)=F(h^{-1}(v_2))=F(10v_2+95),\ 0<v_2<1.$$

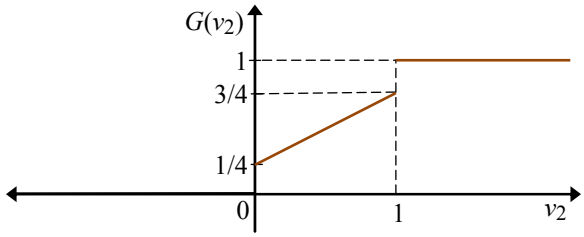
На крај, од функцијата на распределба на v_1 којашто е

$$F(v_1) = \frac{v_1 - 90}{20}, \quad 90 < v_1 < 110,$$

ја добиваме функцијата на распределба на v_2

$$G(v_2) = F(10v_2 + 95) = \frac{10v_2 + 95 - 90}{20} = \frac{2v_2 + 1}{4}, \quad 0 < v_2 < 1.$$

Функцијата на распределба $G(v_2)$ не е непрекината (е мешана), така што и соодветната случајна променлива не е непрекината. Таа е непрекината во интервалот (0,1) и дискретна надвор од него. Нејзиниот облик е даден на следната слика.



Токму таквата ситуација не принуди да работиме со функции на распределба наместо со густини. Забележи дека густина на распределба од 1/2 може да се определи само во интервалот (0,1) што дава плоштина од 1/2 и заедно со двата скока од по 1/4 во дискретниот дел имаме вкупна сума 1. ■

Кога се работи за независност на случајните променливи, таа со трансформација на случајните променливи останува да важи.

Теорема 6.3 Ако X_1 и X_2 се две независни случајни променливи, тогаш се независни и случајните променливи $Y_1 = h_1(X_1)$ и $Y_2 = h_2(X_2)$.

Доказ: Нека B_1 и B_2 се две множества од σ-агебра на реалната оска. Тогаш имаме

$$p(Y_1 \in B_1,\ Y_2 \in B_2) = p(h_1(X_1) \in B_1,\ h_2(X_2) \in B_2) =$$

$$= p(X_1 \in h_1^{-1}(B_1),\ X_2 \in h_2^{-1}(B_2)) = [h_1^{-1}(B_1) \text{ и } h_2^{-1}(B_2) \text{ се од σ-агебрата}] =$$

$$= p(X_1 \in h_1^{-1}(B_1)) \cdot p(X_2 \in h_2^{-1}(B_2)) = p(Y_1 \in B_1) \cdot p(Y_2 \in B_2). \blacksquare$$

6.3. Функции од повеќе случајни променливи*

Случајот на функции од повеќе случајни променливи веројатно е почест во практиката. Типични примери се кога бараме сума или просек на случајни (или функции од случајни) променливи.

Во ваквиот поопшт случај, случајната променлива Y е функција од n заеднички распределени случајни променливи $X_1, X_2, ..., X_n$,

$Y = h(X_1, X_2, ..., X_n)$.

Знаејќи ја заедничката распределба на овие случајни променливи, целта е да се определи распределбата на Y.

Како и во случајот на функции од една случајна променлива, случајот кога $X_1, X_2, ..., X_n$ се дискретни не претставува никаков проблем. Тогаш едноставно Y ги добива сите вредностите што произлегуваат од функцијата $h(\cdot)$, а соодветните веројатности се добиваат со сумирање на веројатностите од заедничкиот закон на распределба за вредности на $X_1, X_2, ..., X_n$, за кои се добиваат соодветните вредности на Y.

Затоа овде се фокусираме на случајот кога случајните променливи $X_1, X_2, ..., X_n$ се непрекинати, со позната заедничка функција или густина на распределба

$F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$, т.е. $f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$.

Како и претходно, за да ја најдеме функцијата на распределба $G(y)$ на Y тргнуваме од

$$G(y) = p(Y < y) = p(h(X_1, X_2, ..., X_n) < y) =$$
$$= F((x_1, x_2, ..., x_n): h(x_1, x_2, ..., x_n) < y) =$$
$$= \int\ldots\int_{(\mathbb{R}^n : h(x_1,x_2,\ldots,x_n) < y)} f(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n$$

каде што интегралот е дефиниран по n-димензионалната област од $\mathbb{R}^n$, за која е исполнето $h(x_1, x_2, ..., x_n) < y$.

Значи густината на распределба $g(y)$ на Y може да се определи преку соодветната функцијата на распределба $G(y)$ која пак се определува од n-димензионалниот интеграл. Критичен чекор во оваа постапка е определувањето на областа на интегрирање од $\mathbb{R}^n$, којашто се разбира зависи од конкретниот проблем. Со растот на n, тоа може да стане многу комплициран проблем.

Да ја разгледаме постапката на определување на распределбата на случајна променлива што е функција од непрекинати случајни променливи низ неколку примери.

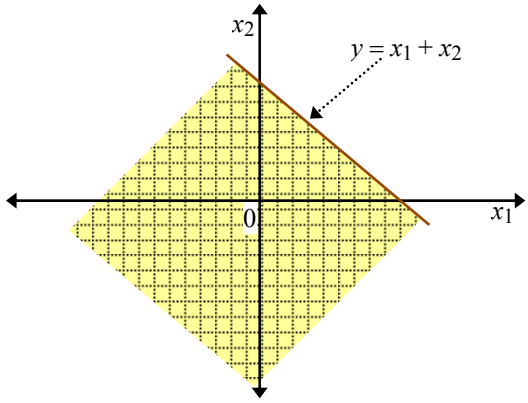
ПРИМЕР 6.5 Нека е $Y = X_1 + X_2$. Најди ја густината на распределба на Y изразена преку заедничката густина $f_{X_1,X_2}(x_1,x_2)$.

Решение

Според погоре опишаната постапка имаме дека

$$G(y) = \iint\limits_{(\mathbb{R}^2 : x_1 + x_2 < y)} f(x_1, x_2)\,dx_1 dx_2 \,.$$

Областа $x_1 + x_2 < y$ е прикажано на следната слика.



Оттука се добива

$$G(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y-x_2} f(x_1, x_2)\,dx_1 dx_2 = \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{y-x_2} f(x_1, x_2)\,dx_1 \,,$$

и сега со диференцирање по y имаме

$$g(y) = \int_{-\infty}^{\infty} f(y - x_2, x_2)\,dx_2 \,.$$

Во случај кога X_1 и X_2 се независни, се добива

$$g(y) = \int_{-\infty}^{\infty} f_{X_1}(y - x_2) f_{X_2}(x_2)\,dx_2 \,. \blacksquare$$

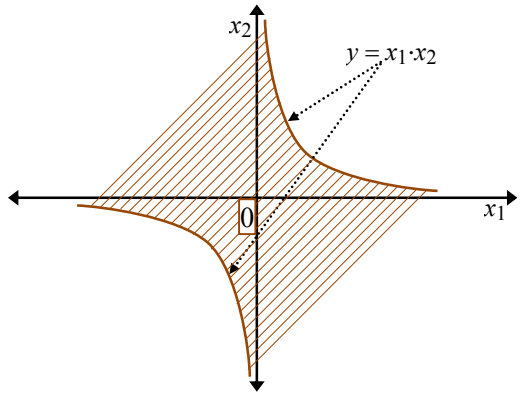
ПРИМЕР 6.6 Нека е $Y = X_1 \cdot X_2$. Најди ја густината на распределба на Y изразена преку заедничката густина $f_{X_1, X_2}(x_1, x_2)$.

Решение

Според погоре опишаната постапка имаме дека

$$G(y) = \iint\limits_{(\mathbb{R}^2 : x_1 \cdot x_2 < y)} f(x_1, x_2)\,dx_1 dx_2 \,.$$

Областа $x_1 \cdot x_2 < y$ се добива од хиперболата $x_1 \cdot x_2 = y$, како што е прикажано на следната слика



Оттука се добива

$$G(y) = \int_0^{\infty} \int_{-\infty}^{y/x_2} f(x_1, x_2) dx_1 dx_2 + \int_{-\infty}^{0} \int_{y/x_2}^{\infty} f(x_1, x_2) dx_1 dx_2 =$$

$$\int_0^{\infty} dx_2 \int_{-\infty}^{y/x_2} f(x_1, x_2) dx_1 + \int_{-\infty}^{0} dx_2 \int_{y/x_2}^{\infty} f(x_1, x_2) dx_1$$

и по диференцирањето по y имаме

$$g(y) = \int_0^{\infty} \frac{1}{x_2} f(\frac{y}{x_2}, x_2) dx_2 - \int_{-\infty}^{0} \frac{1}{x_2} f(\frac{y}{x_2}, x_2) dx_2 = \int_{-\infty}^{\infty} \frac{1}{|x_2|} f(\frac{y}{x_2}, x_2) dx_2 \ .$$

Во случај кога X_1 и X_2 се независни, се добива

$$g(y) = \int_{-\infty}^{\infty} \frac{1}{|x_2|} f_{X_1}(\frac{y}{x_2}) f_{X_2}(x_2) dx_2 \ . \ \blacksquare$$

Примерите што тука ги разгледуваме се секогаш функции со не повеќе од 2 променливи. Се разбира, иако со истиот пристап може да се решат проблеми со функции од повеќе променливи, тогаш е значително потешко да се определи соодветната област на интеграција, а интегралите стануваат гломазни и незгодни за решавање. Сепак, во основа нема некаква суштинска тешкотија истата постапка да се примени на функции од повеќе променливи.

6.4. Повеќе функции од повеќе случајни променливи*

Да го разгледаме и најопштиот случај, m функции од n случајни променливи ($m < n$):

$$Y_1 = h_1(X_1, X_2, ..., X_n),$$
$$Y_2 = h_2(X_1, X_2, ..., X_n),$$
$$...$$
$$Y_m = h_m(X_1, X_2, ..., X_n).$$

Тука проблемот е да се најде заедничката распределба на случајниот вектор $(Y_1, Y_2, ..., Y_m)$ преку распределбата на случајниот вектор $X_1, X_2, ..., X_n)$. Најпрво ќе го разгледаме случајот кога $m = n$, а потоа ќе објасниме како може да се третира случајот $m < n$. Се разбира, како и претходно, и тука се фокусираме на непрекинати случајни променливи бидејќи дискретниот случај е едноставен и се работи по аналогија на функциите од повеќе случајни променливи.

Нека $(Y_1, Y_2, ..., Y_n)$ и $(X_1, X_2, ..., X_n)$ се случајни вектори и нека $h_j(X_1, X_2, ..., X_n), j = 1, 2, ..., n$, се непрекинати по секој аргумент со непрекинати парцијални изводи и нека дефинираат еднозначно пресликување од $\mathbb{R}^n \to \mathbb{R}^n$. Тогаш следува дека инверзните функции $h_j^{-1}(\cdot)$ на $h_j(\cdot)$ дадени со

$$X_j = h_j^{-1}(Y_1, Y_2, ..., Y_n), j = 1, 2, ..., n$$

постојат и се единствени. Тие се исто така со непрекинати парцијални изводи.

За да ја определиме густината на распределба $g(y_1, y_2, ..., y_n)$ на случајниот вектор $(Y_1, Y_2, ..., Y_n)$ преку густината $f(x_1, x_2, ..., x_n)$ на $(X_1, X_2, ..., X_n)$, да согледаме дека кога некоја затворена област D_X од просторот на $(X_1, X_2, ..., X_n)$ се пресликува во затворена област D_Y од просторот на $(Y_1, Y_2, ..., Y_n)$, конзервацијата на веројатноста повлекува дека

$$\int\!\!\dots\!\!\int_{D_Y} f(y_1, y_2, \dots, y_n)dy_1dy_2 \dots dy_n = \int\!\!\dots\!\!\int_{D_X} f(x_1, x_2, \dots, x_n)dx_1dx_2 \dots dx_n .$$

Следејќи го стандардното правило за смена на променливите при повеќекратното интегрирање, горното равенство може да се напише во облик

$$\int\!\!\dots\!\!\int_{D_X} f(x_1, x_2, \dots, x_n)dx_1dx_2 \dots dx_n =$$
$$\int\!\!\dots\!\!\int_{D_Y} f(h_1^{-1}(y_1, y_2, \dots, y_n), \dots, h_n^{-1}(y_1, y_2, \dots, y_n)) \mid J \mid dy_1dy_2 \dots dy_n ,$$

каде што J е јакобијан на смената даден со

$$J = \begin{vmatrix} \frac{\partial h_1^{-1}}{\partial y_1} & \frac{\partial h_1^{-1}}{\partial y_2} & \cdots & \frac{\partial h_1^{-1}}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial h_n^{-1}}{\partial y_1} & \frac{\partial h_n^{-1}}{\partial y_2} & \cdots & \frac{\partial h_n^{-1}}{\partial y_n} \end{vmatrix}.$$

Да забележиме дека вертикалните линии околу J во интегралот означуваат апсолутна вредност, а во дефиницијата на J детерминанта. Од равенството на интегралите ја добиваме бараната формула

$$g(y_1, y_2, \ldots, y_n) = f(h_1^{-1}(y_1, y_2, \ldots, y_n), \ldots, h_n^{-1}(y_1, y_2, \ldots, y_n)) \, |J| .$$

Оваа формула очигледно е обопштен случај на формулата од теоремата 6.1 којашто се добива за $n = 1$. Тука, исто така даваме обопштување на теоремата 6.1 во случај кога пресликувањата $Y_j = h_j(X_1, X_2, \ldots, X_n)$, $j = 1, 2, \ldots, n$ се нееднозначни.

Теорема 6.4 Нека $h_j(X_1, X_2, \ldots, X_n)$, $j = 1, 2, \ldots, n$, се непрекинати функции со непрекинати парцијални изводи што го пресликуваат случајниот вектор $(X_1, X_2, \ldots, X_n)$ со густина на распределба $f(x_1, x_2, \ldots, x_n)$ во случаен вектор $(Y_1, Y_2, \ldots, Y_n)$ со густина на распределба $g(y_1, y_2, \ldots, y_n)$. Нека $X_j = h_j^{-1}(Y_1, Y_2, \ldots, Y_n)$, $j = 1, 2, \ldots, n$ има најмногу пребројво многу корени $(x_{i1}, x_{i2}, \ldots, x_{in}) = (h_{i1}^{-1}(y_1, y_2, \ldots, y_n), \ldots, h_{in}^{-1}(y_1, y_2, \ldots, y_n))$, $i = 1, 2, \ldots, r$, тогаш густината на распределба $g(y_1, y_2, \ldots, y_n)$ на $(Y_1, Y_2, \ldots, Y_n)$ е

$$g(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{r} f(h_{i1}^{-1}(y_1, y_2, \ldots, y_n), \ldots, h_{in}^{-1}(y_1, y_2, \ldots, y_n)) \, |J_i| ,$$

каде што

$$J_i = \begin{vmatrix} \frac{\partial h_{i1}^{-1}}{\partial y_1} & \frac{\partial h_{i1}^{-1}}{\partial y_2} & \cdots & \frac{\partial h_{i1}^{-1}}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial h_{in}^{-1}}{\partial y_1} & \frac{\partial h_{in}^{-1}}{\partial y_2} & \cdots & \frac{\partial h_{in}^{-1}}{\partial y_n} \end{vmatrix},$$

а $h_{i1}^{-1}(y_1, y_2, \ldots, y_n), \ldots, h_{in}^{-1}(y_1, y_2, \ldots, y_n)$ се компоненти на $h_i^{-1}(y_1, y_2, \ldots, y_n)$.

Доказ: Изоставен е поради многу технички детали, но е сосема аналоген на теоремата 6.1. ■

ПРИМЕР 6.7 Да се најде густината на распределба на случајниот вектор (Y_1, Y_2) даден со

$$(Y_1, Y_2) = (\sqrt{X_1^2 + X_2^2}, \frac{X_1}{X_2}),$$

каде што векторот (X_1, X_2) има нормална густина на распределба од облик

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}}.$$

Решение

Имаме дека

$$y_1 = h_1(x_1, x_2) = \sqrt{x_1^2 + x_2^2}, \qquad y_2 = h_2(x_1, x_2) = \frac{x_1}{x_2}.$$

Веднаш забележуваме дека y_1 е секогаш позитивно, па бараната густина на распределба $g(y_1, y_2)$ е 0 кога $y_1 < 0$.

Решението на овој систем равенки по x_1, x_2 ги дава инверзните функции (имаме две решенија). Првото е

$$x_{11} = h_{11}^{-1}(y_1, y_2) = \frac{y_1 y_2}{\sqrt{1+y_2^2}}, \quad x_{12} = h_{12}^{-1}(y_1, y_2) = \frac{y_1}{\sqrt{1+y_2^2}},$$

а второто е истото со минус знак

$$x_{21} = h_{21}^{-1}(y_1, y_2) = \frac{-y_1 y_2}{\sqrt{1+y_2^2}}, \quad x_{22} = h_{12}^{-1}(y_1, y_2) = \frac{-y_1}{\sqrt{1+y_2^2}}.$$

Според теоремата 6.4, имаме $r = 2$ и густината $g(y_1, y_2)$ на случајниот вектор (Y_1, Y_2) е дадена со

$$g(y_1, y_2) = f(h_{11}^{-1}(y_1, y_2), h_{12}^{-1}(y_1, y_2))|J_1| + f(h_{21}^{-1}(y_1, y_2), h_{22}^{-1}(y_1, y_2))|J_2|.$$

Имајќи предвид дека

$$J_1 = \begin{vmatrix} \frac{\partial h_{11}^{-1}}{\partial y_1} & \frac{\partial h_{11}^{-1}}{\partial y_2} \\ \frac{\partial h_{12}^{-1}}{\partial y_1} & \frac{\partial h_{12}^{-1}}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{21}^{-1}}{\partial y_1} & \frac{\partial h_{21}^{-1}}{\partial y_2} \\ \frac{\partial h_{22}^{-1}}{\partial y_1} & \frac{\partial h_{22}^{-1}}{\partial y_2} \end{vmatrix} = J_2 = -\frac{y_1}{1+y_2^2},$$ добиваме

$$g(y_1, y_2) = \frac{y_1}{1+y_2^2} \frac{2}{2\pi} e^{-\frac{y_1^2 y_2^2 + y_1^2}{2(1+y_2^2)}} = \frac{y_1}{(1+y_2^2)\pi} e^{-\frac{y_1^2}{2}},$$ за $y_1 \geq 0$, $-\infty < y_2 < \infty$ и 0 во спротивно. ■

Досега изложените резултати се однесуваат на функции од случајни вектори со иста димензија ($m = n$). Да го разгледаме сега случајот кога имаме m функции од n случајни променливи, т.е. функцијата h: $\mathbb{R}^n \to \mathbb{R}^m$ каде што $m < n$. Во таков случај, m-димензионалниот случаен вектор (Y_1, Y_2, ..., Y_m) го прошируваме со друг ($n-m$)-димензионален случаен вектор вектор ($Y_1^*, Y_2^*, \ldots, Y_{n-m}^*$). Овој вектор може да се конструира како едноставна функција h^*: $\mathbb{R}^n \to \mathbb{R}^{n-m}$ од (X_1, X_2, ..., X_n) што ги задоволува условите за непрекинатост и непрекинатост на парцијалните изводи. Така добиваме функција што пресликува n-димензионален случаен вектор во n-димензионален случаен вектор при што густината на случајниот вектор (Y_1, Y_2, ..., Y_m, $Y_1^*, Y_2^*, \ldots, Y_{n-m}^*$) се добива на претходно разгледаниот начин (теорема 6.4). Бараната густина на распределба на случајниот вектор (Y_1, Y_2, ..., Y_m) потоа се добива со интегрирање на густината на распределба на (Y_1, Y_2, ..., Y_m, $Y_1^*, Y_2^*, \ldots, Y_{n-m}^*$) по компонентите од случајниот вектор ($Y_1^*, Y_2^*, \ldots, Y_{n-m}^*$).

ПРИМЕР 6.8 Нека е $Y = X_1 \cdot X_2$. Најди ја густината на распределба на Y изразена преку заедничка густина $f_{X_1,X_2}(x_1, x_2)$, користејќи ја погоре наведената постапка.

Решение

Според погоре опишаната постапка, функцијата

$$y_1 = h(x_1, x_2) = x_1 \cdot x_2$$

ја прошируваме со некоја едноставна функција со "добри" особини за да добиеме 2-димензионален случаен вектор (Y, Y^*). На пример, ставаме

$$y_2 = h^*(x_1, x_2) = x_2 .$$

Сега, заедничката густина на распределба на (Y, Y^*) е

$$g(y_1, y_2) = f(h^{-1}(y_1, y_2), h^{*-1}(y_1, y_2))|J|,$$

при што

$$x_1 = h^{-1}(y_1, y_2) = \frac{y_1}{y_2},\ x_2 = h^{*-1}(y_1, y_2) = y_2,\ J = \begin{vmatrix} \frac{1}{y_2} & -\frac{y_1}{y_2^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{y_2}.$$

Значи $g(y_1, y_2) = f(\frac{y_1}{y_2}, y_2)\left|\frac{1}{y_2}\right|$, а бараната густина на распределба се добива со интегрирање по y_2,

$$g(y_1) = \int_{-\infty}^{\infty} \frac{1}{|y_2|} f(\frac{y_1}{x_2}, y_2) dy_2 ,$$

што е идентично со резултатот од примерот 6.6. ■

ПРИМЕР 6.9 Нека е $Y = X_1/X_2$. Најди ја густината на распределба на Y изразена преку заедничка густина $f_{X_1,X_2}(x_1, x_2)$.

Решение

Повторно функцијата

$y_1 = h(x_1, x_2) = \dfrac{x_1}{x_2}$ ја прошируваме со $y_2 = h^*(x_1, x_2) = x_2$.

Сега, заедничката густина на распределба е

$$g(y_1, y_2) = f(h^{-1}(y_1, y_2), h^{*-1}(y_1, y_2))|J| ,$$

при што

$$x_1 = h^{-1}(y_1, y_2) = y_1 y_2 , \; x_2 = h^{*-1}(y_1, y_2) = y_2 , \; J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2 .$$

Значи $g(y_1, y_2) = f(y_1 y_2, y_2)|y_2|$, а бараната густина на распределба се добива со интегрирање по y_2,

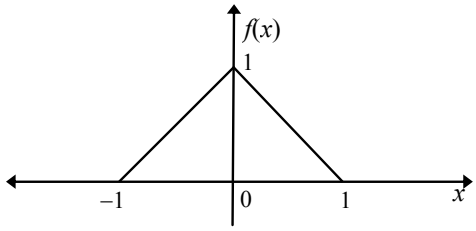
$$g(y_1) = \int_{-\infty}^{\infty} |y_2| f(y_1 y_2, y_2) dy_2 . \; ■$$

ЗАДАЧИ

1. Нека X е случајна променлива со геометриска распределба $f(x) = p(1-p)^{x-1}$, $x = 1, 2, \ldots$. Најди ја распределбата на $Y = X^2$.

2. X е случајна променлива со распределба $f(x) = x/8$, $0 \le x < 4$. Најди ја распределбата на $Y = 2X + 4$.

3. Случајната променлива X има распределба како на следната слика



Определи ја густината на распределба на случајната променлива $Y = 3X + 2$.

4. На некоја локација е определена е распределбата на брзината на ветерот V преку годината (во милји по час):

$$F(v) = \begin{cases} e^{-\left(\frac{v}{36.6}\right)^{-6.96}}, & \text{за } v > 0 \\ 0, & \text{во спротивно} \end{cases}$$

Силата на ветерот W е пропорционална на V^2, т.е. $W = kV^2$.
Определи ја густината на распределба на W.

5. Случајната променлива X е рамномерно распределена во интервалот $(0, \pi/2)$. Најди ја густината на распределба на случајната променлива $Y = \sin X$.

6. Случајната променлива X е со стандардна нормална распределба со густина $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Најди ја густината на распределба на $Y = X^2$.

7. Познато е дека радиусот на една сфера е рамномерно распределен во интервалот $[0.99r_0, 1.01r_0]$. Определи ги густините на распределба на
а) плоштината на сферата и б) волуменот на сферата.

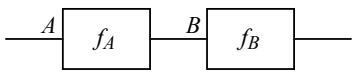
8. Да се најдат законите на распределба на случајните променливи $U = X + Y$ и $V = X \cdot Y$, ако X и Y се независни случајни променливи со распределби

X	10	12	16
$p(X = x_i)$	0.4	0.1	0.5

Y	1	2
$p(Y = y_j)$	0.2	0.8

9. Случајниот вектор (X, Y) е со рамномерна распределба во кругот со центар во (0.0) и радиус r. Да се најде густината на распределба на случајниот вектор (X, Y), маргиналните распределби на X и Y одделно. Дали се X и Y независни?

10. Разгледај го системот со две сериски компоненти A и B:



каде што случајните променливи X_A = "животен век на A" и X_B = "животен век на B" со густини на распределба

$$f_A(t)=\begin{cases} a_1e^{-a_1t}, & \text{за } t>0 \\ 0, & \text{во спротивно} \end{cases}, \qquad f_B(t)=\begin{cases} a_2e^{-a_2t}, & \text{за } t>0 \\ 0, & \text{во спротивно} \end{cases}.$$

Определи ја густината на распределба на $Y = \min(X_A, X_B)$. Обопшти го резултатот за n-сериски врзани компоненти.

11. Густината на распределба на случајниот вектор (X_1, X_2, X_3) е

$$f(x_1,x_2,x_3)=\begin{cases} \dfrac{6}{(1+x_1+x_2+x_3)^4}, & \text{за } (x_1,x_2,x_3)>(0,0,0) \\ 0, & \text{во спротивно} \end{cases}.$$

Најди ја густината на распределба на $Y = X_1 + X_2 + X_3$.

12. Случајните променливи X и Y се независни, и двете со стандардна нормална распределба $f(x)=\dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Најди ја густината на распределба на случајниот вектор $(X + Y, X - Y)$.

7

Гранични теореми и примени

Граничните теореми го сочинуваат математичкиот костур на кој се држи теоријата на веројатност. За практичните цели, тие се во главно корисни поради тоа што обезбедуваат приближни одговори на прашањата и проблемите врзани со определување на однесувањето на статистичките податоци при статистичките оценки и заклучоци. Двете главни гранични теореми, кои постојат во голем број варијанти и околу кои сè се врти, се законот на големите броеви и централната гранична теорема.

Законот на големите броеви е фундаментален резултат на теоријата на веројатност. На еден интуитивен начин ние многу пати досега го дискутиравме овој закон. Имено, интуитивниот начин да се гледа на веројатноста на некој настан е да се пресмета релативната честота на неговото појавување во доволно долго повторување на експериментот. Законот на големите броеви покажува дека математичкиот модел на веројатноста е конзистентен со релативните фреквенции на случување, т.е. е во согласност со стабилноста на експериментите кои теоријата на веројатност ги изучува.

Централната гранична теорема е веројатно најкорисен и најмаркантен резултат во теоријата на веројатност. Во нејзината наједноставна варијанта таа тврди дека *сума* на независни случајни променливи, под одредени доста општи услови, приближно има нормална распределба. Нејзината огромна важност е во тоа што секој примерок во статисти-

ката е низа на случајни променливи, а сумата на распределбите на доволно голем примерок, какви и да се распределбите на неговите елементи, приближно секогаш се потчинува на нормалната распределба. Оваа тенденција случајните променливи сумарно да формираат нормална распределба е аналогно на нешто сосема природно, како тенденцијата каменот да паѓа на земјата или водата да тече надолу. Со други зборови, нормалната распределба е како гравитациона сила што ги собира во себе сумите на независните распределби какви и да се тие. Зачудувачки и спротивен на интуиција изгледа фактот што сума на елементите на примерокот со каква било распределба (или различни распределби) се потчинува на една единствена распределба – нормалната. Уште повеќе, се покажува дека за повеќето распределби, приближувањето кон нормалната распределба се случува многу брзо со зголемувањето на примерокот.

7.1. Некои неравенства со моменти

Во практичните апликации, кога се работи со очекувања и дисперзии, т.е. со моменти, треба да се имаат предвид два аспекта. Првиот, којшто го дискутиравме, се однесува на пресметка на моментите на случајната променлива X со позната распределба. Вториот аспект, што е посилно врзан со практиката каде што расположивите информации се лимитирани, се однесува на проценка на однесувањето на случајната променлива X кога се знаат само некои нејзини моменти. Знаењето на вредностите на очекувањето и дисперзијата не овозможува одговор на прашањата од тип "колку е $p(X \geq \delta)$" и секако е недоволно да се процени распределбата. Сепак, како што сега ќе покажеме, возможно е да се постават некои веројатносни проценки знаејќи ги само очекувањето и дисперзијата.

Следните две неравенства, дадени во следните две теореми, се важни за проценка на отстапувањето на случајната променлива од нејзиното очекување.

Теорема 7.1 (Markov) За секое $\varepsilon > 0$, важи неравенството

$$p(|X| \geq \varepsilon) \leq \frac{E|X|}{\varepsilon}.$$

Доказ: Ако ставиме

$$Y = \begin{cases} 0, & \text{за } X < \varepsilon \\ \varepsilon, & \text{за } X \geq \varepsilon \end{cases}, \text{ очигледно важи } X \geq Y, \text{ па и } EX \geq EY.$$

Очекувањето на случајната променлива Y е

$EY = \varepsilon p(X \geq \varepsilon) \leq EX$, од што веднаш следува дека $p(|X| \geq \varepsilon) \leq \dfrac{E|X|}{\varepsilon}$. ■

Второто неравенство следи од претходното и е познато под име неравенство на Чебишев.

Теорема 7.2 За секое $\varepsilon > 0$, важи

$$p(|X - EX| \geq \varepsilon) \leq \frac{DX}{\varepsilon^2}.$$

Доказ: Ако ставиме $Y = (X - EX)^2$ имаме дека

$$p(|X - EX| \geq \varepsilon) = p((X - EX)^2 \geq \varepsilon^2) \overset{\text{теорема 7.1}}{\leq} \frac{E(X - EX)^2}{\varepsilon^2} = \frac{DX}{\varepsilon^2}.$$ ■

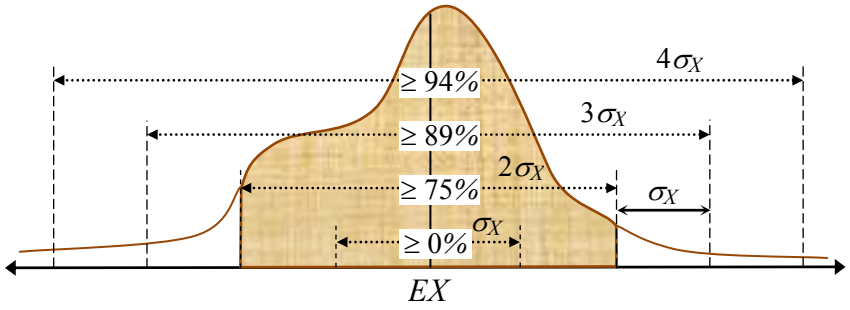
Неравенството на Чебишев покажува дека за мали дисперзии, вредностите на случајната променлива се концентрираат околу нејзиното очекување со веројатност блиска до 1, т.е.

$$p(|X - EX| < \varepsilon) \geq 1 - \frac{DX}{\varepsilon^2}.$$

Ако ставиме $\varepsilon = k\sqrt{DX} = k\sigma_X$, каде што σ_X е стандардната девијација, неравенството го добива добро познатиот облик

$$p(|X - EX| < k\sigma_X) \geq 1 - \frac{1}{k^2}, \quad k \geq 1,$$

што графички е прикажано на сл. 7.1.



Слика 7.1 *% на податоци што паѓаат во околините σ_X, $2\sigma_X$, $3\sigma_X$ и $4\sigma_X$ на очекувањето*

Значи веројатноста случајната променлива X да е во околина на $2\sigma_X$ околу EX е 0.75 (за $k = 2$, 1–1/4 = 0.7500), во околина на $3\sigma_X$ е 0.8999 (1 – 1/9 = 0.8999), во околина на $4\sigma_X$ е 0.9735 (1 – 1/16 = 0.9375) итн.

ПРИМЕР 7.1 Фрламе паричка n пати. Колкава е горната граница на веројатноста да добиме $3n/4$ петки?

Решение

Нека X = "број на петки во n фрлања на паричка" = $X_1 + X_2 + ... + X_n$. Тогаш $EX = n/2$ бидејќи во секое фрлање очекувањето е $EX_i = 1/2$, а X е сума на n независни случајни променливи со очекувања 1/2.

Според неравенството на Марков имаме

$$p(X \geq \frac{3n}{4}) \leq \frac{n/2}{3n/4} = \frac{2}{3}.$$

Според неравенството на Чебишев, имајќи предвид дека $DX_i = E(X_i)^2 - (EX_i)^2 = 1/2 - 1/4 = 1/4$ и $DX = n/4$, добиваме

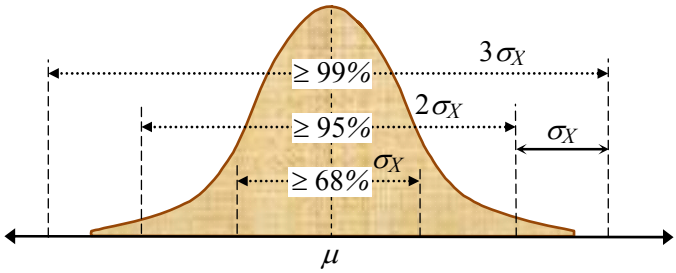
$$p(X \geq \frac{3n}{4}) = p(X - \frac{n}{2} \geq \frac{3n}{4} - \frac{n}{2}) = \frac{1}{2} p(|X - EX| \geq \frac{n}{4}) = \frac{1}{2} \frac{n/4}{(n/4)^2} = \frac{2}{n}$$

што е многу подобра проценка од онаа со неравенството на Марков. Се разбира тоа е сосема логично ако се има предвид дека во неравенството на Чебишев се користи додатна информација, добиена од моментот од 2 ред даден со дисперзијата. ■

Сепак, неравенството на Чебишев е доста слабо. На пример, нека X има стандардна нормална распределба $Z(0,1)$. Тогаш, со неравенството на Чебишев добиваме, на пример,

$$p(|X| \geq 3) \leq \frac{1}{9}$$

што е многу песимистички во однос на вистинската вредност што се добива од распределбата, $p(|X| \geq 3) \approx 2 \cdot 10^{-3}$. Се разбира, проценката на бројот на податоци што паѓаат во околината на очекувањето кога се знае распределбата на X може да се определи многу попрецизно. На пример, за нормалната распределба $Z(\mu, \sigma^2)$, веројатностите X да има вредност во околината на k стандардни девијации околу нејзиното очекување се: во $\mu \pm \sigma$ е 68.3%, во $\mu \pm 2\sigma$ е 95.5% и во $\mu \pm 3\sigma$ дури 99.7% (види слика 7.2).



Слика 7.2 *% на податоци што паѓаат во околините на очекувањето σ_X, $2\sigma_X$ и $3\sigma_X$, кај нормалната распределба*

За да се покаже дека неравенството на Чебишев е најдобро неравенство за $|X - EX| \geq \varepsilon$ кога се знаат само очекувањето и дисперзијата на X, доволно е да се најде случајна променлива за која важи равенство. Таква е случајната променлива Y дадена со законот на распределба

x_i	$-\varepsilon$	0	ε
$p(Y = x_i)$	$\sigma_Y^2/(2\varepsilon^2)$	$1 - \sigma_Y^2/\varepsilon^2$	$\sigma_Y^2/(2\varepsilon^2)$

за која лесно се проверува дека

$$EY = 0, \quad DY = EY^2 = (-\varepsilon)^2 \frac{\sigma_Y^2}{2\varepsilon^2} + \varepsilon^2 \frac{\sigma_Y^2}{2\varepsilon^2} = \sigma_Y^2 \quad \Rightarrow \quad p(|X| \geq \varepsilon) = \frac{\sigma_Y^2}{\varepsilon^2}.$$

Вредноста на неравенството на Чебишев е во неговата општост. Тоа е применливо на сите случајни променливи без разлика на типот или обликот на нивната распределба. Се разбира, кога ја имаме распределбата, неравенството на Чебишев е практично безполезно.

Во практиката понекогаш од корист е едностраното неравенство на Чебишев што го даваме во следната теорема.

Теорема 7.3 (Cantelli) За секое $\varepsilon > 0$, важи неравенството

$$p(X - EX \geq \varepsilon) \leq \frac{DX}{DX + \varepsilon^2}.$$

Доказ: Нека е $z > 0$, тогаш

$$p(X - EX \geq \varepsilon) = p(X - EX + z \geq \varepsilon + z) = p((X - EX + z)^2 \geq (\varepsilon + z)^2) \leq$$

$$\leq \frac{E(X - EX + z)^2}{(\varepsilon + z)^2} = \frac{DX + z^2}{(\varepsilon + z)^2}.$$

Сега ако десната страна ја гледаме како функција од z, нејзинот минимум се добива за $z = DX/\varepsilon$. Така добиваме

$$p(X - EX \geq \varepsilon) \leq \frac{DX + \frac{(DX)^2}{\varepsilon^2}}{(\varepsilon + \frac{DX}{\varepsilon})^2} = \frac{DX}{DX + \varepsilon^2} . \blacksquare$$

Ако ставиме $\varepsilon = kDX = k\sigma$, се добива добро познатиот облик на неравенството

$$p(X - EX \geq k\sigma) \leq \frac{1}{1 + k^2} .$$

Ако направиме замена $Y = -X$, тогаш $EY = -EX$, $DY = DX$ и $EY - Y = X - EX$, тогаш веднаш следува дека и

$$p(EY - Y \geq k\sigma) \leq \frac{1}{1 + k^2} ,$$

т.е. важи и обратното едностране неравенство. Забележи дека секогаш важи $p(X - EX \geq k\sigma) \leq p(|X - EX| \geq k\sigma)$ бидејќи $\{X - EX \geq k\sigma\} \subseteq \{|X - EX| \geq k\sigma\}$, па оттаму и за соодветните горни проценки важи неравенството $\frac{1}{1 + k^2} \leq \frac{1}{k^2}$.

ПРИМЕР 7.2 Да претпоставиме дека коефициентот на интелигенција IQ на 2 милиони луѓе во просек е 100 со стандардна девијација од 15. Да се најде горната граница на бројот на луѓе со IQ > 130?

Решение

Ако X е случајната променлива IQ на човек, според едностраното неравенство имаме

$$p(X - 100 \geq 30 = 2 \cdot 15) \leq \frac{1}{1 + 2^2} = 0.20 ,$$

што значи дека горната граница е 400 илјади луѓе.

Да забележиме дека двостраното неравенство дава

$$p(|X - 100| \geq 30 = 2 \cdot 15) \leq \frac{1}{2^2} = 0.25 ,$$

што е очигледно поголем број бидејќи во проценката ги вклучува и оние со IQ ≤ 70. ■

ПРИМЕР 7.3 За проценка на климата за мал бизнис направено е испитување на 12 трговци со почетен капитал од 50000$. Се покажало дека по една година просечниот капитал на трговците пораснал на 80000$ со стандардна девијација од 30000$. Колку трговци го зголемиле капиталот за 20%, а колку го удвоиле?

Решение

Ако X е новиот капитал на трговците, имаме

$$p(X - 80000 \geq 20000 = \frac{2}{3}30000) \geq \frac{1}{1+(2/3)^2} = 0.6923$$, па се очекува

најмалку [12·0.6923] = 8 трговци да го зголемиле капиталот за 20%.

Бидејќи 100% од 50000 е 50000, а тоа е $\frac{5}{3}\cdot 30000$ што понатаму дава $\frac{1}{1+(5/3)^2} = 0.375$, што значи дека најмалку 4 трговци го зголемиле капиталот за 100%. ■

Со едностраното неравенство лесно се покажува дека медијаната на една случајна променлива X (точка $m(X)$ за која $p(X < m(X)) = 1/2$) е на растојание не поголемо од σ од нејзинот просек, $|m(X) - EX| \leq \sigma$. Имено, од едностраните неравенство за $k = 1$ имаме

$$p(X - EX \geq \sigma) \leq \frac{1}{2}$$, т.е. $p(X \geq EX + \sigma) \leq \frac{1}{2}$ и

$$p(EX - X \geq \sigma) \leq \frac{1}{2}$$, т.е. $p(X \leq EX + \sigma) \leq \frac{1}{2}$

што веднаш повлекува дека $m(X) \in [EX - \sigma, EX + \sigma]$.

Многу често, неравенството на Чебишев се користи за проценка на потребен број мерења, испитувања или примероци од некоја популација. На пример, нека се прават n независни мерења на некоја непозната променлива a. Грешките при мерењата $\delta_1, \delta_2, ..., \delta_n$, можеме да ги сметаме за случајни променливи, при што земаме дека $E\delta_i = 0$, $i = 1, 2, ..., n$. Овој услов може да се смета како отсуство на систематска грешка. Нека $D\delta_i = b^2$, $i = 1, 2, ..., n$. Ако за проценка на a се земе просекот на мерењата, грешката ќе биде

$$Y_n = \frac{\delta_1 + \delta_2 + \ldots + \delta_n}{n} \quad \text{и} \quad DY_n = \frac{1}{n^2}(D\delta_1 + D\delta_2 + \ldots + D\delta_n) = \frac{b^2}{n},$$

а се разбира, $EY_n = 0$. Ние би сакале со висока веројатност грешката Y_n да биде помала од некоја однапред зададена вредност Δ. На пример,

$p(|Y_n| < \Delta) \geq 0.99$, т.е. $p(|Y_n| \geq \Delta) < 0.01$.

Од неравенството на Чебишев имаме дека

$$p(|Y_n| \geq \Delta) \leq \frac{DY_n}{\Delta^2} = \frac{b^2}{n\Delta^2}$$

што значи дека се наоѓаме во дозволената зона на грешка кога

$\frac{b^2}{n\Delta^2} \leq 0.01$, од што веднаш се добива $n \geq 100\frac{b^2}{\Delta^2}$.

На овој начин сме го добиле потребниот број мерења за добивање на зададената точност. Тука останува еден креативен дел од работата, а тоа е да се процени b^2. Оваа проценка може да се подобри ако се земе предвид дека Y_n е сума на независни случајни променливи со конечни дисперзии и тогаш според централната гранична теорема има распределба блиска до нормална. Ако пак за случајната променлива Y_n не се знае ништо друго од очекувањето и дисперзијата, подобрување на проценката за n, не може да се направи.

ПРИМЕР 7.4 Нека p биде непознат процент гласачи што се определени да гласаат за партијата на зелените. За проценка на p треба да се анкетираат n гласачи. За i-тиот гласач, нека случајната променлива

$$\delta_i = \begin{cases} 1 & \text{ако гласа за зелените} \\ 0 & \text{во спротивно} \end{cases}$$

и нека просекот $Y_n = (\delta_1 + \delta_2 + \ldots + \delta_n)/n$ е процентот на гласачи определени да гласаат за зелените. Колкаво треба да се земе n за да се добие проценка на p со грешка не поголема од 1% и сигурност од 95%.

Решение

Треба да се определи такво n, да важи

$p(|Y_n - p| \leq 0.01) \geq 0.95$.

Од неравенството на Чебишев имаме дека

$$p(|Y_n - p| \le 0.01) \ge 1 - \frac{\sigma_\delta}{n \cdot 0.01^2}.$$

Сега имајќи предвид дека σ_δ е случајна променлива од 0/1 тип, таа има приближно Бернулиева распределба па $\sigma_\delta = p(1-p) \le 0.25$. Оттука лесно се добива

$$1 - \frac{\sigma_\delta}{n \cdot 0.01^2} \ge 1 - \frac{0.25}{n \cdot 0.01^2} \ge 0.95 \text{ што дава } n \ge 50000. \blacksquare$$

7.2. Закон на големите броеви

Многу експерименти и набљудувања во врска со природните феномени обично се повторуваат повеќе пати во наизглед идентични услови и вообичаено резултираат во различни исходи. Неконтролираните влијанија се причина за ваквите "случајни" варијации. За надминување на оваа ситуација експериментот или набљудувањето се повторува повеќе пати и резултатите на некој начин се упросечуваат. Во ова поглавје ќе видиме зошто ваков модел на повторување и упросечување работи така добро. Секоја реализација на мерење или набљудување резултира во случајна променлива и така се добива низа независни случајни променливи, се разбира со непозната распределба. Од веројатносен аспект, од таква низа случајни променливи во основа може да се извлечат особините на распределбата што е последица од законот на големите броеви.

Луѓето вклучени во експериментална работа со векови знаеле дека поточни резултати се добиваат кога мерењата или екпериментите се повторуваат, а добиените резултати се упросечуваат. На пример, при мерењето на брзината на светлината во 1879 година, Мајкелсон (Michelson) ги повторувал мерењата за да добие поточна вредност. Денеска, некој би рекол дека Мајкелсон ја искористил моќта на упросечувањето да ја редуцира варијабилноста на мерењата. Упросечувањето има тенденција да ги измазни сите поголеми флуктуации во вредностите и тоа, колку повеќе вредности толку подобра елиминирање на неконтролираните влијанија.

Следната теорема што понекогаш се нарекува "слаб" закон на големите броеви има големо теоретско значење.

Теорема 7.4 Ако $X_1, X_2, ..., X_n, ...$ е низа независни случајни променливи и постои константа $c > 0$ таква што $DX_i \le c$, за секој i, тогаш за секое $\varepsilon > 0$ важи

$$\lim_{n\to\infty} p\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{EX_1+EX_2+\ldots+M\xi_n}{n}\right|<\varepsilon\right)=1$$

Доказ: Го применуваме неравенството на Чебишев на Y_n/n, каде што $Y_n = X_1 + X_2 + \ldots + X_n$. Тоа значи дека за $x > 0$ имаме

$$1\ge p\left(\left|\frac{Y_n}{n}-\frac{EX_n}{n}\right|<x\right)\ge 1-\frac{DY_n}{x^2n^2}\ge 1-\frac{cn}{x^2n^2}\ge 1-\frac{c}{x^2n}\underset{n\to\infty}{\to}1\,.\ \blacksquare$$

Како последица на оваа теорема ги имаме следните поедноставни случаи.

Најпрво, ако имаме $EX_i = \mu$, $DX_i = \sigma^2 \le \infty$, тогаш за секое $\varepsilon > 0$ повторно важи

$$\lim_{n\to\infty} p\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\mu\right|<\varepsilon\right)=1\,.$$

Очигледно ова е специјален случај на теоремата што има голема практична вредност. Имено, земање примерок од некоја популација формира конечна низа од случајни променливи со исто очекување. Законот кажува дека просекот на вредностите на случајните променливи по веројатност се приближува (конвергира) кон нивното очекување.

Како втора последица ќе ја разгледаме Бернулиевата шема на n независни испитувања. Нека s_n е број на поволни случаи во n-те испитувања при што веројатноста на поволен случај е p. Тогаш за секое $\varepsilon > 0$ важи

$$\lim_{n\to\infty} p\left(\left|\frac{s_n}{n}-p\right|<\varepsilon\right)=1\,.$$

Ако се стави $s_n = X_1 + X_2 + \ldots + X_n$, каде што $X_i = 1$ при поволен случај и 0 во спротивно и се има предвид дека $EX_i = p$ и $DX_i = p(1-p)$, веднаш се гледа дека ова е специјален случај на теоремата 7.4. Оваа граница покажува дека релативните честоти на појавување на настаните конвергираат кон веројатностите на настаните, т.е. при аксиоматското градење на теоријата на веројатност, вградена е и стабилност на експериментите. Тоа беше и еден од условите даден експеримент да биде погоден за математичко изучување.

ПРИМЕР 7.5 Да се пресмета коефициентот на варијација за сума од n независни случајни променливи $Y = X_1 + X_2 + \ldots + X_n$ со коефициенти на вари-

јација ν и да се дискутира добиениот резултат од аспект на тоа како n влијае на варијацијата.

Решение

Ако коефициентот на варијација за секоја X_i е ν, тогаш за сумата лесно се добива дека е

$$\nu_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\sqrt{n}\sigma_X}{n\mu_X} = \frac{\nu}{\sqrt{n}}.$$

Добиениот резултат е основа на законот на $\sqrt{n}$ на Шредингер (Schrödinger) кој тврди дека законите на физиката се применливи со веројатност на грешка од ред $1/\sqrt{n}$, каде што n е бројот на молекули што кооперираат во физичкиот процес. Значи ако секој молекул има случајна варијација во однесувањето проценета со ν, тогаш физичкиот процес базиран на заедничка активност на n молекули поседува случајна варијација од $\nu/\sqrt{n}$. Неговата варијација очигледно се намалува како што n се зголемува. Ако се има предвид дека во реалните физички процеси со кои ние секојдневно се судираме n е многу големо, тогаш може да се заклучи дека законите на физиката се егзактни во тие ситуации и покрај локалниот "хаос" што се јавува кога се разгледуваат изолирано мал број молекули. ■

Иако законот на големите броеви од теоремата 7.4 може да се смета за важно и јако тврдење, тој е познат под името слаб закон на големите броеви. Имено, постои посилен закон на големите броеви даден со

$$p\left(\lim_{n\to\infty}\frac{X_1 + X_2 + \ldots + X_n}{n} = \mu\right) = 1.$$

што може да се парафразира со "кога $n \to \infty$, просекот на низата случајни променливи X_1, X_2, ..., X_n, ... конвергира кон $\mu = EX_i$ со веројатност 1", додека слабиот закон се парафразира со "кога $n \to \infty$, веројатноста за просекот на низата случајни променливи X_1, X_2, ..., X_n, ... да конвергира кон $\mu = EX_i$ е 1". Не е лесно да се види, но јакиот закон на големите броеви е навистина посилен од слабиот, бидејќи веројатноста на конвергенција (кај јакиот закон) ја повлекува конвергенцијата по веројатност (кај слабиот закон). Условите кај слабиот закон на големите броеви може да се релаксираат до степен тој сè уште да следи од нив, додека јакиот закон веќе да не важи. Јакиот закон бара појаки услови за појако тврдење. Доволен услов да важи јакиот закон на големите броеви Колмогоров (Andrey Nikolaevich Kolmogorov, 1903-1987) е случајните променливи X_i да се независни и

$$\sum_{i=1}^{\infty} \frac{DX_i}{i^2} < \infty .$$

Тука не треба да нè залажува условот од теоремата 7.4, $DX_i < c$, од кој следи горниот услов, бидејќи условот за конечни дисперзии може комплетно да се испушти од теоремата 7.4, и таа пак да важи. Се разбира, во тој случај доказот е многу покомплициран и користи карактеристични функции (види следно поглавје).

Слабиот закон во основа тврди дека за дадено големо n просекот на X_i веројатно е близок до μ. Тој остава можност

$$\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \mu \right| > \varepsilon$$

да се случи бесконечно пати, но не често. Јакиот закон кажува дека тоа скоро сигурно нема да се случи. Тој повлекува дека со веројатност блиска до 1, имаме дека важи

$$\frac{X_1 + X_2 + \ldots + X_n}{n} = \mu$$

за сите доволно големи n. Така, со земање доволно големо n, ние можеме да ја направиме веројатноста на која било дадена варијација да биде колку што сакаме мала.

Да ги продискутираме конвергенциите врзани со веројатности на едноставниот пример на фрлање паричка. Ако фрламе фер паричка n пати и k_{Π} пати падне петка П, тогаш со зголемување на бројот на фрлања n, веројатноста на настанот

$$\left\{ \left| \frac{k_{\Pi}}{n} - \frac{1}{2} \right| < \varepsilon \right\} \text{ оди кон 1.}$$

За да ја објасниме оваа конвергенција најпрво ќе го појасниме стандардното неразбирање на законот на големите броеви познато како *коцкарска илузија*. Ако на пример ја фрламе паричката 10 пати и добиеме низа {Г, П, Г, П, П, П, П, П, П, П}, во следното фрлање некако изгледа поверојатно да добиеме глава Г поради претходните 7 последователни петки. Оваа идеја потекнува од очекувањето по многу фрлања бројот на петки и глави да биде приближно еднаков. Од друга страна "знаеме" дека паричката нема меморија (фрлањата се независни) па веројатностите за П или Г во секое следно фрлање остануваат еднакви. Законот на големите броеви не кажува ништо за разликата во бројот на петки и глави $|k_{\Pi} - k_{\Gamma}|$, а услов од облик $|k_{\Pi} - k_{\Gamma}| < \varepsilon$ би ја нарушил независноста на фрлањата. Тој само кажува што е со разликата $|k_{\Pi}/n - 1/2|$

и следователно со $|k_\Gamma/n - 1/2|$. Се разбира, некој што не ја прифаќа веројатноста како резултат на релативни честоти може да прифати дека паричката има меморија, т.е. дека реално не постои независност на настани, па за него коцкарската илузија не е илузија.

Нека множеството исходи S при повеќекратно фрлање паричка бидат бесконечните низи од облик $s = \{s_1, s_2, ..., s_n, ...\}$. Ја дефинираме случајната променлива

$$X_k = \begin{cases} 1 \text{ ако } s_k = \Pi \\ 0 \text{ ако } s_k = \Gamma \end{cases}.$$

Ние би сакале да важи, но се разбира тоа не е можно

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{2},$$

бидејќи најпрво лимесот може да непостои, а како второ, постојат низи како што се $s_\Pi = \{\Pi, \Pi, ..., \Pi, ...\}$ или $s_\Gamma = \{\Gamma, \Gamma, ..., \Gamma, ...\}$ чијашто граница сигурно не е 1/2. Од тие причини останува да се разгледуваат конвергенции врзани со веројатноста.

Првиот тип е конвергенција на веројатности, како што е направено во слабиот закон на големите броеви

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{2}\right| < \varepsilon\right) = 1.$$

Ова во суштина е конвергеција на броевите - веројатности на настаните "просекот на X_k се разликува од 1/2 за помалку од ε". Тоа значи дека за доволно големо n, веројатностите на настаните

$$p_n = p\left(s \in S : \left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{2}\right| < \varepsilon\right), \quad \lim_{n\to\infty} p_n = 1$$

се блиски до 1, но може многу од низите да имаат просек што се разликува од 1/2 за повеќе од ε.

Вториот тип е веројатност на конвергенцијата, како што е во јакиот закон на големите броеви

$$p\left(s \in S : \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{2}\right) = 1.$$

Ова во суштина е конвергенција на множества - настани "просекот на X_k да е 1/2". Тоа значи дека просекот е 1/2 скоро за сите низи s. Со дру-

ги зборови, за доволно големо n, веројатноста на подмножеството $S_0 \subseteq S$ дадено со

$$S_0 = \left\{ s : \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k \neq \frac{1}{2} \right\}$$

(S_0 меѓу другите ги содржи низите s_Π и s_Γ) е занемарлива, т.е. како што расте n, множеството S_0 ќе има сè помалку и помалку низи (од тип s_Π и s_Γ), и ќе се сведе на множество со мера (веројатност) 0 (невозможен настан). Интуитивно се чувствува дека веројатноста на конвергенција е помоќна од конвергенцијата на веројатности. И навистина, докажано е дека веројатноста на конвергенција ја повлекува конвергенцијата на веројатноста. Веројатноста на конвергенцијата се нарекува и скоро сигурна конвергенција. Така, понекогаш наместо

$$p\left(\lim_{n\to\infty} Y_n = a \right) = 1, \text{ пишуваме } Y_n \overset{c.c.}{\to} a \quad (c.c. = \text{скоро сигурно}).$$

7.3. Карактеристични функции*

За анализа на случајните променливи и пресметка на некои нивни особини, како што се моментите или обликот на распределбата, како и за докажување важни теореми, од голема полза се карактеристичните функции.

Комплексна случајна променлива е функцијата $X(\omega) + iY(\omega)$, $\omega \in \Omega$, каде што (X, Y) е случаен вектор. По дефиниција се става

$$E(X + iY) = E(X) + iE(Y).$$

За вака дефинираното очекување важат сите особини од теоремата 5.1 од поглавјето 5.1.

Дефиниција 7.1 Карактеристична функција на случајната променлива X е функцијата $\varphi(t) = Ee^{itX}$. Тука t е реална променлива, а i е имагинарната единица, $i^2 = -1$.

Значи карактеристичната функција на X е очекувањето на комплексната функција $e^{itX} = \cos(tX) + i\sin(tX)$. Според дефиницијата на очекувањето имаме

$$\varphi(t) = Ee^{itX} = \sum_{k=1}^{n} e^{itx_k} p(X = x_k), \text{ за дискретна } X \text{ со вредности } x_1, \ldots, x_n$$

$$\varphi(t) = Ee^{itX} = \int_{-\infty}^{\infty} e^{itx} f(x)dx, \text{ за непрекината } X \text{ со густина } f(x).$$

Следните особини на карактеристичните функции се речиси тривијални:

1) $\varphi(0) = 1$ и $|\varphi(t)| \le 1$;

2) $\varphi(-t) = \varphi^*(t)$, каде што * е оператор на конјугираност;

3) ако $Y = aX + b$, тогаш, $\varphi_Y(t) = e^{itb}\varphi_X(at)$;

4) врската (пресликувањето) меѓу множеството карактеристични функции и множеството функции на распределба е взаемно-еднозначна (биекција).

На пример, за првата особина, од $|e^{itx}| \le 1$, следува дека

$$|\varphi(t)| = \left|\int_{-\infty}^{\infty} e^{itx} f(x)dx\right| \le \int_{-\infty}^{\infty} f(x)dx = 1.$$

За третата особина веднаш се добива

$$\varphi_Y(t) = Ee^{it(aX+b)} = e^{itb}Ee^{itaX} = e^{itb}\varphi_X(at).$$

ПРИМЕР 7.6 Да се најде карактеристичната функција на нормалната распределба $Z(0,1)$.

Решение

Треба да се реши интегралот

$$\varphi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\cos x dx + \frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\sin x dx,$$

при што вториот дел е 0 како интеграл на непарна функција по симетричен интервал. Првиот интеграл може да реши со диференцирање

$$\varphi'(t) = -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2}}\sin x dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \sin x de^{-\frac{x^2}{2}},$$

па по примена на парцијална интеграција добиваме

$$\varphi'(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\sin tx\Big/_{-\infty}^{\infty} - \frac{t}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\sin x dx = -t\varphi(t).$$

Решение на диференцијалната равенка е

$$\frac{\varphi'(t)}{\varphi(t)} = -t \;\rightarrow\; \varphi(t) = Ce^{-\frac{t^2}{2}}$$ што со почетниот услов $\varphi(0) = 1$ дава $C = 1$. ■

ПРИМЕР 7.7 Да се најде карактеристичната функција на нормалната распределба $Z(\mu, \sigma^2)$.

Решение

Ќе ја искористиме особината 3) на карактеристичните функции (ако $Y = aX + b$, тогаш, $\varphi_Y(t) = e^{itb}\varphi_X(at)$), бидејќи веќе ја имаме карактеристичната функција на $Z(0,1)$. Нека Y има $Z(\mu, \sigma^2)$, а X нека има $Z(0,1)$ распределба. Тогаш е

$$X = \frac{Y-\mu}{\sigma} \quad \Rightarrow \quad Y = \sigma X + \mu,$$

и сега според особината 3)

$$\varphi_Y(t) = \varphi_{\sigma X+\mu}(t) = e^{it\mu}\varphi_X(\sigma t) = e^{it\mu-\frac{\sigma^2 t^2}{2}}. \blacksquare$$

За да ја илустрираме моќта на карактеристичните функции, ќе покажеме како со нив може да се добијат моментите на случајната променлива X. За таа цел, ја развиваме карактеристичната функција $\varphi(t)$ во ред

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \varphi''(0)\frac{t^2}{2} + \ \dots\ + \varphi^{(n)}(t)\frac{t^n}{n!} + \ \dots$$

каде што примовите се изводи. Коефициентите во овој ред се

$$\varphi(0) = 1 = E(X^0),$$
$$\varphi'(0) = \left.\frac{d\varphi}{dt}\right|_{t=0} = i\int_{-\infty}^{\infty} xf(x)dx = iE(X^1),$$
$$\vdots$$
$$\varphi^{(n)}(0) = \left.\frac{d^n\varphi}{dt^n}\right|_{t=0} = i^n\int_{-\infty}^{\infty} x^n f(x)dx = i^nE(X^n),$$

каде, како што се гледа, од десната страна се добиваат моментите. Сега редот за $\varphi(t)$ го добива следниот облик

$$\varphi(t) = \sum_{k=0}^{\infty}\frac{(it)^k E(X^k)}{k!}.$$

Истиот ред се добива и за дискретна X.

Значи моментите од кој било ред се содржани во развојот на карактеристичната функција $\varphi(t)$ и може да се добијат со нејзино диференцирање. Така имаме

$$E(X^k) = i^{-k}\varphi^{(k)}(0),\ \ k = 1, 2, \ldots .$$

ПРИМЕР 7.8 Да се најде очекувањето и дисперзијата на случајна променлива X со експоненцијална распределба

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{за } x \geq 0 \\ 0, & \text{во спротивно} \end{cases},$$

користејќи карактеристична функција.

Решение

Карактеристичната функција ја добиваме од интегралот

$$\varphi(t) = \int_0^\infty e^{itx}\lambda e^{-\lambda x}dx = \lambda\int_0^\infty \lambda e^{-(\lambda - it)x}dx = \frac{\lambda}{\lambda - it}.$$

Моментите од прв и втор ред се $EX = \frac{1}{i}\frac{d}{dt}\left(\frac{\lambda}{\lambda - it}\right)\Big|_{t=0} = \frac{1}{i}\left(\frac{i\lambda}{(\lambda - it)^2}\right)\Big|_{t=0} = \frac{1}{\lambda}$,

$$E(X^2) = -\frac{d^2}{dt^2}\left(\frac{\lambda}{\lambda - it}\right)\Big|_{t=0} = -\frac{d}{dt}\left(\frac{i\lambda}{(\lambda - it)^2}\right)\Big|_{t=0} = -\left(\frac{-2\lambda}{(\lambda - it)^3}\right)\Big|_{t=0} = \frac{2}{\lambda^2},$$

па за дисперзијата се добива $\frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. ■

Друга важна примена на карактеристичните функции е во можноста да се определи функцијата на распределба на случајната променлива преку нејзината карактеристична функција. Нека X е непрекината случајна променлива. Да забележиме дека дефиницијата на карактеристичната функција $\varphi_X(t)$ во основа е инверзна фуриеова трансформација на $f_X(x)$. Соодветната фуриеова трансформација е

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt .$$

Оваа "инверзна" формула покажува како се добива густината на распределба од карактеристичната функција. Уште повеќе, теоријата на фуриеови трансформации кажува дека $f_X(x)$ е еднозначно определена, т.е. нема две различни густини на распределба што може да се добијат од иста карактеристична функција. Ова е веќе содржано во особината 4) на карактеристичните функции.

Во многу физички проблеми често е позгодно да се определува густината на распределбата со најпрво определување на карактеристич-

ната функција, а потоа да се направи нејзина фуриеова трансформација. Исто така, карактеристичните функции се погодни за определување распределби на случајни променливи што се функции од случајни променливи, особено сумите на независните случајни променливи.

Иако "инверзна" формула следува директно од теоријата на фуриеви трансформации, ние тука ќе ја изведеме користејќи веројатносни концепти, како што тоа е направено во [Soong 2004].

Во секоја таблица интегрални може да се најде дека

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin at}{t}dt=\begin{cases}-1, & \text{за } a<0\\ 0, & \text{за } a=0\\ 1, & \text{за } a>0\end{cases}$$

што дава и дека

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin at-i(1-\cos at)}{t}dt=\begin{cases}-1, & \text{за } a<0\\ 0, & \text{за } a=0\\ 1, & \text{за } a>0\end{cases}$$

бидејќи $i(1-\cos at)/t$ е непарна функција по t на симетричен интервал (интегралот е 0). Сега го вадиме i пред интегралот и целото равенство го множиме со $-1/2$ и додаваме $1/2$ од двете страни, а за a ставаме $X-x$ што дава

$$\frac{1}{2}-\frac{i}{2\pi}\int_{-\infty}^{\infty}\frac{1-e^{i(X-x)t}}{t}dt=\begin{cases}1, & \text{за } X<x\\ 1/2, & \text{за } X=x\\ 0, & \text{за } X>x\end{cases}.$$

За фиксно x, горното равенство е функција од случајната променлива X и може да се смета за дефиниција на нова случајна променлива Y. Случајната променлива Y може да се смета за дискретна и зема вредности 1, 1/2 и 0 со веројатности $p(X<x)$, $p(X=x)$ и $p(X>x)$ соодветно. Очекувањето на случајната променлива Y е

$$EY=(1)p(X<x)+(1/2)p(X=x)+(0)p(X>x).$$

Од друга страна, X е непрекината и x е само една точка на непрекинатата распределба на X што дава $p(X=x)=0$. Оттука, само првиот член во EY е ненулти што понатаму дава

$$EY=p(X<x)=F_X(x)=$$
$$\frac{1}{2}-\frac{i}{2\pi}\int_{-\infty}^{\infty}\frac{1-Ee^{i(X-x)t}}{t}dt=\frac{1}{2}-\frac{i}{2\pi}\int_{-\infty}^{\infty}\frac{1-e^{-ixt}\varphi_X(t)}{t}dt.$$

Сега ја имаме функцијата на распределба $F(x)$ на X. Со нејзино диференцирање ја добиваме соодветната густина на распределба

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt\,,$$ а тоа е точно "инверзна"-та формула

што следува од фуриеовите трансформации.

Во случај кога X е дискретна, "инверзната" формула е

$$p(X=x) = \lim_{u\to\infty}\frac{1}{2u}\int_{-u}^{u} e^{-itx}\varphi_X(t)dt\,,$$

и таа може да се изведе по истата линија на расудување како во непрекинатиот случај. Имено, тука тргнуваме од познатиот интеграл

$$\frac{1}{2u}\int_{-u}^{u} e^{iat}dt = \begin{cases} \dfrac{\sin au}{au}, & \text{за } a\neq 0 \\ 1, & \text{за } a=0 \end{cases},$$

пуштаме $u\to\infty$ и повторно го заменуваме a со $X-x$. Така добиваме нова случајна променлива Y, дефинирана со

$$Y = \lim_{u\to\infty}\frac{1}{2u}\int_{-u}^{u} e^{i(X-x)t}dt = \begin{cases} 0, & \text{за } X\neq x \\ 1, & \text{за } X=x \end{cases},$$

чиешто очекување е

$$EY = (1)p(X=x) + (0)p(X\neq x) = p(X=x).$$

Оттука веднаш следува

$$p(X=x) = \lim_{u\to\infty}\frac{1}{2u}\int_{-u}^{u} Ee^{i(X-x)t}dt = \lim_{u\to\infty}\frac{1}{2u}\int_{-u}^{u} e^{-itx}\varphi_X(t)dt\,.$$

Изведените резултати за "инверзните" функции ќе ги формулираме во следната теорема.

Теорема 7.5 Ако $f(x)$, $F(x)$ и $\varphi(t)$ се густина на распределба, функција на распределба и карактеристична функција на една случајна променлива X, тогаш важи

$$p(X=x) = \lim_{u\to\infty}\frac{1}{2u}\int_{-u}^{u} e^{-itx}\varphi(t)dt\,,$$ за дискретна X,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dt\,,$$ за непрекината X, и

$$F(x_2)-F(x_1)=\frac{1}{2\pi}\lim_{u\to\infty}\int_{-u}^{u}\frac{e^{-itx_1}-e^{-itx_2}}{it}\varphi(t)dt$$

за точки x_1, x_2 во кои $F(x)$ е непрекината. ■

Карактеристичните функции се посебно згодни за изучување на распределбите на суми на независни случајни променливи. Следната, релативно едноставна теорема, покажува зошто е тоа така.

Теорема 7.6 Ако случајните променливи X_1, X_2, ..., X_n, се независни, тогаш карактеристичната функција на $X_1 + X_2 + ... + X_n$ е еднаква на производот на индивидуалните карактеристчни функции X_i.

Доказ: Ако ставиме $Y = X_1 + X_2 + ... + X_n$, тогаш по дефиниција

$$\varphi_Y(t)=Ee^{itY}=Ee^{it(X_1+X_2+\ldots+X_n)}=Ee^{itX_1}e^{itX_2}\cdots e^{itX_n}\overset{\text{поради независност}}{=}$$

$$Ee^{itX_1}Ee^{itX_2}\cdots Ee^{itX_n}=\varphi_{X_1}(t)\varphi_{X_2}(t)\cdots\varphi_{X_n}(t)\,.\ \blacksquare$$

Теоремата 7.6 заедно со "инверзните" формули од теоремата 7.5 овозможуваат да се најде распределбата на суми на независни случајни променливи без барање на области на интегрирање како што работевме во поглавјето 6.3.

ПРИМЕР 7.9 Нека X_1 и X_2 се две независни случајни променливи со експоненцијална распределба со параметар λ. Најди ја распределбата на $Y = X_1 + X_2$.

Решение

Карактеристичната функција на експоненцијалната распределба веќе ја имаме од примерот 7.2

$$\varphi_{X_1}(t)=\varphi_{X_2}(t)=\int_0^\infty e^{itx}\lambda e^{-\lambda x}dx=\lambda\int_0^\infty\lambda e^{-(\lambda-it)x}dx=\frac{\lambda}{\lambda-it}\,.$$

Карактеристичната функција на Y е едноставно

$$\varphi_Y(t)=\varphi_{X_1}(t)\varphi_{X_2}(t)=\frac{\lambda^2}{(\lambda-it)^2}\,.$$

Сега, густината на распределба на Y според "инверзната" формула е

$$f_Y(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}\varphi_Y(t)dt=\frac{\lambda^2}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itx}}{(\lambda-it)^2}dt=\begin{cases}\lambda^2xe^{-\lambda x},\ \text{за}\ x\ge 0\\ 0,\ \text{во спротивно}\end{cases},$$

што е гама распределба со параметри $\alpha = 2$, $\beta = 1/\lambda$. ■

За понатамошните разгледувања поврзани со граничните распределби (распределби што се резултат на конвергенција на низи распределби), ќе ја формулираме следната теорема. Таа во основа ја покажува непрекинатоста на биекцијата меѓу карактеристичните функции и функциите на распределба.

Теорема 7.7 Нека $\varphi_1(t)$, $\varphi_2(t)$, ..., $\varphi_n(t)$, ... е низа карактеристични функции и нека $F_1(x)$, $F_2(x)$, ..., $F_n(x)$, ... е низа на соодветните функции на распределба. Тогаш, ако

$\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$ за секое t и $\varphi(t)$ е непрекината за $t = 0$, важи

1) $\varphi(t)$ е карактеристична функција што соодветствува на некоја функција на распределба $F(x)$, таква што

$\lim_{n\to\infty} F_n(x) = F(x)$ за секое x;

2) и обратно, ако $\lim_{n\to\infty} F_n(x) = F(x)$ за секое x, тогаш

$\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$, каде што $\varphi(t)$ е карактеристична функција што соодветствува на функцијата на распределба $F(x)$.

И во двата случаи, ако граничната функција е непрекината, конвергенцијата е рамномерна. ■

Теоремата овозможува во граничните случаи да можеме по потреба да работиме со карактеристични функции или алтернативно, со функции на распределба.

7.4. Централна гранична теорема

Законот на големите броеви дава информација за веројатноста на настанот една посебна функција од низа случајни променливи, т.е. нивната скалирана (упросечена) сума, да се разликува од очекуваната вредност за некој позитивен реален број ε. Оваа конвергенција формира база за многу други слични резултати што помагаат во решавањето на општиот проблем, каква е приближно распределбата на некоја "добра" функција $h(X_1, X_2, ..., X_n)$. На пример, земајќи ја сумата, вадејќи го од неа нејзиното очекување и делејќи го тоа со стандардната девијација, ние често пати можеме да го изведеме нејзиното асимптотско однесување, т.е. нејзината распределбата кога $n \to \infty$. Оваа група резултати обично колективно се нарекува централна гранична теорема. Таа во ос-

нова разгледува конвергенции на не-дегенеративни случајни променливи по рескалирање на растојанието од законот на големите броеви.

Централната гранична теорема е без сомнение најважниот пронајдок во областа на теоријата на веројатност и статистиката. Таа тврди дека сумата на доволно голем број независни случајни променливи при некои "слаби" услови, приближно има нормална распределба.

Постојат многу варијации и верзии на централната гранична теорема познати по имињата на авторите, на пример, теорема на Муавр-Лаплас (De Moivre–Laplace), Чебишев (Chebyshev), Љапунов (Lyapunov), Бери-Есен (Berry–Esseen), Линберг-Фелер (Lindeberg–Feller), итн. Ние тука најпрво даваме една стандардна верзија на теоремата (мала варијација на Муавр-Лаплас) што најчесто се сретнува во литературата, а понатаму ќе ја дадеме и поопштата верзија на Љапунов.

Теорема 7.8 Ако случајните променливи X_1, X_2, ..., X_n, ... се независни, со иста распределба и конечна дисперзија, $\mu = EX_k$, $\sigma^2 = DX_k$, тогаш важи

$$p\left(\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} < x\right) \underset{n\to\infty}{\overset{\text{рамномерно}}{\rightarrow}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du\,, \quad \text{за } x \in \mathbb{R}.$$

Доказ: Да ставиме $Y_n = \dfrac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} = \sum_{k=1}^{n} \dfrac{X_k - \mu}{\sigma\sqrt{n}}$.

Карактеристичната функција на $X_k - \mu$, ја разложуваме во ред

$$\varphi_{X_k-\mu}(t) = 1 + itE(X_k - \mu) - \frac{t^2}{2} E(X_k - \mu)^2 + t^2 R(t)$$

каде што $R(t) \to 0$ кога $t \to 0$. Од $E(X_k - \mu) = 0$ и $E(X_k - \mu)^2 = \sigma^2$, следи

$$\varphi_{X_k-\mu}(t) = 1 - \frac{t^2\sigma^2}{2} + t^2 R(t)\,.$$

Сега, од особината карактеристична функција на сума на случајни променливи да е производ на нивните карактеристични функции, имаме

$$\varphi_{\sum_{k=1}^{n} X_k - \mu}(t) = \left(1 - \frac{t^2\sigma^2}{2} + t^2 R(t)\right)^n, \text{ од каде што}$$

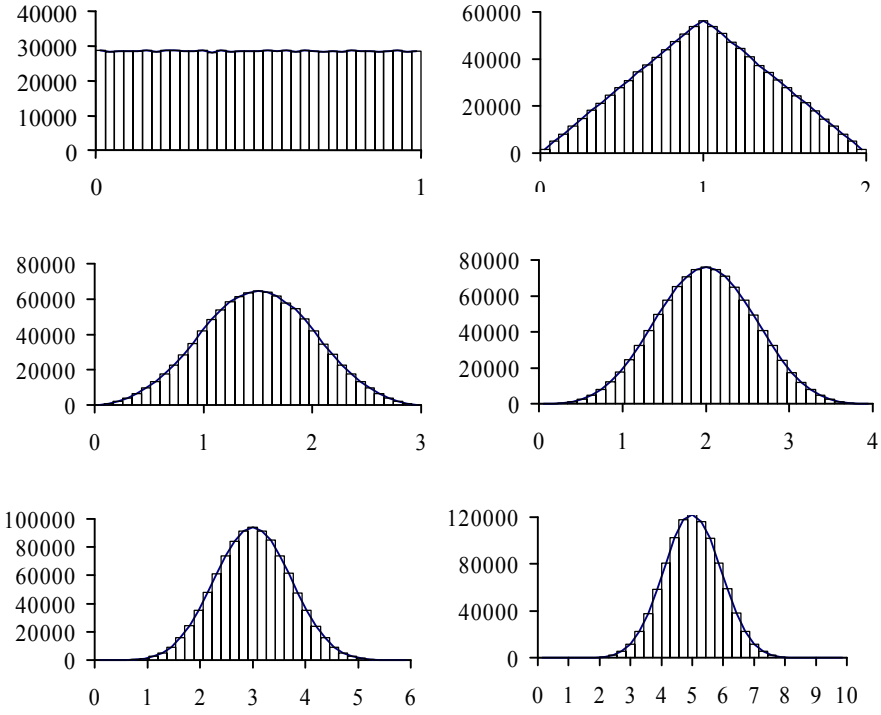
$$\varphi_{Y_n}(t) = \varphi_{\sum_{k=1}^{n} X_k - \mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = \left(1 - \frac{t^2}{2n} + \frac{t^2}{\sigma^2 n} R\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n,$$

Ако пуштиме $n \to \infty$ и имаме предвид дека тогаш $n\left(\frac{t^2}{\sigma^2 n} R(\frac{t}{\sigma\sqrt{n}})\right) \to 0$ конечно за секое фиксно t добиваме $\lim_{n\to\infty} \varphi_{Y_n}(t) = e^{-\frac{t^2}{2}}$.

Оттука, според теоремата 7.7 и непрекинатоста на граничната карактеристична функција (на нормалната распределба $Z(0,1)$) следи рамномерна конвергенција на соодветните функции на распределба кон нормалната $Z(0,1)$ распределба. ■

Со други зборови, теоремата кажува дека сума на независни случајни променливи што имаат иста распределба со очекување μ и стандардна девијација σ има приближно нормална распределба $Z(n\mu, n\sigma^2)$ кога n е доволно големо.

Следните 6 слики покажуваат како сума на случајни променливи со рамномерна распределба конвергира кон нормална распределба. За таа цел направени се по 1 000 000 симулации за сума на n = 1, 2, 3, 4, 6 и 10 случајни променливи со рамномерна распределба и добиените распределби се нацртани со хистограм (за хистограм види поглавје 10).



Од сликите се гледа дека за ваква симетрична, недеформирана распределба како што е рамномерната, веќе за сума од $n = 3$ случајни променливи, распределбата е приближно нормална.

Теоремата 7.8 е доста "слаба" варијанта на централната гранична теорема бидејќи во условот се бара случајните променливи да се со иста распределба. Без доказ ја даваме "јаката" варијанта позната како теорема на Љапунов во која случајните променливи може да имаат произволна (разнородна) распределба.

Теорема 7.9 Нека случајните променливи $X_1, X_2, ..., X_n, ...$ се независни со конечен апсолутен момент од 3-ти ред. Да ставиме

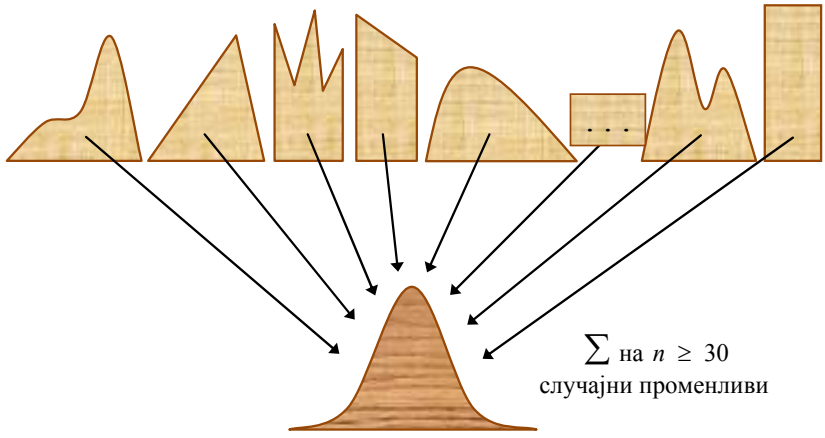
$$\mu_n = EX_n, \qquad b_n^2 = DX_n, \qquad c_n^3 = E\left|X_n - \mu_n\right|^3$$

$$A_n = \sum_{k=1}^{n} \mu_n, \quad B_n = \sum_{k=1}^{n} b_n^2, \quad C_n = \sum_{k=1}^{n} c_n^3 .$$

Ако $\lim\limits_{n\to\infty} \dfrac{C_n}{B_n} = 0$, тогаш

$$p\left(\frac{X_1 + X_2 + \ldots + X_n - A_n}{B_n} < x \right) \underset{n\to\infty}{\overset{\text{рамномерно}}{\to}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du \text{ , за } x \in \mathbb{R}. \blacksquare$$

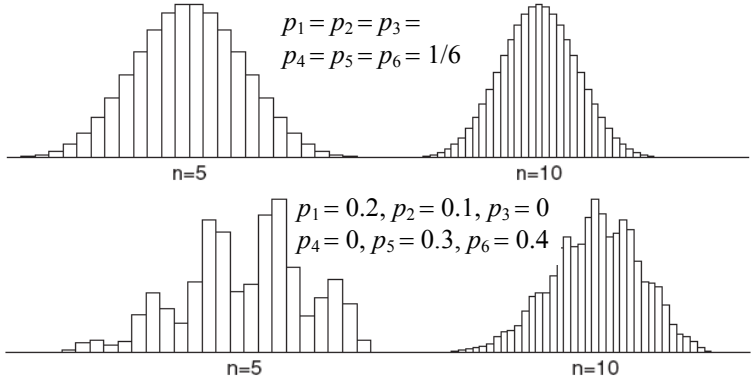
Значи суштинското барање во централната гранична теорема е случајните променливи X_k да се независни, а распределбите може да бидат сосема произволни. Сл. 7.3 симболички го илустрира овој факт.



Слика 7.3 *Сума на произволни случајни променливи води приближно кон нормална распределба*

Едно важно прашање за практичната примена на централната гранична теорема е *колку големо треба да биде n* за да се добие добра апроксимација на нормалната распределба. Со други зборови, колку е брза конвергенцијата кон нормалната распределба? Ова е прашање за кое е тешко да се даде општ одговор. Во голем дел од литературата може да се најде како правило дека за "магичниот број" $n \geq 30$ апроксимацијата со нормална распределба е доволно добра. За примерок $n < 30$, треба да се користи студентовата распределба. Некои автори сметаат дека овој број треба да биде $n \geq 50$, $n \geq 100$ или дури $n \geq 250$. Има размислувања дека магичниот број од 30 примероци (случајни променливи) потекнува од пред-компјутерското време кога за апроксимација се користела студентова распределба со соодветен број степени на слобода. Тука треба да се има предвид дека студентовата распределба станува вистински нормална распределба акко бројот на степени на слобода е ∞. Од причина што не изгледало практично да се работи со долги листи на студентова распределба (вредностите за даден степен на слобода се во еден ред), направен е компромис, табелите на студентовата распределба да одат до 30 степени на слобода што може убаво да се смести на еден лист хатија. Значи можно е од ваква бесмислена причина да потекнува нашироко прифатениот број од 30 примероци за добра апроксимација со нормалната распределба.

Веројатно најправилно е да се прифати дека нема некоја дефинитивна граница за апроксимацијата со нормална распределба да биде добра, поради бесконечниот број на "ненормалности" што дадена популација може да ги поседува. На пример, добро е познато дека сума на симетрични распределби многу побрзо се приближува кон нормалната распределба од сума на несиметрични распределби (види сл. 7.4).



Слика 7.4 *Симулација на сума од 5 и 10 фрлања на хомогена и нехомогена коцка; кај нехомогената, конвергенцијата е многу побавна*

Колку е поголемо отстапувањето од нормалната распределба толку поголем број примероци (случајни променливи) од популацијата треба да се земат за апроксимацијата да биде "доволно" добра.

Централната гранична теорема го разоткрива фактот зошто во пракса толку многу случајни феномени, како на пример заработувачката на берза, нивото на холестерол во крвта, висината или тежината на луѓето или животните, траењето на бременоста, животниот век на живите суштества, итн., се приближно нормално распределени. Сите вакви случајни вредности може да се разгледуваат како резултат на голем број мали независни случајни ефекти собрани заедно. На пример, трошењето на бензин (или нафта) од сите автомобили од одреден бренд, модел и фабрика, иако се произведени во идентичен процес, се разликува од автомобил до автомобил. Оваа случајност (случајна променлива) веројатно се должи на многу причини како што се: непрецизностите во процесот на производство, разлики во употребениот материјал, мали разлики во тежината и други спецификации, разлики во користеното гориво, однесувањето на возачот, бројот на лица и товар што се вози итн. Ако прифатиме дека сите овие мали разлики придонесуваат во случајната променлива на трошење гориво, централната гранична теорема тврди дека таа има тенденција да се потчинува на нормалната распределба. Оваа сума на мали ефекти се јавува насекаде и е во основа последица на тоа што секој покомплексен објект или суштество во природата е составен од помали елементи со случајно "однесување" во однос на целината. Во екстремни случаи такви се на пример популациите на луѓето и животните составени од многу единки, биолошките системи составени од клетки, физичките појави што се резултат на "однесувањето" на електроните, атомите или молекулите итн. Генерално, денеска сите веруваат во точноста на централната гранична теорема – математичарите бидејќи таа е експериментално потврден факт, а останатите бидејќи таа е математички докажан факт.

Историјата на централната гранична теорема е доста интересна. Првата верзија на теоремата (далеку од денешните верзии) била изложена од познатиот математичар Моавр (Moivre) во 1733, каде што нормалната распределба била користена за апроксимација на распределбата на бројот на петки при многукратно фрлање на паричка. Тој пристап тогаш бил далеку пред своето време, и бил практично заборавен сè додека познатиот француски математичар Лаплас (Laplace) повторно не го реанимирал во 1812. Тој ја проширил работата на Моавр користејќи ја нормалната распределба за апроксимација на биномната распределба. Но како и со Моавр, резултатите на Лаплас привлекле многу малку внимание во неговото време. Дури по речиси 100 години била вистин-

ски согледана важноста на централната гранична теорема, кога рускиот математичар Љапунов (Lyapunov) во 1901 ја дефинирал теоремата во општи рамки и прецизно докажал како таа функционира од математички аспект. Денеска, централната гранична теорема се смета за неофицијална "кралица" на теоријата на веројатност.

Без доказ ја даваме класичната теорема на Моавр-Лаплас.

Теорема 7.10 Ако p е веројатноста, $0 < p < 1$, на некој настан во Бернулиевата шема (p е постојано) тогаш

$$p\left(a \le \frac{\mu_n - np}{\sqrt{np(1-p)}} < b \right) \underset{n\to\infty}{\overset{\text{рамномерно}}{\to}} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{u^2}{2}}\, du \text{ , за } -\infty \le a \le b \le \infty. \blacksquare$$

Интересно е да се забележи дека горната верзија на централната гранична теорема бара фиксно p и обезбедува разумна апроксимација на биномните веројатности кога p не е блиско до 0 или 1. За случаите кога p е блиско до 0, покажано е дека нормалната распределба дава слаба апроксимација. Во тој случај, $p \to 0$, $n \to \infty$, Пуасоновата распределба дава многу подобра апроксимација (види теорема 4.2). Овој резултат понекогаш се нарекува *закон на малите броеви*.

Во следните неколку параграфи (а-г) ќе појасниме некои од "популарните" грешки во врска со централната гранична теорема.

а) Најчестата грешка е нејзиното користење во случаи кога таа е јасно несоодветна. Тоа се случува кога шаблонски се користи "сите распределби конвергираат кон нормалната со растење на големината на примерокот". Централната гранична теорема не го тврди тоа. Не! Природата на распределбата обично не се менува со зголемувањето на примерокот. Она што се менува е природата на распределбата на некои скалирани *суми на случајните променливи*, но секогаш со некои рестрикции. Грубо кажано, овие рестрикции се направени да се осигура ниедна случајна променлива од низата да нема доминантно влијание во сумата. Условот што обезбедува незначајност на индивидуална случајна променлива во сумата може да се искаже во многу форми и тоа е основната причина зошто имаме многу верзии на теоремата.

б) Еден "посуптилен" проблем се јавува при собирање на податоци со упросечување што влијае на ефектот што се очекува од централната гранична теорема. На пример, често погрешно се смета дека дневно собраните податоци со упросечување од кои понатаму се добиваат неделни или месечни податоци $X_{нед} = \sum_{k=1}^{n} X_k$, повлекуваат неделните или месечните да имаат распределба што гравитира кон нормална без

разлика на распределбата на дневните набљудувања. Ако се согледа фактот дека неделата има 7 дена, месецот 4 недели, итн., тогаш не може да се очекува ефектот на централната гранична теорема да дојде до израз. Со други зборови, n во горното упросечување не се зголемува доволно, и така недостасува најважниот елемент на централната гранична теорема.

в) Можеби најизразениот проблем што се јавува кај статистичките оценки е лажната "сигурност" кога се работи со мал или не-нормално распределен примерок. Проблемот е во тоа што некои автори сметаат дека централната гранична теорема "ќе се погрижи" за ваквите случаи. На пример, често се смета дека малиот примерок при тестирањето хипотези не е проблем бидејќи студентовата (t) или Фишеровата (F) распределба со соодветните степени на слобода "ќе се справат" со малиот примерок, без потреба распределбата на популацијата да биде нормална. Се чини дека честа конфузија предизвикува важноста на распределбата на примерокот (како при тестирање хипотези) и барањата на конкретната статистика (како t-тестот). На пример, во некоја книга може да се најде забелешка "Бидејќи распределбата на примерокот, а не на популацијата, се користи за тестирање на хипотезите, тоа значи дека секогаш кога примерокот е голем ние можеме комплетно да го релаксираме барањето за нормалност на популацијата, а сеуште да ги користиме особините на нормалната крива за тестирање на хипотезите". Некој читател ова би го перцепирал како тврдење дека самиот примерок ги одразува особините на популацијата без да земе предвид дека тестирањето хипотези е веројатносен концепт базиран на теоретски бескраен примерок, т.е. на долго "трчање" на релативните фреквенции.

г) Нормалната распределба не е единствената кон која конвергираат сумите на случајните променливи. Постои "централна гранична теорема" за секој член од така наречените Леви-Кинчине (Levy–Khintchine) распределби што покрај нормалната, Пуасоновата и Кошиевата, вклучуваат цела група распределби познати под име бесконечно деливи (infinitely divisible) распределби. Уште повеќе, суми на функции од случајни променливи некогаш конвергираат кон добро познати распределби. Таква е сумата на квадрати на случајни променливи со нормална распределба што конвергира кон хи-квадрат распределба.

д) Треба да се има предвид кога се разгледуваат други функции од низи случајни променливи, како на пример максимум (а не скалирани суми), граничната распределба никогаш не е нормална.

Од многуте постоечки обопштувања на централната гранична теорема, тука ќе го наведеме најочигледното. Имено, обопштувањето е

во насока на испитување на граничните теореми во случај на низи од случајни вектори $(X_{1k}, X_{2k}, ..., X_{mk})$, $k = 1, 2, ..., n, \ldots$. Тогаш законот на големите броеви е тривијално исполнет бидејќи кога тој важи за секој елемент тогаш важи и за случајниот вектор. Кај централната гранична теорема ситуацијата е различна, бидејќи граничната распределба е дефинирана преку моментите од прв и втор ред (μ и σ^2), што вклучува и ковариjанса меѓу елементите од случајниот вектор. Се покажува дека и во ваков случај важи централната гранична теорема под услов ниеден случаен вектор да не доминира во сумата, а конвергенцијата е

$$p\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_{1k}, X_{2k}, \ldots, X_{mk}) - (\mu_1, \mu_2, \ldots, \mu_m) < x\right) \underset{n\to\infty}{\to} Z(0, \Sigma)$$

каде што $(\mu_1, \mu_2, ..., \mu_n) = E(X_{1k}, X_{2k}, ..., X_{mk})$, $k = 1, 2, ..., n, \ldots$, а Σ е ковариjансата ($m \times m$ матрица) на $(X_{1k}, X_{2k}, ..., X_{mk})$, за секое $k = 1, 2, ..., n$.

ЗАДАЧИ

1. Нека случајната променлива X има рамномерна распределба на интервалот $[0, 2]$. Определи ја долната граница за $p(|X - 1| \leq 0.75)$ користејќи го неравенството на Чебишев и спореди го овој резултат со точната вредност на оваа веројатност.

2. Врнежите на снег во некој регион (годишно) е случајна променлива со очекување од 70 инчи.
   а) Што може да се каже за веројатноста дека оваа година снежните врнежи ќе бидат меѓу 55 и 85 инчи?
   б) Дали проценката може да се подобри ако додатно се знае дека стандардната девијација е 10 инчи?

3. Во просек 90% од производството на една машина одговара на стандардите. На машината се изработени 18000 предмети. Колкава е веројатноста дека бројот на предмети што одговара на стандардите се разликува од просечниот број вакви предмети за повеќе од 200?

4. Процесот на дупчење дупки на одреден тип електронски плочки прави дијаметри со стандардна девијација од 0.01 милиметри. Колку дијаметри на дупките треба да се проверат така што со веројтност од најмалку 8/9, про-

секот на измерените дијаметри да биде во околина од 0.005 на просечниот дијаметар μ?

5. Бројот X на авиони што пристигнуваат на некој аеродром во еден период на време е со распределба

$$p(k) = \frac{100^k e^{-100}}{k!}, \quad k = 0, 1, 2, \ldots .$$

Користејќи го неравенството на Чебишев определи ја долната граница за веројатноста $p(80 \le X \le 120)$.

6. Определи ги карактеристичните функции на следните распределби:

$$f(x) = \begin{cases} 0, x < 5 \\ a, x \ge 5 \end{cases}, \quad f(x) = \begin{cases} 0, & x < 0 \\ x^2, & 0 \le x < 1 \cdot \\ 1, & x > 1 \end{cases}$$

7. Определи ги карактеристичните функции на случајната променлива X со

а) рамномерна распределба во интервалот $[-b, b]$;

б) распределба по Лапласовиот закон $f(x) = 0.5e^{-|x|}$.

8. Коцкар игра рулет на две полиња, Црвено или Црно. Последните 6 пати топчето застанало на Црвено поле, па ако законот на големите броеви е точен, веројатноста дека при следното фрлање топчето ќе застане на Црно мора да биде поголема отколку да застане на Црвено. Дали е тоа така? Образложи.

9. Законот на големите броеви и централната гранична теорема важат за случајни променливи над кои е потребно наметнување на рестрикции од 3 типа: а) Распределба; б) Зависност и в) Хомогеност. Образложи.

10. Објасни интуитивно зошто

а) конвергенцијата по веројатност е посилен тип конвергенција од конвергенцијата по распределба;

б) конвергенцијата "скоро сигурно" е посилен тип конвергенција од конвергенцијата по веројатност.

11. Нека p биде непознат процент гласачи што се определени да гласаат за партијата на зелените. За проценка на p треба да се анкетираат n гласачи. За i-тиот гласач, нека случајната променлива

$$\delta_i = \begin{cases} 1 \text{ ако гласа за зелените} \\ 0 \text{ во спротивно} \end{cases}$$

и нека просекот $Y_n = (\delta_1 + \delta_2 + ... + \delta_n)/n$ е процентот на гласачи опеделени да гласаат за зелените. Колкаво треба да се земе n за да се добие проценка на p со грешка не поголема од 1% и сигурност од 95%. Направи ја проценката користејќи ја централната гранична теорема.

12. Каков е односот меѓу законот на големите броеви и централната гранична теорема?

8

Од веројатност до статистика

Тргнувајќи од првата глава, ние поминавме долг пат низ областа од математиката позната како теорија на веројатност, во обид да се постават математички рамки за моделирање на случајните феномени. Предмет на изучување беа случајните експерименти (феномени) што поседуваат својство на *стабилност*, т.е., со други зборови, *регуларност на шансите* или *регуларност на случајноста*. Математичката рамка за теоријата на веројатност е мотивирана со формализирање на концептот на регуларност на шансите на исходите од случајните експерименти. Основната математичка структура што произлегува од овој формализам е просторот на веројатност $(\Omega, \mathcal{F}, p)$ каде што Ω е множество на елементарните настани, $\mathcal{F}$ е σ-алгебрата на настани, а p: $\mathcal{F} \to \mathbb{R}$ е функција што ги задоволува 3-те аксиоми а1) – а3). Понатаму структурата е проширена со воведувањето на случајните променливи, чијашто главна улога беше да се префрлиме од настани во броеви што го направи достапен апаратот на математичката анализа. Дискусиите во теоријата на веројатност во главно се вртеа околу техниките за пресметка на веројатностите (probability calculus).

Статистиката во некоја смисла е обратна од теоријата на веројатност. Во теоријата на веројатност, врз база на веројатносниот модел $(\Omega, \mathcal{F}, p)$ со зададени веројатности на елементарните настани, проблемот е да се пресметаат веројатностите на произволните настани од $\mathcal{F}$. Во статистиката, врз база на емпириските податоци од кои може да се проце-

нат веројатностите на некои настани, проблемот е да се дефинира веројатносниот модел $(\Omega, \mathcal{F}, p)$.

Оваа глава е воведот во вториот дел на книгата, посветен на статистичките модели и оценки. Главна цел на оваа глава е да се воспостави *мост* меѓу математичкиот модел што ја дефинира *регуларноста на шансите* на настаните, наречен теорија на веројатност, и статистиката претставена со статистичките модели.

8.1. Мост што недостасува*

Би можело да се каже дека веројатноста засега има посебно место во науката бидејќи таа презентира идеи со повеќе интерпретации од кој било друг концепт во историјата на науката. Секој обид да се направи класификација на сите овие интерпретации е практично невозможна работа бидејќи не постојат јасни граници меѓу нив. Фокусирајќи се на мал дел од овие интерпретации од аспект на статистички оценки, ние накусо ќе ги разгледаме следните три интерпретации на веројатноста базирани на:

1) Класичната дефиниција,

2) Релативни фреквенции,

3) Степен на верување.

Класичната интерпретација е дадена преку класичниот простор на веројатност користен во игрите на среќа и е базиран на доделување еднакви веројатности на елементарните настани според претпоставката за физичка симетрија. Интерпретацијата преку *релативните фреквенции* потекнува од морталитетот и наталитетот на податоците изразени со статистичка стабилност на случувања на настаните при повторувањата на експериментите. *Степенот на верување* потекнува од обидите да се квантифицира релацијата меѓу презентираните согледувања, докази и убедувањето во свеста на индивидуата.

8.1.1. Класична интерпретација на веројатноста*

Генерално е прифатено дека, историски, теоријата на веројатност била воспоставена од коцкарските игри т.е. со разгледување на игрите со шанси како што се фрлањето коцки или парички.

Ако се разгледува случаен експеримент со N еднаквоверојатни исходи (исходи со исти шанси) и во нив настанот A се случува N_A пати, тогаш според класичната дефиниција на веројатноста

$$p(\mathrm{A}) = \frac{N_{\mathrm{A}}}{N},$$

каде што N_{A} обично се нарекува поволен број случаи, а N се сите можни случаи. Иако имплицитно, во пресметките, корените на овој пристап датираат од 17-тиот век, првата експлицитна дефиниција на овој пристап била дадена од Лаплас на почетокот од 19-тиот век.

Прва важна особина на ваквата дефиниција е дека таа се базира на природата на експлицитни шанси на настаните во врска со конкретен експеримент. Втората, суштинска особина е дека таа ја користи "очигледната" физичка симетрија на исходите од експериментот преку што ги одредува поволниот број и бројот на сите можни случаи (исходи).

Ваквата дефиниција на веројатноста може да се смета за несоодветна од повеќе причини. Таа се базира на експлицитен механизам на пресметка на шансите изграден околу физичката симетрија на експериментите со рамноправни исходи. Се разбира, во литературата, најголеми критики се упатуваат кон концептот на еднакви шанси. Што значи терминот "еднаква шанса" и како да ги препознаеме таквите исходи? Лаплас се обидел да воведе принцип на препознавање на еднаквите шанси што подоцна е наречен принцип на "неразличност", кој се врти околу идејата - ако нема причина да се фаворизира некој исход во однос на друг, тие се со исти шанси. Ваквиот принцип води до некои парадокси дискутирани во литературата [Hacking 1975]. Друг проблем е што да се прави ако експериментот нема симетрични исходи. Оттука, класичната дефиниција на веројатноста не може да биде мост меѓу статистичкиот модел и веројатносниот модел на регуларност на шансите. Сепак класичната дефиниција на веројатноста е корисен концепт за пресметка на веројатноста во случаите кога експериментите резултираат во исходи со експлицитни шанси и физичка симетрија.

8.1.2. Интерпретација преку фреквенции*

Интерпретацијата на веројатноста преку фреквенции потекнува од поимот статистичка регуларност воспоставен во текот на 18-тиот и 19-тиот век. Уште тогаш е забележено дека собраните податоци, демографски, економски и социјални, и покрај непредвидливоста на индивидуално ниво, имаат зачудувачка стабилност на нивните фреквенции на групно ниво во подолг временски период. Ваквата социјална статистика (нумеричка наука за општеството) била главна област на примена на статистиката. Најопштиот заклучок од таквите студии бил дека *регуларност може да произлезе од неред и ирационалност*. Општеството може да се карактеризира со релативно стабилни рати на висина, тежи-

на, образование, интелигенција, бракови, криминал, самоубиства, умирања, итн., на луѓето. Идејата дека нередот на индивидуално ниво води до ред на колективно ниво е добро воспоставена во физиката, што веќе беше предмет на дискусија преку законот на Шредингер (пример 7.5).

Во контекст на фреквенциите, веројатноста на некој настан A е гранична вредност на *релативните фреквенции* во кои се случува A ако соодветниот експеримент се повторува многу пати, при исти услови. Ако повторуваме n пати експеримент при исти услови и притоа n_{A} пати се случува A, тогаш веројатноста би била границата на релативните фреквенции на случувањата на A, т.е.

$$p(\mathrm{A}) = \lim_{n\to\infty}\left(\frac{n_{\mathrm{A}}}{n}\right).$$

Се разбира, оваа граница математички не е добро дефинирана и кажува само што се случува во просек кога замислиме бесконечна низа на повторувања на експериментот. Поради тоа, не изненадува што најпрво се мислело дека математичката основа на ваквата дефиниија е во основа слабиот закон на големите броеви

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum\nolimits_{k=1}^{n} X_k - p(\mathrm{A})\right| < \varepsilon\right) = 1, \text{ за секој } \varepsilon > 0, \text{ каде што}$$

$$X_k = \begin{cases} 1 & \text{ако A се случил во } k\text{- тото повторување} \\ 0 & \text{во спротивно} \end{cases} \quad \text{и} \quad \frac{1}{n}\sum\nolimits_{k=1}^{n} X_k = \left(\frac{n_{\mathrm{A}}}{n}\right).$$

Во поопшт случај, ако се дозволи веројатностите да флуктуираат од едно до друго повторување на експериментот, на пример $p_k = p(\mathrm{A})_k$, $k = 1, 2, \ldots$, Пуасоновиот закон на големите броеви

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum\nolimits_{k=1}^{n} X_k - \frac{1}{n}\sum\nolimits_{k=1}^{n} p_k\right| < \varepsilon\right) = 1, \text{ за секој } \varepsilon > 0,$$

бил интепретиран како математичка демонстрација на фактот дека повторување на експериментот секогаш води до константна просечна вредност. Од друга страна, конвергенцијата на просекот кон некоја константна вредност

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum\nolimits_{k=1}^{n} p_k\right) = p\,,$$

не е резултат на некоја "невидлива сила" што воведува ред, туку консеквенца на обичните математички рестрикции

$$0 < p_k < 1 \quad \text{и} \quad (1 - p_k)p_k \leq 1/4.$$

Овие рестрикции имплицитно наметнуваат асимптотска хомогеност на низата веројатности, но сепак не е јасно дека вредноста p коинцидира со $p(\mathrm{A})$.

Кога законот на големите броеви се повикува како доказ за коректноста на дефиницијата на веројатноста преку релативните фреквенции, често се согледува дека аргументот страда од еден тип на циркуларност

> *Се користи конвергенција по веројатност да се дефинира веројатност!*

Овој проблем е отфрлен од некои математичари, како Борел (Borel) или Рени (Renyi) со образложение дека

"Дефиницијата" на веројатноста како нумеричка вредност околу која се вртат релативните фреквенции не е математичка дефиниција, туку само еден интуитивен опис на реалната позадина на поимот веројатност. Од друга страна, законот на големите броеви е теорема изведена од математичкиот поим на веројатност (аксиоматски изграден), па тука нема проблематична циркуларност во дефиницијата [Renyi 1970, стр. 159].

8.1.3. Интерпретација преку степен на верување*

Интересот за ваквата интерпретација на веројатноста е во тоа што таа води кон пристап на статистичко расудување познато како Баесов пристап. Степенот на верување ги покрива двата пристапи, субјективниот и објективниот.

Кај субјективниот степен на верување, веројатноста на некој настан A се базира на субјективна проценка на оној што ги доделува веројатностите. Ако се има предвид дека субјективната проценка на веројатноста е зависна од интелектуалното и психолошкото искуство, веројатностите мора да бидат условни, каде што условот е индивидуалното искуство. За проценка на веројатноста на настан A, се користи Баесовата формула

$$p(\mathrm{A} \mid e) = \frac{p(\mathrm{A})P(e \mid \mathrm{A})}{p(e)}, \ p(e) > 0.$$

Тука $p(\mathrm{A} \mid e)$ е веројатност на A при согледувања e, $p(e \mid \mathrm{A})$ е тежина (вредност, доказ) на согледувањето за даден A, а $p(e)$ е тежина на согледувањето. Баесовата формула претставува кохерентен начин да се добие степенот на верување за веројатноста на настанот A, во светло на согледувањето e.

За прашањето дали субјективните веројатности се однесуваат во согласност со математичката дефиниција (трите аксиоми) одговорот е потврден.

Кај објективниот степен на верување, веројатноста на некој настан A се воспоставува да биде фер, според верувањето на секоја *рационална особа*, под претпоставка да се има дефиниција за терминот *рационално*. Тоа е обид субјективната димензија на интерпретацијата на веројатноста да се придвижи кон објективна. Ваквата интерпретација на веројатноста понекогаш се нарекува и *логичка веројатност*.

8.1.4. Која интерпретација ?*

Генерално може да се заклучи дека од филозофски аспект, двете интерпретации, со фреквенција и степен на верување се подеднакво релевантни и дискусиите околу нивните предности и недостатоци веројатно ќе продолжат и понатаму. Од методолошки аспект, во оваа книга е користена интерпретацијата на веројатноста преку релативните фреквенции. Причината за тоа лежи во фактот што за емпириско моделирање на собрани податоци, иитерпретацијата на веројатноста со фреквенции е најпогодна за објективна статистичка анализа. При работа со собрани податоци врз кои се нема контрола, интерпретациите преку степенот на верување е тешко да се оправдаат. Податоците што во себе содржат систематски информации треба да се моделират независно од индивидуалните или разумните верувања на оној што ги моделира. Интерпретацијата на веројатноста преку степен на верување е попогодна за анализа на несигурностите и ризиците и се користи во ситуациите кога:

а) Нема друга алтернатива од користење на априорни претпоставки поради немање податоци за некој аспект на моделот или голема несигурност на податоците. На пример, ваква ситуација имаме кога мора да се донесе полициска или медицинска одлука врз база на комбинација од несигурни докази или наоди;

б) Моделот се базира на огромна заедничка распределба со возможно стотици параметри и тогаш единствен начин да се дојде до заклучок за непознатите параметри е Баесовиот пристап. На пример, ваква ситуација има при обработка на слики и сигнали во информатиката, во генетиката итн;

в) Има несигурност во одредени параметри на детерминистичкиот модел. На пример, кај анализа на ризици, економија на здравството, климатски анализи итн.

г) Посебна приказна е информатиката каде што Баесовото резонирање е основа кај машинското учење, препознавањето на облици, софтвери за филтрирање на спамови, разни рангирни шеми на преферирани пребарувања итн.

Интерпретацијата на веројатноста преку фреквенциите има суштинска импликација на статистичките методи погодни за анализа на собраните податоци. Понатаму секогаш ќе работиме со "сигурни" податоци што содржат објективни информации, па мостот што недостасува меѓу веројатноста и статистичките податоци ќе го бараме во интерпретацијата на веројатноста преку релативните фреквенции.

8.2. За регуларност на случајноста*

Теоријата на веројатност ја воведовме преку формализација на поимот *случаен експеримент* дефиниран со:

а) Исходите се нееднозначни, но сите се познати однапред;

б) Можност за повторување при исти услови;

в) Стабилност што значи дека постои одредена регуларност на појавување на исходите.

Овие особини дефинираат случаен феномен со регуларност на шансите искажана во третата особина. Тука мора да нагласиме дека втората особина претпоставува дека ќе работиме со интерпретација на веројатноста преку фреквенции. За интерпретацијата преку степенот на верување, втората особина не е неопходна.

Крајната формализација на горните услови што го дефинираат случајниот експеримент може да резултира во едноставен *параметарски* статистички модел

$$\Phi = \{f(x_1, x_2, \ldots, x_n ; \Theta) \mid x_i \in \mathbb{R}\},$$

каде што Φ е фамилија густини (или функции) на распределба што зависат од множество непознати параметри Θ. Статистичкиот модел го претставува механизмот на шанси што не е потполно дефиниран сè додека не се одредат вредностите на непознатите параметри. Во некоја смисла изворната нееднозначност (несигурност) во врска со настаните на случајниот експеримент се сведува на нееднозначност (несигурност) на непознатите параметри. Еднаш кога вредноста на параметрите Θ е некако определена (од собраните податоци), статистичкиот модел може да се користи за изведување многу пошироки заклучоци од оние што ги нудат собраните податоци. Тука треба да се нагласи дека статистичкиот

модел е чисто математичка конструкција чиј емпириски партнер е регуларноста на шансите скриена во податоците.

Што подразбираме под терминот *регуларност на шансите*?

Суштината содржана во оваа фраза доаѓа од двете основни карактеристики на набљудуваниот случаен експеримент:

случајност: неедназначност во појавувањето на исходите,
регуларност: вградена шанса – стабилност на појавување на поединечните исходи.

Дефинирање на поимот шанса (одговара на доделување почетна веројатност на некој елементарен настан) може да се спореди со дефинирање на некој природен и секому јасен, но индивидуален поим како што на пример е убавината. Слично на убавината, шансата не е дефинирана, но ние мислиме дека ќе ја препознаеме кога ќе ја видиме. Од друга страна, за разлика од убавината, препознавањето на шансите во науката не треба да биде оставено на окото на набљудувачот. Понатаму ќе се обидеме да ги формализираме поимот шанса, т.е. поимите случајност и регуларност тргнувајќи од поширок, филозофски аспект.

8.2.1. Што е случајност ?*

Најпрво ќе го разгледаме проблемот преку еден пример. Имено, да се обидеме да ја препознаеме шансата на појавување на број во следните низи на цели брови од 0 до 9:

s_1 = {4,1,4,2,1,3,5,6,2,3,7,3,0,9,5,0,4,8,8,0,1,6,8,8,7,2,4,2,0,9,6,9,8,0,7,8,5,
6,9,6,7,1,8,7,5,3,7,6,9,4,8,0,7,3,1,7,6,6,7,9,7,3,7,9,9,0,7,3,2,4,7,8,4,6,
2,1,0,7,0,3,8,8,5,0,3,8,7,5,3,4,3,2,7,6,4,1,5,7,2,7,…};

s_2 = {1,4,1,5,9,2,6,5,3,5,8,9,7,9,3,2,3,8,4,6,2,6,4,3,3,8,3,2,7,9,5,0,2,8,8,4,1,
9,7,1,6,9,3,9,9,3,7,5,1,0,5,8,2,0,9,7,4,9,4,4,5,9,2,3,0,7,8,1,6,4,0,6,2,8,
6,2,0,8,9,9,8,6,2,8,0,3,4,8,2,5,3,4,2,1,1,7,0,6,7,9,…}.

Детален преглед на овие низи покажува дека не може да се најде некоја препознатлива шема во низата што би овозможила да се "предвидат" следните членови. Така, на прв поглед, се чини дека овие две низи се случајни.

Од друга страна, низата

s_3 = {5,4,8,3,8,7,0,9,6,7,7,4,1,9,3,5,4,8,3,8,7,0,9,6,7,7,4,1,9,3,
5,4,8,3,8,7,0,9,6,7,7,4,1,9,3,5,4,8,3,8,7,0,9,6,7,7,4,1,9,3,
5,4,8,3,8,7,0,9,6,7,7,4,1,9,3,5,4,8,3,8,7,0,9,6,7,7,4,1,9,3,
5,4,8,3,8,7,0,9,6,7,…}

не е случајна бидејќи по секои 15 елементи се повторува истата шема, т.е. истите 15 броеви. Се разбира, за да се "открие" евентуалната шема на повторување, треба да се испитаат доволно голем број елементи (можеби "бесконечно") што во општ случај е невозможна работа. На пример, следната секвенца

s_4 = {3,1,9,5,8,7,6,2,8,8,6,5,9,7,9,3,8,1,4,4,3,2,9,8,9,6,9,0,7,2,1,6,4,
9,4,8,4,5,3,6,0,8,2,4,7,4,2,2,6,8,0,4,1,2,3,7,1,1,3,4,0,2,0,6,1,8,
5,6,7,0,1,0,3,0,9,2,7,8,3,5,0,5,1,5,4,6,3,9,1,7,5,2,5,7,7,3,1,9,5,…}

изгледа дека е случајна, но ако ги разгледаме уште следните 30 елементи

s_4 = {3,1,9,5,8,7,6,2,8,8,6,5,9,7,9,3,8,1,4,4,3,2,9,8,9,6,9,0,7,2,1,6,4,
9,4,8,4,5,3,6,0,8,2,4,7,4,2,2,6,8,0,4,1,2,3,7,1,1,3,4,0,2,0,6,1,8,
5,6,7,0,1,0,3,0,9,2,7,8,3,5,0,5,1,5,4,6,3,9,1,7,5,2,5,7,7,
3,1,9,5,8,7,6,2,8,8,6,5,9,7,9,3,8,1,4,4,3,2,9,8,9,6,9,0,7,2,1,6,4,9,…}

испаѓа дека по 96-тиот елемент низата се повторува.

Имајќи ги предвид горните примери, можеме да заклучиме дека суштината на случајноста е непостоењето на препознатлива шема што го прави предвидувањето на следниот елемент на секое место во низата со која било сигурност невозможно. Заедничкото за низите s_3 и s_4 е тоа што според она што може да се согледа, некој може со сигурност "лесно" да ги погоди останатите броеви од низата. Од друга страна, за низите s_1 и s_2 до бројот до кој се дадени, не може да се согледа некаква шема што би ни овозможила со каква било сигурност да ги погодиме следните броеви. Како да се операционализира овој поим на случајност? Интуитивно, може да замислиме коцкарска машина што фабрикува вакви низи броеви и ние треба да ставиме облог на појавата на следниот број според правилата (шансите) определени од машината. Ако не постои победничка стратегија, т.е. при доволно долга игра машината не може да биде победена, може да сметаме дека фабрикуваните броеви се случајни и стабилни, т.е. поседуваат регуларност на шансите. Ова вклучува не само непредвидливост, т.е. случајност на појавување на следниот број во низата, туку и фактот дека шансите ги рефлектираат релативните фреквенции на појавување на броевите. Од друга страна, какви било шанси на обложување да постави машината за низите од тип s_3 и s_4, таа секогаш ќе губи.

Ваквото интуитивно појаснување на случајноста и стабилноста не ни дава начин за формализација на овие поими. Овој проблем на препознавање на случајност и стабилност, т.е. на регуларност на шансите (ако такво нешто навистина постои) има длабоко филозофско значење и

како проблем се јавува во многу научни дисциплини. Филозофските аспекти на случајноста како поим, накусо ќе ги дискутираме во поглавјата 8.2.3 и 8.2.4.

Различни идеи биле и се во оптек за тоа како да се воведе поимот на случајност, т.е. на непостоење шема. Многу големи математичари низ историјата го разгледувале проблемот на случајност, но дури неодамна, во ерата на компјутерите и информатичката технологија, станало јасно колку овој проблем е комплициран и тежок за дефинирање.

Најлесно би било случајноста да ја дефинираме описно како серија настани без какво и да било значење, без правило, без причина и последица и слично. Сепак тука треба да се биде внимателен и да не се меша нашето субјективно незнаење (човековата неспособност да ја разбере реалноста) со објективното непостоење на значење или правило. Ние лесно би можеле погрешно да заклучиме дека низа броеви е случајна кога не можеме да го препознаме правилото по кое низата е добиена. На пример, низите s_1 и s_2 што ги окарактеризиравме како случајни се всушност децималите на ирационалните броеви $\sqrt{2}$ и π, што е далеку од случајно. Некој би можел да размислува во насока што води до заклучок дека евентуалната дефиниција на случајност во основа води до парадокс. Имено, вистински случајна низа не се потчинува на ниедно правило, а од друга страна, на непостоењето на правило може да се гледа како на стриктно правило, што значи постоење на правило со "негативен" предзнак.

Една важна насока во обидите за дефинирање на случајноста се надоградува над идејата на Колмогоров (Kolmogorov) од 60-тите којашто непостоењето на шема, т.е. правило, го квантифицира со можноста низата броеви да може да се компримира (алгоритамски) [Li, Vitanyi 1997]. Така, низите s_3 и s_4 многу лесно се компримираат (како и секоја периодична низа), додека за низите s_1 и s_2 не се гледа како тоа би можело да се направи без да се знае нивното потекло.

Попрецизно, според Колмогоров, комплексноста на секоја работа може да се дефинира како најкус рецепт, т.е. алгоритам со кој таа работа може да се генерира. Колачите се покомплексни од лебот, бидејќи бројот на инструкции за нивното правење е поголем. Слично, комплексноста на некоја низа броеви е должината на најкусата компјутерска програма што неа ја генерира. На пример, низата 0101010101010101 може да се редуцира на "9 пати 01" и така да се меморира на покус начин. Од друга страна, за низата 010001010000111010 веднаш не се гледа некое правило за редукција, па неа тогаш би морале целосно да ја запомниме. Телефонските броеви 02 234 567 или 02 125 125 се очигледно

"лесни за помнење" бидејќи правилото за нивно компримирано запомнување е едноставно. Понекогаш компримираното запомнување е субјективно, кога датата на раѓање и/или некои други "субјективни" податоци се дел од низата.

Се чини дека концептот на комплексност на алгоритмите за дефинирање на случајноста како нешто за кое нема покус опис од деталното наведување на целината е доста разумен пристап. Ако се ограничиме на низи броеви, низата е случајна ако не може да се редуцира во каква било покуса форма или алгоритам од наведувањето на целата низа. Бидејќи нема правило со кое би ја редуцирале, таа може да се меморира само целосно. Бинарните броеви што се јазикот на денешните компјутери се случајни кога нивната комплексност е еднаква со бројот на нивните цифри. Програмата што во компјутерот се сместува како бинарен код (низа) не може да биде покуса од случајниот број што таа го генерира. Имено, не постои покус начин да се генерира случајниот број од задавањето на сите негови цифри.

Сите денешни програми за генерирање случајни броеви генерираат само "псевдо-случајни броеви". Алгоритамот што нив ги креира е секогаш покус од низата генерирани броеви, па тие низи не го задоволуваат критериумот за случајност базиран на комплексноста на алгоритамот. Уште во 1951 година, познатиот математичар и татко на концептот на работа на денешните компјутери Нојман (John Von Neumann) го сумирал проблемот во една реченица "Секој што се обидува да генерира случајни броеви со аритметички методи прави грешка". Сепак, модерната наука тешко би напредувала без користењето на генераторите на (псевдо) случајните броеви. Тие денеска се користат практично во сите научни области за симулации и комплексни пресметки.

Типичен пример на секојдневно користење на степен на случајност е компресијата на датотеките на компјутер. Ваквите програми бараат повторувања низ податоците и креираат интерни "речници" за податоците да ги запишат во скратена форма. Може да сметаме дека колку поголема редукција на датотеката е направена, толку помалку биле случајни податоците во неа. Вистински случајна низа податоци не може да биде компримирана со никаков алгоритам.

8.2.2. Случајност наспроти непредвидливост*

Случајноста за разлика од непредвидливоста е објективна категорија. Постои цел правец во филозофијата – детерминизам, што тврди дека случајноста не постои, барем не како објективна, туку само како субјективна категорија. На пример, една низа криптирани броеви за не-

кој што не го знае "клучот" му изгледа случајна што не е точно, таа за него само е непредвидлива. Се разбира, оној што го знае клучот лесно ја трансформира низата во читлива информација.

Интригантен аспект на ваквата ситуација е тоа што никогаш не можеме да знаеме дали некоја работа е навистина случајна, бидејќи секогаш постои сомнение дека постои некој "клуч" што ја отклучува навидум случајната информација. Ова е една од основите на суеверието, но исто така и мотивација за откритијата во науката.

Според универзалната хипотеза на детерминизмот, не постои случајност во космосот, туку само непредвидливост, бидејќи постои само еден можен исход за секој експеримент. Се разбира, нашето незнаење ја повлекува непредвидливоста што понатаму води до потребата од користење на случајност, т.е. теорија на веројатност да го покрие недостатокот на знаење за експериментот и исходите.

Некои математички низи, како што се децималите на ирационалните броеви, на пример $\sqrt{2}$ или π, поседуваат некои карактеристики да се третираат како случајни низи, но бидејќи тие се генерираат со куси алгоритми (опишлив механизам), тие се само псевдо-случајни. За некој што не го знае механизмот на нивното добивање псевдо-случајните низи се непредвидливи.

Хаотичните системи се непредвидливи во пракса поради нивната екстремна сензитивност на почетните услови, но тие не се случајни. Детерминираноста е таа што нив ги прави хаотични. Системите со случајност по дефиниција не може да бидат хаотични.

Генерално е прифатено дека постојат 3 механизми за (очигледно) непредвидливо (не може да се тврди случајно) однесување на некој систем:

1) Непредвидливост што доаѓа од околината, како Брауновото движење на честичките или хардверски генератор на случајни броеви;
2) Непредвидливост што доаѓа од почетните услови, како во теориите на хаос и системите чувствителни на мали варијации во почетните услови (ефект на пеперутка). Тука спаѓаат сите проблеми од веројатноста, од фрлање коцки до комплексни природни и општествени експерименти;
3) Интерно генерирана непредвидливост, т.е. псевдо - случајност, како во генераторите на случајни броеви. Обично ваквите непредвидливости се генерираат многу побрзо од непредвидливостите генерирана од околината.

Според некои размислувања, сè додека сме свесни за околината, не можеме да имаме вистинска случајност. Тоа може да се смета и за светот на квантната механика кој се смета за случаен можеби токму поради нашата неможност да го перцепираме и бидеме свесни за него. Единствен начин да имаме случајност е на местата каде што свеста нема влијание, т.е. при перцепција без свесност. Случајноста може да се случи само природно, а не синтетички.

За да се разбере што значи нешто да е случајно, некој може да се обиде да го спореди со спротивното. Но што е спротивното на случајност, ред? Кој го прави редот? Случајност може да се смета за збор со кој го означуваме нашиот недостаток на знаење, т.е. ситуација во која работите не сме ги дефинирале и шематизирале. Ништо не е случајно освен во споредба со тоа што сме му дале ред, а ништо нема ред додека ние не му го дадеме.

8.2.3. Детерминизам наспроти недетерминизам*

Во тесна врска со прашањето на постоење/непостоење на случајноста се недетерминизмот/детерминизмот – правци и движења во филозофијата и општо во науката. Постоењето на случајност обично се поистоветува со валидноста на недетерминизмот. Како да го дефинираме детерминизмот (недетерминизмот)?

Со веројатносна терминологија, користејќи поими на експеримент, исход и настан може да ги дефинираме детерминизмот и недетерминизмот со

Детерминизам: ако при повторување на експеримент под исти услови секогаш имаме ист исход;

Недетерминизам: ако при повторување на експеримент под исти услови може да имаме нееднозначни исходи.

Но како да се направат исти услови? Тоа е невозможно бидејќи времето не може да се сопре. Дали неможноста да се направат исти услови е суштина на недетерминизмот? Изгледа логично и разумно да се смета дека нешто повторено при исти услови дава ист резултат. При исти услови сè би било детерминистичко, но не постојат исти услови (иста положба на атомите во космосот во различно време). Имаме само приближно исти услови и тоа според нашите перцептивни можности. Што би можеле да заклучиме? Поради немање исти услови реалноста е недетерминистичка, а поради имање замислени исти услови во свеста: (апстракцијата) епистемологијата, т.е. човековата мисла, сознанието и науката може да се сметаат за детерминистички.

Да ја примениме горната дискусија на низи цифри. Ако секоја низа цифри се децимали на некој реален (ирационален) број, тогаш не постои недетерминистичка низа цифри 0–9, од што следува дека генерално нема недетерминистичка низа броеви, од што некој би можел да заклучи дека нема недетерминизам туку само псевдо-недетерминизам. Но погоре заклучивме дека реалноста е можеби недетерминистичка. Дали имаме контрадикција? Не, ако согледаме дека реалните (ирационалните) броеви не се реална работа туку само производ на човековата мисла. Затоа таквите низи не може да бидат недетерминистички. Тие само може да бидат непредвидливи, т.е. псевдо-случајни, кога не знаме од каде потекнуваат. Од овој аспект може да се заклучи дека детерминизмот е производ на човековата мисла да се справи со непознатото. Сè е во ставот (верувањето) и ширината на дозволената апроксимација на реалноста во свеста на луѓето. Поборниците на детерминизмот имаат свои аргументи и го тврдат обратното, недетерминизмот е измислица за справување со непознатото.

Во литературата може да се најде и следната филозофско-физичка дефиниција на детерминизмот.

> *Детерминизам*: филозофска доктрина што секоја работа, вклучувајќи секој настан, активност и одлука е неизбежена последица на претходната состојба на работите. Детерминизмот тврди дека секоја работа е детермирирана со претходните работи и е проширување на концептот причина и последица.

Ваквата дефиниција имплицитно претпоставува временска димензија во која минатите настани се причина за идните настани. Некој би можел ова да го критизира сметајќи дека минатото, сегашноста и иднината коегзистираат во една безвременска и конзистентна реалност каде што не само што минатото е причина за иднината туку и обратно, иднината е причина за минатото. Каузалноста е тогаш само макроскопска илузија.

Треба да се биде многу внимателен кога се користат термините "случаен", "недетерминистички", "непредвидлив", "хаотичен". Тие понекогаш се користат како синоними што е погрешно, бидејќи сите имаат свое значење. Терминот "непредвидлив" значи дека е надминат лимитот на нашата способност за предвидување, т.е. лимитот на нашето знаење. Непредвидливост е епистемолошка (субјективна) особина, додека недетерминистички треба да биде онтолошка (објективна) особина. Непредвидливоста е далеку од тоа да повлекува недетерминизам.

Терминот "случаен" е близок до недетерминистички и ние ги користевме како синоними кога под случајност подразбираме "вистинска"

случајност што ја немаме во експериментите (на пример, фрлање коцка или генерирање случаен број).

Хаотичните системи се детерминистички, непредвидливи (епистемолошки) и не "вистински" случајни. Нивната чувствителност на почетните услови, т.е. неможноста да се предвиди нивното однесување се нарекува ефект на пеперутка (Butterfly Effect).

Прашањето дали во природата имаме само недетерминизам, само детерминизам, двата, или нешто трето, е отворено. Голем дел од ова прашање потекнува од варијациите во разбирањето на поимите детерминизам и недетерминизам и нивната врска со други термини, како и начинот на нивното разгледување, епистемолошко или онтолошко.

Квантната физика, физиката на субатомско ниво во главно се смета за недетерминистичка, и како таква важи за многу успешна наука. Принципот на неопределеност (Хајзенберг) поставува лимит на тоа што ние можеме да го знаеме за реалноста и го дава епистемолошкиот хоризонт зад кој науката не може да продре. Ова е основа на Копенхагенската интерпретација на квантната механика којашто базично кажува дека е бесмислено да се прашува "што вистински се случува" во интеракциите на ова ниво. Сепак, ништо од тоа, ниту принципот на неопределеност не имплицира недетерминизам.

Епистемолошкиот недетерминизам (непредвидливоста, неопределеноста) доаѓа од дуализмот што е секогаш наметнат од "набљудувачот" и "објектот на набљудување". Од овој дуализам нема бегање. Класичната физика претпоставува пасивен набљудувач што во квантниот свет е погрешно, и тоа е изворот на епистемолошкиот недетерминизам. Но тој никако не повлекува и онтолошки недетерминизам.

Од друга страна, теоријата на релативност, како и речиси сите "макроскопски" науки се сметаат за детерминистички. Се чини логично секојдневната реалност делумно да се согледува како недетерминистичка бидејќи се судираме со мали броеви (мал број примероци) за да би можела да се "открие" детерминистичката шема на случувањата. Да се потсетиме на законот на Шредингер (пример 7.5), кој тврди дека мал број (молекули на субатомско ниво) го оневозможува детерминистичкото однесување. Во друг, поголем дел од реалноста, законите на физиката се детерминистички (егзактни) и покрај локалниот "хаос" што се јавува кога се разгледуваат изолирано мали броеви (молекули). Се чини дека целокупната наука е едно бавно движење од недетерминизам кон детерминизам со зголемувањето на човековото искуство и знаење, т.е. акумулацијата на поголемиот број информации. Нема конечен одговор на прашањата поврзани со детерминизмот и недетерминизмот, ниту

пак одговор на прашањето дали во природата постои (само) недетерминизам, постои детерминизам или нешто трето. И едното и другото се погледи на реалноста - алатки (засега се чини доста успешни) со кои човекот се обидува да ја моделира и изведува заклучоци за реалноста.

Некои размислувања за недетерминизмот – И она што го сметаме за детерминизам, на пример ако фрлиме камен ќе падне на земја, е специјален случај на недетерминизам каде што предвидувањата се скоро сигурни. Недетерминизам е во основа потрага по знаење и начин да се придвижиме кон детерминизмот. Самата теорија на веројатност е обид на детерминизација (веројатности, распределби и др.) на недетерминизам што опишува непознат детерминизам. Таа изучува недерминистички феномени што се скоро детерминистички (повторување при исти услови, стабилност).

Некои размислувања за детерминизмот – Недетерминизмот доаѓа на сцена кога не можеме детерминистички да објасниме некоја појава. На пример фрлање коцка би бил детерминистички феномен и кога би ја знаеле механиката на раката и сите други физички влијанија и секогаш би можеле да го предвидиме резултатот. Човековата неможност да ги согледа и вкалкулира сите можни влијанија го "тера" да прибегне кон недетерминизмот и така да се обиде донекаде да го детерминира непознатиот детерминизам.

Несомнена вистина е дека не може да се докаже дека ниту детерминизмот ниту недетерминизмот е погрешен или помалку вреден. Верувањето во детерминизмот може да се оправдува со принципот кога нема причина да се преферира некое решение во однос на други, тогаш треба да се прифати наједноставното (Occam's Razor). Со други зборови, нема причина да се повикува недетерминизмот во инаку детерминистичкиот свет бидејќи со него не се добива ништо повеќе во објаснувањето на реалноста.

8.3. Формализација на регуларност на случајноста

За случајноста нашироко дискутиравме во претходните поглавја. Сега ќе се фокусираме на регуларност на случајноста на еден прагматичен начин.

8.3.1. Кон функција на распределба

Како да откриеме дали некоја низа случувања е погодна за статистичко моделирање и ако е, како тоа да се направи. На пример, да ги раз-

гледаме одново низите броеви s_1 и s_2. За да ја испитаме регуларноста на случајноста, т.е. на шансите, не се гледа што друго би можело да се направи од пребројувањето на случувањата, т.е. броевите и така да се проценат шансите за секој од нив. Ако како порано замислиме коцкарска машина што фабрикува вакви низи броеви, тогаш шансите определени од машината (δ_i, i=1, 2, …, n) ги рефлектираат веројатностите (p_i, i=1, 2, …, n) на начин

$$\delta_i = \left(\frac{p_i}{1 - p_i} \right), \; i=1, 2, \ldots, n.$$

За вака понудените опклади нема победничка стратегија. Но дали постојат и кои се веројатностите p_i. За доволно голема низа можеме да ги разгледуваме релативните фреквенции и дали тие со зголемувањето на елементите во низата се стабилизираат околу некоја фиксна вредност. За низата s_1 имаме

рел.фр. за	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
n = 1000	0.108	0.098	0.108	0.083	0.100	0.104	0.091	0.103	0.113	0.092
n = 10000	0.095	0.100	0.100	0.098	0.102	0.100	0.103	0.096	0.103	0.102
n = 50000	0.099	0.101	0.100	0.100	0.101	0.103	0.099	0.099	0.099	0.100

додека за низата s_2 добиваме

рел.фр. за	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
n = 1000	0.093	0.116	0.103	0.102	0.094	0.095	0.094	0.095	0.101	0.105
n = 10000	0.097	0.102	0.102	0.097	0.101	0.105	0.102	0.096	0.099	0.101
n = 50000	0.101	0.101	0.097	0.099	0.100	0.101	0.100	0.100	0.101	0.100

Со овие вредности, лесно се заклучува дека двете низи поседуваат регуларност на случајностите што се моделира со дискретна рамномерна распределба

$$f(x) = \frac{1}{10}, \quad F(x) = \frac{x+1}{10}, \quad \text{за } x = 0, 1, \ldots, 9.$$

Да се потсетиме дека s_1 и s_2 се децималите на ирационалните броеви $\sqrt{2}$ и π. Периодичните низи s_3 и s_4 се децималите на рационалните броеви 17/31 и 31/97. Од теоријата на броеви добро е познато дека децималите на рационалните броеви прават конечна или периодична низа, додека кај ирационалните броеви низата е непериодична. Уште повеќе, децималите на ирационалните броеви се рамномерно распределени, како што веќе видовме на примерите за $\sqrt{2}$ и π.

Тука повторно се наметнува прашањето како со користење на детерминистичка математичка формула (алгоритам) се генерира случајна

низа. Во некоја смисла, не постои објективна непредвидливост на појавување на следната цифра бидејќи алгоритамот ја генерира со сигурност. Па каде е тука случајноста, а потоа и регуларноста на случајноста? Во светлото на претходните (филозофски) дискусии, тука немаме онтолошка случајност (којашто можеби и не постои), но имаме епистемиолошка случајност (псевдо случајност). Имено, ако ја немаме информацијата дека низата е децимална експанзија на конкретен ирационален број, тогаш нема начин да се открие правилото за нејзино генерирање. Ако пак се испуштат неколку стотини или илјади цифри, тоа станува уште поневозможно. Ова е слично на собирањето на статистички податоци, коешто вообичаено не почнува од почетокот на постоењето на податоците ниту оди до крајот.

Како што гледаме, регуларност на случајноста имаме не само во природните процеси, туку и во човековата свест чијшто производ се ирационалните броеви. Врската меѓу регуларноста на случајноста и ирационалните броеви прв ја забележал Веил (Hermann Weyl, 1885 - 1955), кој предложил и соодветен алгоритам за генерирање низи (псевдо) случајни броеви меѓу 0 и 1 со

$$\{u_k = \text{децимален дел од } a \cdot k,\ k = 1, 2, \ldots\}$$

за даден ирационален број a.

Последното прашање што останува да се одговори е ако конвергенцијата на релативните фреквенции не е докажлива, посебно за помал примерок, како да ја формализираме регуларноста во сите низи што сметаме дека поседуваат регуларност на случајноста. Од повеќето пристапи, најкористен е концептот на емпириска (кумулативна) функција на распределба. Наместо барање за конвергенција на релативните фреквенции во различни точки во обид да се дефинираат веројатностите по точки, ние ги разгледуваме кумулативните фреквенции преку целата реална оска.

8.3.2. Емпириска функција на распределба

Основен метод во статистиката е методот базиран на случаен примерок. Од множество објекти (генералното множество) или како што вообичаено се нарелува *популација*, се избираат n-објекти што формираат случаен *примерок*. Примерокот се подвргнува на анализа и врз основа на добиените резултати се изведуваат заклучоци за целата популација.

Популацијата може да смета за случајна променлива. Тогаш, анализата се сведува на определување на распределба на соодветната слу-

чајна променлива, а понекогаш само некои нејзини бројни карактеристики како: просек, дисперзија, моменти итн. Ако X е популацијата од која сме земале примерок со вредности $x_1, x_2, \ldots, x_n$, тогаш примерокот треба да биде репрезентативен, т.е. тој треба некако да ги одразува особините на популацијата. Но ние не ја познаваме популацијата, туку за неа треба да судиме според примерокот. Во ваквата "незгодна" ситуација, единствено на што можеме да се потпреме е случајноста, т.е. регуларноста на случајноста вградена во примерокот. Тоа значи дека примерокот треба да биде случаен, т.е. секој елемент од популацијата треба да има исти шанси да биде избран во примерокот и вообичаено, секој избор треба да биде независен од претходните. Таквиот примерок е случаен примерок и заклучоците изведени од него ќе треба да имаат веројатносен карактер што се однесува на популацијата.

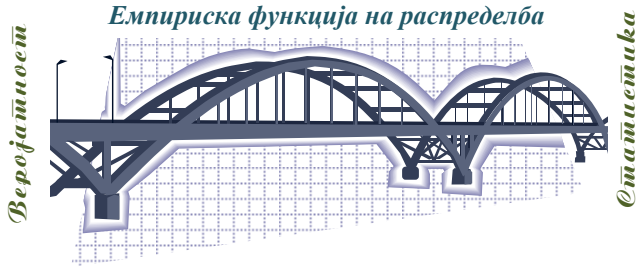
ПРИМЕР 8.1 Да претпоставиме дека сакаме да ја определиме просечната тежина μ на пастрмката во Охридското езеро. Како тоа би го направиле?

Решение

Во овој случај *популацијата* се пастрмките во езерото, т.е. нивните тежини. Случајната променлива на популацијата е X = "тежина на пастрмка во езерото". За да ја најдеме просечната тежина на пастрмките, се разбира, не можеме да ги уловиме сите пастрмки во езерото, и да им ја измериме тежината. Единствено што ни останува е да уловиме одреден број пастрмки (на пример 100), т.е. земеме случаен *примерок*, да ги измериме нивните тежини и преку нив некако да ја оцениме просечната тежина μ на пастрмката во езерото (на популацијата). За примерокот да биде навистина случаен, уловените пастрмки треба да бидат од различни места на езерото, да има од плитки и подлабоки места, покрај населени и надвор од населенеи места, од места со повеќе и помалку храна, итн. Секоја упецана пастрмка, како елемент од примерокот е случајна променлива X_k бидејќи може да зема различни вредности (тежини) со некои веројатности. Откако ќе ја уловиме и измериме пастрмката, добиваме конкретна вредност x_k, една вредност на случајната променлива X_k. Распределбата на тежините на популацијата, како и на примерокот се непознати, но за оценка на просечната тежина μ на сите пастрмки (популацијата) може приближно да ја искористиме просечната тежина од примерокот. Така, $\mu \approx (X_1 + X_2 + \ldots + X_{100})/100$ би бил оценувач на μ, додека $(x_1 + x_2 + \ldots + x_{100})/100$ е оценка на μ. За друг земен примерок оценувачот останува ист, но оценката секако ќе биде друга. Затоа и оценувачот е случајна променлива како функција од случајни променливи. ■

Токму во преминот од карактеристиките на случајниот примерок кон веројатносните карактеристки на популацијата се наоѓа мостот што

недостасува меѓу веројатноста и статистиката. Овој мост ќе го изградиме со така наречената емпириската функција на распределба.



Случајниот примерок ќе го разгледуваме како дискретна случајна променлива

X	x_1	x_2	...	x_n
$p(X=x_i)$	$1/n$	$1/n$	...	$1/n$

(вредностите $x_1, x_2, \ldots, x_n$ се подредени) со соодветна функција на распределба $F_n(x)$ дадена со

$$F_n(x)=\begin{cases} 0 & \text{за } x < x_1 \\ \dfrac{k}{n} & \text{за } x_{k-1} \le x < x_k \\ 1 & \text{за } x \ge x_n \end{cases}.$$

$F_n(x)$ се нарекува емпириската функција на распределба. Таа ја дава релативната честота на настанот $X < x$, додека соодветната функција на распределба $F(x)$ на популацијата треба да ја искажува веројатноста на истиот настан. Тука суштинско прашање со кое се соочувале многу познати математичари во историјата е, дали $F_n(x)$ е добра апроксимација на $F(x)$. Се разбира, според законот на големите броеви (верзија на Бернули), следува дека $F_n(x) \to F(x)$ по веројатност кога $n \to \infty$ за $x \in (-\infty, \infty)$. Оваа конвергенција по веројатност не е доволно добар резултат што би обезбедувал емпириската функција на распределба да биде доволно добра апроксимација на $F(x)$. Следната теорема го обезбедува многу посилно ова барање.

Теорема 8.1 (Гливенко-Кантели (Glivenko-Canteli)) Нека $F_n(x)$ е низа на емпириски функции на распределба во врска со некој случаен примерок и нека $F(x)$ е функцијата на распределба на популацијата. Тогаш важи

$$p\left(\lim_{n\to\infty}\left(\sup_{-\infty<x<\infty}\left|F_n(x)-F(x)\right|\right)=0\right)=1. \blacksquare$$

Не е воопшто чудно што Рени [Renyi 1970], којшто е еден оние што со "задоволство" ја користел оваа теорема како мост меѓу регуларноста на случајноста од примерокот и веројатносниот модел, ја нарекол *фундаментална теорема на математичката статистика*.

Асимптотското однесување во случај на непрекината $F(x)$ не зависи од природата на $F(x)$. За дискретна $F(x)$ тоа не е секогаш случај.

За брзината на конвергенцијата (непрекината $F(x)$) се добива

$$p\left(\lim_{n\to\infty}\left(\sup_{-\infty<x<\infty}\frac{\sqrt{n}}{\sqrt{2\ln\ln n}}\left|F_n(x)-F(x)\right|\right)\le\frac{1}{2}\right)=1 \text{ и}$$

$$p\left(\sup_{-\infty<x<\infty}\left|F_n(x)-F(x)\right|>\varepsilon\right)\le 2e^{-2n\varepsilon^2}.$$

Тука би нагласиле дека брзината на конвергенцијата секако зависи од природата на $F(x)$. За дискретна $F(x)$, супремумот во првото неравенство наместо $\le 1/2$ станува ≤ 1.

Во врска со оваа теорема се многу други гранични резултати.

Така, на пример, централната гранична теорема тврди дека точкасто, $F_n(x)$ асимптотски има нормална распределба со стандарден степен на конвергенција од $\sqrt{n}$, т.е. $\sqrt{n}(F_n(x)-F(x))\to N(0,F(x)(1-F(x)))$.

Директната врска меѓу емпириската функција на распределба и емпириската густина на распределба е дадена со

$$f_n(x)=\frac{1}{2c_n}(F_n(x+c_n)-F_n(x-c_n)) \text{ за } c_n\underset{n\to\infty}{\to}0,\ nc_n\underset{n\to\infty}{\to}\infty,$$

каде што c_n е низа броеви.

Од теоремата 8.1 следува дека бројните карактеристики од случајниот примерок, како очекувањето EX или дисперзијата DX, конвергираат кон очекувањето и дисперзијата на популацијата.

Секоја вредност x_i од случајниот примерок може да се разгледува како вредност на случајна променлива X_i (земање еден елемент од примерокот X) со ист закон на распределба како и X. Оттука, $x_1, x_2, \ldots, x_n$ може да се сметаат за вредности на низа (независни) случајни променливи т.е. вредности на случајниот вектор $(X_1, X_2, \ldots, X_n)$. Очекувањето, дисперзијата, како и секоја друга функција $h(X_1, X_2, \ldots, X_n)$ од случајниот примерок може исто така да се разгледува како случајна променлива.

На крај да забележиме дека опсегот на важење на наведените гранични тврдења може да се прошири и на случаи кога $F(x)$ не е непрекината бидејќи зависноста на конвергенцијата од природата на непознатата $F(x)$ не прави непремостливи тешкотии при параметарската статистика, каде што обликот на $F(x)$ секогаш се задава однапред. Суштинската карактеристика на регуларноста на случајноста е секогаш запазена бидејќи конвергенцијата на кумулативните релативни фреквенции дадени во $F_n(x)$ е стабилен закон.

8.4. За природата на статистичките модели

Во обид да се олесни моделирањето на статистичките податоци, веројатносните концепти за формализирање на регуларноста на случајноста ќе ги поделиме во 3 широки категории: *распределба*, *зависност* и *хетерогеност*. Овие категории овозможуваат на еден кохерентен начин да се разгледуваат статистичките информации во градењето на моделот. Тие се база на секој статистички модел во смисла што секој таков модел е мешавина на состојки од овие 3 категории.

Прво што треба да се има предвид при емпириското моделирање на статистичките податоци е дека

статистичкиот модел е само едно множество веројатносни претпоставки од трите категории: распределба, зависност и хетерогеност.

Статистичкиот модел го опишува механизмот на случајност и шанси со кој се обидуваме да ја досегнеме систематската информација скриена во емпириските податоци (регуларноста на случајноста). Тој се разликува од други модели по тоа што процесите ги искажува преку веројатносни структури како распределба, независност, моменти итн. Примарната задача на статистичкиот модел е да обезбеди статистички адекватен опис на набљудуваниот случаен феномен, но не претендира да понуди објаснување.

Статистичките модели може да бидат разгледувани на *описен /семантички/релативен/емпириски/нереален* начин. *Описната природа* (не објаснувањето) на случајните феномени е примарната особина на статистичките модели. Статистичкиот модел не е лингвиетичка категорија, туку е дефиниран *семантички* во терминологија на множества, тргнувајќи од множества настани па до функции и апарат на математичката анализа. Неговата *релативност* е во однос на расположивите веројатносни структури, т.е. тој е само механизам на шанси дефиниран во термини на веројатносните претпоставки. Валидноста на статистич-

киот модел се проценува според неговата способност да ги фати систематските информации во емпириските податоци. Крајната вистина е се разбира само релативна, искажана во однос на концептуалната рамка што ја нарекуваме теорија на веројатност. Оваа рамка се менува со текот на времето (се надеваме прогресивно) и нема причина она што денеска се смета за вистина да не биде и крајна вистина (што тоа и да значи). Затоа наместо терминот вистина вообичаено се користи валидност. Сепак, не треба да се смета дека, на пример, денеска познатите распределби се сите важни распределби потребни за моделирање на емпириските податоци. Во иднина би можело да се очекуваат иновации во насока на нови распределби како и нови форми на зависност и хетерогеност. Емпириската димензија на статистичките модели се базира на конструктивниот емпиризам што отфрла "каква било објективност во природата". Примарниот критериум за проценка на статистичкиот модел е неговата емпириска адекватност, т.е. неговата адекватност според емпириските податоци. Според тоа, статистичкиот модел не претендира да биде "реален", туку неговата задача е да опишува и предвидува.

Горната дискусија за природата на статистичките модели е предмет на дебата во филозофијата на науките. За нас од поголем интерес е класификацијата на статистичките модели на параметарски и непараметарски.

Параметарските модели Φ се задаваат со фамилија густини (или функции) на распределба што зависат од множество непознати параметри Θ,

$$\Phi = \{f(x\,;\Theta) \mid x \in \mathbb{R}\}.$$

Значи кај параметарските модели обликот (типот) на распределбата $f(\cdot)$ се задава однапред, и останува според емпириските податоци да се определат непознатите параметри Θ.

Спротивно, терминот *непараметарски* се користи во многу различни контексти, но најчесто означува статистички модел со веројатносна компонента дефинирана со фамилија непознати распределби

$$\Phi = \{f(x) \mid f(\cdot) \text{ е множество соодветни распределби}\}.$$

Значи кај непараметарските модели немаме однапред определен специфичен облик на распределба, туку само индиректно правиме претпоставки за особините на распределбата (соодветност) како што се: мазност (дискретна, непрекината, диференцијабилна), постоење моменти или на некој друг начин проценета соодветност на фамилијата распределби. Непараметарските модели само прават имплицитни (наместо експлицитни) претпоставки за непознатата распределба.

На прв поглед изгледа дека непараметарскиот пристап има одредени предности во однос на параметарскиот бидејќи не бара така ограничувачка претпоставка како што е обликот на распределбата. Со тоа се чини дека се избегнува можната статистичка несоодветност на моделот. Од друга страна, непараметарскиот модел е често спакуван во претпоставки што не може да се проверат. Така, статистичките заклучоци губат на прецизност и воопшто, на валидност.

Проблемот на избор на погрешен модел има повеќе димензии, отколку само претпоставката за обликот на распределбата. Кај поедноставните статистички модели, валидноста на претпоставките за независност и еднаква распределеност се посериозен проблем од обликот на распределбата. Минимизацијата на претпоставките во однос на распределбата што би соодвествувала на емпирските податоци многу често води до непрецизност и грешки во статистичките заклучоци. Како општо правило би нагласиле дека поспецифични веројатносни претпоставки за статистичкиот модел водат до попрецизни статистички оценки и појаки статистички тестови.

Индиректните претпоставки за распределбата како што се моментите и мазноста на густината не се доволни за градење на "добар" статистички модел. Постоењето моменти индиректно повлекува распределба, бидејќи тие се определени со густината на распределба

$$E(X^n) = \int_{-\infty}^{\infty} x^n f(x)dx < \infty, \ n = 1, 2, \ldots .$$

Попрецизно, постоењето на моментите зависи од "дебелината" на краевите на соодветната густина на распределба. Тоа го покажува следниот резултат што тврди дека за случајна променлива X и позитивен број k важи

$$\lim_{x\to\infty} x^k p(|X| > x) = 0 \ \Rightarrow \ E(|X|^r) < \infty \ \text{ за } \ 0 \le r < k.$$

Нормалната распределба има потенки краеви од студентовата и нејзините моменти постојат, додека за студентовата, моментите по одредено k не постојат.

Моментите во општ случај не ја определуваат распределбата еднозначно (дури и ако користиме бесконечно од нив). Сепак, ако се лимитираме на одредена класа распределби, проблемот е решлив. Одговорот на прашањето зошто моментите, како поопшт концепт од претпоставката за обликот на распределбата не се користат за статистичко моделирање е едноставен. Параметарскиот пристап дава многу попрецизни и почисти статистички заклучоци. Моментите имплицираат користење не-

равенства коишто се доста груби во споредба со соодветните аналози добиени со параметарскиот пристап.

ПРИМЕР 8.2 Да го разгледаме експериментот фрлање коцка, соодветната случајна променлива X и веројатностите во врска со настанот $\{|X - EX| > 2.5\}$ користејќи ја а) точната распределба; б) рамномерна распределба; и в) геометриска распределба.

Решение

Точната распределба е

x_i	1	2	3	4	5	6
$p(X = x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

$EX = 1(1/6) + 2(1/6) + \ldots + 6(1/6) = 3.5$

$DX = 1^2(1/6) + 2^2(1/6) + \ldots + 6^2(1/6) - 3.5^2 = 2.9166$

а) Користејќи ја распределбата добиваме

$p(|X - EX| > 2.5) = p(|X - 3.5| > 2.5) = p(1 > X > 6) = 0.$

Неравенството на Чебишев дава горна граница од

$$p(|\xi - 3.5| > 2.5) < \frac{2.9166}{2.5^2} = 0.467 .$$

б) Ако сега примениме параметарски пристап и претпоставиме дека X е рамномерно распределена со непознат параметар θ, тогаш

$X \sim$ рамномерна во $(-\theta, \theta)$, т.е. $f(x,\theta) = \frac{1}{2\theta}$, $EX = 0$, $DX = \frac{\theta^2}{3}$.

Поради споредба со пристапот преку моменти (неравенство на Чебишев), бараната веројатност ја решаваме во облик

$$p(|X - 0| > 2.5) = p(|X| > 1.5\sqrt{DX}) = p(|X| > \frac{1.5\theta}{\sqrt{3}}) = 2\,p(X > 0.866\theta) =$$

$$= 2\left(0.5 - \frac{0.866\theta}{2\theta}\right) = 0.134$$

Некој би можел да го добие истото одбирајќи го θ така што $1.5\sqrt{DX} = 2.5$.

Од друга страна, неравенството на Чебишев дава

$$p(|X| > 1.5\sqrt{DX}) = p(|X| > 1.5\sqrt{\frac{\theta^2}{3}}) \le \frac{1}{1.5^2} = 0.444 .$$

в) Ако сега се обидеме параметарски со геометриска распределба имаме

$$f(x,\theta)=\theta(1-\theta)^{x-1},\ 0\le\theta\le 1,\ x=1,2,3,\ldots,\quad EX=\frac{1}{\theta},\ DX=\frac{1-\theta}{\theta^2}.$$

Поради споредба со неравенство на Чебишев, ќе земеме на пример $DX = 1.25$, што дава $\theta = 0.58$, $EX = 1.725$, и бараната веројатност ја решаваме во облик

$$p(|X-1.725|>2.5)=p(|X-1.725|>\sqrt{5}\sqrt{DX})=\sum_{x=5}^{\infty}\theta(1-\theta)^{x-1}=$$
$$=0.58\cdot 0.42^4(1+0.42+0.42^2+\ldots)=0.031$$

Од друга страна, неравенството на Чебишев дава

$$p(|X-MX|>\sqrt{5}\sqrt{DX})\le\frac{1}{(\sqrt{5})^2}=0.2\ .\ \blacksquare$$

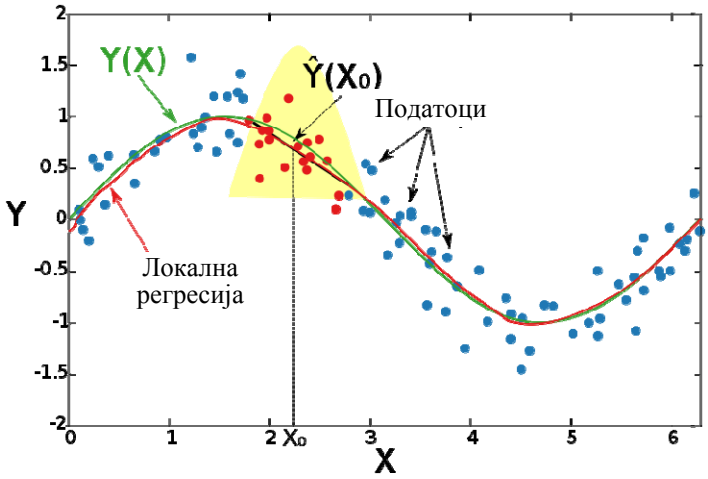
Горниот пример покажува колку проценките на веројатноста направени со неравенството на Чебишев (користење моменти) се инфериорни (груби) во однос на параметарските претпоставки за обликот на распределбата, макар тие да биле погрешни. Додека во "вистинската" распределба, веројатноста на разгледуваниот настан е 0, кај двете претпоставки (погрешни) за обликот на распределбата, рамномерната и геометриската, оваа веројатност е 0.134 и 0.031. Тоа е многу подобро од проценките со моменти дадени преку неравенството на Чебишев. Тука би напоменале дека овие примери се типични, а не екстремни случаи.

Сепак, со некои додатни претпоставки, можно е донекаде да се подобрат проценките базирани на моменти. Имено, нека X е случајна променлива со конечно очекување и дисперзија. Ако додатно претпоставиме дека X е непрекината и унимодална со медијана m_0, тогаш може да се добие подобра проценка од неравенството на Чебишев ($\frac{4}{9\varepsilon^2}<\frac{1}{\varepsilon^2}$) [Biswas 1991].

Покрај екстензивното користење неравенства, во непараметарската статистика многу често е користењето на подредените емпириски податоци и нивните особини. Таквата можност доаѓа од *веројатносната интегрална трансформација* што се базира на тврдењето: во случај на непрекината функција на распределба $F(x)$ на случајната променлива X, без разлика на нејзиниот облик, случајната променлива $Y = F(X)$ има рамномерна распределба на (0, 1). Се разбира тука непрекинатоста на $F(x)$ е индиректно претпоставка за распределбата, но без прецизирање на нејзиниот облик како што се прави кај параметарските методи. Незгодата со овој пристап е во тоа што точниот облик на $F(x)$ мора да биде познат (што не е реално) па повторно мора да се оди на асимптотски пристапи преку емпириската функција на распределба. Во иста група

спаѓаат и методите на "мазнење" на емпириската функција на распределба со цел да се откријат некои особини на непознатата распределба [Silverman 1986, Hardle 1990]. Така повторно се доаѓа до поопшти, но помалку прецизни модели во однос на параметарските.

Денеска се доста популарни таканаречените кернел методи во непараметарските модели. Кернел функциите се користат за мазнење на околините на точките добиени од емпириските податоци со цел да се дојде до густина на распределба (види слика).



Обично моделот претпоставува некои особини на густината на распределба $y = f(x)$ од тип,

$f(x)$ е ограничена на интервал $[a, b]$;

$f(x)$ има непрекинати изводи до трет ред освен на краевите.

Се разбира овие претпоставки не може да се проверат користејќи ги емпириските податоци. Ова е типичен пример на непараметарски модел чијашто цел е да

> *жртвува мал процент од оптималноста на параметарскиот модел за да ја намали чуствителноста на погрешниот избор на обликот на распределбата* [Scott 1992, стр. 33].

Непараметарските тестови претпоставуваат помалку знаење за емпириский материјал во однос на стандардните параметарски тестови. Ова е атрактивно за многумина поради општоста што овозможува да се оди на експерименти без познавање на материјалот.

Тука треба да се нагласи дека иако досегашната дискусија е критички ориентирана кон непараметарските модели, тие се сепак многу

корисни и имаат важна улога во емпириското моделирање. Непараметарските модели обично:

а) зависат (имплицитно) од веројатносни претпоставки што често не може да се тестираат;

б) бараат голем број емпириски податоци;

в) "нескромни" се, премногу се општи;

г) не обезбедуваат мост за да се поврзат со теоретските модели;

д) даваат понепрецизни статистички заклучоци.

Користењето на непараметарските модели со цел да се избегне погрешниот избор на обликот на распределбата не може убедливо да се оправда од следните причини. Како прво, тргнувањето од претпоставките за моделот кон самиот модел може да се направи поефективно во контекст на специфицирање и респецифицирање (поправка) на параметарски модел. Како второ, секогаш мора да се плати цена кога се избираат поопшти, т.е. понепрецизни претпоставки за моделот. Непрецизни претпоставки често водат кон апсурдни статистички заклучоци. Трето, користењето на непараметарските модели често се оправдува во случаите кога е јасно дека нормалната распределба е несоодветна. Ова е слабо оправдување бидејќи постојат бројни други распределби што може да се користат за градење на моделот.

Тука е природно да се постави прашање за улогата на непараметарските модели во статистиката. Еден логичен заклучок би бил дека најважната улога на непараметарските техники со своите кернел функции е во прелиминарната анализа на податоците и во други ситуации кога треба да се тестира валидноста на претпоставките врзани за параметарскиот модел. Во оваа книга, непараметарските модели ги сметаме за комплемент, но не за алтернатива на параметарските. Статистичките техники што понатаму ќе ги разгледуваме се базираат секогаш на параметарско моделирање.

8.5. Вовед во параметарски статистички модел

Секој метод во статистиката директно или индиректно се базира на *случаен примерок*. Веројатносните претпоставки за креирање на статистички модел, генерално може да се поделат во три широки категории [Spanos 1999]:

а) Распределба,

б) Зависност,

в) Хетерогеност.

Почетниот, едноставен статистички модел што се гради над веројатностниот модел се состои од фамилија густини на распределби што зависат од некои множества параметри Θ и случаен примерок,

1) Веројатносен модел: $\Phi = \{f(x; \Theta) \mid x \in \mathbb{R}\}$,

2) Модел на примерок: $(X_1, X_2, \ldots, X_n)$ е случаен примерок.

Бројот на параметри вообичаено е мал. На пример, за нормалната распределба имаме $\Theta = (\theta_1, \theta_2) = (\mu, \sigma)$.

Мотивот за вака дефиниран модел е фактот што стабилните експерименти најчесто имаат исходи што се набљудуваат како нумерички податоци. Од тие причини, овој статистички модел е зададен исклучиво во термини на случајни променливи. Од аспект на веројатносните претпоставки, овој едноставен, но нашироко користен модел се категоризира со:

а) Распределба: *произволна од даден облик*,

б) Зависност: *независни случајни променливи во примерокот*,

в) Хетерогеност: *идентично распределени случајни променливи во примерокот*.

Главната улога на статистичкиот модел е да обезбеди сумарна слика на систематските информации содржани во податоците. Заа таа цел се користи стабилноста, т.е. регуларноста на шансите скриена во податоците.

Се поставува прашање што е тоа што го прави случајниот примерок така фундаментално важен поим. Краток одговор е дека претпоставките за независност и идентична распределеност ги поедноставуваат и моделот, и статистичките заклучоци и оценки. Ова огромно поедноставување е вградено во редукцијата на заедничката распределба на примерокот. Ако со $f_k(x_k; \boldsymbol{\theta}_k)$ ја означиме индивидуалната распределба на X_k, каде што $\boldsymbol{\theta}_k$ се непознатите параметри на X_k, тогаш да се потсетиме дека во таков случај имаме:

независност,

$$f(x_1, x_2, \ldots, x_n; \Theta) = \prod_{k=1}^{n} f_k(x_k; \boldsymbol{\theta}_k), \text{ за сите } (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n;$$

идентична распределба,

$f_k(x_k; \boldsymbol{\theta}_k) = f(x_k; \boldsymbol{\theta})$, за сите $k = 1, 2, \ldots, n$.

Оттука, заедничката распределба едноставно се редуцира на производ на идентичните маргинални распределби

$$f(x_1, x_2, \ldots, x_n; \Theta) = \prod_{k=1}^{n} f(x_k; \boldsymbol{\theta}),$$ за сите $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Значи претпоставките за независност и еднаква распределеност на случајниот примерок драстично ја поедноставува заедничката распределба во два важни аспекта:

1) Редукција на димензионалноста,

Распределбата $f(x_1, x_2, \ldots, x_n; \Theta)$ е јасно n-димензионална, додека $\prod_{k=1}^{n} f_k(x_k; \boldsymbol{\theta})$ е 1-димензионална;

2) Редукција на параметрите,

Бројот на непознати параметри во $\boldsymbol{\theta}$ е најчесто значително помал од оној во Θ.

ПРИМЕР 8.3 Да се разгледа случајот кога распределбата на примерокот, т.е. на случајниот вектор $(X_1, X_2, \ldots, X_n)$ е нормална

$$f(x_1, x_2, \ldots, x_n; \Theta) = Z\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{bmatrix}\right)$$

каде што $\mu_i = EX_i$ се очекувањата, а $\sigma_{ij} = K_{X_i, X_j} = K_{X_j, X_i}$ се ковариј ациите на елементите на случајниот вектор. Разгледај како се намалува бројот на параметрите со претпоставките за независност и еднаква распределеност на примерокот?

Решение

Бројот на непознати параметри $\Theta = \{\mu_i, \sigma_{ij}, i,j = 1, 2, \ldots, n\}$ е $n(n+1)/2$ поради симетријата на ковариј ациите.

Ако се наметне условот за независност, ковариј ациите на различните случајни променливи стануваат 0,

$$\sigma_{ij} = \begin{cases} \sigma_{ii}, \text{ за } i = j \\ 0, \quad \text{за } i \neq j \end{cases}, \; i,j = 1, 2, \ldots, n$$

па почетната распределба се редуцира на

$$f(x_1, x_2, \ldots, x_n; \Theta) = Z\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \begin{bmatrix} \sigma_{11} & 0 & \ldots & 0 \\ 0 & \sigma_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{nn} \end{bmatrix}\right).$$

Значи со условот за независност, индивидуалните (маргиналните) густини на распределба на случајните променливи од примерокот стануваат

$$f_{X_k}(x) = Z(\mu_k, \sigma_{kk}),\ k = 1, 2, \ldots, n,$$

а редукцијата на параметрите, иако драстична, не помага моделот да стане оперативен бидејќи остануваат $2n$ непознати параметри

$$\boldsymbol{\theta}_k = \{\mu_k,\ \sigma_{kk},\ k = 1,\ 2,\ \ldots,\ n\}$$

чиј број расте со зголемување на примерокот.

Сега, ако дополнително го примениме условот за идентична распределеност на случајните променливи од примерокот:

$$\mu_1 = \mu_2 = \ldots = \mu_n = \mu, \quad \text{и} \quad \sigma_{11} = \sigma_{22} = \ldots = \sigma_{nn} = \sigma, \ \text{т.е.}\ \boldsymbol{\theta} = \{\mu,\ \sigma\}$$

заедничката распределба се сведува на производ на маргиналните распределби $Z(\mu,\ \sigma^2)$.

На крај заклучуваме дека претпоставката за независност и идентична распределеност доведе до соодветната редукција на непознатите параметри во насока

$$\Theta = \{\mu_i,\ \sigma_{ij},\ i,j=1,\ 2,\ \ldots,\ n\} \rightarrow \boldsymbol{\theta}_k = \{\mu_k,\ \sigma_{kk},\ k=1,\ 2,\ \ldots,\ n\} \rightarrow \boldsymbol{\theta} = \{\mu,\ \sigma\},$$

а ова понатаму води до едноставен нормален модел,

1) Веројатностен модел

$$\{f(x;\ \Theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mid x \in \mathbb{R},\ \Theta = \{\mu,\ \sigma\} \in \mathbb{R}\times\mathbb{R}^+\},$$

2) Модел на примерок: $(X_1,\ X_2,\ \ldots,\ X_n)$ е случаен примерок. ■

Горниот пример убаво покажува колку е драстична редукцијата и на двете - димензионалноста и бројот на непознати параметри кога се направат претпоставки за независност и еднаква распределеност на случајниот примерок. Од друга страна, примерот дава јасна слика на тешкотиите што се јавуваат кога една или двете претпоставки не се исполнети. Ако не се наметнат рестрикции на независноста и хетерогеноста, се јавуваат два суштински проблема,

а) Проклетство на димензионалноста: $f(x_1, x_2, \ldots, x_n; \Theta)$ е n-димензионална;

б) Проклетство на параметрите: бројот на непознати параметри во Θ расте со зголемувањето на примерокот n.

Проблемите со зависност и хетерогеност веќе беа индиректно разгледувани во претходните глави и се сведуваат на комплицираните концепти (особено за изведување оценки и заклучоци во статистиката) на: распределби на случајни вектори, функции од случајни променливи и условни распределби.

ЗАДАЧИ

1. Класичната дефиниција на веројатноста денеска се смета за реликвија од минатото. Образложи.

2. Зошто би било важно која од двете интерпретации на веројатноста се користи, преку фреквенции или степен на верување?

3. Формулирај две слабости на интерпретацијата на веројатноста преку степенот на верување.

4. Објасни зошто интерпретацијата на веројатноста преку фреквенции е погодна при работа со емпириски податоци, додека интерпретацијата на веројатноста преку степенот на верување е погодна за работа со експериментални податоци.

5. Објасни зошто децималите на ирационалните броеви се погодни за користење како случајни броеви. Дали тие и навистина се користат како генератор на случајни броеви?

6. Според убедувањето на читателот кој пристап во согледување на реалноста е: поедноставен, поелегантен, посодржаен, попреспективен; детерминизмот или недетерминизмот?

7. Испитувана е чувствителноста на некој канал на примерок од 40 телевизори, при што се добиени следните податоци (групни) во микроволти:

Интервал	75-124	125-174	175-224	225-274	275-324	325-374	375-424
Бр.тел.	0	1	5	9	6	8	6

425-474	475-524	525-574	575-624	625-674	675-725
2	2	0	0	1	0

Состави емпириска функција на распределба за овие податоци.

8. Дискутирај го погледот на Рени на законот на големите броеви како поместување од емпирија кон веројатност и евентуалната циркуларност на аргументите.

9. Различните гранични теореми што се однесуваат на асимптотското однесување на емпириската функција на распределба го поместуваат емпирискиот кон математичкиот концепт на функција на распределба $F(x)$. Зошто ги има толку многу?

10. Непараметарските статистички модели може да се разгледуваат како несоодветен обид да се справиме со проблемот на погрешен избор на параметарскиот модел (misspecification). Дали е тоа точно?

9

Вовед во статистика

Мостот што беше воведен во претходната глава овозможи врска меѓу математичкиот концепт на статистички модел и емпирискиот концепт на регуларност на случајноста. Централен столб на овој мост е емпириската функција на распределба $F_n(x)$ којашто е емпириски пандан на функцијата на распределба $F(x)$. Вака дефинираниот мост понатаму го детерминира моделот на статистиката познат како *класичен* (или фреквенционен) пристап и тој е соодветен на интерпретацијата на веројатноста преку релативни фреквенции. Понатамошните дискусии во оваа книга се главно во полза на параметарскиот статистички модел прилагоден за анализа на неекспериментални (набљудувани) податоци. Оправдувањето за ваквиот пристап: класичен, параметарски со неекспериментални податоци зафаќа добар дел од дискусиите во оваа глава. Главните дилеми се околу прашањата:

а) класично наспроти Баесово статистичко заклучување;

б) анализа на експериментални наспроти неекспериментални податоци;

в) распределба на примерокот.

9.1. За статистички модел

Статистиката, според Фишер [Fisher 1956], се состои од поставување (параметарски) статистички модел што обезбедува соодветен (веројатносен) опис на случајниот феномен преку обезбедените емпириски

податоци. Како што веќе видовме, наједноставниот статистички модел се состои од

1) Веројатностен модел, даден со фамилија густини распределби што зависат од неке множество параметри Θ,
$$\Phi = \{f(x\,;\Theta) \mid x \in \mathbb{R}\},$$
2) Модел на примерок, даден со случајниот примерок
$$(X_1, X_2, \ldots, X_n).$$

Емпириските податоци $(x_1, x_2, \ldots, x_n)$ претставуваат една реализација на случајниот феномен опишан со статистичкиот модел. Попрецизно, податоците може да се разгледува како низа специфични вредности на примерокот, т.е. случајните променливи $X_1, X_2, \ldots, X_n$. Така, примерокот може да се разгледува како пресликување

$$(X_1, X_2, \ldots, X_n) : \Omega \to \mathrm{V} \subseteq \mathbb{R}^n,$$

каде што V е множеството дозволени вредности, т.е. простор на примерокот (sample space). Податоците $(x_1, x_2, \ldots, x_n)$ може да се инерпретираат како точка во просторот на примерокот. Дедуктивниот аргумент на овој концепт е едноставен,

ако премисите се точни, одредени валидни резултати секако следуваат.

Премисите не се ништо друго од поставениот статистички модел. Оттука следува дека суштинскиот проблем кај параметарската статистика е сигурноста за валидноста на премисите, т.е. изборот на статистичкиот модел. При погрешно избран модел, заклучоците и резултатите што од него следуваат се нормално сомнителни, т.е.

лош влез $\to$ *лош излез* (*garbage in* $\to$ *garbage out*).

Премисите, т.е. претпоставките за моделот, како што се: обликот на распределбата, независноста и идентичната распределба на примерокот се критични за успешноста на моделот, т.е. за валидноста на изведените резултати. Откако параметрите Θ се определени од податоците, статистичкиот модел е определен и може да биде користен за изведување бројни заклучоци во врска со случајниот феномен.

Досега во текстот, се трудевме да бидеме внимателни и терминот примерок го користевме за случајниот вектор $(X_1, X_2, \ldots, X_n)$, додека за податоците $(x_1, x_2, \ldots, x_n)$ користевме термин вредност или реализација на примерокот. Понатаму во текстот, често пати ќе користиме само тер-

мин примерок, а од контекстот ќе биде јасно дали се работи за случаен вектор или за обични податоци.

ПРИМЕР 9.1 Да го разгледаме едноставниот Бернулиев модел:

1) Веројатностен модел, $\Phi = \{f(x; \Theta) = \theta^x(1-\theta)^{1-x} \mid 0 \le \theta \le 1, x = 0, 1\}$,
2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0, 1\}^n$.

Во Бернулиевиот модел X_n се независни и со иста (Бернулиева) распределба.

На пример, еден примерок со големина $n = 30$ би можел да биде

(0,0,1,0,1,1,0,0,1,0,0,0,1,1,0,1,0,1,0,0,1,0,1,0,0,0,1,1,0,0),

каде што секој елемент (0 или 1) е вредност на соодветната случајна променлива во векторот $(X_1, X_2, \ldots, X_{30})$.

Како да се оцени θ? Во овој случај тој претставува непозната веројатност што може (точкасто) да се оцени преку релативната честота на 1-те во примерокот

$\theta \approx 12/30 = 0.4$.

Така моделот се сведува на распределбата $f(x) = 0.4^x \cdot 0.6^{1-x}$ од која понатаму може да изведуваат бројни заклучоци за настаните. ■

Поставувањето однапред на статистички модел е примарна особина на статистичкото изведување заклучоци и така тоа се разликува од описната статистика што е предмет на следната глава. Значи *првиот* чекор во овој процес е поставувањето на статистичкиот модел преку фамилија распределби што зависат од множество непознати параметри.

Во *вториот* чекор треба да се определи заедничката распределба $f(X_1, X_2, \ldots, X_n; \Theta)$ на случајниот вектор $(X_1, X_2, \ldots, X_n)$. Да забележиме дека означувањето $f(X_1, X_2, \ldots, X_n; \Theta)$ наместо $f(x_1, x_2, \ldots, x_n; \Theta)$ не е вообичаено. Оваа распределба вообичаено се нарекува распределба на примерокот и ваквото означување го користиме да ја нагласиме разликата меѓу примерокот како случаен вектор и реализација на примерокот како вектор од реални вредности. Во овој чекор во игра влегуваат претпоставките за независност и/или еднаква распределеност на случајните променливи X_i.

Понатаму, во *третиот* чекор, се комбинираат априорните информации од распределбата на примерокот и самиот примерок (набљудуваните податоци) за да се определат вредностите на параметрите. На пример, еден модерен пристап е да се дефинира функцијата на подобност $L(\theta)$ (likelihood function). Таа го искажува степенот на подобност при-

дружена на различните вредности за $\theta \in \Theta$ да бидат вистински параметри на моделот во светло на поедина реализација на примерокот x_1, x_2, …, x_n,

$$L(\theta; x_1, x_2, \ldots, x_n) : \theta \to [0, \infty).$$

ПРИМЕР 9.2 Во Бернулиевиот модел

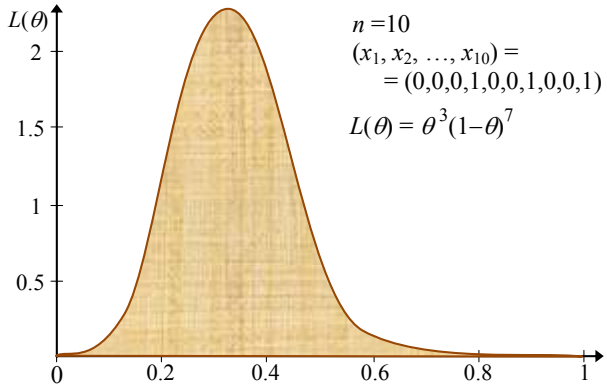
1) Веројатносен модел, $\Phi = \{f(x; \Theta) = \theta^x(1-\theta)^{1-x} \mid 0 \le \theta \le 1, x = 0, 1\}$,
2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0, 1\}^n$,

распределбата на примерокот е од облик

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{k=1}^{n} f(x_k, \theta) = \prod_{k=1}^{n} \theta^{x_k}(1-\theta)^{1-x_k} =$$

$$\theta^{x_1+x_2+\ldots+x_n}(1-\theta)^{n-(x_1+x_2+\ldots+x_n)}$$

На следната слика е прикажана функцијата на подобност за примерок од 10 елементи.



Да забележиме дека $L(\theta; 0,0,0,1,0,0,1,0,0,1)$ е непрекината функција (од θ), и можеме да и бараме максимум, т.е. оценка со максимална веројатност. Од $L'(\theta) = 0$, лесно се добива $\theta = 3/10$. ■

Статистичките процедури, како оценките на непознатите параметри, тестирањето хипотези или предвидувањата се базираат на информациите сумирани во $f(x_1, x_2, \ldots, x_n; \Theta)$. Тоа значи дека успешноста на овие процедури (критично) зависи од претпоставките за статистичкиот модел, т.е. од обликот на распределбите во Φ и добриот избор на примерокот.

9.2. Проширен статистички модел*

Едноставниот статистички модел што досега го дискутиравме има само две компоненти, веројатносен модел и модел на примерокот. Иако тој е стандардно добар за изучување на регуларностите на случајноста во најголемиот број експерименти, сепак кога се испитуваат експерименти со примероци што покажуваат зависност и хетерогеност едноставниот модел не е доволен. Од тие причини, тука ќе воведеме и трета компонента во статистичкиот модел што ќе ја наречеме статистички генератор [Spanos 1999].

Нека $(\Omega, \mathcal{F}, p)$ е простор на веројатност. Имајќи предвид дека настаните од интерес се елементи на $\mathcal{F}$, ние интутивно може да го дефинираме *информација* како подмножество (подалгебра) $\mathcal{D} \subseteq \mathcal{F}$, т.е. $\mathcal{D}$ е информација во однос на веројатносниот простор $(\Omega, \mathcal{F}, p)$. Така $\mathcal{D}$ може да варира од неинформативен случај кога $\mathcal{D} = \{\Omega, \varnothing\}$ (ова е секогаш априори познато, т.е. не знаеме ништо) до комплетно информативен случај кога $\mathcal{D} = \mathcal{F}$ (знаеме сè). Ние секогаш може да дефинираме случајна променлива X таква што најмалата σ-алгебра генерирана од X (тоа е X^{-1}: $\sigma((-\infty, x], x \in \mathbb{R}) \to \mathcal{F}$) коинцидира со $\mathcal{D}$ ($\sigma(X) = \mathcal{D}$) и така да ја разгледуваме информацијата како рестрикција на просторот на настани $\mathcal{F}$ во однос на некој набљудувачки механизам (примерок) од експериментот што е предмет на изучување. Ова овозможува да се операционализираат изразите од облик $E(Y \mid \mathcal{D})$, што може да се интерпретираат како условно очекување на случајната променлива Y за дадена информација $\mathcal{D}$ (како познато множество настани). Може да сметаме дека со каква било трансформација на информацијата, таа не може да се зголеми, туку само постои можност да се намали. Формално, за секоја "добра" борелова функција $g(\cdot)$, важи

$$\sigma(g(X)) \subseteq g(X),$$

а равенство $\sigma(g(X)) = g(X)$ важи само во случаите кога $g(\cdot)$ е еден-еден (инјективна) функција.

ПРИМЕР 9.3 Да го разгледаме експериментот фрлање 2 парички со Ω = {ПП, ПГ, ГП, ГГ}. $\mathcal{F}$ нека биде партитивното множество на Ω, $|\mathcal{F}| = 16$. Дефинираме случајни променливи X и Y со

$$X(\text{ГГ}) = 0,\ X(\text{ПГ}) = X(\text{ГП}) = 1,\ X(\text{ПП}) = 2;$$
$$Y(\text{ГГ}) = Y(\text{ПП}) = 2,\ Y(\text{ПГ}) = Y(\text{ГП}) = 1.$$

Да ја најдеме σ-алгебрата генерирана од X. Прасликите од X се:

$B_0 = X^{-1}(0) = \{TT\}$, $B_1 = X^{-1}(1) = \{HT, TH\}$, $B_2 = X^{-1}(2) = \{HH\}$,

при што овие множества се дисјунктни и нивната унија е Ω. Оттука, σ-алгебрата генерирана од X е

$\sigma(X) = \{S, \varnothing, B_0, B_1, B_2, B_0 \cup B_1, B_0 \cup B_2, B_1 \cup B_2\}$. ■

Сега, $E(Y \mid \mathcal{D} = \sigma(X))$ стандардно може да се изрази со

$$E(Y \mid X = x_k) = \sum_{i=1}^{n} y_i p(Y = y_i \mid X = x_k) = \sum_{i=1}^{n} y_i \frac{p(Y = y_i, X = x_k)}{p(X = x_k)},$$

при што условното очекување може да се интерпретира како случајна променлива

$$Z \equiv E(Y \mid \sigma(X)): \Omega \to \mathbb{R},$$

таква што кога $X(\omega) = x_k$ тогаш $Z(\omega) = z_k$. Случајната променлива Z може да се разгледува во светло на настаните $B \in \sigma(X)$ како

$$Z \equiv E(Y \mid \sigma(X)) = \\ = \sum_{B \in \sigma(X)} E(Y \mid \sigma(X)) \cdot p(B) = \sum_{B \in \sigma(X)} \sum_{i=1}^{n} y_i \cdot p(Y = y_i, B)$$

од што јасно се гледа дека $E(Y \mid \sigma(X))$ е случајна променлива во однос на $\sigma(X)$.

ПРИМЕР 9.4 Да ја разгледаме следната заедничка распределба:

X \ Y	−1	0	1	
−1	0.1	0.2	0.1	0.4
1	0.2	0.1	0. 3	0.6
	0.3	0.3	0.4	

Условните распределби $(Y \mid X = -1)$ и $(Y \mid X = 1)$ се

Y	−1	0	1
$p(Y \mid X = -1)$	1/4	1/2	1/4

Y	−1	0	1
$p(Y \mid X = 1)$	1/3	1/6	1/2

Оттука, за условните очекувања добиваме

$E(Y \mid X = -1) = (-1)(1/4) + 0(1/2) + 1(1/4) = 0$

$E(Y \mid X = 1) = (-1)(1/3) + 0(1/6) + 1(1/24) = 1/6.$

Очекувањето $E(Y \mid \sigma(X))$ е случајна променлива во смисла што таа зема две вредности 0 и 1/6 со веројатности 0.4 и 0.6 соодветно. ■

Условното очекување ги има следните поважни особини:

1) Линеарност:
$E(a \cdot Y + b \cdot Z \mid \sigma(X)) = aE(Y \mid \sigma(X)) + bE(Z \mid \sigma(X))$, a и b константи;

2) $EY = E(E(Y \mid \sigma(X)))$;
На пример во 9.4 имаме дека $EY = (-1)0.3 + 0 \cdot 0.3 + 1 \cdot 0.4 = 0.1 = 0 \cdot 0.4 + (1/6) \cdot 0.6 = E(E(Y \mid \sigma(X)))$;

3) Извлекување на познатото:
$E(h(Y) \cdot g(X) \mid \sigma(X)) = g(X)E(h(Y) \mid \sigma(X))$;

4) Најдобар апроксиматор по најмали квадрати:
$E(Y - E(Y \mid \sigma(X)))^2 \le E(Y - g(X))^2$ за секоја функција $g(\cdot)$.

9.2.1. Статистички генератор

Статистички генератор на условниот момент од прв ред на случајната променлива Y (при претпоставка $EY^2 < \infty$), во однос на информацијата $\mathcal{D}$ е декомпозицијата

$$Y = E(Y \mid \mathcal{D}) + u,$$

каде што $u = Y - E(Y \mid \mathcal{D})$ се нарекува несистематска компонента или почесто грешка т.е. деформација (disturbance term). Постоењето на вакватa декомпозиција е гарантирано со постоењето на моментот од втор ред на случајната променлива Y.

Деформацијата ги задоволува следните особини

1) $E(u \mid \mathcal{D}) = 0$,

2) $E(u^2 \mid \mathcal{D}) = D(Y \mid \mathcal{D}) < \infty$,

3) $E(u \cdot E(Y \mid \mathcal{D})) = 0$.

Статистичкиот генератор станува оперативен кога условната информација $\mathcal{D}$ потекнува од случајниот примерок, обично $\mathcal{D} = \sigma(x_1, x_2, \ldots, x_n)$. Статистичкиот генератор може да се прошири и на условни моменти од повисок ред во облик

$$u^k = E(u^k \mid \mathcal{D}) + v_k, \text{ за } k = 2, 3, \ldots,$$

каде што $u = Y - E(Y \mid \mathcal{D})$. Од посебен интерес се неколкуте условни централни моменти од низок ред.

ПРИМЕР 9.5 Да разгледаме некои модели со статистички генератор.

а) Едноставен модел

$\mathcal{D}$ е неинформативно, $\mathcal{D} = \{\Omega, \varnothing\}$, што значи $E(Y \mid \mathcal{D}) = EY$. Тогаш едноставниот нормален модел е:

1) Статистички генератор $Y_k = EY_k + \varepsilon_k$, $k = 1, 2, 3, ...$;

2) Веројатностен модел

$\Phi = \{f(x; \Theta) = Z(\mu, \sigma^2)\}$, $x \in \mathbb{R}$, $\mathbb{R}_\Theta = \mathbb{R} \times \mathbb{R}_+$, $\Theta = \{\mu, \sigma\}$,

3) Модел на примерок: $(X_1, X_2, \ldots, X_n)$.

б) Регресионен модел

$\mathcal{D}$ вклучува зависносни информации во облик, $\mathcal{D} = (X = x_k)$ што води до општ облик на статистички генератор

$Y_k = E(Y_k \mid X = x_k) + u_k, \quad k = 1, 2, 3, ...$

Најчесто користен е нормалниот линеарен регресионен модел:

1) Статистички генератор: $Y_k = \beta_0 + \beta_1 x_k + u_k$, $k = 1, 2, 3, ...$;

2) Веројатностен модел:

$$\Phi = \{f(Y_k \mid x_k; \Theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_k - \beta_0 - \beta_1 x_k)^2}{2\sigma^2}}\},$$

$x \in \mathbb{R}$, $\mathbb{R}_\Theta = \mathbb{R}^2 \times \mathbb{R}_+$, $\Theta = \{\beta_0, \beta_1, \sigma\}$,

$$\beta_0 = EY_k + \beta_1 EX_k, \quad \beta_1 = \frac{K_{X_k, Y_k}}{DX_k}, \quad \sigma^2 = DY_k - \frac{(K_{X_k, Y_k})^2}{DX_k};$$

3) Модел на примерок: $(X_1, X_2, \ldots, X_n)$ е примерок секвенцијално извлечен од $f(Y_k \mid x_k; \Theta)$. ■

Да забележиме дека веројатносниот дел на регресиониот модел е даден преку условна нормална распределба и дека моделот на примерокот повеќе не е случаен, т.е. тој е само независен, но повеќе не е идентично распределен. Последното доаѓа оттаму што условните распределби $f(Y_k \mid x_k)$ се менуваат со k (условните очекувања се менуваат со x_k).

Нормалниот линеарен регресионен модел има одредени слабости коишто се избегнуваат со други модели како што е линеарниот студентов регресионен модел што тука нема да го разгледуваме.

Сите параметри во статистичките модели од примерот 9.5 имаат веројатносна интерпретација преку моменти на набљудуваните случајни променливи, т.е. статистичка интерпретација. Така во нормалниот линеарен регресионен модел, деформацијата u_k е изведена и ограничена од веројатносната структура на (Y_k, X_k), а истото важи и за параметрите $\{\beta_0, \beta_1, \sigma\}$. Y_k се генерираат за дадени двојки (x_k, u_k).

Статистичкиот генератор за стандарден нормален модел има облик $Y_k = \mu + \sigma^{1/2}\varepsilon_k$, каде што $\varepsilon_k \sim Z(0,1)$ се независни и идентично распределени. Ова може да се замисли како механизам кој кога се симулира со генератор на случајни броеви, дава податоци со иста веројатносна структура како примерокот (нормални, независни и идентично распределени). Кај нормалниот линеарен регресионен модел $Y_k = \beta_0 + \beta_1 x_k + \sigma\cdot\varepsilon_k$, каде што $\varepsilon_k \sim Z(0,1)$ се независни и идентично распределени.

Ваквите модели играат важна улога во анализата на инженерските (и други) проблеми. На пример, да претпоставиме дека се мери протокот на струја (јачината) низ тенка бакарна жица. Моделот за овој феномен може да биде Омовиот закон:

$W = V/R$, (W е јачина на струјата, V е волтажа, а R е отпорот).

Ваквиот модел е изграден врз база на претходно знаење. Ако се направат повеќекратни мерења на струјата, вредностите ќе се разликуваат од предвидувањата на моделот поради многу фактори (неможност да се повторат истите услови). Поради тоа, пореалистичен модел би бил

$W = V/R + \varepsilon$,

каде што ε е додадена променлива на моделот што го одразува фактот дека вредностите на јачината на струјата не може перфектно да се сложуваат со моделот. За ε размислуваме како променлива што ги вклучува ефектите од сите немоделирани извори на варијабилност што влијаат на системот.

9.2.2. Степен на зависност

Статистичкиот генератор дава еден природен начин за проценка на зависноста меѓу две случајни променливи Y и X. Тргнуваме од декомпозицијата на Y,

$Y = E(Y \mid X) + u$,

и сега имајќи ја предвид особината 2) на деформацијата u, $E(u\cdot E(Y \mid X)) = 0$, равенството важи и за дисперзиите на декомпозицијата на Y,

$DY = D(E(Y \mid X)) + D(u)$.

Степенот на зависност $Dep(Y \mid X)$ може да се дефинира со

$$Dep(Y \mid X) = \frac{D(E(Y \mid X))}{DY} = 1 - \frac{D(u)}{DY}.$$

Овој израз првично бил воведен од Колмогоров (Andrey Nikolaevich Kolmogorov 1903–1987), кој го нарекувал степен на корелација.

Степенот на зависност има некои пожелни особини како што се:

1) $0 \le Dep(Y \mid X) \le 1$;

2) $Dep(Y \mid X) = 0$ ако X и Y се независни;

3) $Dep(Y \mid X) = 1$ акко $Y = h(X)$ скоро секаде.

Да забележиме дека според особината 3), $Dep(Y \mid X)$ може да се интерпретира како мера на веројатносна зависност аналогно на математичката функционална зависност $Y = h(X)$, каде што $h(\cdot)$ соодвествува на регресиона функција за Y за дадено $X = x$. Така, $Dep(Y \mid X)$ може да се интерпретира како максимална корелација меѓу Y и сите возможни функции $h(X)$, со максимум што се добива за регресионата функција $h_0(X) = E(Y \mid X)$,

$$Dep(Y \mid X) = \max_{h(\cdot)} \rho^2(Y, h(X)) = \rho^2(Y, h_0(X)).$$

Степенот на зависност не е симетрична функција по Y и X, како што е коефициентот на корелација. Интуитивно, две регресиони функции, на пример $E(Y \mid X) = h(X)$ и $E(X \mid Y) = g(Y)$ не мора да имаат иста форма. Во случај кога регресионата функција е линеарна по X, $E(Y \mid X) = \beta_0 + \beta_1 X$, степенот на зависност коинцидира со квадратот на коефициентот на корелација, $Dep(Y \mid X) = \rho^2(Y, X)$. Ова доаѓа оттаму, што во овој случај $D(E(Y \mid X)) = \rho^2(X, Y) / DX$.

9.3. Статистички оценки

Статистиката во основа се состои од множество процедури за изведување заклучоци за регуларноста на случајноста скриена во набљудуваните податоци и користи

а) априорна информација за формата на веројатносниот модел, и

б) (случаен) примерок $(X_1, X_2, \ldots, X_n)$.

9.3.1. Оценки на непознати параметри

Откако сме поставиле параметарски статистички модел, прв проблем што се наметнува е определувањето на непознатите параметри од Θ. Информациите за тоа се во примерокот (X_1, X_2, …, X_n), т.е. во една конкретна вредност на овој случаен вектор. Во основа ние бараме оценувач на θ од Θ (поединечно) којшто е нешто најдобро што може да се извлече од примерокот. Оценувачот на θ може да се разгледува како пресликување (функција) $h(\cdot)$ од просторот на примерокот што е подмножество $V \subseteq \mathbb{R}^n$ во множеството параметри Θ,

$h(\cdot): V \to \Theta$.

Ова пресликување вообичаено се означува со $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ и притоа $\hat{\theta}$ е оценувач на θ. Да забележиме дека $\hat{\theta}$ е случајна променлива, како функција од случајните променливи X_1, X_2, …, X_n. Ако за случајните променливи земеме конкретни вредности x_1, x_2, …, x_n и ставиме $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$, тогаш $\hat{\theta}$ е обична вредност – оценка на непознатиот параметар θ. И во двата случаја користиме иста ознака, а од контекстот е јасно дали $\hat{\theta}$ е оценувач, или $\hat{\theta}$ е оценка на θ.

ПРИМЕР 9.6 Во Бернулиевиот модел

1) Веројатностен модел, $\Phi = \{f(x; \Theta) = \theta^x (1-\theta)^{1-x} \mid 0 \le \theta \le 1, x = 0, 1\}$,
2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0, 1\}^n$,

бидејќи знаеме дека $\theta = EX$ кога X има Бернулиева распределба, за оценувач $\hat{\theta}$ на θ е природно да се земе

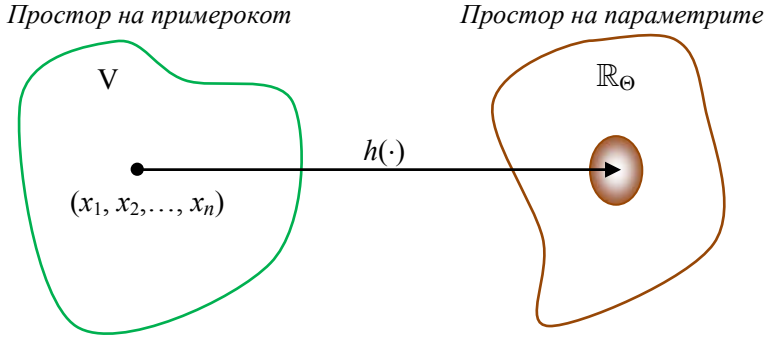
$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} X_k \,. \blacksquare$$

$\hat{\theta}$ како случајна променлива може да земе многу различни вредности во зависност од податоците. Така, ако земеме примерок m пати, добиваме m оценки $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ на непознатиот параметар θ. Овие оценки може понатаму да се комбинираат со цел да се добие "подобрена" оценка на θ, т.е. во крајна инстанца на $f(x, \hat{\theta})$. Понекогаш може да биде подобро сите примероци да се соберат во еден "голем" примерок што исто така дава подобра оценка на θ. Понатаму ќе видиме дека големината на примерокот е многу битен фактор за статистичките оценки. Се разбира, треба да се има предвид дека во многу ситуации не е воз-

можно или е неисплатливо примерокот да се зголемува. На пример, кај археолошките ископувања, број на жртви при несреќи, итн., т.е. кога податоците се набљудувани (над кои немаме никакво влијание, види поглавје 9.5) примерокот често не може да го зголемуваме.

Инерпретацијата на податоците од примерокот како една од многу различни реализации коишто претпоставуваме дека се случајни, овозможува да одиме подалеку од податоците со кои располагаме и изведуваме заклучоци за самиот механизам на случајноста на експериментот. Тоа е поради тоа што кога еднаш на θ и е дадена конкретна вредност (со оценката), механизмот на случајноста дефиниран преку однапред избраниот статистичкиот модел станува еден идеализиран опис на експериментот што е предмет на анализа.

Дефинирањето на еднозначна функција $h(\cdot)$: $\mathrm{V} \to \Theta$ каде што од обликот $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ вообичаено се нарекува *точкаста* оценка на непознат параметар. Друга форма на оценки се *интервалните* оценки, каде што се бара повеќезначна функција што дефинира област во просторот на параметрите $\mathbb{R}_\theta$ во која со висока веројатност се наоѓа вредноста на θ.



Ако се има предвид дека параметрите често се обични реални броеви, не е чудно што најчесто се користат области – интервали за оценки на непознатите параметри. Обично интервалот се задава со две значења на $h(\cdot)$ во облик $(\hat{\theta}_1, \hat{\theta}_2)$, каде што

$$\hat{\theta}_1 = h_1(x_1, x_2, \ldots, x_n), \quad \hat{\theta}_2 = h_2(x_1, x_2, \ldots, x_n)$$

при што обично се бара интервалот да го содржи непознатиот параметар θ со висока веројатност, на пример,

$$p(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) = 0.95 = 95\%.$$

Тоа значи дека при долги повторувања на оценката, интервалот $(\hat{\theta}_1, \hat{\theta}_2)$ ќе го содржи θ во 95% од случаите. Се разбира, во секоја поединечна оценка, немаме гаранција дека θ е во интервалот.

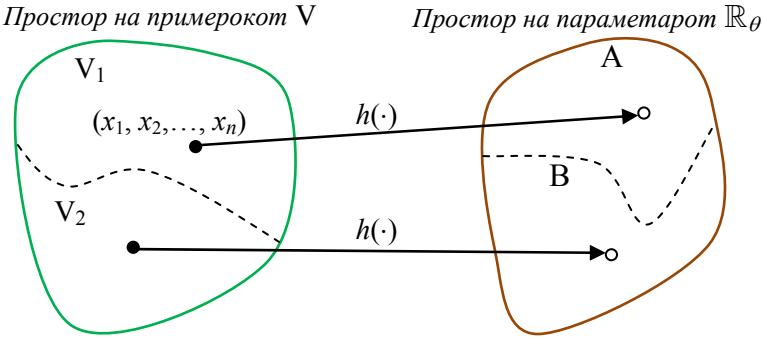
9.3.2. Тестирање хипотези

Друга форма на изведување заклучоци за непознатите параметри е тестирањето хипотези, коешто дава одговор (по веројатност) на прашањата од тип:

а) $\theta = 0.8$;
б) $\theta \leq 0.4$;
в) $\theta \geq 1.2$.

Како што понатаму ќе видиме, сите овие хипотези се базираат на поделба на параметарскиот простор (вообичаено $\mathbb{R}$ или $\mathbb{R}^n$) на 2 дела (дисјунктни подмножества) A и B. Понатаму, користејќи го примерокот, проблемот е да се направи оценка која од двете хипотези (претпоставки) за θ под а) $\theta = 0.8$ или $\theta \neq 0.8$, под б) $\theta \leq 0.4$ или $\theta > 0.4$ или под в) $\theta \geq 1.2$ или $\theta < 1.2$ е точна. Поточно, ако соодветната функција од примерокот (во врска со θ) припаѓа на A се прифаќа едната хипотеза, а ако пак таа припаѓа на $\mathrm{B} = \mathbb{R}_\theta / \mathrm{A}$ се прифаќа алтернативата, т.е. другата хипотеза.

Вака дефинираната постапка е позната и под името Нојман-Пирсонов (Neyman-Pearson) тест.



Функцијата $h(\cdot)$ го дели просторот на примерокот V на две подмножества V_1 и V_2 што соодвествуваат на подмножествата A и B на $\mathbb{R}_\theta$, т.е. $V_1 = h^{-1}(\mathrm{A})$ и $V_2 = h^{-1}(\mathrm{B})$. Тука главен проблем е определувањето на подмножествата A и B како и функцијата $h(\cdot)$. Се разбира, како што по-

натаму ќе видиме, овој проблем многу се поедноставува кога однапред се знае обликот на распределбата на примерокот.

Да забележиме дека пресликувањето $h(\cdot)$ е функција од примерокот, па следователно секој веројатносен заклучок во врска со хипотезата што се испитува се базира на распределбата на примерокот. Според тоа, ние никогаш не сме сигурни дека заклучокот изведен врз база на конкретниот примерок е точен или погрешен, и сме присилени да направиме веројатносен заклучок за тоа дали хипотезата да ја прифатиме или отфрлиме со конкретни веројатности за двата случаја.

9.3.3. Предвидувања

Предвидувањата во статистиката се бават со определување на соодветна функција од примерокот $X_1, X_2, \ldots, X_n$ што овозможува "поглед позади" податоците, т.е. предвидување за идните податоци како што е оценката за непознатиот податок X_{n+1}. Формално, треба да се дефинира оптимална функција $q(\cdot)$ таква што

$$\hat{X}_{n+1} = q(X_1, X_2, \ldots, X_n) .$$

Природен избор за $q(\cdot)$ е таа да биде оптимална во смисла на најмали квадрати, т.е. таа треба да биде таква што ќе го минимизира просекот на квадратната грешка

$$E(X_{n+1} - q(X_1, X_2, \ldots, X_n))^2 .$$

Како што веќе видовме (поглавје 9.2.1) $q(X_1, X_2, \ldots, X_n)$ не е ништо друго, од условното очекување

$$q(X_1, X_2, \ldots, X_n) = E(X_{n+1} \mid X_1, X_2, \ldots, X_n).$$

ПРИМЕР 9.7 Во случај на Бернулиевиот модел, едноставен начин за да се изведе предвидувач X_{n+1} е да се искористи статистички генератор

$$X_{n+1} = \theta + u_{n+1}.$$

Со оглед на тоа што θ е непознато и $E(u_{n+1}) = 0$, природен предвидувач е

$$\hat{X}_{n+1} = \hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} X_k = q(X_1, X_2, \ldots, X_n) .$$

Генерално, $q(\cdot)$ може да се разгледува како композиција на пресликувања од просторот на примерокот V во просторот на параметрите $\mathbb{R}_\theta$, и потоа од $\mathbb{R}_\theta$ во просторот на предвидувања којшто е дел од просторот на примерокот, да речеме V_p, т.е.

$q(h(\cdot)):\mathrm{V} \to \mathbb{R}_\theta \to \mathrm{V}_p$.

Оттука е јасно дека $q(X_1, X_2, \ldots, X_n)$ е исто така случајна променлива со распределба што зависи од онаа на $\hat{\theta}$. Оттука, секое веројатносно тврдење за прецизноста на $\hat{X}_{n+1}$ се базира на распределбата на примерокот $\hat{\theta}$. ■

9.4. Класичен наспроти Баесов пристап*

Тука накратко ќе го опишеме Баесовиот пристап споредeн со класичниот пристап што е предмет на оваа книга. Балансирана споредба на овие пристапи може да се најде, на пример во [Poirier 1995].

Генерално, различните пристапи кон статистиката може да се класифицираат според три основни карактеристики:

а) како ја интерпретираат веројатноста;

б) кои се релевантните информации за статистиката;

в) дали улогата на статистиката е за изведување заклучоци (inferential) или за донесување одлуки (prescriptive, decision-making).

9.4.1. Класичен фреквентен пристап*

Ова е пристапот се користи во најголемиот дел од литературата, во практиката како и во оваа книга. Во класичниот пристап:

а) интерпретацијата на веројатноста е преку релативните фреквенции што нашироко го дискутиравме во претходните глави;

б) во контекст на фреквентната интерпретација на веројатноста, примерокот, т.е. набљудуваните податоци се единствена релевантна информација за статстиката;

в) класичниот пристап е наменет примарно за изведување заклучоци (inferential), но има и проширувања во насока на донесување одлуки.

Како што веќе видовме, класичниот пристап тргнува од поставување на статистички модел, а набљудуваните податоци (вредност на примерокот) $(x_1, x_2, \ldots, x_n)$ се гледа како една реализација на механизмот на регуларност на случајноста, претставен преку статистичкиот модел. Оваа интерпретација природно води до статистички анализи што го нагласува "долгорочното" однесување под суштински исти услови.

9.4.2. Баесов пристап*

Веднаш би нагласиле дека Баесовиот пристап има многу верзии и варијации и затоа тука накусо ќе ги наведеме карактеристиките заеднички за повеќето верзии. Во Баесовиот пристап:

а) интерпретацијата на веројатноста е базирана на степенот на верување (хипотеза), при што доминантен е субјективниот степен на верување;

б) релевантните информации ги вклучуваат набљудуваните податоци како и претходното верување за обликот на распределбата. Попрецизно, набљудуваните податоци се единствениот исход од единствен експеримент, а не исход од многуте можни реализации на експериментот;

в) Баесовиот пристап е примарно во насока на донесување одлуки (decision-making).

Со оглед на тоа што субјективниот степен на верување игра битна улога, не е изненадување дека кај Баесовиот модел статистичките процедури се базираат на ревизија на почетниот степен на верување во светло на набљудуваните податоци. Кај параметарската статистика примарната улога на податоците е да го ревидира верувањето за вредноста на параметрите од множеството Θ. Почетното верување (хипотеза) може да се претстави со почетна густина на распределба

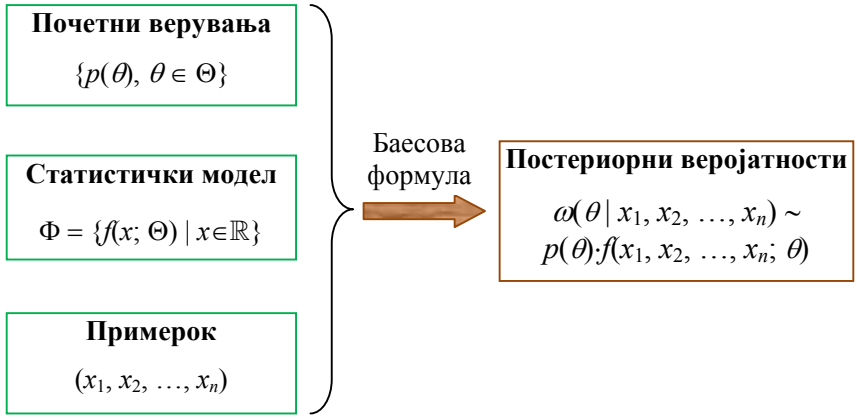
$$Pr(\cdot): \mathbb{R}_\theta \to [0, 1],$$

што е една почетна оценка за тоа како варираат по веројатност вредностите на $\theta \in \Theta$. На пример, ако веруваме дека $\theta = 0.9$ е поверојатно од $\theta = 1.1$, на $\theta = 0.9$ ќе му доделиме повисока веројатност.

Користејќи ја информацијата од примерокот (податоците) сумирана во заедничката распределба на примерокот $f(x_1, x_2, \ldots, x_n; \Theta)$, почетната распределба се ревидира во постериорна распределба $\omega(\Theta \mid x_1, x_2, \ldots, x_n)$ со Баесовата формула

$$\omega(\Theta \mid x_1, x_2, \ldots, x_n) = \frac{p(\Theta) \cdot f(x_1, x_2, \ldots, x_n \mid \Theta)}{\int f(x_1, x_2, \ldots, x_n \mid \Theta) d\Theta}.$$

Бидејќи именителот не зависи од параметрите Θ, може да ставиме пропорционалност меѓу левата и десната страна, т.е. $\omega(\Theta \mid x_1, x_2, \ldots, x_n) \sim p(\Theta) \cdot f(x_1, x_2, \ldots, x_n; \Theta)$.



Концептот на Баесовиот пеистап може да се сумира во една линија,

Почетно верување + Примерок = Ново, подобрено верување.

9.5. Експериментални наспроти набљудувани податоци

Важен аспект од статистиката е анализата на набљудуваните податоци и согледување дали ние имаме или не некоја активна улога во одредувањето на нивните нумерички вредности. Од една крајна страна, може да правиме експеримент во контролирана околина (да речеме лабораторија), и со контрола на одредени влијанија, т.е. фактори (ги нарекуваме влез) да го испитуваме нивниот ефект на други фактори (ги нарекуваме излез), воспоставувајќи причинско-последична врска меѓу влезот и излезот. Од друга крајна страна, имаме набљудувани податоци над кои немаме никакво влијание, т.е. вредностите на податоците вклучени во влезот и излезот се комплетно без наше влијание (ако такво нешто постои, бидејќи самото набљудување евентуално влијае на податоците). Меѓу овие крајности имаме експерименти и податоци со помало или поголемо влијание на набљудувачот.

9.5.1. Експериментални податоци

Во почетокот на XX век, експериментите заедно со причинско-последичните објаснувања биле практично синоним за научен метод. Податоците од експериментите спроведувани во "идеални" услови и каде што истражувачите имаат комплетна контрола на возможните влијанија, вообичаено немаат потреба од статистичка анализа. Многу често, ваквите причинско-последични врски што се предмет на истра-

жување користат математички апроксимативни техники. Повеќето експерименти од модерната физика, хемија, биологија и другите науки што се изведуваат во лабораториски услови се од ваква природа. Клучот на успешноста на ваквите експерименти е во изолацијата на феноменот од интерес од други (неконтролирани) влијанија. Ако тоа не се обезбеди, заклучоците базирани на добиените податоци ќе бидат неадекватни или дури погрешни.

Се разбира, во најголем број случаи, задоволувачката контрола на споредните влијанија не може да се направи. Тоа значи едно поместување од полна кон делумна контрола на несаканите влијанија и за такви случаи се развиени повеќе (статистички) техники, како рандомизација, блокирање, репликација, за неутрализирање овие влијанија. Со други зборови, се прави обид за изолација од несаканите влијанија не со директна контрола, туку со други средства.

Статистичкиот модел и експериментот се две страни од иста монета. Експериментот има за цел да ја изолира причинско-последичната врска меѓу влезот и излезот, а она што е вон контрола е несистематска (често бел шум) грешка. Ако таа содржи систематска информација што може да се детектира со статистички модел, тогаш веројатно експериментот игнорира важно влијание и најмалку што треба да се направи е тоа влијание некако да се неутрализира.

Во некои случаи кога реализацијата на контролиран експеримент не е возможна, а предмет на истражување е фиксна популација, постојат некои техники на збирна анализа (survey sampling) што може да се користат. Кај лабораториското експеримент се обидуваме да го изолираме феноменот од интерес преку контрола или неутрализација на сите вклучени влијанија. Збирната анализа го изолира феноменот од интерес земајќи ги предвид сите влијанија преку внимателно осмислена селекција на примерокот и придружените информации.

ПРИМЕР 9.8 Широко користен пример на земање примерок со влијание е кај проценката на рејтингот на политичарите, т.е. волја на гласачите при изборите. Во таков случај, вообичаено се прави анкета на мала пропорција од гласачката популација. За добиените резултати да бидат реални, потребно е внимателно да се избере примерокот на гласачи со цел тој да ја одразува волјата на целата популација. Исто така, потребно е внимателно да се одберат прашањата за анкетата што е проблем кој нема посебно да го разгледуваме во оваа книга. При изборот на примерокот некои од техниките на сумарна анализа се од голема полза:

1) *Слоевит примерок* (*Stratified sampling*). Овој метод на земање примерок може да се користи кога постои однапред позната информација за

хетерогеноста на популацијата што е предмет на анализа. Хетерогеноста значи дека популацијата може да се подели на групи, т.е. слоеви. И сега, земајќи случајни примероци од слоевите може да се подобри репрезентативноста на примерокот. На пример, се покажува дека прецизноста на проценката на просекот на популацијата (според дисперзијата) расте со разликите на просеците меѓу слоевите. Во случај на проценка на волјата на гласачите, слоевит примерок би бил кога би се земале случајни примероци засебно според: степенот на приход или образование, според возраст, место на живеење итн;

2) *Примерок по групи* (*Cluster sampling*). Овој метод се користи кога популацијата е природно веќе поделена во групи, а потребна е одредена економичност во трошоците при земањето примерок. Притоа од секоја група се зема случаен примерок сразмерен на големината на групата. Во случај на проценка на пулсот на гласачите би можеле да се земаат случајни примероци по изборните единици, општини, градови итн;

3) *Примерок по квоти* (*Qouta sampling*). Овој метод се користи кога треба да се испита како некои фактори влијаат на карактеристиките на популацијата што е предмет на анализа. На пример, при испитување на јавното мислење често пати е важно да се знае какво е тоа од страна специфична група луѓе одбрана според возраст, заработувачка, пол итн. Во случај на испитување на пулсот на гласачите, целта би била да се испитаат факторите што влијаат на нивната одлука, игнорирајќи ја случајноста на примерокот. ■

Збирните податоци се слични на експерименталните податоци каде што статистичкиот модел и експериментот се две страни од иста монета. Како што веќе дискутиравме погоре, целта на експериментот е да се изолира врската меѓу влезот и излезот, а во овој случај да се идентификуваат сите влијателни фактори со внимателно одбирање на збирните податоци. Ако изолацијата е успешна, тоа што не е земено предвид со збирните податоци треба да е несистематско влијание. Се разбира, идентификацијата на причинско-последичната врска меѓу влезот и излезот е многу потешко преку збирните податоци отколку во лабораториски контролирана околина.

9.5.2. Набљудувани податоци

Кога на колекцијата податоци во врска со експериментот што се испитува немаме никакво влијание, нив ги сметаме за набљудувани податоци. Тоа значи дека набљудувачот во овој случај е пасивен и не може да влијае на нумеричките вредности на набљудуваните променливи. Ова е спротивно во однос на експерименталните и збирните пода-

тоци каде што набљудувачот има активна улога во определувањето на овие нумерички вредности.

Тука е природно да се постави прашањето дали истите техники за работа со експерименталните податоци може да се користат и кај набљудуваните податоци. Историски гледано, една од посилните страни на статистиката е леснотијата со која техниките користени во контекст на една дисциплина може да се користат во други дисциплини. На некој начин статистиката со своите методи и техники се користи на речиси идентичен начин во различните науки, било да се природни или општествени. На статистичките методи може да се гледа како на тројански коњи што се уфрлуваат во другите дисциплини без да се води доволно сметка за нивната соодветност. Дури и во една иста дисциплина, обично треба да се води сметка за секое индивидуално испитување и направи соодветно прилагодување. На пример, да претпоставиме дека собираме астрономски податоци, т.е набљудувани податоци за движењето на планетите со цел да се процени вториот Кеплеров закон и тоа

r - растојание на планетата до сонцето,

φ - аголот меѓу линијата од планетата до сонцето и главната оска на елипсата (патеката).

Во случај кога движењето би го разгледувале во приближно изолиран систем, би можеле да ги користиме истите статистички техники како и кај експерименталните податоци. Од друга страна, за некои од планетите е практично невозможно да се определи елипсата на движење (веројатно не е елипса) поради надворешни влијанија. Така, венера е преблиску до земјата, и затоа влијанието на земјата не може да се игнорира што понатаму води до проблем на 3 тела за којшто сеуште нема решение. Во случај на јупитер и марс, чиишто растојанија до другите планети се значителни, природата била многу "покоректна" и овозможила користење на методите разработени во контекст на експерименталните податоци. Значи во случај на набљудувани податоци, коишто не потекнуваат од приближно изолиран систем, методите и техниките што се користат за експерименталните податоци често пати се неадекватни.

Разгледување на набљудуваните податоци како тие да се мерења од контролиран експеримент може да биде несоодветно. Исто така, терминте *популација* и *примерок* не се секогаш соодветни за набљудуваните податоци бидејќи тие носат конотација на набљудување на изолиран систем. Вообичаено тоа што ние го набљудуваме е некој активен процес што не може да се изолира од околните влијанија, а не некоја популација од која земаме репрезентативен примерок. За несреќа, терминот примерок е толку интегриран во статистиката што тој се-

когаш има исто значење без разлика на типот на податоците. Правилна дефиниција на терминот примерок би била, примерок е *множество случајни променливи со специфична веројатносна структура*. При статистичкото моделирањето на експерименталните податоци, проблемот на избор на статистичкиот модел е релативно едноставен и затоа во литературата тој обично не се дискутира. За набљудуваните податоци овој проблем може да биде деликатен и да бара додатни активности.

9.6. Одново за лицето на статистичката анализа

Традиционалното лице на статистичката анализа е базирана на 3 компоненти: статистички оценки, тестирања и предвидувања, но тоа не дава комплетна слика на процесот на статистичко моделирање. Според таткото на модерната статистика Фишер (Fisher, 1890 – 1962), статистиката има за цел да направи редукција на огромниот квантитет на податоци во неколку нумерички вредности (параметри) коишто соодветно ги сумираат сите релевантни информации од податоците. Тој ги класифицирал проблемите на статистичката анализа во 3 широки области:

1) Спецификација – избор на соодветен статистички модел;

2) Оценки – избор на статистики за оценка на параметрите;

3) Распределби – изведување распределби за примерокот од 2).

Тестирањето хипотези е вклучено во 2).

Од горните 3 области, најмалку внимание и се посветува на 1) (спецификацијата). Генерално, сите се "среќни" со прифаќањето на абстрактниот простор на веројатност со тројката $(\Omega, \mathcal{F}, p)$ како да тој расте под грмушка, а потоа статистичкиот модел едноставно следува. Спецификацијата, т.е. изборот на моделот треба секогаш да биде тестиран на валидност, и доколку се покаже несоодветност треба да се направи респецификација, т.е. повторен избор на моделот.

Статистичката анализа би можеле да ја поделиме на следните неколку фази:

1) Спецификација;

2) Оценки;

3) Респецификација;

4) Интервали на доверба и тестирања;

5) Предвидувања.

Спрцификацијата е првата фаза во статистичката анализа и може да се формулира како

поставување на соодветен статистички модел за податоците што се обработуваат.

Ова е суштинска фаза и во случај на набљудувани податоци е релативно "незгодна". Несоодветниот избор на статистичкиот модел понатаму води до повеќе или помалку погрешни резултати во сите следни фази.

Да го разгледаме примерот на гласачките анкети и прашањето

Дали на следните избори ќе гласате за партијата А *или* Б?

Со оглед на тоа да има само 2 можни одговора, може да се дефинира случајна променлива X, со вредности $X(\text{А}) = 0$ и $X(\text{Б}) = 1$ што сугерира користење на Бернулиевиот модел со претпоставките:

Распределба: Бернулиева,

Зависност: Независен,

Хетерогеност: Идентична распределба.

Дали ваквиот модел е соодветен, т.е. валиден? Тестирањето на валидноста на претпоставките на моделите, посебно веројатносните претпоставки ќе бидат дискутирани во следните глави. Тука за даденият пример ќе направиме неформална, интуитивна дискусија. Со оглед на тоа што случајната променлива зема 2 вредности, претпоставката за Бернулиева распределба е секако валидна, но другите две претпоставки за независност и идентична распределба треба да се проценат. Случајниот примерок од популацијата гласачи треба да биде избран така што секој гласач треба да има *еднакви шанси да биде избран*. На пример, ако се работи за избори на државно ниво, примерок земен од еден град, регион, република или држава (како САД) не може да се смета за случаен примерок. Примерок земен со случаен избор на телефонски броеви повторно не е случаен бидејќи некои гласачи може да немаат телефон, а евентуален избор на гласачи од исто семејство веројатно повлекува зависност поради меѓусебните влијанија. Додатно, разумно е да се претпостави дека гласачите од различни области во државата често имаат различни намери (хетерогеност). Ова повлекува дека статистичкиот модел е веројатно недоволно соодветен бидејќи не ја зема предвид систематската информација дека рејтингот на партиите, т.е. намерите на гласачите се различни во различните области (во Македонија тоа би биле изборните единици).

Несоодветноста на едноставниот Бернулиев модел бара респецификација на моделот со надеж дека "поправениот" модел ќе биде посоодветен. Кога се работи за изборите, би требало во моделот да се вклучи хетерогеност во смисла во различните изборни единици да имаме различни рејтинзи (θ се разликува од една до друга изборна единица). Така едноставниот Бернулиев модел може да се замени со проширен нехомоген Бернулиев модел:

1) Статистички генератор, $X_{ik} = \theta_k + u_{ik}$, $k = 1, 2, \ldots, n$, $i \in \mathbb{N}$;

2) Веројатностен модел;

$$\Phi = \{f(x_k; \Theta) = \theta_k^{x_k} (1-\theta_k)^{1-x_k} \mid 0 \le \theta_k \le 1, x_k = 0, 1, \ k = 1, 2, \ldots, n \},$$

3) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0, 1\}^n$, $n = \sum_{k=1}^{N} n_k$,

е независен примерок.

Понатаму, случајниот примерок треба да биде *слоевит* (*stratified*), каде што слоевите се изборните единици. За добивање добри резултати примерокот од секоја изборна единица треба да е доволно голем за секое θ_k да биде оценето адекватно.

Горната дискусија сугерира дека изборот на статистичкиот модел и проверка на неговата валидноста е најважен и често најтежок дел во статистичките анализи. Тоа доаѓа оттаму што таквите одлуки не се шаблонски туку треба да се донесат според бараните анализи и расположивите податоци.

Проблемот на поставување на веројатносниот модел, т.е. на распределбата на примерокот е веројатно најбазичната фаза при статистичката анализа. Овој проблем вообичаено бара математичка дедукција за природата на распределбата на примерокот што често се базира на искуствени елементи. Дали постои систематски начин да се дојде до распределбата на примерокот? Одговорот е, не! Тоа е и основната причина зошто овој проблем е тежок и бара одредено ниво на креативност, искуство и суптилност во размислувањето. Сепак, постојат одреден број искуствени елементи - насоки, а се разбира, од голема полза се граничните теореми (централната гранична теорема) што при одредени општи услови овозможува добивање на приближно добри статистички резултати.

Оценките на параметрите во моделот, креирањето интервали на доверба и тестирањата хипотези ќе бидат предмет на изучување во следните глави.

ЗАДАЧИ

1. Објасни од веројатносен аспект што значи примерок, а што е реализација на примерокот.

2. Објасни го поимот "распределба на примерок".

3. Која е разликата меѓу експерименталните и набљудуваните податоци од аспект на статистичката анализа?

4. Најди ги условните очекувања $E(X \mid \mathcal{D})$ во екстремните случаи кога $\mathcal{D} = \{\Omega, \varnothing\}$ и $\mathcal{D} = \mathcal{F}$.

5. Заедничката распределба на X и Y (пример 9.4) е:

		Y			
		-1	0	1	
X	-1	0.1	0.2	0.1	0.4
	1	0.2	0.1	0.3	0.6
		0.3	0.3	0.4	

   Најди го очекувањето EX, преку условното очекување $E(X \mid Y)$.

6. Објасни ги накусо поимите: спецификација и респецификација.

7. Зошто распределбата на примерокот е суштински концепт во статистичката анализа?

8. Зошто е практично многу тешко да се најде распределбата на примерокот?

10

Описна статистика

Сумарното прикажување на податоците од примерокот е важен чекор во секоја статистичка анализа бидејќи нè фокусира на суштинските карактеристики на податоците и обезбедува информации што помагаат во избор на моделот што ќе се користи за решавање на проблемот. Описната статистика вообичаено се дели на две широки области:

а) пресметки на сумарните нумерички карактеристики на податоците; и

б) претставување на податоците користејќи визуелни техники како што се дијаграмите и графиконите.

Повеќето статистички анализи денеска се прават на компјутер, користејќи некој од многуте програмски пакети за статистички пресметки.

10.1. Нумерички карактеристики на податоци

Тука накусо ќе ги дадеме основните нумерички карактеристики на податоците од примерокот. Тие во главно се однесуваат на мерите на локација, варијабилност, релативни локации, итн.

Да забележиме дека голем дел од овие нумерички карактеристики во малку друга форма веќе ги разгледувавме како бројни карактеристики на случајните променливи. Исто така, од малку друг аспект, дел од нив ќе ги разгледуваме во следната глава како "добри" оценки на непознати параметри.

10.1.1. Мери за локацијата

Основна мера за локацијата на податоците е средната вредност или *просекот*. Ако $x_1, x_2, \ldots, x_n$ се вредности на примерокот, општо познато е дека просекот $\bar{x}$ е

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Покрај просекот, понекогаш се користат уште две други мери за локација на податоците, *медијаната* и *модот*.

Медијана е "средниот" податок, кога податоците се сортирани во растечки редослед. Попрецизно, ако податоците во растечки редослед се $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, тогаш медијаната $\tilde{x}$ е

$$\tilde{x} = \begin{cases} x_{[n/2]+1}, \\ (x_{n/2} + x_{[n/2]+1})/2, \end{cases}$$ каде што $[\ldots]$ е цел дел.

Медијаната е помалку осетлива од просекот на евентуалните екстремно високи или ниски вредности, и во таквите случаи таа понекогаш се преферира како мера на централната тенденција на податоците.

Модот е податокот со најголема фреквенција на појавување. Најголемата фреквенција на појавување може да се појави на две или повеќе различни вредности и тогаш податоците имаат два или повеќе мода. Податоци со 2 мода се нарекуваат бимодални, а со повеќе мода мултимодални.

Процентил (percentil) е вредност (во проценти) што ни дава информација за распределеноста на податоците во интервалот меѓу најмалиот и најголемиот податок. Попрецизно, p-ти процентил е оној податок за кој најмалку p проценти од податоците се исти или помали од него и најмалку $(100 - p)$ проценти од податоците се исти или поголеми од него. Постапката за определување на p-тиот процентил е следната:

1) Сортирај ги податоците во растечки редослед;
2) Пресметај го индексот j, позицијата на p-тиот процентил како $j = (p/100)n$;
3) Ако j не е цел број, заокружи го и тогаш p-тиот процентил е податокот на j-тата позиција.

   Ако j е цел број, p-тиот процентил е просекот од податокот на j-тата и j+1-та позиција.

Во светлото на процентилите, медијаната може да се дефинира како 50-ти процентил.

Некои специфични процентили имаат посебно име. На пример, *квартали* се процентилите: 25-ти (прв квартал), 50-ти (втор квартал) и 75-ти (трет квартал).

ПРИМЕР 10.1 Во следната табела е даден примерок на цени (во долари, во растечки редослед) за закуп на еднособни апартмани во некој град во САД:

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Пресметај ги: просекот, медијаната, модот, како и 90-тиот процентил и 3-тиот квартал.

Решение

Просекот е $\bar{x} = 34356/70 = 490.80$.

Медијаната е $\tilde{x} = (475 + 475)/2 = 475$.

Модот е 450, бидејќи оваа цена се појавува најмногу (7 пати).

За 90-тиот процентил најпрво пресметуваме $j = (90/100)70 = 63$, и сега бидејќи j е цел број 90-тиот процентил е (580 + 590)/2 = 585.

Третиот квартал е 75-ти процентил па имаме $j = (75/100)70 = 52.5$ (се заокружува на 53), па третиот квартал е 525 (вредноста на 53-тата позиција). ■

10.1.2. Мери за варијабилност

Основни мери за варијабилноста на податоците се рангот (опсегот), меѓукварталниот ранг, дисперзијата, стандардната девијација и коефициентот на варијација.

Рангот на податоците е едноставно разликата меѓу најголемиот и најмалиот податок. Ова е, се разбира, наједноставната мера за варијабилноста на податоците.

Меѓукварталниот ранг е разликата меѓу третиот и првиот квартал. Ова во основа е рангот на "средните" 50% од податоците и тој го надминува проблемот на чуствителност на рангот од екстремните вредности.

Дисперзијата на податоците s^2 е просекот на квадратите на разликите меѓу секој податок и просекот

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Зошто ставаме $1/(n-1)$ наместо $1/n$ ќе биде објаснето во следната глава.

Позитивниот квадратен корен на дисперзијата $s = \sqrt{s^2}$ се нарекува *стандардна девијација*. Тој се изразува во истите единици како и податоците и затоа е подобро споредлив со просекот, како и со самите податоци.

Коефициент на варијација v на податоците дава информација за тоа колку е голема стандардната девијација на податоците во однос на просекот. Тој се пресметува (во проценти) со

$$v = \frac{s}{\bar{x}}100 .$$

ПРИМЕР 10.2 За примерокот на цени за закуп на еднособни апартмани во некој град во САД од примерот 10.1 пресметај ги мерите на варијабилност.

Решение

Рангот на податоците е 615 – 425 = 190.

Меѓуквартални ранг е 3-ти квартал – 1-ви квартал = 525 – 445 = 80.

Дисперзијата е $s^2 = 2996.16$.

Стандардната девијација е $s = 54.74$.

Коефициент на варијација е (54.74/490.80)100 = 11.15. ■

10.1.3. Мери за релативна локација

Како што самото име кажува, мерите за релативната локација даваат информации за локација на податоците релативно, во однос на некоја друга мера како просекот или дисперзијата.

Стандардизираната вредност (z-скор) мери колку стандардни девијации секој податок е далеку од просекот со

$$z_j = \frac{x_j - \bar{x}}{s}.$$

Јасно е дека секој податок помал од просекот има негативен z-скор и обратно, секој податок поголем од просекот има позитивен z-скор.

Теоремата на Чебишев тврди дека најмалку $(1 - 1/k^2)$ податоци од кој било примерок паѓаат во k стандардни девијации околу просекот, (k

> 1). Така на пример, најмалку 75% од податоците се во околина на k = 2 стандардни девијации на просекот, 89% од податоците се во околина на k = 3 стандардни девијации на просекот и 94% од податоците се во околина на k = 4 стандардни девијации на просекот. Овие проценти се поголеми ако распределбата на податоците е нормална, и соодветните вредности приближно се најмалку 68% за 1 стандардна девијација, 95% за 2 стандардни девијации, 99.7% за 3 стандардни девијации и практично 100% за 4 стандардни девијации.

ПРИМЕР 10.3 За примерокот на цени за закуп на еднособни апартмани во некој град во САД од примерот 10.1 пресметај ги z-скоровите за првиот и последниот податок, како и бројот на податоци што паѓаат во 1, 2 и 3 стандардни девијации околу просекот.

Решение

z-скорот за првиот податок е $z_1 = (425 - 490.80)/54.74 = -1.2$, а за последниот $z_{70} = (615 - 490.80)/54.74 = 2.27$.

$\bar{x} \pm k \cdot s$	Интервал	% во интервалот
Во $\bar{x} \pm s$	$490.80 \pm 54.74 = [436.06, 545.54]$	48/70 = 68.57%
Во $\bar{x} \pm 2s$	$490.80 \pm 2 \cdot 54.74 = [381.32, 600.28]$	68/70 = 97.14%
Во $\bar{x} \pm 3s$	$490.80 \pm 3 \cdot 54.74 = [326.58, 655.02]$	70/70 = 100%

Забележи дека во теоремата за бројот на податоци во околините на просекот стои зборот "најмалку". Во конкретните примери процентите се секогаш поголеми. ■

Релативно често се случува податоците од примерокот да не се дадени експлицитно, туку само групно по класи каде што во секоја класа j паѓаат f_j податоци (фреквенција на класата). Тогаш сме приморани нумеричките карактеристики да ги пресметуваме на друг начин. На пример, просекот логично би бил

$$\bar{x} = \frac{\sum_{j=1}^{n} f_j M_j}{n}$$, каде што M_j е средината на класата j.

Дисперзијата би се пресметувала соодветно со

$$s^2 = \frac{\sum_{j=1}^{n} f_j (M_j - \bar{x})^2}{n-1}.$$

ПРИМЕР 10.4 Да претпоставиме дека податоците од примерокот на цени за закуп на едноспобни апартмани од примерот 10.1 се дадени групно во табелата:

Класа ($)	Фреквенција	Класа ($)	Фреквенција
420 - 439	8	520 - 539	4
440 - 459	17	540 - 559	2
460 - 479	12	560 - 579	4
480 - 499	8	580 - 599	2
500 - 519	7	600 - 619	6

Пресметај го просекот, дисперзијата и стандардната девијација.

Решение

Просекот е $\bar{x} = (8 \cdot 429.5 + 17 \cdot 449.5 + 12 \cdot 469.5 + \ldots)/70 = 34525/70 = 493.21$. Спореди го ова со вистинскиот просек на примерокот 490.80.

За дисперзијата повторно со обична пресметка добиваме $s^2 = (8(429.5 - 493.21)^2 + 17(449.5 - 493.21)^2 + 12(469.5 - 493.21)^2 + \ldots)/69 = 3017.8$. Стандардната девијација е $S = 54.94$. Спореди го ова со вистинскиата стандардна девијација на примерокот 54.74. ■

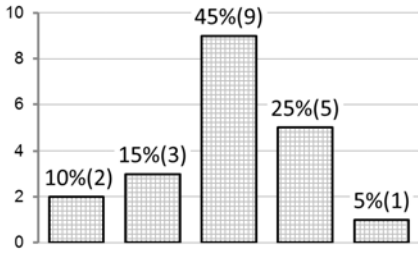
10.2. Визуелно претставување на податоци

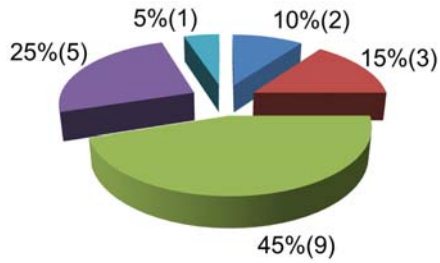
За визуелно претставување на податоците вообичаено се користат дијаграми со барови или пити во 2 или 3 димензии кои обично ги изразуваат процентуалните (или други) односи извлечени од податоците.

ПРИМЕР 10.5 Гостите на еден хотел имале прилика да се изјаснат за квалитетот на услугата преку понудени 5 можности: слаба (1), подпросечна (2), просечна (3), надпросечна (4) и одлична (5). Изјаснувањето на примерок од 20 гости било: 2, 3, 3, 4, 3, 4, 4, 3, 4, 3, 2, 1, 5, 3, 4, 3, 3, 2, 1, 3, 4. Состави дујаграм со барови и пита за уценките на услугата во хотелот.

Решение

Баровите ги даваме во 2Д, а питата во 3Д:





Како што знаеме постојат огромен број варијации на ваквите дијаграми како во 2Д, така и во 3Д. ■

Многу помалку во секојдневната употреба се појавуваат таканаречените точкасти и стебло-лисја дијаграми. Од друга страна, тие често се користат во статистиката за добивање на глобална слика за податоците.

Точкастиот дијаграм дава графичка сумарна слика на податоците во случаите кога нивниот број е разумно мал. Во овој приказ секој податок се претставува со точка на соодветна локација на хоризонтална мерна оска. Ако некоја вредност се повторува повеќе пати, за секое појавување на вредноста се црта точка вертикално на истата локација. Сликата за податоците што се добива од точкастиот дијаграм опфаќа информации за локациите, раштрканоста, екстремите и празнините.

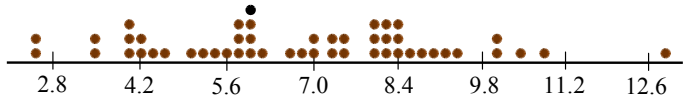
ПРИМЕР 10.6 Во следната табела се дадени податоци за процент од даноците што оди на високото образование за поедини државо во САД (по азбучен редослед на државите):

10.8	6.9	8.0	8.8	7.3	3.6	4.1	6.0	4.4	8.3	8.1	8.0	5.9	5.9	7.6	8.9	8.5
8.1	4.2	5.7	4.0	6.7	5.8	9.9	5.6	5.8	9.3	6.2	2.5	4.5	12.8	3.5	10.0	9.1
5.0	8.1	5.3	3.9	4.0	8.0	7.4	7.5	8.4	8.3	2.6	5.1	6.0	7.0	6.5	10.3	

Состави точкаст дијаграм за овие податоци.

Решение

Точкастиот дијаграм би можел да изгледа вака



Како што се гледа, процентот од даноците во државите е во главно меѓу 4 и 9%. Екстремите се под 2.8% (две држави) и над 11 (една држава). ■

Точкаст дијаграм може да се користи и во 3 димензии, но така не се добива ништо во прегледноста како некој би очекувал.

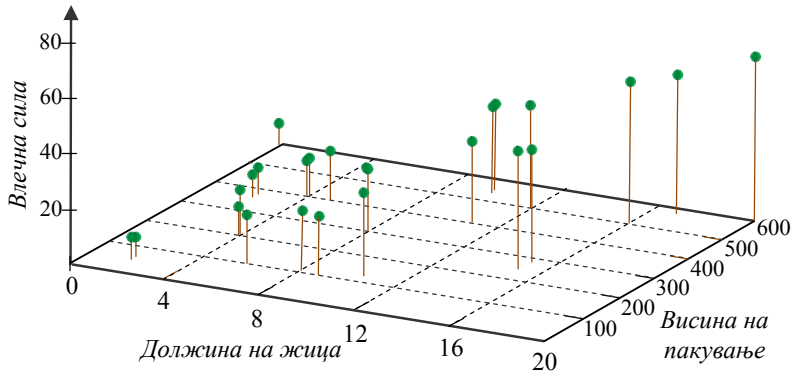
ПРИМЕР 10.7 Во следната табела се дадени податоци за 3 променливи собрани од студија за полупроводници што се изработуваат во некоја фабрика. Секој полупроводник од земен примерок од 25 полупроводници е поврзан на рамка со жица. Трите променливи во табелата се: влечна сила (силата потребна да се скрши обвивката), должина на жицата и висината на пакувањето на полупроводникот.

Влечна сила	Должина на жица	Висина на пакување	24.35	9	100			
9.95	2	50	27.50	8	300	10.30	1	585
24.45	8	110	17.08	4	412	34.93	10	540
31.75	11	120	37.00	11	400	46.59	15	250
35.00	10	550	41.95	12	500	44.88	15	290
25.02	8	295	11.66	2	360	54.12	16	510
16.86	4	200	21.65	4	205	56.63	17	590
14.38	2	375	17.89	4	400	22.13	6	100
9.60	2	52	69.00	20	600	21.15	5	400

Состави точкаст дијаграм за влечната сила како функција од должината на жицата и висината на пакувањето. Објасни што сугерира добиениот дијаграм?

Решение

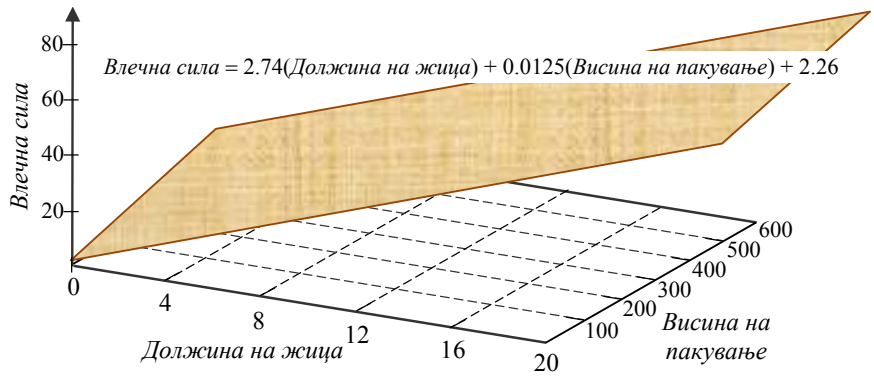
Од вака дадените податоци не може да се забележат или брзо одговорат битни прашања, на пример, "Кој е процентот на примероците со висина на пакувањето под 120?". Поради умерено големиот број примероци, точкастиот дијаграм е недоволно ефективен.



Од дијаграмот се гледа дека влечната сила се зголемува со зголемување на должината на жицата и висината на пакувањето. Уште повеќе, оваа зависност изгледа доста линеарна, што овозможува по методот на најмали квадрати да се најде рамнина што релативно добро ги опишува податоците. Моделот за ова е од облик

$$\textit{Влечна сила} = \beta_0 + \beta_1(\textit{Должина на жица}) + \beta_2(\textit{Висина на пакување}) + \varepsilon$$

каде што параметрите β_0, β_1 и β_2 се определуваат од податоците (види пример 15.6). Рамнината се бара така што сумата на квадратите на растојанијата од неа до точките да биде најмала. Ваквата рамнина што интуитивно поминува "централно" низ податоците е дадена на сликата подолу.



Вака добиениот емпириски модел понатаму може да се користи за предвидување на вредностите на влечната сила за кои било вредности на должината на жицата и висината на пакувањето. ■

Точкастиот дијаграм е корисен во случаи на мал примерок, да речеме неколку десетини податоци. Кога бројот на податоци е умерено голем, покорисни се некои други графички прикази, како што е дијаграмот стебло-лисја.

Нека x_1, x_2, …, x_n се податоци од примерокот. Конструкцијата на дијаграмот се прави со следните чекори:

1) Подели го секој од податоците x_i на два дела, *стебло* што се состои од еден или повеќе водечки знаци или цифри од податокот, и *лист* што се состои од останатоте знаци или цифри;
2) Прикажи ги вредностите на стеблата на податоците во вертикална колона;
3) Прикажи ги вредностите на лисјата во хоризонтални редици покрај своите стебла.

ПРИМЕР 10.8 Во следната табела се дадени цврстините на компресија во фунти по квадратен инч (psi) на 80 примероци од нова алуминиум-литиумска легура што би требало да се користи во авионската индустрија:

105	221	183	186	121	181	180	143	97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110	163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123	134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169	199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135	196	201	200	176	150	170	118	149

Состави дијаграм стебло-лисја за овие податоци.

Решение

Од вака дадените податоци не може да се забележат или брзо одговорат битни прашања, на пример, "Кој е процентот на примероците со цврстина под 120 psi?". Поради умерено големиот број примероци, точкастиот дијаграм е недоволно ефективен.

За стебло ќе ги земеме првите две цифри од броевите (една, ако бројот е двоцифрен). Дијаграмот е даден во следната табела (даден е во 3 колони):

Стебло	Лисја	Фрекв.						
7	6	1	13	4 1 3 5 3 5	6	19	9 6 0 9 3 4	6
8	7	1	14	2 9 5 8 3 1 6 9	8	20	7 1 0 8	4
9	7	1	15	4 7 1 3 4 0 8 8 6 8 0 8	12	21	8	1
10	5 1	2	16	3 0 7 3 0 5 0 8 7 9	10	22	1 8 9	3
11	5 8 0	3	17	8 5 4 4 1 6 2 1 0 6	10	23	7	1
12	1 0 3	3	18	0 3 6 1 4 1 0	7	24	5	1

Сега процентот на примероците со цврстина под 120 psi може веднаш да го најдеме собирајќи ги фреквенциите тргнувајќи од стеблото 12 па надолу, 3 + 2 + 1 + 1 + 1 = 8, што дава процент 8/80 = 10.00%. ■

Бројот на стебла и лисја во дијаграмот треба да биде "избалансиран". На пример, ако бројот на стебла е мал, секогаш може да се додадат нови стебла. Така, следниот дијаграм (количество супстанца добиено од хемиски процес) има малку стебла (во однос на лисјата),

Стебло	Лисја
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

и со воведување "нови" стебла може да се "избалансира" во

Стебло	Лисја	Стебло	Лисја
6д	1 3 4	8д	1 3 4 4
6г	5 5 6	8г	7 8 8
7д	0 1 1 3	9д	2 3
7г	5 7 8 8 9	9г	5

каде што цифрите на лисјата се поделени во две групи, 0 до 4 и 5 до 9 (д - долна, г - горна цифра). На соодветен начин може да се направи намалување на бројот на стеблата (нивно групирање) ако нивниот број е сразмерно голем во однос на бројот на лисјата.

Постојат многу варијации на ваквите дијаграми. На пример кога дијаграмот се базира на процентили обично се добиваат цртежи со правоаголници што даваат информации за кварталите (box plots). Некои дијаграми се специфични за типот на податоците со кои се работи како што е претставувањето на временските серии. Софтверските пакети од статистиката, вообичаено обезбедуваат визуелно претставување на податоците со различни техники слични на точкастите или стебло-лисја дијаграми.

10.3. Распределба на фреквенции и хистограм

Распределбата на фреквенции е компактна сумарна информација за податоците што грубо ги опишува или густината или функцијата на распределба. За конструкција на распределбата на фреквенции, најпрво треба рангот во кој се наоѓаат податоците да се подели на интервали – класи. Ако е тоа возможно, класите треба да се со иста широчина за визуелната информација за фреквенциите да биде поизразена. Потоа само останува да се избројат бројот на податоци што паѓаат во секоја класа и тоа да се прикаже графички, обично со барови.

Веројатно најважното прашање при дизајн на хистограм е одлуката за бројот на класи што би требало да се користи за поделба на рангот на податоците. Се разбира, бројот на класи треба да зависи од рангот во кој се наоѓаат податоците како и од бројот на податоците. Ако бројот на класи е преголем или премал, бројот на податоци во секоја класа ќе биде мал (може некаде и 0) или голем и тогаш хистограмот ќе биде значително "рамен" што нема да ја одразува скриената закономерност во податоците. Генерално, бројот на класи меѓу 5 и 20 е задоволителен за помал број податоци. Во литературата се предложени многу формули за определување на приближен број на класи во зависност од бројот на податоците (n), на пример: $2\sqrt[3]{n}$, $\sqrt{n}$, $\log_2 n + 1$ (за $n \geq 30$), итн.

Постапката за добивање хистограм може да се сумира во следните чекори:

1) Најди го рангот (*rang*) на податоците, како разлика на најголемиот (*max*) и најмалиот (*min*) податок (*rang* = *max* – *min*);

2) Подели го рангот на класи k_1, k_2, k_3, … според бројот на податоци, и тоа:
   < 50 податоци $\Rightarrow$ 5 до 7 класи,
   50 до 99 податоци $\Rightarrow$ 6 до 10 класи,
   100 до 250 податоци $\Rightarrow$ 7 до 12 класи,
   > 250 податоци $\Rightarrow$ примени некоја од горенаведените формули.

3) Најди ја фреквенцијата на појавување на податоците $f_1, f_2, f_3, \ldots$ во секоја од класите k_1, k_2, k_3, …, а потоа најди ги релативните фреквенции $f_1/n, f_2/n, f_3/n, \ldots$;

4) Нацртај го хистограмот со барови со широчина на класата, и височина според релативните фреквенции (или фреквенции).

   За разлика од графиконите со барови, хистограмот вообичаено нема растојанија меѓу соседните класи (барови).

Графикот добиен со поврзување на точките на скок на фреквенциите со отсечка се нарекува оџаиве (ogive). Додека вдолж x-оската се класите на податоци, на y-оската се стават кумулативните фреквенции или кумулативните релативни фреквенции. Како што фреквенцијата на податоците грубо ја прикажува густината на распределба, кумулативните фреквенции грубо ја прикажуваат финкцијата на распределба.

ПРИМЕР 10.9 Менаџерот на автомобилски сервис сака да добие идеја за распределбата на трошокот за деловите за подесување на работата на моторите. Земен е примерок од 50 сметки што муштериите ги платиле за таа намена. Вредностите заокружени до поблиската цела вредност во долари биле: 91, 78, 93, 57, 75, 52, 99, 80, 97, 62, 71, 69, 72, 89, 66, 75, 79, 75, 72, 76, 104, 74, 62, 68, 97, 105, 77, 65, 80, 109, 85, 97, 88, 68, 83, 68, 71, 69, 67, 74, 62, 82, 98, 101, 79, 105, 79, 69, 62 и 73. Состави хистограм за дадените трошоци.

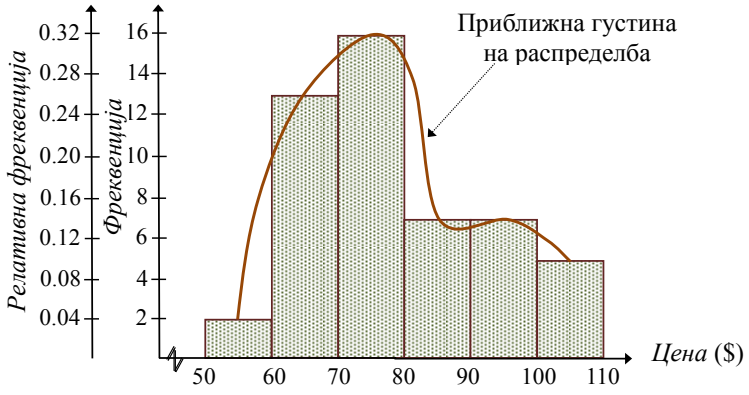
Решение

Рангот на податоците е 109 – 52 = 57.

За 50 податоци, може да земаме 6 класи, па ширината на секоја класа е 57/6 = 9.5 што ќе го заокружиме на 10. Фреквенциите по класи се:

Класи (Цена во $)	Фреквенција	Релативна фреквенција
50 – 59	2	0.04
60 – 69	13	0.26
70 – 79	16	0.32
80 – 89	7	0.14
90 – 99	7	0.14
100 – 109	5	0.10
Вкупно	50	1.00

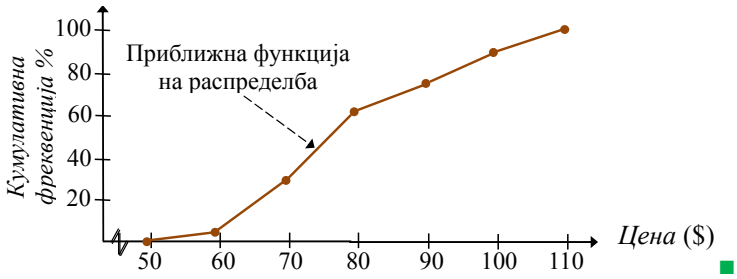
Хистограмот, преку фреквенции е даден на следната слика



Кумулативните фреквенции и релативни фреквенции се дадени во следната табела:

Класи (Цена во $)	Фреквенција	Релативна фреквенција
≤ 59	2	0.04
≤ 69	15	0.30
≤ 79	31	0.62
≤ 89	38	0.76
≤ 99	45	0.90
≤ 109	50	1.00

Графикот на кумулативните фреквенции изгледа вака



Во многу ситуации еднаквата ширина на класите не е добар избор ако податоците се концетрирани во едни региони, а во други се многу раштркани. Во такви случаи, по определувањето на фреквенциите и релативните фреквенции во секоја класа, висината на секој правоаголник се пресметува со

$$\text{висина на правоаголник} = \frac{\text{релативна фреквенција на класата}}{\text{ширина на класата}}.$$

Оваа висина на правоаголникот обично се нарекува *густина*. Се разбира овој концепт на густина функционира и за правоаголници со иста ширина (специјален случај). Имено, користењето густина се оправдува со фактот што вкупната плоштина на правоаголниците е 1, што е во согласност со густината на распределба. Кога ширината на класите е различна, не користењето на густината води до барови со дисторзирани плоштини. Хистограмот со густини ја има добрата особина што плоштината на секој правоаголник е еднаква со релативната фреквенција на соодветната класа. Тоа се добива од

Релативна фреквенција = (ширина на класа)(густина на класа) =
= (ширина на правоаголник)(висина на правоаголник) = плоштина.

ПРИМЕР 10.10 Корозијата на челикот е сериозен проблем при негово користење во структури изложени на атмосферски влијанија. Од тие причини, се испитуваат различни композитни материјали како алтернатива. Следните 48 податоци ја даваат цврстината на еден композитен материјал:

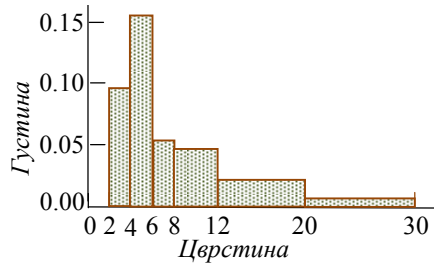
11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Подели ги податоците во класи. Определи ги фреквенциите и релативните фреквенции на цврстините по класи и состави хистограм.

Решение

Рангот на податоците е 25.5 – 3.4 = 22.1, но поради нерамномерното групирање на податоците по потенцијалните класи и нивната раштрканост кај поголемите вредности (25.6, 20.6, 17.1) треба да се користат класи со различна ширина. Тоа е направено во следната табела:

Класа	Фрекв.	Рел. Фрекв.	Густина
2 - 4	9	0.1875	0.094
4 - 6	15	0.3125	0.156
6 - 8	5	0.1042	0.052
8 - 12	9	0.1875	0.047
12 - 20	8	0.1667	0.021
20 - 30	2	0.0417	0.004



Од хистограмот се гледа големата несиметричност на распределбата и големото издолжување со одење кон повисоките вредности. ■

Хистограмите се користат во многу апликации во случаи кога се потребни информации за непознати распределби. Во многу ситуации податоците природно се поделени на класи, а главниот проблем - недоволната големина на примерокот за добивање на добра апроксимација на распределбата, кога тоа е можно, се решава со додавање на симулирани податоци.

10.4. Веројатносни дијаграми

Како да се оцени дали одредена распределба е соодветна (е соодветен модел) за опишување на податоците? Еден начин е да се состави хистограм и од него да се направи обид за согледување дали претпоставената распределба е соодветна или пак хистограмот "сугерира" некоја

друга како посоодветна. Сепак, хистограмите не се доволно доверливи индикатори за обликот на непознатата распределба освен ако примерокот не е навистина голем. Едноставен графички начин за приближна проверка дали податоците се "согласуваат" со претпоставента распределба се веројатносните дијаграми.

Целата постапка во основа се базира на процентили (percentiles) на примерокот. Од аспект на случајните променливи, $(100 \cdot p)$-иот процентил на распределбата со функција $F(\cdot)$ е бројот $b(p)$ таков што $F(b(p)) = p$. Со други зброви, $b(p)$ е број на x-оската таков што плоштината под густината на распределба налево од него е точно p. Така, за 0.4 процентил важи $F(b(0.4)) = 0.4$ или за 0.9 процентил $F(b(0.9)) = 0.9$. На пример, за стандардната нормална распределба имаме дека 0.4 процентил е –0.2533, а 0.9 процентил е 1.2816. Во случај кога располагаме само со примерок, нема подобар начин од тоа процентилот $100 \cdot p$ да се дефинира како број на примероци чиишто вредности (по големина) паѓаат во тој процент. На пример, 40-ти процентил е бројот на податоци што се во групата од оние со 40% најмали вредности. Нека податоците од примерокот со големина n се подредени во растечки редослед. Тогаш j-тиот најмал податок е $100(j - 0.5)/n$-ти процентил на примерокот.

Сега, ако примерокот е земен од претпоставената распределба, процентилите на примерокот (подредените вредности на примерокот) би требало да бидат разумно блиски до процентилите на претпоставената распределба на популацијата. Тоа значи дека за $j = 1, 2, \ldots, n$, би требало да има разумно согласување меѓу j-тата најмала вредност на примерокот и $100(j - 0.5)/n$ процентил на претпоставената распределба. Ако ги разгледаме паровите (процентил на примерокот, процентил на популацијата), т.е. истото

> (j-та најмала вредност од примерокот, $100(j - 0.5)/n$ процентил на популацијата) за $j = 1, 2, \ldots, n$,

тогаш кога претпоставената распределба одговара на примерокот и двете вредности во парот треба да се приближно еднакви. Нацртани како точки во координатен систем, тие треба да бидат блиски до симетралата на првиот квадрант (правата со наклон од $\pi/4$). Позначајна девијација на точките од оваа права значи оправдан сомнеж во коректноста на претпоставената распределба на популацијата.

Вообичаено е да се испитува дали претпоставената распределба е од некој тип без да се води сметка за вредноста на параметрите. На пример, нема многу смисла да се проверува дали претпоставената распределба е експоненцијална со $\lambda = 0.1$ или стандардна нормална ($\mu = 0$

и $\sigma = 1$), туку се испитува дали таа генерално е експоненцијална или нормална. Оценките на непознатите параметри како λ, μ или σ, ќе ги разгледуваме во следната глава.

Поради раширеноста и важноста, веројатносните дијаграми најчесто се користат за груба проценка дали податоците се во согласност со нормалната распределба. Тука клучна улога игра фактот дека односот на процентилите на стандардната и општата нормална распределба е едноставно

проценил на $Z(\mu, \sigma) = \mu + \sigma \cdot$ *проценил на* $Z(0,1)$.

Тоа значи дека кога податоците се од општа нормална распределба, наместо да бидат блиски до симетралата на првиот квадрант, тие треба да се блиски до права линија со наклон определен со σ и поткрената за вредноста на μ.

Дефиниција 10.1 Цртежот на n-те точки

(j-та најмала вредност на примерокот, $100(j - 0.5)/n$-ти процентил)

во дводимензионален координатен систем се нарекува нормален веројатносен дијаграм.

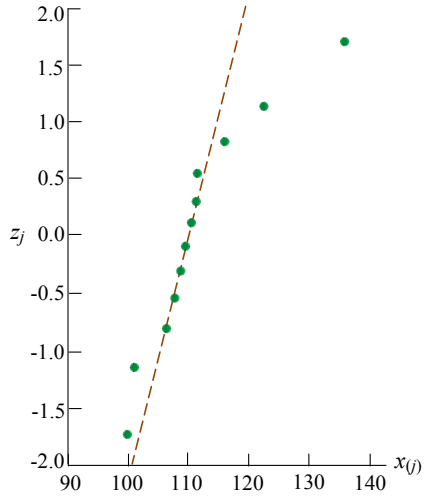
Ако точките од нормалниот веројатносен дијаграм се приближно вдолж права линија, тогаш има индиции дека распределбата на популацијата е приближно нормална. Ако отстапувањето од права линија е значително, може да сметаме дека податоците се од некоја друга распределба. Оценката дали податоците се сложуваат или отстапуваат од права линија е субјективно. За степенот на отстапување на податоците од правата линија постои таканаречен Андерсон-Дарлинг-ов тест (Anderson-Darling – "distance" test) што се сведува на тестирање хипотеза за согласност на податоците со нормалната распределба. Сепак во општ случај, воведувањето на некакви математички концепти за аналитички да се проверува степенот на сложување на податоците со правата линија е доста непоуздано и несоодветно поради непрецизноста на целата постапка.

ПРИМЕР 10.11 Испитувано е дејството на тоа како додатоци во исхраната со калциум влијаат на крвниот притисок. Како и во други медицински испитувања, испитаниците се поделени на група што зема калциум и плацебо група. Резултатите на првата група се: 108, 110, 123, 129, 112, 111, 107, 112, 136, 102, 116, 100; додека кај плацебо групата измерени се вредностите: 123, 109, 112, 102, 98, 114, 119, 112, 110, 117, 130, 112. Познато е дека распределбата на вредностите на притисокот е приближно нормална (плацебо групата). Дали таа останува приближно нормална и по земање на калциум?

Решение

Резултатите се дадени во следната табела и придруженият график:

j	$x_{(j)}$	$(j - 0.5)/12$	z_j
1	100	0.042	–1.728
2	102	0.125	–1.150
3	107	0.208	–0.813
4	108	0.292	–0.548
5	109	0.375	–0.319
6	110	0.458	-0.105
7	111	0.542	0.105
8	112	0.625	0.319
9	112	0.708	0.548
10	116	0.792	0.813
11	123	0.875	1.150
10	136	0.958	1.728



Субјективен впечаток е дека овие податоци се приближно вдолж права линија (освен последните), што значи може да сметаме (со резерва) дека вредностите на притисокот кај лицата суплементирани со калциум има приближно нормална распределба. ■

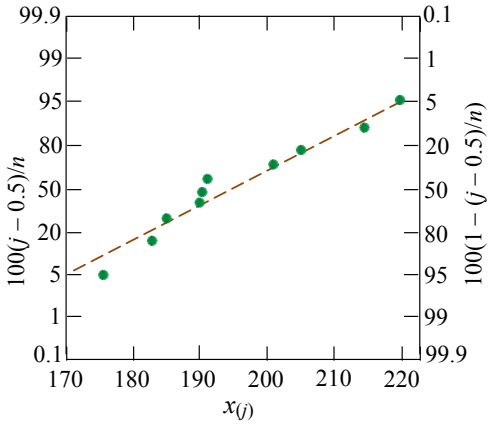
Алтернативен (поедноставен) начин за конструкција на нормален веројатносен дијаграм е z-процентилите на вертикалната оска да се заменат со нелинеарно претставени веројатности $(j - 0.5)/n$. Скалингот на оската се прави таков што точките повторно паѓаат на права линија кога распределбата е нормална. На пример, често користени вредности за градација на вертикалната оска се: 0.001, 0.01, 0.05, 0.2, 0.5, 0.8. 0.95, 0.99 и 0.999. Значи постапката би одела на сосема одентичен начин. Најпрво податоците од примерокот $x_1, x_2, \ldots, x_n$ ги сортираме во растечки редослед, добивајќи $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, каде што $x_{(1)}$ е најмалиот, а $x_{(n)}$ е најголемиот податок. Потоа се црта секој податок $x_{(j)}$ со својата фреквенција-веројатност $(j - 0.5)/n$ (може и со проценти $100(j - 0.5)/n$), т.е. се цртаат точките $(x_{(j)}, (j - 0.5)/n)$, $j = 1, 2, ..., n$. Ако претпоставената нормална распределба адекватно ги опишува податоците, точките ќе бидат приближно вдолж права линија.

ПРИМЕР 10.12 Испитано е времетраењето (во минути) на 10 батерии за лаптоп компјутер и добиени следните вредности: 176, 191, 214, 220, 205, 192, 201, 190, 183 и 185. Има индиции дека времетраењето на батериите е со приближно нормална распределба. Провери го тоа користејќи веројатносен дијаграм.

Решение

Резултатите се дадени во следната табела и придруженият график:

j	$x_{(j)}$	$(j-0.5)/10$	z_j
1	176	0.05	–1.64
2	183	0.15	–1.04
3	185	0.25	–0.67
4	190	0.35	–0.39
5	191	0.45	–0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64



Очигледно податоците приближно паѓаат на права линија, па оттука заклучуваме дека времетраењето на батериите има приближно нормална распределба. ■

Не-нормална распределби на популацијата често паѓа во следните три категории:

а) Таа е симетрична, но краевите се "пострмни" отколку кај нормалната распределба;

б) Таа е симетрична, но краевите се "помалку стрмни" отколку кај нормалната распределба;

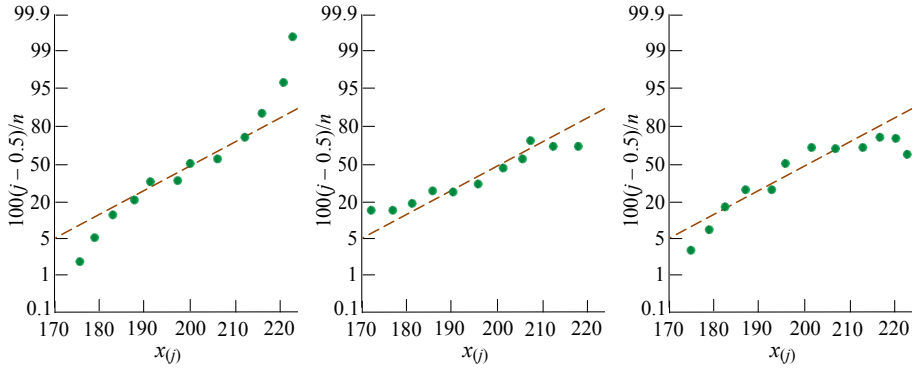
в) Таа е "искривена" и нема симетричен облик.

На пример, рамномерната распределба е со "пострмни" краеви бидејќи таа паѓа на 0 надвор од конечен интервал. Од друга страна, распределбата $f(x) = 1/(\pi(1+x^2))$ е со "помалку стрмни" краеви во однос на $e^{-x^2/2}$.

Кога точките кај нормалниот веројатносен дијаграм не се приближно вдолж права линија, тоа често значи дека распределбата на популацијата е во една од трите категории.

Ако краевите на распределбата на популацијата се "пострмни" (случај а)), тогаш најмалите и најголемите податоци ќе бидат помалку "екстремни" отколку кај нормалната распределба. Визуелно, тоа значи дека податоците од средината на дијаграмот ќе "следат" некоја права линија, но на левиот крај ќе имат тенденција да бидат под линијата (по-

датоците < процентилот на z), додека на десниот крај тенденцијата е да бидат над линијата (податоците > процентилот на z). Ова резултира во дијаграм со точки во облик на C. Ако краевите на распределбата на популацијата се "помалку стрмни" отколку кај нормалната распределба (случај б)), тогаш најмалите и најголемите податоци ќе бидат "поекстремни" отколку кај нормалната распределба, па добиваме дијаграм но сега обратен, во облик на S. Во случај на "искривена" распределба, податоците често пати имаат (конвексно или конкавно) заоблен облик. Овие три случаи последователно се прикажани на сл. 10.1.



Слика 10.1 *Три примери на веројатносни дијаграми што индицираат ненормална распределба*

Дури и во ситуации кога распределбата на популацијата е точно нормална, точките на веројатносниот дијаграм нема да лежат точно на права линија. Од таа причина е потребно одредено ниво на искуство и субјективно знаење точно да се процени дијаграмот. Генерално, ако големината на примерокот е $n < 30$, тој може да покаже значителни девијации од линеарност иако распределбата на популацијата е нормална. Во таков счучај, само сериозни отстапки од линеарност би требало да се интерпретираат како силна индикација за не-нормалност. Со зголемување на n, линеарноста станува "поевидентна" и интерпретацијата на дијаграмот поедноставна и посигурна. Генерално зборувајќи, мал примерок од нормална распределба има поголеми шанси да покаже "нелинеарно" однесување од голем примерок.

Кога се работи за веројатносни дијаграми за проверка дали податоците се согласуваат со некоја друга (не-нормална) распределба, проблемот на нивното составување не е така едноставен. За добивање ефикасна постапка потребен е индивидуален пристап за секоја распределба (види задача 11).

ЗАДАЧИ

1. Вредноста *pH* на некој раствор е мерена 8 пати со ист инструмент при што се добиени следните податоци: 7.15, 7.20, 7.18, 7.19, 7.21, 7.20, 7.16 и 7.18.

   а) Пресметај го просекот, дисперзијата и стандардната девијација;

   б) Пресметај го 0.45-тиот процентил и третиот квартал.

2. Следните податоци се измерени температури (°F) на одредена компонента во авионски мотор: 84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.

   а) Пресметај го просекот, стандардната девијација и првиот квартал;

   б) Состави точкаст дијаграм за податоците;

   в) Отстрани ја најмалата вредност и пресметај ги одново вредностите од а).

3. Група од ентузијасти по вино го тестирале "pinot noir" од Орегон, САД и давале оценка од 0 до 100 бода. Резултатите се следните:

   94 90 92 91 91 86 89 91 91 90 90 93 87 90 91 92 89 86 89 90
   88 95 91 88 89 92 87 89 95 92 85 91 85 89 88 84 85 90 90 83

   а) Состави дијаграм стебло-лисја и дај коментар;

   б) Пресметај го просекот, стандардната девијација и медијаната;

   в) Ако вино со оценка најмалку 90 е изузетно квалитетно, која е пропорцијата од групата што го смета виното "pinot noir" за изузетно?

4. Испитувани се механичките особини на метал што се користи во воздухопловството, при што за 153 примероци добиени се јачина на растегнување (ksi) дадени во следната табела:

   122.2 124.2 124.3 125.6 126.3 126.5 126.5 127.2 127.3 127.5 127.9 128.6 128.8 129.0
   129.2 129.4 129.6 130.2 130.4 130.8 131.3 131.4 131.4 131.5 131.6 131.6 131.8 131.8
   132.3 132.4 132.4 132.5 132.5 132.5 132.5 132.6 132.7 132.9 133.0 133.1 133.1 133.1
   133.1 133.2 133.2 133.3 133.3 133.5 133.5 133.5 133.8 133.9 134.0 134.0 134.0
   134.0 134.1 134.2 134.3 134.4 134.4 134.6 134.7 134.7 134.7 134.8 134.8 134.8 134.9
   134.9 135.2 135.2 135.2 135.3 135.3 135.4 135.5 135.5 135.6 135.6 135.7 135.8 135.8
   135.8 135.8 135.8 135.9 135.9 135.9 135.9 136.0 136.0 136.1 136.2 136.2 136.3 136.4
   136.4 136.6 136.8 136.9 136.9 137.0 137.1 137.2 137.6 137.6 137.8 137.8 137.8 137.9
   137.9 138.2 138.2 138.3 138.3 138.4 138.4 138.4 138.5 138.5 138.6 138.7 138.7 139.0
   139.1 139.5 139.6 139.8 139.8 140.0 140.0 140.7 140.7 140.9 140.9 141.2 141.4 141.5
   141.6 142.9 143.4 143.5 143.6 143.8 143.8 143.9 144.1 144.5 144.5 147.7 147.7

   а) Состави стебло-лисја дијаграм со отстранување на децималната цифра (не заокружување) и повторувајќи го секое стебло 5 пати (за листови 0 и

1, па за листови 2 и 3, итн. за листови 8 и 9). Зошто лесно се наоѓа секоја вредност за јачината на растегнување?

б) Состави хистограм со еднаква широчина на класи. Првата класа да тргне од 122 и оди до 124, итн. Каква е приближната густина на разпределба?

5. Дали времетраењето на американските филмови се разликува од времетраењето на француските? За оговор на ова прашање од двете кинематографии случајно се избрани по 25 филма чиишто времетраења се:

   Американски: 94, 90, 95, 93, 128, 95, 125, 91, 104, 116, 162, 102, 90, 110, 92, 113, 116, 90, 97, 103, 95, 120, 109, 91, 138;
   Француски: 123, 116, 90, 158, 122, 119, 125, 90, 96, 94, 137, 102, 105, 106, 95, 125, 122, 103, 96, 111, 81, 113, 128, 93, 92.

   Состави споредбен стебло-лисја дијаграм со зеднички стебла (лево стави ги лисјата од американските, а десно од француските) и коментирај ги односите на времетраењата од добиениот дијаграм.

6. Медицински термометри од одреден тип се испорачуваат во пакувања од 50. Земен е примерок од 60 пакувања при што во секое пакување бројот на термометри што не ги задоволува спецификациите бил:

2	1	2	4	0	1	3	2	0	5	3	3	1	3	2	4	7	0	2	3
0	4	2	1	3	1	1	3	4	1	1	6	0	3	3	3	6	1	2	3
2	3	2	2	8	4	5	1	3	1	5	0	2	3	2	1	0	6	4	2

   а) Определи ги фреквенциите и релативните фреквенции на бројот на термометри што не ги задоволува спецификациите по пакување;

   б) Која пропорција на пакувања имаат најмногу 5, помалку од 5, најмалку 5 термометри надвор од спецификациите?

   в) Нацртај хистограм и коментирај некои негови карактеристики.

7. Следните податоци (во растечки редослед) се примерок од животниот век на микро-дупчалка даден со број на дупки пред откажувањето кога се дупчи одреден композитен материјал:

   11, 14, 20, 23, 31, 36, 39, 44, 47, 50, 59, 61, 65, 67, 68, 71, 74, 76, 78, 79, 81, 84, 85, 89, 91, 93, 96, 99, 101, 104, 105, 105, 112, 118, 123, 136, 139, 141, 148, 158, 161, 168, 184, 206, 248, 263, 289, 322, 388 и 513.

   а) Зошто не е содветно да се користат класи, како на пример, 0–50, 50–100, 100–150, итн?

   б) Состави табела на фреквенции и хистограм и коментирај ги;

   в) Состави табела на фреквенции и хистограм на природен логаритам на податоците ($\ln(x)$) и коментирај некои негови карактеристики.

   г) Која пропорција на податоци има животен век помал од 100 дупки, а која најмалку 200?

8. Конструирај нормален веројатносен дијаграм за следниот примерок на дебелината на покривката што се добива со бои со низок вискозитет: 0.83, 0.88, 0.88, 1.04, 1.09, 1.12, 1.29, 1.31, 1.48, 1.49, 1.59, 1.62, 1.65, 1.71, 1.76 и 1.83. Дали дебелината на наносот на бојата има приближно нормална распределба?

9. Направени се 20 набљудувања на прекини на диелектрична волтажа на парчиња прицврстени со термостатичка смола. Вредностите $(i - 0.5)/n$ за кои се потребни z процентилите се $(1 - 0.5)/20 = 0.025$, $(2 - 0.5)/20 = 0.075$, итн.

Набљ.	24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
z – проц.	–1.96	–1.44	–1.15	–0.93	–0.76	–0.60	–0.45	–0.32	–0.19	–0.06
Набљ.	27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88
z – проц.	0.06	0.19	0.32	0.45	0.60	0.76	0.93	1.15	1.44	1.96

Состави нормален веројатносен дијаграм за овие податоци.

10. Откажувањето поради замор на материјалот на различни делови на авионите е предмет на интензивно проучување. За одредена компонента на воените авиони измерени се следните животни векови до откажување (поради замор на материјал) дадени во (часови на летање)/10^4: 0.736, 0.863, 0.865, 0.913, 0.915, 0.937, 0.983, 1.007, 1.011, 1.064, 1.109, 1.132, 1.140, 1.153, 1.253, 1.394. Состави нормален веројатносен дијаграм за овие податоци и оцени дали тие се со приближно нормална распределба.

11. Следните податоци се времиња до откажување (животен век) во 1000-и часови на 16 чипови: 82.8, 11.6, 359.5, 502.5, 307.8, 179.7, 242.0, 26.5, 244.8, 304.3, 379.1, 212.6, 229.9, 558.9, 366.7 и 204.6. Конструирај веројатносен дијаграм користејќи соодветни процентили на експоненцијална распределба за $\lambda = 1$. Потоа објасни зошто дијаграмот може да се користи за оценка на шансите дека примерокот е земен од популација со која било експоненцијална распределба (произволно λ).

12. Во 1789, Хенри Кевендиш (Henry Cavendish) ја пресметал густината на земјата користејќи торзионо нишало. Неговите 29 мерења, изразени како мултипликација на густината на водата се:

5.50 5.55 5.57 5.34 5.42 5.30 5.61 5.36 5.53 5.79 5.47 5.75 4.88 5.29 5.62
5.10 5.63 5.86 4.07 5.58 5.29 5.27 5.34 5.85 5.26 5.65 5.44 5.39 5.46

а) Пресметај го просекот, стандардната девијација и медијаната на податоците;

б) Дали медијаната на податоците подобро ја оценува густината на земјата од просекот?

в) Состани нормален веројатносен дијаграм и дај соодветен коментар.

11

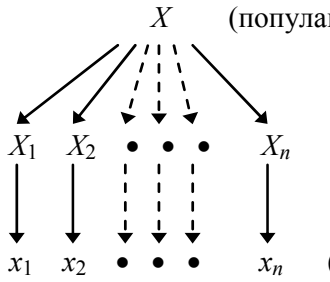
Оценки на непознати параметри

Секоја статистичка анализа се базира на две основни компоненти, статистички модел даден со двојката $(\Phi, (X_1, X_2, \ldots, X_n))$, каде што Φ е веројатносниот модел, а $(X_1, X_2, \ldots, X_n)$ е модел на примерокот и множество податоци $(x_1, x_2, \ldots, x_n)$. Податоците се интерпретираат како реализација на механизмот на случајноста зададен со веројатносниот модел. Оценките ги користат информациите од податоците за добивање на вредности за параметрите θ од Θ, што се непознати во веројатносниот модел $\Phi = \{f(x; \Theta) \mid x \in \mathbb{R}\}$. Еднаш кога параметарите θ_1, θ_2, …, θ_k ќе бидат оценети со оценките $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, ние комплетно го добиваме веројатносниот модел $\hat{\Phi} = \{f(x; \hat{\theta}_1, \hat{\theta}_1, \ldots, \hat{\theta}_k) \mid x \in \mathbb{R}\}$. Вообичаено секој од параметрите се оценува индивидуално.

Оценката на параметарот θ е пресликување $h(\cdot)$ од просторот на примерокот што е подмножество $V \subseteq \mathbb{R}^n$ во множеството параметри Θ,

$h(\cdot)$: $V \to \Theta$.

Пресликување вообичаено се означува со $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ и притоа $\hat{\theta}$ е оценката на θ. Ако се стави $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$, тогаш $\hat{\theta}$ е случајна променлива како функција од случајните променливи X_1, X_2, …, X_n и ја нарекуваме оценувач. Оваа ситуација е прикажана на следниот дијаграм.



Да забележиме дека и за оценувачот и за оценката користиме иста ознака $\hat{\theta}$, а од контекстот е јасно дали $\hat{\theta}$ е случајна променлива т.е. *оценувач* или конкретна вредност т.е. *оценка*.

ПРИМЕР 11.1 Во Бернулиевиот модел:

1) Веројатностен модел, $\Phi = \{f(x;\, \theta) = \theta^x(1-\theta)^{1-x} \mid 0 \le \theta \le 1,\, x = 0,\, 1\}$,

2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0,1\}^n$,

следните неколку функции би можеле да бидат "разумни" оценувачи на параметарот θ:

а) $\hat{\theta}_1 = X_1$, б) $\hat{\theta}_2 = \frac{1}{2}(X_1 + X_2)$, в) $\hat{\theta}_3 = \frac{1}{3}(X_1 + X_2 + X_n)$

г) $\hat{\theta}_4 = \frac{1}{n}\sum_{i=1}^{n} X_i$, д) $\hat{\theta}_5 = \frac{1}{n+1}\sum_{i=1}^{n} X_i$, ѓ) $\hat{\theta}_6 = \frac{1}{n+2}\sum_{i=1}^{n} X_i$.

Од друга страна, $\hat{\theta}_7 = X_1 - X_n$ не може да биде оценувач на θ бидејќи е надвор од интервалот [0,1] кога $X_1 = 0$ и $X_n = 1$. Исто така, $\hat{\theta}_8 = 0.6$ не е оценувач бидејќи доменот не е во просторот на примерокот.

Кај едноставниот нормален модел

1) Веројатностен модел,

$\Phi = \{f(x;\, \Theta) = Z(\mu,\, \sigma)\},\, x \in \mathbb{R},\, \mathbb{R}_\Theta = \mathbb{R}\times\mathbb{R}_+,\, \Theta = \{\mu,\, \sigma\}$,

2) Модел на примерок, $(X_1, X_2, \ldots, X_n)$,

истите функции би можеле да бидат оценувачи на параметарот μ:

а) $\hat{\mu}_1 = X_1$, б) $\hat{\mu}_2 = \frac{1}{2}(X_1 + X_2)$, в) $\hat{\mu}_3 = X_1 - X_n$

г) $\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n} X_i$, д) $\hat{\mu}_5 = \frac{1}{n+1}\sum_{i=1}^{n} X_i$, ѓ) $\hat{\mu}_6 = \frac{1}{n+2}\sum_{i=1}^{n} X_i$.

Ако се има предвид дека параметарот μ зема вредности од целото $\mathbb{R}$, не е возможно да се дефинира функција $h(X_1, X_2, \ldots, X_n)$ што не е оценувач на параметарот μ. ■

Секоја функција од примерокот $X_1, X_2, \ldots, X_n$ што не зависи од непознатиот параметар често се нарекува *статистика* и како таква таа е случајна променлива. Така оценувачот $h(X_1, X_2, \ldots, X_n)$ на непознат параметар е статистика. Одредени статистики играат важна улога во анализата на податоците и се користат за: оценки на непознати параметри, тестирање хипотези, предвидувања, регресии и други статистички анализи.

Со оглед на тоа што со леснотија може да се дефинираат многу "разумни" оценувачи на непознат параметар, се поставува прашањето: Кој од нив да се избере? Интиутивниот одговор е јасен. Треба да се избере оној оценувач што најдобро можно го оценува непознатиот параметар. Математички, $\hat{\theta}$ едноставно би требало да се избере така што да ја минимизира разликата $|\hat{\theta} - \theta|$. Проблемот е што оваа разлика не може да се пресмета бидејќи,

1) таа зависи од непознат параметар θ,

2) $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е случајна променлива, па таа прима многу различни вредности (со одредени веројатности).

Фактот што $\hat{\theta}$ е случајна променлива сугерира дека секоја формализација на квалитетот на оценувачот покрај други елементи, треба да ја вклучи и неговата распределба.

Генерално, распределбата на оценувачот е дадена со заедничката густина на распределба на примерокот $f(x_1, x_2, \ldots, x_n, \hat{\theta})$. Оценувачот е функција од случајни променливи, а изведувањето на нивните распределби го дискутиравме во поглавјето 6.3. Соодветната функција на распределба е дадена со

$$G(y) = p(\hat{\theta} < y) = \int\ldots\int_{(\mathbb{R}^n : h(x_1, x_2, \ldots, x_n) < y)} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n .$$

ПРИМЕР 11.2 Да ги најдеме распределбите на различните оценувачи $\hat{\theta}$ од примерот 11.1.

Ако случајните променливи $X_1, X_2, \ldots, X_n$ се независни со бернулиева распределба, тогаш $\sum_{i=1}^{n} X_i$ има биномна распределба $\mathrm{Bin}(n\theta, n\theta(1-\theta))$, со очекување $n\theta$ и дисперзија $n\theta(1-\theta)$. Ова доаѓа од $p\left(\sum_{i=1}^{n} X_i = k\right) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ бидејќи сумата едноставно го дава бројот на случувања на настан (број на единици) во n повторувања на експеримент. Имајќи го тоа предвид, за оценувачите на θ ги добиваме следните распределби:

а) $\hat{\theta}_1 \sim \mathrm{Bin}(\theta, \theta(1-\theta))$, б) $\hat{\theta}_2 \sim \mathrm{Bin}(\theta, \frac{1}{2}\theta(1-\theta))$, в) $\hat{\theta}_3 \sim \mathrm{Bin}(\theta, \frac{1}{3}\theta(1-\theta))$,

г) $\hat{\theta}_4 \sim \mathrm{Bin}(\theta, \frac{1}{n}\theta(1-\theta))$, д) $\hat{\theta}_5 \sim \mathrm{Bin}(\frac{n}{n+1}\theta, \frac{n}{(n+1)^2}\theta(1-\theta))$,

ѓ) $\hat{\theta}_6 \sim \mathrm{Bin}(\frac{n}{n+2}\theta, \frac{n}{(n+2)^2}\theta(1-\theta))$.

Од овие распределби веднаш се гледа дека оценувачите а) – г) имаат распределба со просек θ што е еднаков на параметарот што се оценува – центрираност, но дисперзиите се различни. Од нив најмала дисперзија има г) (за $n > 3$) и интуитивно тој оценувач е подобар од другите што ја имаат особината на центрираност (просекот е ист со параметарот). Се разбира, најмала дисперзија има оценувачот ѓ), но тој не е центриран.

Кај нормалниот модел, ако се има предвид дека сума на случајни променливи со нормална распределба е случајна променлива со нормална распределба, распределбите на оценувачите се следните:

а) $\hat{\mu}_1 \sim \mathrm{Z}(\mu, \sigma^2)$, б) $\hat{\mu}_2 \sim \mathrm{Z}(\mu, \frac{1}{2}\sigma^2)$ в) $\hat{\mu}_3 \sim \mathrm{Z}(0, 2\sigma^2)$

г) $\hat{\mu}_4 \sim \mathrm{Z}(\mu, \frac{1}{n}\sigma^2)$, д) $\hat{\mu}_5 \sim \mathrm{Z}(\frac{n}{n+1}\mu, \frac{n}{(n+1)^2}\sigma^2)$,

ѓ) $\hat{\mu}_6 \sim \mathrm{Z}(\frac{n}{n+2}\mu, \frac{n}{(n+2)^2}\sigma^2)$.

И во овој случај, се чини дека $\hat{\mu}_4$ е најдобар бидејќи тој е центриран и од сите други центрирани оценувачи има најмала дисперзија, $D(\hat{\mu}_4) = \sigma^2/n$. ■

Овој пример покажува дека и во двата случаја, на бернулиев и нормален модел, најдобрите оценувачи (интуитивно) коинцидираат. Се разбира постојат добри причини за ваквиот резултат. И во двата случаја параметарот што се оценува е просекот на распределбата EX, а најдобриот оценувач е просекот на примерокот $\frac{1}{n}\sum_{i=1}^{n} X_i$. Идејата за оцену-

вање на моментите на распределбата преку соодветните моменти на примерокот има долга историја во статистиката што датира од XIX-тиот век. Во тоа време не постоела јасна дистинкција меѓу релативната фреквенција и веројатноста, па следователно не се водело сметка и за разликата меѓу моментите на примерокот и моментите на распределбата. Оттука се и корените на таканаречениот *принцип на изедначување на моментите* што гласи:

> *дефинирај го оценувачот на моментот на распределбата како соодветен момент на примерокот*.

Принципот на изедначување на моментите се имплементира така што најпрво непознатиот параметар θ се изразува како функција од моменти на распределбата (на пример $\theta = h(\mu,\sigma^2, \ldots)$). Потоа, во функцијата едноставно моментите на распределната се заменуваат со моментите на примерокот $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, ... , добивајќи оценувач на θ даден со $\hat{\theta} = h(\hat{\mu},\hat{\sigma}^2,\ldots)$. Ваквата процедура е на некој начин обратна од методот на моменти за оценка на параметри што ќе го разгледаме понатаму.

11.1. Некои статистики за оценки на параметри

Како што некои распределби се од посебно значење (на пример нормалната, студентовата или $\chi 2$), така и некои статистики за оценки на непознатите параметри се користат во огромен број случаи и со тоа заслужуваат посебно внимание.

11.1.1. Просек на примерокот

Статистиката

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

вообичаено се нарекува просек на примерокот земен од популацијата X. Нека просекот и дисперзијата на популацијата бидат $EX = \mu$ и $DX = \sigma^2$. Просекот и дисперзијата на статистиката $\overline{X}$ се добиваат едноставно кога се претпостави независноста на примерокот

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}(n\mu) = \mu,$$

$$D\overline{X} = E(\overline{X} - \mu)^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Со зголемување на примерокот (раст на n), дисперзијата на $\overline{X}$ опаѓа и $E\overline{X}$ праволиниски се приближува до μ. Интуитивно е јасно дека $\overline{X}$ е добар оценувач на μ. Да се потсетиме дека во прилог на ова се "изјасни" и законот на големите броеви (поглавје 7.2).

Кога се работи за распределбата на $\overline{X}$, според централната гранична теорема распределбата на $\overline{X}$ се приближува кон нормална кога $n \to \infty$. Попрецизно, распределбата на случајната променлива

$$(\overline{X} - \mu)\left(\frac{\sqrt{n}}{\sigma}\right)$$

конвергира кон Z(0,1) кога $n \to \infty$.

11.1.2. Дисперзија на примерокот

Статистиката

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

се нарекува дисперзија на примерокот земен од популацијата X. Очигледно оценувачот S^2 го "мери" просекот на раштрканоста на податоците околу просекот. Но зошто ставаме $1/(n-1)$ наместо $1/n$? Се покажува дека во тој случај очекувањето на S^2 е токму σ^2, што понатаму ќе видиме дека е пожелна особина при оценувањето на непознатите параметри. За ова да го покажеме, статистиката ќе ја напишеме во облик

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left((X_i - \mu) - (\overline{X} - \mu)\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left((X_i - \mu) - \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n}(X_i - \mu)(X_j - \mu). \end{aligned}$$

Сега ако го пресметаме просекот на S^2 користејќи ја взаемната независност на $X_1, X_2, \ldots, X_n$, добиваме

$ES^2 = \sigma^2$.

Дисперзијата на статистиката (оценувачот) S^2 може да се добие со пресметка на просекот член по член,

$$DS^2 = E(S^2 - \sigma^2)^2 = \ldots = \frac{1}{n}\left(E(X-\mu)^4 - \frac{n-3}{n-1}\sigma^4 \right),$$

што покажува дека дисперзијата на S^2 е инверзна функција од n.

Распределбата на S^2 би можела во принцип да се изведе со помош на техниките глава 6, но тоа би било прекомплицирано поради комплексната природа на изразот за S^2. Во специјален случај, кога S^2 е статистика земена од популација со нормална распределба $Z(\mu, \sigma^2)$, тогаш случајната променлива $(n-1)S^2/\sigma^2$ има $\chi 2$ (хи-квадрат) распределба со $n - 1$ степени на слобода. Во поглавјето 6.4 веќе покажавме дека сума на квадрати на случајни променливи со нормална распределба дава случајна променлива со $\chi 2$ распределба.

11.1.3. Моменти на примерокот

Статистиката за k-ти момент земен од популацијата X е

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k .$$

Слично како за обичниот просек и дисперзија, и за овој оценувач може да се покаже дека

$$Em_k = EX^k, \quad Dm_k = \frac{1}{n}\left(EX^{2k} - (EX^k)^2 \right),$$

каде што EX^k е k-ти момент на популацијата X.

11.1.4. Подредени статистики*

Примерокот $X_1, X_2, \ldots, X_n$ земен од популацијата X секогаш може да се подреди, на пример во растечки редослед. Нека $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ биде вака преуреден примерок, каде што $X_{(1)}$ е најмалиот, а $X_{(n)}$ најголемиот елемент. Тогаш вредноста $X_{(k)}$ се нарекува статистика од k–ти ред. Екстремните вредности $X_{(1)}$ и $X_{(n)}$ се од посебен практичен интерес бидејќи тие може да претставуваат минимално или максимално дозволено оптоварување на некој систем. На пример, кај надежноста на системите, распределбата на животниот век на систем со n сериски компоненти (откажува кога која било компонента ќе откаже) е функција од минималното време на откажување на компонентите. Од друга страна, кај

систем со n паралелни компоненти (откажува кога која секоја компонента ќе откаже) распределбата на животниот век е определена од максималното време на откажување на компонентите. Во ваквите случаи од интерес се случајните променливи

$$Z = \max(X_1, X_2, \ldots, X_n) \text{ и}$$

$$W = \min(X_1, X_2, \ldots, X_n).$$

Претпоставувајќи независност и еднаква распределеност на примерокот со функција на распределба $F(x)$ и густина $f(x)$, функцијата на распределба на Z е

$$F_Z(x) = p(Z < x) = p(\text{сите } X_i < \text{x}) = p(X_1 < x \text{ и } X_2 < x \text{ и ... и } X_n < x) =$$
$$= F_{X_1}(x) \cdot F_{X_2}(x) \cdot \ldots \cdot F_{X_n}(x) = F(x) \cdot F(x) \cdot \ldots \cdot F(x) = F(x)^n.$$

Слично за функцијата на распределба на W добиваме

$$F_W(x) = p(W < x) = p(\text{барем едно } X_i < x) = p(X_1 < x \text{ или } X_2 < x \text{ или ...}$$
$$\text{или } X_n < \text{x}) = 1 - p(X_1 \geq x \text{ и } X_2 \geq x \text{ и ... и } X_n \geq x) =$$
$$= 1 - \left(1 - F_{X_1}(x)\right) \cdot \left(1 - F_{X_2}(x)\right) \cdot \ldots \cdot \left(1 - F_{X_n}(x)\right)$$

$$= 1 - \left(1 - F(x)\right)^n.$$

Соодветните густини на распределба и во двата случаи се добиваат со диференцирање

$$f_Z(x) = \frac{dF_Z(x)}{dx} = n\left(F(x)\right)^{n-1} f(x) \text{ и}$$

$$f_W(x) = \frac{dF_W(x)}{dx} = n\left(1 - F(x)\right)^{n-1} f(x).$$

Јасно е дека распределбите на Z и W зависат од распределбите на X_i, но и во асимптотскиот случај кога $n \to \infty$ нема резултат што би бил комплетно независен од обликот на $F(x)$ [Soong 2004].

Просекот и дисперзијата на подредените статистики би можеле да се добијат со интегрирање, но тие не може да се изразат како функција од моментите на популацијата X.

11.2. Критериуми за квалитетот на оценките

Како што може да се види од примерот 11.1, доста е лесно да се дефинира оценувач на непознат параметар. Главен проблем е како да се избере најдобриот од многуте можни. Секој оценувач е функција од

случајни променливи (примерокот), па следователно и тој е случајна променлива. Оттука, секоја одлука за избор на најдобар оценувач ќе биде базирана на распределбата на примерокот.

Проблемот на дефинирање на добар оценувач е сличен, на пример, на ситуацијата кога некој ловец пука на дивеч што не го гледа, бидејќи тој и дивечот се на спротивни страни од една планина. Ловецот мора да направи стратегија (правила) со тоа што му стои на располагање, како што е аголот под кој пука или јачината на истрелот, за да истрелот биде колку е можно поблиску до целта. Слично, и ние треба да избереме правила што ќе овозможат максимално можно погодување на непознатиот параметар θ.

Идеален оценувач $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ би бил таков што тој би земал само една вредност (онаа на параметарот θ) со веројатност 1, независно од кој било реализиран примерок. Ваквиот случај $p(\hat{\theta} = \theta) = 1$ води кон дегенеративна распределба на примерокот. За конечен примерок со големина n, таков оценувач не постои. Секој оценувач добива различни вредности за различни вредности на примерокот. Како максимално да се доближиме до идеален оценувач? Некој би можел интуитивно тоа да го постави преку првите 2 момента, т.е. да бара

$$M\hat{\theta} = \theta \text{ и } D\hat{\theta} = 0 .$$

Тоа би значело дека оптималниот оценувач треба да има просек во вистинската вредност на параметарот и дисперзија 0. За конечен примерок со големина n, второто барање е невозможно да се достигне, но кога $n \to \infty$ тоа е остварливо. Оттука следува потребата да се воведат особини на оценувачите што се однесуваат на конечен примерок (исполнети за секој n) и асимптотски особини (исполнети кога $n \to \infty$).

Постојат повеќе критериуми по кои може да се евалуира квалитетот на оценките на непознатите параметри. Овие критериуми обично ги дефинираат пожелните особини за оценувачот како и начинот на кој квалитетите на различните оценувачи би можеле да се споредуваат.

11.2.1. Центрираност

Дефиниција 11.1 Оценувачот $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е центриран оценувач за θ ако

$$E\hat{\theta} = \theta ,$$

т.е. просекот од распределбата на функцијата од примерокот е еднаков на параметарот што се оценува. Во спротивно, оценувачот е нецентри-

ран со отстапување $E\hat{\theta} - \theta$. Пожелноста на оваа особина е јасна, бидејќи секако би сакале во просек $\hat{\theta}$ да биде блиско до θ.

Да се потсетиме дека $\overline{X}$ и S^2 беа центрирани оценувачи на просекот μ и дисперзијата σ^2. S^2 беше малку "неприроден" бидејќи наместо $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ ние би очекувале $S^{2*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ да биде центриран. Но поради $ES^{2*} = \frac{n-1}{n}\sigma^2$, оценувачот S^{2*} има отстапување од центрираноста за $(1/n)\sigma^2$. Ова покажува дека центрираноста не мора да биде одлучувачки критериум за избор на оценувачот, и дека некои други критериуми би можеле во одредени ситуации да превладаат при изборот.

Додатно, интересно е да се одбележи дека центриран оценувач не постои секогаш. На пример, за експоненцијалниот модел со веројатносен модел $\{f(x;\theta) = \theta \cdot e^{-\theta x},\ \theta > 0,\ x > 0\}$ не постои центриран оценувач за θ [Schervish 1995, стр. 297].

Центрираноста не е инваријантна во однос на трансформациите на непознатиот параметар. Имено, ако $E\hat{\theta} = \theta$ и ако $\nu = g(\theta)$ и $\hat{\nu} = g(\hat{\theta})$ за некоја трансформација $g(\cdot)$, тогаш во општ случај $E\hat{\nu} \neq \nu$.

11.2.2. Ефикасност

Покрај критериумот просекот на оценувачот $\hat{\theta}$ да биде близок до параметарот (центрираност), природно е да се бара вредностите на $\hat{\theta}$ да бидат со висока веројатност блиски до θ. Ова барање води кон критериумот $\hat{\theta}$ да има што е можно помала дисперзија.

Дефиниција 11.2 $\hat{\theta}$ е центриран оценувач на θ со најмала дисперзија ако за секој друг центриран оценувач θ^* на θ важи

$$D\hat{\theta} \leq D\theta^*.$$

Од два центрирани оценувачи, секако би го преферирале оној со помала дисперзија бидејќи тогаш оценките се поблиски до нивниот просек, т.е. до вистинската вредност на параметарот. Често пати наместо дисперзијата $D\hat{\theta}$, за проценка на ефикасноста на оценувачот се користи стандардната девијација $\sigma_{\hat{\theta}} = \sqrt{D\hat{\theta}}$ и таа вообичаено се нарекува стандардна грешка на $\hat{\theta}$.

ПРИМЕР 11.3 Веќе видовме дека $\overline{X}$ е центриран оценувач на просекот на популација μ. Дали ефикасноста на $\overline{X}$ се подобрува со зголемување на примерокот n?

Решение

Веќе видовме дека дисперзијата на оценувачот $\overline{X}$ е

$D\overline{X} = \frac{\sigma^2}{n}$, т.е. стандардната грешка е $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$,

што очигледно опаѓа со растењето на n. ■

ПРИМЕР 11.4 Нека популацијата X има просек μ_0 и дисперзија σ_0^2. Земаме примерок $X_1, X_2, \ldots, X_5$ и го оценуваме просекот со

$$\hat{\mu}_0 = \frac{1}{5}(X_1 + X_2 + X_3) + \frac{3}{10}X_4 + \frac{1}{10}X_5$$

Каков е оценувачот $\hat{\mu}_0$?

Решение

Од $EX_i = \mu_0$ добиваме дека

$$E\hat{\mu}_0 = \frac{3}{5}\mu_0 + \frac{3}{10}\mu_0 + \frac{1}{10}\mu_0 = \mu_0,$$

т.е. $\hat{\mu}_0$ е центриран оценувач на μ_0. За дисперзијата добиваме

$$D\hat{\mu}_0 = \frac{3}{25}\sigma_0^2 + \frac{9}{100}\sigma_0^2 + \frac{1}{100}\sigma_0^2 = \frac{22}{100}\sigma_0^2 \quad > \quad \frac{20}{100}\sigma_0^2 = \frac{1}{5}\sigma_0^2 = D\overline{X},$$

што покажува дека центрираниот оценувач $\hat{\mu}_0$ е помалку ефикасен оценувач на μ_0 од $\overline{X}$. ■

Природно е да се постави прашањето дали $\overline{X}$ (за фиксно n) е центриран оценувач на μ со најмала дисперзија? Директното докажување дека дисперзијата на $\overline{X}$ е помала од дисперзиите на сите други центрирани оценувачи на μ е секако тешко да се направи. За одговор на вакви прашања поврзани со наоѓање оценувачи со налмала дисперзија, од огромна полза е теоремата на Крамер-Рао.

Теорема 11.1 **(**Cramér-Rao**).** Нека $X_1, X_2, \ldots, X_n$ е примерок земен од популацијата X со густина на распределба $f(x_1, x_2, \ldots, x_n; \theta)$, каде што θ е

непознат параметар, и нека $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е центриран оценувач на θ. Тогаш, за дисперзијата на $\hat{\theta}$ важи

$$D\hat{\theta} \geq \left(nE\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right)^{-1} = \left(E\left(\frac{\partial \ln f(x_1, x_2, \ldots, x_n;\theta)}{\partial \theta} \right)^2 \right)^{-1},$$

ако даденото диференцирање и очекување постојат. Аналоген резултат важи и кога X е дискретна.

Доказ: Зедничката гистина на распределба на X_1, X_2, …, X_n поради нивната независност е $f(x_1;\theta)f(x_2;\theta) \cdots f(x_n;\theta)$. Очекувањето на статистиката $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е $E\hat{\theta} = E(h(X_1, X_2, \ldots, X_n))$. Бидејќи $\hat{\theta}$ е центриран, имаме дека

$$\theta = E\hat{\theta} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, x_2, \ldots, x_n) f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n ,$$

што со диференцирање (по θ) дава

$$1 = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, x_2, \ldots, x_n) \left(\sum_{j=1}^{n} \frac{1}{f(x_j,\theta)} \frac{\partial f(x_j,\theta)}{\partial \theta} \right) f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n$$

$$= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, x_2, \ldots, x_n) \left(\sum_{j=1}^{n} \frac{\partial \ln f(x_j,\theta)}{\partial \theta} \right) f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n .$$

Сега воведуваме нова случајна променлива Y со

$$Y = \sum_{j=1}^{n} \frac{\partial \ln f(X_j,\theta)}{\partial \theta},$$

за која според горното равенство (по дефиниција) имаме дека

$$1 = E(\hat{\theta} \cdot Y),$$

и од $E(\hat{\theta} \cdot Y) = E(\hat{\theta})E(Y) + \rho_{\hat{\theta},Y}\sigma_{\hat{\theta}}\sigma_Y$ следува дека

$$1 = \theta EY + \rho_{\hat{\theta},Y}\sigma_{\hat{\theta}}\sigma_Y .$$

Сега ќе покажеме дека $EY = 0$.

Од равенствата

$$1 = \int_{-\infty}^{\infty} f(x_i;\theta) dx_i . \quad i = 1, 2, \ldots, n ,$$

со диференцирање добиваме

$$0 = \int_{-\infty}^{\infty} \frac{\partial f(x_i;\theta)}{\partial \theta} dx_i = \int_{-\infty}^{\infty} \frac{\partial \ln f(x_i;\theta)}{\partial \theta} f(x_i;\theta) dx_i , \quad i = 1, 2, \ldots, n ,$$

што веднаш дава $EY = 0$. Оттука равенството $1 = \theta EY + \rho_{\hat{\theta},Y}\sigma_{\hat{\theta}}\sigma_Y$ веднаш се сведува на

$$1 = \rho_{\hat{\theta},Y}\sigma_{\hat{\theta}}\sigma_Y ,$$

и сега ако се потсетиме дека за коефициентот на корелација важи $\rho^2 \le 1$, добиваме дека

$$\sigma_{\hat{\theta}}^2 \ge \frac{1}{\sigma_Y^2} .$$

На крај, бидејќи Y е сума на n независни случајни променливи со очекување 0 и дисперзија $E(\partial \ln f(X;\theta)/\partial\theta)^2$, дисперзијата на Y е едноставно сумата на n-те дисперзии

$$\sigma_{\hat{\theta}}^2 = nE\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 ,$$

со што доказот е комплетиран. ■

Во горниот доказ, имлицитно е претпоставено дека диференцирањата по θ под интегралите е дозволено. Неравенството на Крамер-Рао ја дава долната граница на дисперзијата за кој било центриран оценувач и го изразува лимитот на точноста со која еден параметар може да биде оценет. Да забележиме дека долната граница е функција од θ, т.е. од параметарот што се оценува.

Понатаму даваме некои поважни забелешки во врска со неравенството на Крамер-Рао:

а) Неравенството може да се напише во еквивалентен облик

$$D\hat{\theta} \ge -\left(E\left(\frac{\partial^2 \ln f(x_1, x_2, \ldots, x_n;\theta)}{\partial \theta^2} \right) \right)^{-1} ,$$

што често пати е пресметковно позгоден за работа;

б) Неравенството може да се прошири на случај на произволни оценувачи $\hat{\theta}$, (не мора да бидат центрирани) и тогаш

$$D\hat{\theta} \ge \left(\frac{\partial E\hat{\theta}}{\partial \theta} \right)^2 \left(E\left(\frac{\partial \ln f(x_1, x_2, \ldots, x_n;\theta)}{\partial \theta} \right)^2 \right)^{-1} ;$$

в) Неравенството може лесно да се прошири на случај на повеќе непознати параметри $\theta_1, \theta_2, \ldots, \theta_m$ во $f(x;\theta_1, \theta_2, \ldots, \theta_m)$ што се оценуваат со $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ врз база на примерок големина n. Во таков случај важи

$$K_{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_m} \geq \frac{\Lambda^{-1}}{n},$$

каде што K е коваријацијата, а Λ^{-1} е инверзна матрица на Λ дадена со елементите

$$\Lambda_{ij} = E\left(\frac{\partial \ln f(X;\theta_1,\theta_2,\ldots,\theta_m)}{\partial \theta_i}\frac{\partial \ln f(X;\theta_1,\theta_2,\ldots,\theta_m)}{\partial \theta_j}\right),\ i,j = 1,2,\ldots,m.$$

Ова повлекува дека

$$D\hat{\theta}_j \geq \frac{(\Lambda^{-1})_{jj}}{n} \geq \frac{1}{n\Lambda_{jj}},\quad j = 1,2,\ldots,m;$$

г) За даден центриран оценувач $\hat{\theta}$ на θ, односот на неговата Крамер-Рао-ва долна граница со неговата дисперзија обично се нарекува *целосна ефикасност* на оценувачот. Значи ефикасноста на секој центриран оценувач е помала или еднаква на 1, и ако е 1 тогаш оценувачот се нарекува *целосно ефикасен*. Потребен и доволен услов еден центриран оценувач $\hat{\theta}$ на θ да биде целосно ефикасен е разликата $\hat{\theta} - \theta$ да може да се претстави во облик

$$\hat{\theta} - \theta = h(\theta)\left(\frac{\partial \ln f(x_1,x_2,\ldots,x_n;\theta)}{\partial \theta}\right)$$

за некоја функција $h(\theta)$.

Да се вратиме на прашањето дали $\overline{X}$ (за фиксно n) е центриран оценувач на μ со најмала дисперзија? Да забележиме дека за примена на неравенството на Крамер-Рао, мора да биде позната густината на распределба на популацијата $f(x_1, x_2, \ldots, x_n;\theta)$.

ПРИМЕР 11.5 Нека популацијата X има нормална распределба $Z(\mu, \sigma^2)$. Определи ја долната граница на дисперзијата за центрираните оценувачи на μ и σ^2. (За σ^2 земи $\mu = 0$, т.е. распределба $Z(0,\sigma^2)$).

Решение

За μ имаме дека

$\ln f(X;\mu) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(X-\mu)^2/2\sigma^2}\right) = \ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{(X-\mu)^2}{2\sigma^2}$, и сега

$\frac{\partial \ln f(X;\mu)}{\partial \mu} = \frac{X-\mu}{\sigma^2}$, од што следува

$E\left(\frac{\partial \ln f(X;\mu)}{\partial \mu}\right)^2 = \frac{1}{\sigma^4} E(X-\mu)^2 = \frac{1}{\sigma^2}$.

Значи долната граница на дисперзијата на секој центриран оценувач на μ е σ^2/n, што е еднакво на дисперзијата на $\overline{X}$. Заклучуваме дека $\overline{X}$ има најмала дисперзија од сите центрирани оценувачи на μ ($\overline{X}$ е целосно ефикасен оценувач на μ) кога популацијата е со нормална распределба. Да забележиме дека поради

$f(x_1, x_2, \ldots, x_n; \mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$, имаме

$\ln f(x_1, x_2, \ldots, x_n; \mu) = n \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2$ и

$\frac{\partial \ln f(x_1, x_2, \ldots, x_n; \mu)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu)$, па следува дека

$\overline{X} - \mu = \frac{1}{n}\sum_{i=1}^n (x_i - \mu) = \frac{\sigma^2}{n}\left(\frac{\partial \ln f(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta}\right)$,

што влегува во шемата на потребен и доволен услов еден оценувач да биде целосно ефикасен (според забелешката г)), бидејќи може да се смета дека $h(\mu) = \sigma^2/n$.

За σ^2 ќе воведеме ознака $\theta = \sigma^2$ (поради диференцирањето) и понатаму добиваме

$\ln f(X;\theta) = \ln\left(\frac{1}{\sqrt{2\pi\theta}} e^{-X^2/2\theta}\right) = -\frac{1}{2}\ln 2\pi\theta - \frac{X^2}{2\theta}$, што дава

$\frac{\partial \ln f(X;\theta)}{\partial \theta} = \frac{X^2}{2\theta^2} - \frac{1}{2\theta}$. Понатаму одиме со еквивалентниот облик на неравенството од забелешката а) и добиваме

$\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2} = -\frac{X^2}{\theta^3} + \frac{1}{2\theta^2}$, што конечно дава

$E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right) = -\frac{\theta}{\theta^3} + \frac{1}{2\theta^2} = -\frac{1}{2\theta^2}$. Оттука (види забелешката а)), долната граница на секој центриран оценувач на θ е $2\theta^2/n$.

За споредба со центрираниот оценувач S^2 (на дисперзијата), од поглавјето 11.1.2, дисперзијата на S^2 е

$$DS^2 = \frac{1}{n}\left(MX^4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{1}{n}\left(3\sigma^4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{2\sigma^4}{n-1} = \frac{2\theta^2}{n-1},$$

бидејќи $MX^4 = 3\sigma^4$ кога X има нормална распределба. Според тоа, ефикасноста на S^2 означена со $ef(S^2)$ е

$$ef(S^2) = \frac{\textit{долна граница}\text{ на Крамер - Рао}}{DS^2} = \frac{n-1}{n},$$

што значи дека во овој случај, S^2 не е целосно ефикасен оценувач на θ (т.е. σ^2). Бидејќи $ef(S^2) \to 1$ кога $n \to \infty$, оценувачот S^2 може да се смета за *целосно асимптотски ефикасен*. ■

Секогаш треба да се има предвид дека (квалитетот на) оценувачот зависи од распределбата на популацијата. За илустрација на оваа ситуација ќе претпоставиме дека е земен примерок $X_1, X_2, \ldots, X_n$ од популација со една од трите распределби:

$$f(x) = Z(\mu, \sigma^2), \qquad f(x) = \frac{1}{\pi(1+(x-\mu)^2)} \quad \text{и} \quad f(x) = \begin{cases} 1/(2c), & -c \le x-\mu \le c \\ 0, & \textit{во спротивно} \end{cases},$$

т.е. нормалната, Кошиевата и рамномерната. Сите тие се симетрични околу μ, а Кошиевата е со "поголеми" веројатносни опашки во споредба со нормалната. Да разгледаме 4 оценувачи на μ:

а) $\overline{X}$ - обичен просек, б) $\widetilde{X}$ - медијана,

в) $\overline{X}_e = (X_{min} + X_{max})/2$; г) $\overline{X}_{10\%}$ - просек на најголемите 10%.

Сега најдобрата оценка на μ зависи од распределбата, и тоа:

1) Ако примерокот доаѓа од нормална распределба, $\overline{X}$ е најдобриот оценувач, бидејќи веќе знаеме дека е со најмала дисперзија од сите центрирани оценувачи;

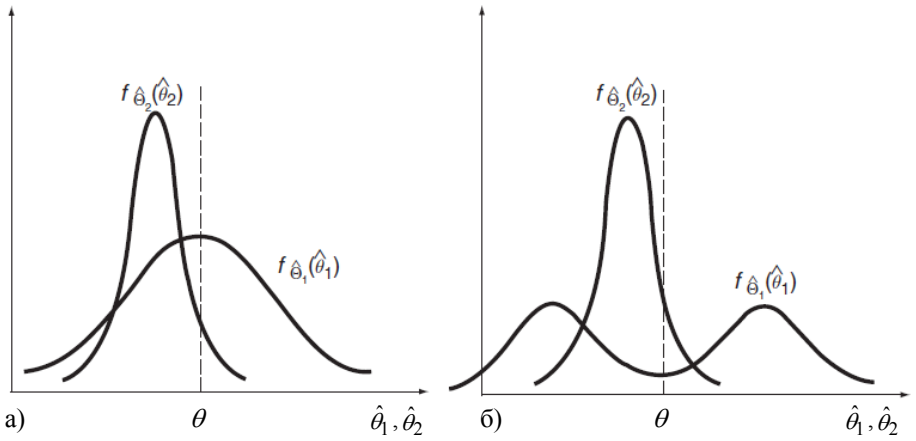
2) Ако примерокот доаѓа од Кошиева распределба, $\overline{X}$ и $\overline{X}_e$ се лоши оценувачи бидејќи се пречуствителни на екстремните вред-

ности, а Кошиевата распределба со "тешките" опашки секогаш има такви. Од друга страна, медијаната е добар оценувач на μ;

3) Ако примерокот доаѓа од рамномерна распределба, $\overline{X}_e$ е најдобриот оценувач бидејќи распределбата нема "опашки".

4) Оценувачот $\overline{X}_{10\%}$ не е најдобар во ниеден од трите случаи, но е разумно добар кај сите три распределби.

Тука повторно би забележиле дека иако центрираноста како и малата дисперзија се пожелни критериуми, нецентрираните оценувачи не би смеело однапред да се сметаат како инфериорни. Да ги разгледаме распределбите на оценувачите $\hat{\theta}_1$ и $\hat{\theta}_2$ на непознатиот параметар θ, дадени на сл. 11.1. И двете слики, а) и б), покажуваат пример на распределби $f_{\hat{\theta}_1}(x)$ и $f_{\hat{\theta}_2}(x)$ при што и во двата случаи $\hat{\theta}_1$ е центриран оценувач на θ, додека $\hat{\theta}_2$ не е.



Слика 11.1 *Густини на распределба на два оценувача на* θ

Од сликите јасно се гледа дека вредностите на $\hat{\theta}_2$ се поблиски (по веројатност) до вредноста на θ. Оваа ситуација е многу подраматична на сликата б), каде што веројатноста $\hat{\theta}_1$ да биде блиску до θ е многу мала. Следниот пример уште еднаш ја илустрира ситуацијата кога центрираноста не треба да се преферирана за сметка на поголема дисперзија.

ПРИМЕР 11.6 Во Бернулиевиот модел:

1) Веројатностен модел, $\Phi = \{f(x; \Theta) = \theta^x(1-\theta)^{1-x} \mid 0 \le \theta \le 1, x = 0, 1\}$,

2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0, 1\}^n$,

знаеме дека дисперзијата на центрираниот оценувач $\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ на θ, е $D\hat{\theta} = \theta(1-\theta)/n$. Дали $\hat{\theta}$ е целосно ефикасен оценувач?

Решение

Со директна пресметка добиваме

$$\ln f(X;\mu) = X \ln\theta + (1-X)\ln(1-\theta),$$

$$\frac{\partial \ln f(X;\mu)}{\partial\theta} = X\frac{1}{\theta} - (1-X)\frac{1}{1-\theta}, \text{ што дава}$$

$$M\left(\frac{\partial \ln f(X;\mu)}{\partial\theta}\right)^2 = \left(-\frac{1}{1-\theta}\right)^2 p(X=0) + \left(\frac{1}{\theta}\right)^2 p(X=1) = \frac{1}{1-\theta} + \frac{1}{\theta} = \frac{1}{\theta(1-\theta)}.$$

Оттука, според Крамор-Рао, долната граница на дисперзијата е $(n/\theta\,(1-\theta))^{-1}$ што е еднакво на $D\hat{\theta}$. Заклучуваме дека $\hat{\theta}$ е целосно ефикасен оценувач на θ.

Да го разгледаме оценувачот

$$\hat{\theta}_1 = \frac{n\overline{X}+1}{n+2},$$

којшто не е центриран бидејќи

$$E\hat{\theta}_1 = \frac{n\theta+1}{n+2},$$

но има дисперзија $D\hat{\theta}_1 = \frac{n^2}{(n+2)^2} D\hat{\theta} = \frac{n\theta(1-\theta)}{(n+2)^2} \le D\hat{\theta}$.

Значи иако $\hat{\theta}_1$ е нецентриран оценувач, неговата дисперзија е помала од онаа на $\hat{\theta}$, посебно кога n има "умерена" вредност. Ова, во одредени случаи, може да биде доволна причина за избор на $\hat{\theta}_1$ како подобар оценувач од $\hat{\theta}$, иако $\hat{\theta}_1$ е нецентриран. ■

Како што веќе кажавме, неравенството на Крамер-Рао не е секогаш применливо. Попрецизно, условите под кои тоа може да се разгледува се следните:

1) Просторот на параметарот е отворено множество во $\mathbb{R}$ (за m параметри во $\mathbb{R}^m$);
2) Подршката на распределбата, множеството $\{(x_1, x_2, \ldots, x_n) \mid f(x_1, x_2, \ldots, x_n; \theta) > 0\}$ е исто за сите вредности на θ,

3) $\dfrac{\partial \ln f(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta}$ постои и е конечно;

4) За $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ може да се менува редоследот на диференцирање и интегрирање, т.е.

$$\frac{\partial}{\partial \theta} \int \ldots \int h(x_1, x_2, \ldots, x_n) f(x_1, \cdots, x_n; \theta) dx_1 \cdots dx_n =$$

$$= \int \ldots \int h(x_1, x_2, \ldots, x_n) \frac{\partial}{\partial \theta} f(x_1, \cdots, x_n; \theta) dx_1 \cdots dx_n < \infty .$$

Овие услови наметнати на распределбата на примерокот, обезбедуваат да имаме таканаречен *регуларен веројатносен модел*. *Првиот* услов исклучува гранични точки за да се овозможи диференцирање од сите страни на точките. За да ја илустрираме важноста на *вториот* услов, ќе го разгледаме веројатносниот модел $\{f(x;\ \theta) = 1/\theta \mid 0 < \theta < \infty,\ 0 < x < \theta\}$, за кој според Крамер-Рао

$$\frac{\partial \ln f(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \ln\left(\frac{1}{\theta^n}\right) = \frac{\partial}{\partial \theta}(-n \ln \theta) = -\frac{n}{\theta},$$ од каде што

следува дека долната граница на дисперзијата е $(n/\theta)^2$ што е целосно неприменливо. Третиот и четвртиот услов се јасни и тие директно се користени при доказот на теоремата 11.1.

11.2.3. Редукција на нецентрираност*

Понекогаш се јавуваат случаи кога би сакале да ја намалиме нецентрираноста на некој оценувач. Нека $\hat{\theta}_n = h(X_1, X_2, \ldots, X_n)$ е нецентриран оценувач на θ (индексот n е воведен да нагласи дека статистиката е базирана на примерок со големина n) и нека нецентрираноста е изразена во облик

$$E\hat{\theta}_n - \theta = \frac{a_1(\theta)}{n} + \frac{a_2(\theta)}{n^2} + \ldots + \frac{a_k(\theta)}{n^k} + \ldots .$$

Ја разгледуваме низата оценувачи $\widetilde{\theta}_{n-1,i} = h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, $i = 1, 2, .., n$ на θ, каде што за секој елемент на низата се користи истата статистика (формула), но во секој елемент е испуштена една случајна променлива (примерокот е со големина $n - 1$). Понатаму оценувачот $\widetilde{\theta}_n$ го дефинираме како просек на низата оценувачи $\widetilde{\theta}_{n-1,i}$ со

$$\widetilde{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\theta}_{n-1,i} .$$

Конвексната комбинација на $\hat{\theta}_n$ и $\tilde{\theta}_n$ го дава конечниот оценувач $\bar{\theta}_n$ на θ даден со

$$\bar{\theta}_n = n\hat{\theta}_n - (n-1)\tilde{\theta}_n,$$

познат под име ножот на Џек (Jack-knife). Уште во 50-тите години од минатиот век [Quenouille 1956] има покажано дека кај новиот оценувач $\bar{\theta}_n$, нецентрираноста од прв ред $a_1(\theta)/n$ исчезнува. Тоа значи дека во случаите кога нецентрираноста е само од прв ред (а таква е многу често), оценката добиена со ножот на Џек станува центрирана (нецентрирана оценка ја трансформира во центрирана).

ПРИМЕР 11.7 Примени го ножот на Џек за оценка на дисперзијата σ^2 од нормалниот модел, дадена со

$$S^{2*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

Решение

Нецентрираноста на S^{2*} е од прв ред бидејќи

$$ES^{2*} - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Понатаму имаме:

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)^2,$$

$$\tilde{\theta}_{n-1,k} = \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq k}}^{n}X_i^2 + \frac{1}{(n-1)^2}\left(\sum_{\substack{i=1\\ i\neq k}}^{n}X_i\right)^2 \text{ и}$$

$$\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_{n-1,i} = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n(n-1)^2}\left(\sum_{i=1}^{n}X_i^2 + (n-2)\sum_{i=1}^{n}X_i^2\right).$$

Заменувајќи ги $\hat{\theta}_n$ и $\tilde{\theta}_n$ во оценувачот $\bar{\theta}_n = n\hat{\theta}_n - (n-1)\tilde{\theta}_n$ добиваме

$$\bar{\theta}_n = \sum_{i=1}^{n}X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2 - \frac{n-1}{n}\sum_{i=1}^{n}X_i^2 + \frac{1}{n(n-1)}\sum_{i=1}^{n}X_i^2 + \frac{n-2}{n(n-1)}\left(\sum_{i=1}^{n}X_i\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}X_i\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

Заклучуваме дека оценувачот со ножот на Џек е идентичен со S^2- центрираната оценка на дисперзијата. ■

Иако идејата позади оценките со ножот на Џек изгледа природно и логично, нејзината вградена закономерност интуитивно е тешко дофатлива. Дискусијата би заклучиле со напомената дека истиот пристап може лесно да се прошири за намалување на нецентрираноста не само од прв, туку и од повисок ред.

11.2.4. Средна квадратна грешка

Горенаведените мери на ефикасност овозможуваат избор меѓу центрираните оценувачи, но не нуди начин на избор меѓу нецентрирани и центрирани оценувачи. Ова е важно прашање бидејќи центриран оценувач може и да не постои или некогаш тој не е добар од аспелт на асимптотски или некои други карактеристики. На пример, кој оценувач би избрале од $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_5, \hat{\mu}_6$ од примерот 11.1, каде што првите 3 се центрирани, а задните два не. Дисперзиите на центрираните оценувачи $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3$ се значително поголеми од оние на $\hat{\mu}_5, \hat{\mu}_6$. Како да споредиме центриран и нецентриран оценувач?

Ако би сакале да ја пенализираме нецентрираноста на оценувачот $\hat{\theta}$, не би требало да ја користиме неговата дисперзија ($E\hat{\theta} \neq \theta$), туку нејзините варијации околу θ. Стандардна мера за ваквите ситуации е средната квадратна грешка (*MSE* = Mean Square Error) дефинирана со

$$MSE(\hat{\theta},\theta) = E(\hat{\theta} - \theta)^2 .$$

Во случај на центриран оценувач $\hat{\theta}$, очигледно $MSE(\hat{\theta},\theta) = D\hat{\theta}$. За нецентрирани оценувачи $\hat{\theta}$ со директна пресметка имаме

$$MSE(\hat{\theta},\theta) = E(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2 = D\hat{\theta} + (E\hat{\theta} - \theta)^2 ,$$

каде што $E\hat{\theta} - \theta$ е степенот на нецентрираност на оценувачот $\hat{\theta}$.

Оценувачот $\hat{\theta}$ е со најмала средна квадратна грешка ако

$$MSE(\hat{\theta},\theta) \leq MSE(\theta^*,\theta),$$

за секој друг оценувач θ^* на θ (и сите вредности на θ).

ПРИМЕР 11.8 Спореди ги *MSE*-та на оценувачите $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ со $\hat{\theta}_5$ и $\hat{\theta}_6$ кај Бернулевиот модел од примерот 11.1.

Решение

Според средната квадратна грешка, $\hat{\theta}_5$ и $\hat{\theta}_6$ се подобри оценувачи од $\hat{\theta}_1, \hat{\theta}_1$ и $\hat{\theta}_3$ (за $n > 3$) бидејќи

$$MSE(\hat{\theta}_5)=\frac{n}{(n+1)^2}\theta(1-\theta)+\left(\frac{-\theta}{n+1}\right)^2=\frac{n\theta(1-\theta)+\theta^2}{(n+1)^2}\le\begin{cases}1\\1/2\\2\end{cases}=\begin{cases}MSE(\hat{\theta}_1)\\MSE(\hat{\theta}_2)\\MSE(\hat{\theta}_3)\end{cases},$$

$$MSE(\hat{\theta}_6)=\frac{n}{(n+2)^2}\theta(1-\theta)+\left(\frac{-\theta}{n+2}\right)^2=\frac{n\theta(1-\theta)+\theta^2}{(n+2)^2}\le\begin{cases}1\\1/2\\2\end{cases}=\begin{cases}MSE(\hat{\theta}_1)\\MSE(\hat{\theta}_2)\\MSE(\hat{\theta}_3)\end{cases}$$

за речиси сите вредности на паранетарот θ. Додатно, очигледно е дека важи и $MSE(\hat{\theta}_6) < MSE(\hat{\theta}_6)$. ■

Претходниот пример укажува дека $\hat{\theta}_5$ и $\hat{\theta}_6$ се веројатно многу подобри оценувачи од $\hat{\theta}_1, \hat{\theta}_1$ и $\hat{\theta}_3$ не само поради помалата средна квадратна грешка, туку и поради фактот што кај нив таа се намалува со зголемувањето на големината на примерокот n. Ова понатаму води кон асимптотските каракатеристики на оценувачите што се на некој начин обопштување на законот на големите броеви и централната гранична теорема од (глава 7).

11.2.5. Конзистентност

Веќе кажавме дека идеален оценувач θ^* за кој важи $p(\theta^* = \theta) = 1$ не е возможен за примерок со конечна големина n, па оттука е логично од оценувачот да бараме ваква карактеристика кога тоа е можно, т.е. кога n расте до бесконечност.

Дефиниција 11.3 Оценувачот $\hat{\theta}$ е конзистентен оценувач на θ ако за секој $\varepsilon > 0$ важи

$$\lim_{n\to\infty} p(|\hat{\theta}-\theta|<\varepsilon)=1 \text{ (со спротивниот настан } \lim_{n\to\infty} p(|\hat{\theta}-\theta|\ge\varepsilon)=0\text{).}$$

Ова се чита како веројатностите на настаните "$\hat{\theta}$ се разликува од θ за помалку од $\varepsilon > 0$ се приближува до 1, кога n оди кон бескрајност". Со ваква карактеристика, големината на примерокот станува многу важен елемент што ја одредува блискоста на оценките до параметарот.

Следуваат некои поважни забелешки за конзистентноста на оценувачите:

а) $\hat{\theta}$ како функција од примерокот секако зависи од n, па горниот лимес е добро дефиниран;

б) Конзистентноста во некоја смисла е проширување на законот на големите броеви за други функции од примерокот $h(X_1, X_2, \ldots, X_n)$, а не само неговата сума;

в) Во случај кога $\hat{\theta}$ е со ограничена дисперзија, за проверка на конзистентноста од голема полза е неравенството на Чебишев.

ПРИМЕР 11.9 Провери дали S^2 е конзистентен оценувач на σ^2.

Решение

Користејќи го неравенството на Чебишев, добиваме

$$p(|S^2 - \sigma^2| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} E(S^2 - \sigma^2)^2 .$$

Поради центрираноста $ES^2 = \sigma^2$, па имаме дека $E(S^2 - \sigma^2)^2 = DS^2$, а веќе покажавме дека $DS^2 = 2\sigma^4/(n-1)$. Оттука следува

$$\lim_{n\to\infty} p(|S^2 - \sigma^2| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{1}{\varepsilon^2}\left(\frac{2\sigma^4}{n-1}\right) = 0 .$$

Значи S^2 е конзистентен оценувач на σ^2. ■

Овој пример ја инспирира следната теорема.

Теорема 11.2 Нека $\hat{\theta}$ е оценувач на θ базирана на примерок со големина n. Тогаш, ако

$\lim_{n\to\infty} E\hat{\theta} = \theta$ и $\lim_{n\to\infty} D\hat{\theta} = 0$, оценувачот $\hat{\theta}$ е конзистентен.

Доказ: Според неравенството на Чебишев,

$$p(|\hat{\theta} - \theta| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} E(\hat{\theta} - \theta)^2 = \frac{1}{\varepsilon^2} E(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2 =$$

$$= \frac{1}{\varepsilon^2}\left(D\hat{\theta} + 0 + E(E\hat{\theta} - \theta)^2\right) = \frac{1}{\varepsilon^2}\left(D\hat{\theta} + (E\hat{\theta} - \theta)^2\right).$$

Оттука, и од условите на теоремата, веднаш следува конзистентноста на $\hat{\theta}$. ■

Тука би нагласиле дека конзистентноста е минимална карактеристика во смисла што кога оценувачот е неконзистентен, тој не е вреден за понатамошно разгледување. Се разбира, конзистентноста не значи дека оценувачот е добар. Има многу примери на конзистентни, но бескорисни оценувачи.

ПРИМЕР 11.10 Спореди ги карактеристиките на оценувачите на μ кај нормалниот модел од примерот 11.1.

Решение

Основните карактеристики на овие оценувачи се сумирани во следната табела:

Оценувач на μ	Очекување	Дисперзија	Карактеристики
$\hat{\mu}_1 = X_1$	μ	1	центриран, голема дисперзија, неконзистентен
$\hat{\mu}_2 = \frac{1}{2}(X_1 + X_2)$	μ	1/2	центриран, голема дисперзија, неконзистентен
$\hat{\mu}_3 = (X_1 - X_n)$	0	2	нецентриран, голема дисперзија, неконзистентен
$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n} X_i$	μ	$1/n$	центриран, мала дисперзија, конзистентен
$\hat{\mu}_5 = \frac{1}{n+1}\sum_{i=1}^{n} X_i$	$\frac{n\mu}{n+1}$	$\frac{n}{(n+1)^2}$	нецентриран, помала дисперзија, конзистентен
$\hat{\mu}_6 = \frac{1}{n+2}\sum_{i=1}^{n} X_i$	$\frac{n\mu}{n+2}$	$\frac{n}{(n+2)^2}$	нецентриран, најмала дисперзија, конзистентен

Кој оценувач од горните 6 би избрале? Првите 3 оценувачи се неконзистентни и веднаш ги елиминираме. Од останите 3 конзистентни оценувачи, само $\hat{\mu}_4$ е центриран (тој е и целосно ефикасен), па така е и најдобриот избор. Да забележиме дека изборот меѓу $\hat{\mu}_5$ и $\hat{\mu}_6$ би бил доста нејасен бидејќи $\hat{\mu}_6$ има помала дисперзија но е подалеку од центрираност во споредба со $\hat{\mu}_5$. Очигледно е дека колку повеќе го зголемуваме именителот во $\hat{\mu}_6$ (на пример n + 1000) толку повеќе ја намалуваме дисперзијата, но истовремено се оддалечуваме од центрираноста. ■

Во горните примери, а најчесто и во практиката, при одлуката за избор на оценувач се користат само првите 2 момента (очекувањето и дисперзијата) на примерокот. Теоретски е поправилно за статистичките одлуки да се користи распределбата на примерокот, бидејќи само така може целосно да се искористи регуларноста на случајноста скриена во податоците.

Конзистентноста би можела да се појача со барањето веројатноста на конвергенцијата да биде скоро сигурна,

$$p(\lim_{n\to\infty} \hat{\theta} = \theta) = 1, \text{ т.е. } \hat{\theta} \xrightarrow{c.c.} \theta \ (c.c. = \text{скоро сигурно}).$$

Ваквата конзистентност вообичаено се нарекува *јака конзистентност*. Скоро сигурната конвергенција ја повлекува конвергенцијата на веројатности и неа веќе ја разгледувавме во поглавјето 7.2.

Користејќи ги варијантите на јакиот закон на големите броеви, може да се покаже дека кај Бернулевиот модел важи $\hat{\theta} \overset{c.c.}{\to} \theta$, а исто така и кај нормалниот модел $\hat{\mu} \overset{c.c.}{\to} \mu$.

11.2.6. Други критериуми*

Како што погоре кажавме, конзистентноста (слаба или јака) е едно проширување на законот на големите броеви во насока што од простата сума $\sum_{i=1}^{n} X_i$ од законот на големите броеви одиме кон поопшти функции од примерокот $h(X_1, X_2, \ldots, X_n)$. Следниот критериум, *асимптотската нормалност* во одредена смисла е проширување на централната гранична теорема.

Оценувачот $\hat{\theta}$ е *асимптотски нормален* оценувач на θ ако постои нормализирачка низа c_n, $n = 1, 2, \ldots$, таква што

$$c_n(\hat{\theta} - \theta) \overset{a}{\sim} Z(0, D_\infty\theta),$$ каде што

$\overset{a}{\sim}$ означува асимптотска распределба, $D_\infty\theta$ е асимптотската дисперзија на $\hat{\theta}$, додека низата c_n е функција од n и кај случаен примерок таа е дефинирана со $c_n = \sqrt{n}$.

На пример, оценувачите $\hat{\theta}_4$ и $\hat{\theta}_5$ кај Бернулиевиот модел се асимптотски нормални бидејќи

$$\sqrt{n}(\hat{\theta}_4 - \theta) \overset{a}{\sim} Z(0, \theta(1-\theta)) \text{ и } \sqrt{n}(\hat{\theta}_5 - \theta) \overset{a}{\sim} Z(0, \theta(1-\theta)).$$

Слично, $\hat{\mu}_4$ и $\hat{\mu}_5$ кај нормалниот модел се исто така асимптотски нормални бидејќи

$$\sqrt{n}(\hat{\mu}_4 - \mu) \overset{a}{\sim} Z(0,1) \text{ и } \sqrt{n}(\hat{\mu}_5 - \mu) \overset{a}{\sim} Z(0,1).$$

За конзистентните и асимптотски нормалните оценувачи може асимптотската дисперзија да биде одлучувачка при изборот на оценувачот. Нејзината долна граница е дефинирана со

$$\lim_{n\to\infty}\left(\frac{1}{c_n}\right)^2 nE\left(\frac{\partial \ln f(X;\mu)}{\partial\theta}\right)^2.$$

На пример, кај Бернулиевиот модел имавме $nE\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2 = \frac{n}{\theta(1-\theta)}$, па долната граница е $\lim_{n\to\infty}\frac{1}{n}\frac{n}{\theta(1-\theta)} = \theta(1-\theta)$. Кај нормалниот модел од примерот 11.1 (распределбата е Z(μ, 1)), имаме $nE\left(\frac{\partial \ln f(X;\mu)}{\partial\mu}\right)^2 = n$, па долната граница на асимптотската дисперзија е $\lim_{n\to\infty} n(1/n) = 1$.

Во директна врска со асимптотската нормалност е критериумот на *асимптотска ефикасност*. Еден асимптотски нормален оценувач $\hat{\theta}$ на θ се нарекува *асимптотски ефиксен* ако

$c_n(\hat{\theta}-\theta) \overset{a}{\sim} Z(0, \textit{асимптотска дисперзија – долна граница})$, т.е.

ако асимптотската дисперзија $D_\infty\theta$ е минимална, т.е. еднаква со нејзината долна граница.

Кај Бернулиевиот и нормалниот модел од примерот 11.1, асимптотските дисперзии $\theta(1-\theta)$ и 1 се очигледно еднакви на нивните долни граници, па оценувачите $\hat{\theta}_4$, $\hat{\theta}_5$, $\hat{\mu}_4$ и $\hat{\mu}_5$ се асимптотски ефикасни.

Генерално зборувајќи, дискусијата околу оценувачите на непознатите параметри во главно се потпира на првите два момента од примерокот. Ова може да даде погрешен впечаток дека распределбата на примерокот не е директно потребна, туку само индиректно за определување на првите два момента. Иако дефиницијата на конзистентност

$$\lim_{n\to\infty} p(|\hat{\theta}-\theta|<\varepsilon) = 1,$$

ја враќа улогата на распределбата за определување на низата веројатности, најпогодниот начин за покажување на конзистентноста повторно се базираше на првите два момента (теорема 11.2). Слично и кај другите асимптотски карактеристики на оценувачите (на пример асимптотската ефикасност) улогата на распределбата на примерокот не е експлицитно очигледна. Нашето фокусирање на првите два момента примарно е базирано на концепциски и пресметковни погодности. Постојат и други карактеристики на оценувачите што не се базираат на моментите туку на некои други бројни карактеристики.

Оценувачот $\hat{\theta}$ на θ е *центриран по мод* ако распределбата на $\hat{\theta}$ има мод што коинцидира со θ, $Mode(\hat{\theta}) = \theta$.

Оценувачот $\hat{\theta}$ на θ е *центриран по медијана* ако распределбата на $\hat{\theta}$ има медијана што коинцидира со θ, $Median(\hat{\theta}) = \theta$.

На пример, во случај на нормалниот модел, стандардниот оценувач $\hat{\mu}_4 = \overline{X} = \sum_{i=1}^{n} X_i$ е центриран со нормална распределба на примерокот. Оттука, според особините на модот и медијаната (поглавје 5.1.3), веднаш следува дека $\hat{\mu}_4$ е исто така центриран по мод и медијана.

Покрај со бројните карактеристики на распределбата на примерокот, постојат и други начини да се дефинира блискоста на оценките до вистинската вредност на параметарот коишто поексплицитно ја користат распределбата на примерокот. На пример, споредбата на блискоста на два оценувачи $\hat{\theta}$ и $\widetilde{\theta}$ до θ би можела да се направи со следната мера на *концетрација*,

$$p(|\hat{\theta} - \theta| \le c) \ge p(|\widetilde{\theta} - \theta| \le c), \text{ за сите } c > 0.$$

Кога горниот услов е исполнет сметаме дека $\hat{\theta}$ е *поконцетриран* оценувач околу θ отколку $\widetilde{\theta}$. Да забележиме дека во споредбата директно се вклучени распределби на два оценувача. Ваков тип оценувачи нема да користиме во оваа книга, а тука ги воведовме со цел да ја нагласиме улогата на распределбата на примерокот при споредба на квалитетот на оценувачите.

11.2.7. Доволност*

Досегашните дискусии за квалитетот на оценувачите се вртеа околу условите под кои може да се дојде до најдобриот. Ако се вратиме на аналогијата со ситуацијата кога некој ловец пука на дивеч од другата страна на планината (не го гледа), ние досега ги разгледувавме само прашањата за близината на истрелот до дивечот откако тој е веќе направен. Факторите како што се аголот под кој пука или јачината на истрелот за тој да се приближи до целта, не беа разгледувани. Доволноста на оценувачот ги адресира ваквите прашања на подесувањето на оружјето за подобро покривање на областа на целта.

Идејата за доволност се однесува на можноста на редукција на димензионалноста на набљудуваните податоци (примерокот) без губење на која било информација. Таа за прв пат била формализирана некаде во 30-тите години на минатиот век.

Нека $X_1, X_2, \ldots, X_n$ е примерок од популацијата X чијашто распределба $f(x_1, x_2, \ldots, x_n; \theta)$ зависи од непознат параметар θ. Ако $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е статистика таква што за секоја друга статистика $\tilde{\theta} = g(X_1, X_2, \ldots, X_n)$, условната распределба на $\tilde{\theta}$, при дадено $\hat{\theta} = \theta_0$ не зависи од θ, тогаш $\hat{\theta}$ е *доволна статистика* за θ. Додатно, ако $E\hat{\theta} = \theta$, тогаш $\hat{\theta}$ е *доволен оценувач* на θ.

На пример, во Бернулиевиот модел, примерокот $X_1, X_2, \ldots, X_n$ е составен од случајни променливи со распределба $p(X_k = 1) = \theta$ и $p(X_k = 0) = 1 - \theta$. Секоја реализација на примерокот е низа од 0-и и 1-ци, на пример, 0, 1, 0, 0, 1, …, 1. Сега статистиката $h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$ ги содржи сите релевантни информации за параметарот, т.е. распределбата, бидејќи позициите на 1-ците во реализациите на примерокот не се важни. Значи наместо да водиме сметка за n броеви во реализациите на примерокот, ние водиме сметка само за еден – нивната сума, што е значителна редукција на димензионалноста. Интуитивно, $\hat{\theta} = \sum_{i=1}^{n} X_i$ е доволна статистика.

Дефиниција 11.4 Статистиката $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е доволна статистика за θ акко условната распределба $f(x_1, x_2, \ldots, x_n; \theta)$ на примерокот $X_1, X_2, \ldots, X_n$ при $h(x_1, x_2, \ldots, x_n) = \theta_0$ не зависи од θ,

$$f(x_1, x_2, \ldots, x_n \mid h(x_1, x_2, \ldots, x_n) = \theta_0; \theta) = q(x_1, x_2, \ldots, x_n)$$

за секоја точка $(x_1, x_2, \ldots, x_n)$ од доменот на распределбата.

Интуитивно, ваквата дефиниција тврди дека, ако $\hat{\theta}$ е доволна статистика за θ, сите информации во примерокот во врска со θ се содржани во $\hat{\theta}$. Уште повеќе, ако располагаме со центриран оценувач $\tilde{\theta}$ на θ што не е функција од доволен оценувач, секогаш е можно да се најде центриран оценувач базиран на доволна статистика со дисперзија помала од онаа на $\tilde{\theta}$. Доволните оценувачи имаат дисперзии што се помали од дисперзиите на сите други центрирани оценувачи (што не зависат од доволна статистика).

Горната дискусија може поформално да се искаже со следното тврдење [Rao 1949].

Теорема 11.3 Ако $\tilde{\theta}$ е центриран оценувач на θ и $h(x_1, x_2, \ldots, x_n)$ е доволна статистика за θ, тогаш статистиката

$$\hat{\theta} = E(\tilde{\theta} \mid h(X_1, X_2, \ldots, X_n))$$

ги има следните особини:

а) $\hat{\theta}$ е оценувач на θ;

б) центрираност: $E\hat{\theta} = \theta$;

в) поголема ефикасност од $\tilde{\theta}$: $D\hat{\theta} \le D\tilde{\theta}$. ■

Ова тврдење овозможува подобрување на центриран оценувач со друг поефикасен центриран оценувач, но не кажува ништо за евентуалната целосна ефикасност на оценувачот. За тоа понатаму би требало да се искористи теоремата на Крамер-Рао, но ако новиот оценувач не кореспондира со добиената долна граница, повторно го немаме целосно ефикасниот оценувач.

Дефиницијата за доволна статистика е доста "интуитивна" и не дава директен начин за проверка или наоѓање на таква статистика. Следната корисна теорема, што ја даваме без доказ, значително го олеснува овој проблем.

Теорема 11.4 Статистиката $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ е доволна статистика за θ акко постојат функции $g(h(X_1, X_2, \ldots, X_n); \theta)$ и $v(X_1, X_2, \ldots, X_n)$ каде што првата зависи од $X_1, X_2, \ldots, X_n$ само преку $h(\cdot)$, а втората не зависи од θ, такви што распределбата на примерокот $f(x_1, x_2, \ldots, x_n; \theta)$ може да се претстави (факторизира) со

$$f(x_1, x_2, \ldots, x_n;\theta) = g(h(x_1, x_2, \ldots, x_n); \theta) \cdot v(x_1, x_2, \ldots, x_n),$$

за сите $(x_1, x_2, \ldots, x_n)$ од доменот на распределбата. ■

Испитувањето на доволност со оваа теорема се сведува на испитување на распределбата на примерокот, и со малку имагинација уочување на можната факторизација.

ПРИМЕР 11.11 Провери ја доволноста на статистиката $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ во Бернулиевиот модел $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$, $0 \le \theta \le 1$, $x = 0, 1$.

Решение

Густината на распределба на примерокот (поради независноста) е

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{j=1}^{n} f(x_j;\theta) = \theta^{\sum x_j}(1-\theta)^{n-\sum x_j} =$$

$$= \theta^{n\frac{1}{n}\sum x_j}(1-\theta)^{n(1-\frac{1}{n})\sum x_j}.$$

Оттука веднаш следува факторизацијата во која

$$g(h(x_1, x_2, \ldots, x_n);\ \theta) = \theta^{n\frac{1}{n}\sum x_j}(1-\theta)^{n(1-\frac{1}{n})\sum x_j} \quad \text{и} \quad v(x_1, x_2, \ldots, x_n) = 1. \blacksquare$$

ПРИМЕР 11.12 Нека примерокот $X_1, X_2, \ldots, X_n$ е земен од популацијата X со Пуасонова распределба $f(k;\theta) = \dfrac{\theta^k e^{-\theta}}{k!}$, $k = 0, 1, 2, \ldots$, каде што θ е непознат параметар. Обиди се да најдеш доволна статистика за θ.

Решение

Густината на распределба на примерокот е

$$f(x_1, x_2, \ldots, x_n;\ \theta) = \prod_{j=1}^{n} f(x_j;\theta) = \frac{\theta^{\sum x_j} e^{-n\theta}}{\prod x_j!},$$

што може да биде факторизирано така што

$$g(h(x_1, x_2, \ldots, x_n);\ \theta) = \theta^{\sum x_j} e^{-n\theta} \text{ и } v(x_1, x_2, \ldots, x_n) = \frac{1}{\prod x_j!}.$$

Значи статистиката $\hat{\theta} = \sum_{j=1}^{n} X_j$ е доволна статистика за θ. $\blacksquare$

Резултатите од ова поглавје сугерираат дека најдобрата стратегија за барање на најдобриот центриран оценувач е да се провери постоењето на доволна статистика, а потоа да се продолжи барањето со помош на теоремата 11.3 или едноставно со инспирација. Ова барање може да води кон комплицирани но практично бескорисни статистики и затоа е потребно тоа да се базира само на доволни статистики што се "економични".

Една доволна статистика $h(X_1, X_2, \ldots, X_n)$ е минимална ако секоја друга доволна статистика $g(X_1, X_2, \ldots, X_n)$ е функција од неа, т.е.

$g(X_1, X_2, \ldots, X_n) = q(h(X_1, X_2, \ldots, X_n))$ за некоја функција $q(\cdot)$.

Да забележиме дека за секој статистички модел што досега го разгледувавме, постои минимална доволна статистика. За разлика од тешкотиите што се јавуваат при испитувањето на статистика на доволност (било по дефиниција или со факторизација), за испитување на статистика на минимална доволност постои релативно лесна постапка дадена со следното тврдење.

Нека постои статистика $h(X_1, X_2, \ldots, X_n)$ таква што за две различни реализации на примерокот $(x_1, x_2, \ldots, x_n)$ и $(z_1, z_2, \ldots, z_n)$, односот

$\dfrac{f(x_1,x_2,\ldots,x_n;\theta)}{f(z_1,z_2,\ldots,z_n;\theta)}$ не зависи од θ акко

$$h(x_1, x_2, \ldots, x_n) = h(z_1, z_2, \ldots, z_n),$$

тогаш $h(X_1, X_2, \ldots, X_n)$ е минимална доволна статистика за θ.

ПРИМЕР 11.13 Провери ја минималната доволност на $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ во бернулиевиот модел.

Решение

Во овој случај односот

$$\frac{f(x_1,x_2,\ldots,x_n;\theta)}{f(z_1,z_2,\ldots,z_n;\theta)} = \frac{\theta^{\sum x_j}(1-\theta)^{n-\sum x_j}}{\theta^{\sum z_j}(1-\theta)^{n-\sum z_j}} = \left(\frac{\theta}{1-\theta}\right)^{\sum(x_j - z_j)},$$

не зависи од θ акко $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} z_j$. Така, статистиката $\overline{X}$ не е само доволна, туку е и минимално доволна. ■

11.2.8. Комплетност*

Ако поставиме цел да најдеме оптимален оценувач користејќи доволна статистика, треба да забележиме дека минималната доволна статистика не гарантира еднозначност на оценувачот бидејќи секоја функција (еден-еден) од минимална доволна статистика е исто така минимално доволна. За добивање еднозначност на оптималниот оценувач, потребно е да се воведе поимот за комплетност.

Една доволна статистика $\tilde{\theta} = h(X_1, X_2, \ldots, X_n)$ е комплетна ако фамилијата на густини $\{f_{\tilde{\theta}}(x_1,x_2,\ldots,x_n;\theta)\}$ е комплетна т.е. за секоја функција $g(x)$ важи

$$Mg(\tilde{\theta}) = 0 \text{ повлекува } g(\tilde{\theta}) \overset{c.c.}{=} 0 \text{ (с.с. = скоро секаде),}$$
$$\text{за сите } (x_1, x_2, \ldots, x_n) \text{ за кои } f_{\tilde{\theta}}(x_1,x_2,\ldots,x_n;\theta) > 0.$$

Комплетноста е особина на фамилија густини (густините зависат од θ) што интуитивно кажува дека единствена центрирана оценка на 0 е самата 0. Корисноста на оваа особина се состои во тоа што ако $h(X_1, X_2, \ldots, X_n)$ е комплетна и доволна статистика, и $\hat{\theta} = g(h((X_1, X_2, \ldots, X_n))$ е центриран оценувач на θ, т.е. $Eg(h((X_1, X_2, \ldots, X_n)) = \theta$, тогаш оценувачот $\hat{\theta}$ е единствен.

Односот меѓу комплетна доволна статистика и минимална доволна статистика е во тоа што комплетната доволна статистика е и минимално доволна. Ова не води до крајот на барањето на најдобриот центриран оценувач преку користење на доволни статистики. Крајниот резултат е даден во следната теорема.

Теорема 11.5 Нека $h((X_1, X_2, \ldots, X_n)$ е комплетна доволна статистика за θ. Ако постои центриран оценувач $\hat{\theta}$ на θ којшто е функција од $h((X_1, X_2, \ldots, X_n)$, т.е. $\hat{\theta} = g(h((X_1, X_2, \ldots, X_n))$, тогаш оценувачот $\hat{\theta}$ е најдобар и единствен. ■

ПРИМЕР 11.14 Нека примерокот $X_1, X_2, \ldots, X_n$ е земен од популацијата X со рамномерна распределба $f(x, \theta) = 1/\theta$, $\theta \in (0, \infty)$, $x \in [0, \theta]$, којашто патем речено е проблематична од аспект на барање оптимален оценувач на параметарот. Најди оптимален оценувач за θ.

Решение

За рамномерната распределба веќе видовме дека долната граница на дисперзијата за центрираните оценувачи според Крамер-Рао не може да се определи. Ако се има предвид дека $EX = \theta/2$, некој би можел да се обиде да го форсира оценувачот

$$\tilde{\theta} = 2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right).$$

Ова секако не е лош оценувач бидејќи тој е центриран

$$E\tilde{\theta} = 2\left(\frac{1}{n}\sum_{i=1}^{n} EX_i\right) = \frac{2}{n}\frac{n\theta}{2} = \theta, \text{ и конзистентен}$$

$$D\tilde{\theta} = \frac{\theta^2}{3n} \to 0, \text{ кога } n \to \infty.$$

Сепак тој не е најдобриот оценувач. За да дојдеме до најдобриот оценувач ќе ја разгледаме статистиката

$$X_{(n)} = \max\{X_1, X_2, \ldots, X_n\},$$

и дополнително ќе означиме $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$. $X_{(n)}$ е доволна статистика за θ, бидејќи имаме

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\theta^n} = \left(\frac{1}{\theta^n} I(\theta - x_{(n)})\right) I(x_{(1)}) \text{ (факторизација за } x_{(n)}),$$

каде што $I(x)$ е индикатор функција дадена со $I(x) = \begin{cases} 0, \text{за } x < 0 \\ 1, \text{за } x \geq 1 \end{cases}$. Факторизацијата покажува дека $X_{(n)}$ е доволна статистика за θ.

Користејќи ја теоремата 11.4, продолжуваме понатаму во барање на подобар центриран оценувач

$$\hat{\theta} = E(\tilde{\theta} \mid X_{(n)}) = E\left(\frac{2}{n}\sum_{i=1}^{n} EX_i \right) \mid X_{(n)}) = \frac{n}{n+1} X_{(n)}.$$

Фамилијата густини на рамномерната распределба е комплетна, па оттука оценувачот $\hat{\theta}$ е најдобриот центриран оценувач на θ, којшто исто така е и единствен (според теоремата 11.5). ■

Во светло на комплетноста, стратегијата да се дојде до оптимален центриран оценувач би била во случај кога постои комплетна доволна статистика $\bar{\theta}$, да се појде од произволна центриран оценувач $\tilde{\theta}$ и потоа да се изведува условното очекување $E(\tilde{\theta} \mid \bar{\theta})$.

Интересно е да нагласиме дека за важната фамилија експоненцијални распределби, проблемот на наоѓање минимална статистика што истовремено е и доволна статистика е релативно лесно. Густина на распределба што припаѓа на експоненцијалната фамилија распределби е дадена со

$$f(x;\theta_1,\theta_2,\ldots,\theta_k) = c(\theta_1,\theta_2,\ldots,\theta_k)\cdot h(x)\cdot e^{\sum_{i=1}^{k} g_i(\theta_1,\theta_2,\ldots,\theta_k)\cdot t_i(x)},$$

каде што $c(\theta_1, \theta_2, \ldots, \theta_n) \geq 0$, $h(x) \geq 0$, и $g_i(\theta_1, \theta_2, \ldots, \theta_n)$, $t_i(x)$, $i = 1, 2, \ldots, k$ се реални функции. Многу добро-познати распределби се во оваа група: нормална, гама, бета, биномна и Пуасонова. За оваа фамилија распределби, статистиката

$$\sum_{j=1}^{n} t_i(x_j),\ i = 1, 2, \ldots, k$$

е минимална доволна статистика за $g_i(\theta_1, \theta_2, \ldots, \theta_n)$, $i = 1, 2, \ldots, k$, и во случај кога $n = k$, статистиката е и комплетна.

11.3. Методи на оценување

Откако ги разгледавме пожелните особини: центрираност, ефикасност, конзистентност итн., што една оценка би требало да ги поседува, се поставува прашањето како да се конструира добар оценувач. Во ова поглавје ќе разгледаме неколку методи за конструкција на оценувачи и тоа: методот на *моменти*, методот на *максимална подобност* и методот на *најмали квадрати*.

11.3.1. Метод на моменти

Ова е најстариот систематски метод за оценка непознати параметри предложен од Пирсон (Karl Pearson, 1857 – 1936) и екстензивно користен од него и неговите соработници. Последните години тој е доста занемарен поради недостаток на оптималност и поради растечката популарност на методот на максимална подобност. Сепак, популарноста на методот на моменти останува, делумно поради неговата пресметковна ефикасност, а делумно поради можноста во одредени ситуации да биде подобрен.

Методот на моменти е концепциски доста едноставен. Нека $f(x;\ \theta_1,\ \theta_2,\ \ldots,\ \theta_m)$ е густина на распределба што зависи од непознатите параметри θ_k, $k = 1,\ 2,\ \ldots,\ m$ што треба да бидат оценети според примерокот $X_1, X_2, \ldots, X_n$ земен од популацијата X. Моментите на популацијата X се дадени со

$$EX^k = \int_{-\infty}^{\infty} x^k f(x;\theta_1,\theta_2,\ldots,\theta_m)dx, \quad k = 1, 2, \ldots .$$

Од друга страна, моментите од кој било ред може да се најдат од примерокот

$$EX^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \ldots .$$

Изедначувајќи доволен број моменти од двете равенства и решавајќи ги равенствата по θ_k, $k = 1,\ 2,\ \ldots,\ m$, ги добиваме бараните оценки.

Да забележиме дека не е неопходно да разгледуваме m последователни моменти, туку е доволна која било група од m равенки со моменти што имаат решение. Се разбира, моментите од понизок ред вообичаено се позгодни поради тоа што бараат помалку манипулации со податоците. Атрактивноста на методот на моменти е во тоа што равенките се добиваат праволиниски и нивното решавање обично не претставува некој проблем. Недостатокот на методот е во неможноста во општ случај да се обезбедат пожелните особини, како центрираност и ефикасност, додека конзистентноста е секогаш исполнета.

ПРИМЕР 11.15 Најди оценувачи на непознатите параметри μ и σ^2 во нормалната распределба користејќи го методот на моменти.

Решение

Треба да ги изедначиме моментите на популацијата (од распределбата) со моментите пресметани од примерокот. Моментите на популацијата се

$$EX^k = \int_{-\infty}^{\infty} x^k f(x;\mu,\sigma^2)dx = \int_{-\infty}^{\infty} \frac{x^k}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} dx, \quad k = 1, 2,$$

и тие релативно лесно се пресметуваат од карактеристичната функција на нормалната распределба $\varphi(t) = e^{it\mu-\sigma^2t^2/2}$ (поглавје 7.3),

$$EX = i^{-1}\frac{d\varphi}{dt}\bigg|_{t=0} = i^{-1}e^{it\mu-\sigma^2t^2/2}(i\mu-\sigma^2t)\bigg|_{t=0} = \mu \quad \text{и}$$

$$EX^2 = i^{-2}\frac{d^2\varphi}{dt^2}\bigg|_{t=0} = -e^{it\mu-\sigma^2t^2/2}(i\mu-\sigma^2t)^2 + e^{it\mu-\sigma^2t^2/2}\sigma^2\bigg|_{t=0} = \mu^2+\sigma^2.$$

Ги изедначуваме овие моменти со моментите на примерокот

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu \quad \text{и} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2+\sigma^2,$$

од каде што веднаш се добиваат оценувачите (ги означуваме со $\hat{\mu}$ и $\hat{\sigma}^2$)

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{и} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2.$$

Како што знаеме, $\hat{\mu}$ е центриран, комплетно ефикасен и конзистентен оценувач на μ, додека $\hat{\sigma}^2$ е нецентриран, ефикасен и конзистентен оценувач на σ^2. ■

ПРИМЕР 11.16 Користејќи го методот на моменти, најди оценувач на непознатиот параметар θ кај рамномерната распределба

$$f(x;\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0, & \text{во спротивно} \end{cases}.$$

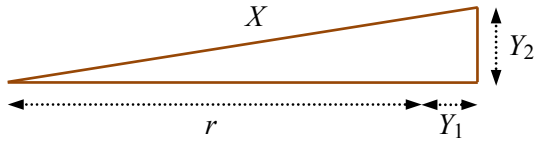
Решение

Моментот од прв ред е $EX = \frac{\theta}{2}$, од што следува $\hat{\theta} = 2EX = \frac{2}{n}\sum_{i=1}^{n} X_i$.

Оценувачот $\hat{\theta}$ е центриран и конзистентен, но не е најдобар. Оценувачот $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$ е подобар, а најдобриот центриран оценувач којашто е и единствен е $\frac{n}{n+1}X_{(n)}$. (види пример 11.14). ■

ПРИМЕР 11.17 Мериме должина r на некој објект користејќи инструмент што, како и секогаш, има вградена непрецизност. Поради тоа, она што се мери е случајна променлива X (види слика), а Y_1 и Y_2 се идентично и нормално рас-

пределени случајни променливи со очекување 0 и дисперзија σ^2. Со методот на моменти, најди оценувач на непознатиот параметар $\theta = r^2$ врз база на примерок со големина n.



Решение

Случајната променлива X е дефинирана со

$$X = \sqrt{(r + Y_1)^2 + Y_2^2} \ .$$

Распределбата на X не ни е потребна бидејќи моментите може да се пресметаат директно знаејќи ја распределбата на Y_1 и Y_2. Да забележиме дека иако не се бара оценувач на σ^2, тој како непознат параметар мора да биде разгледан заедно со θ. Таков параметар, чијашто вредност не е од интерес, понекогаш во литературата се нарекува досаден (nuisance) параметар.

За два параметра вообичаено потребни се две равенки со моменти. Имајќи го предвид обликот на равенката за X (сé е под корен), позгодни за работа се моментите од парен ред. Оттука за вториот и четвртиот момент имаме

$$EX^2 = E(r + Y_1)^2 + EY_2^2 = \theta + 2\sigma^2 ,$$

$$EX^4 = E\left((r + Y_1)^2 + Y_2^2\right)^2 = \theta^2 + 8\theta\sigma^2 + 8\sigma^4 .$$

Од овој систем равенки по θ и σ^2 ја добиваме оценката $\hat{\theta}$

$$\hat{\theta} = \sqrt{2\left(EX^2\right)^2 + EX^4} \ ,$$

каде што останува уште да се заменат вториот и четвртиот момент со соодветните моменти од примерокот. Оценката за σ^2 е

$$\hat{\sigma}^2 = \frac{1}{2}(EX^2 - \hat{\theta}) \ . \ \blacksquare$$

11.3.2. Метод на максимална подобност

Методот на максимална подобност (Maximum Likelihood) бил воведен во 1922 и од теоретски аспект тој е најважниот општ метод за добивање оценувачи. Тој се базира на разгледување на примерокот како функција од непознатите параметри дефинирајќи функција на подобност (likelihood function) пропорционална на распределбата на примерокот

$$L(\theta_1, ..., \theta_m; X_1, X_2, \ldots, X_n) \sim f(x_1, x_2, \ldots, x_n; \theta_1, ..., \theta_m),$$

Функцијата на подобност го искажува степенот на соодветност придружен на различните вредности за θ_i да бидат вистински параметри на случајниот процес во светло на поедина реализација на примерокот X_1, X_2, …, X_n. Да забележиме дека $L(\cdot)$ е функција од θ_1, ..., θ_m, така што таа има различна димензија од $f(\cdot)$ којашто е функција од x_1, x_2, …, x_n. За поедноставно, понатаму ќе разгледуваме случаи со само еден параметар θ.

Формално, функцијата на подобност може да се дефинира како

$L(\theta; X_1, X_2, \ldots, X_n) : \theta \to [0, \infty)$, а

целта е да се определи конкретната вредност $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ таква што

$$L(\hat{\theta}; X_1, X_2, \ldots, X_n) = \max_{\theta} L(\theta; X_1, X_2, \ldots, X_n).$$

Кога $L(\cdot)$ е диференцијабилна, овој максимум може да се најде со диференцирање по θ и изедначување на 0,

$$\frac{\partial L(\theta; X_1, X_2, \ldots, X_n)}{\partial \theta} = 0, \text{ при што } \frac{\partial^2 L(\theta; X_1, X_2, \ldots, X_n)}{\partial \theta^2} < 0.$$

При претпоставка за независност на примерокот функцијата $L(\cdot)$ ја дефинираме со

$$L(\theta; X_1, X_2, \ldots, X_n) = f(X_1; \theta) f(X_2; \theta) \cdots f(X_n; \theta),$$

а во случај кога популацијата X е дискретна

$$L(\theta; X_1, X_2, \ldots, X_n) = p(X_1; \theta) p(X_2; \theta) \cdots p(X_n; \theta).$$

Често пати е позгодно да се максимизира логаритамот на $L(\cdot)$ со оглед на тоа што максимумот е ист ($L(\cdot)$ е позитивна, а логаритамот е монотона функција).

Проширувањето во случај на повеќе параметри е праволиниско. Имено, ако имаме m непознати параметри θ_i, $i = 1, 2, \ldots, m$, оценките со максимална подобност се добиваат од системот равенки

$$\frac{\partial \ln L(\theta_1, \theta_2, \ldots, \theta_m; X_1, X_2, \ldots, X_n)}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, m.$$

Оценките со максимална подобност имаат повеќе добри особини на кои ќе се навратиме понатаму.

ПРИМЕР 11.18 Најди оценки на непознатите параметри μ и σ^2 во нормалната распределба користејќи го методот на максимална подобност.

Решение

Логаритамот на функцијата на подобност дава

$$\ln L(\mu,\sigma^2;X_1,X_2,\ldots,X_n)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{1}{2}n\ln\sigma^2-\frac{1}{2}n\ln 2\pi\,.$$

Означуваме $\theta_1=\mu$ и $\theta_2=\sigma^2$ и го добиваме системот равенки од

$$\frac{\partial \ln L}{\partial\theta_1}=\frac{1}{\theta_2}\sum_{i=1}^{n}(X_i-\theta_1)=0\,,$$

$$\frac{\partial \ln L}{\partial\theta_2}=\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i-\theta_1)^2-\frac{n}{2\theta_2}=0\,.$$

Од равенките веднаш се добива дека

$$\theta_1=\frac{1}{n}\sum_{i=1}^{n}X_i\,,\qquad \text{т.е. } \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$$

$$\theta_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\theta_1)^2\,,\quad \text{т.е. } \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2=\frac{n}{n-1}S^2\,.$$

Ова се совпаѓа со оценките добиени со методот на моменти. ■

ПРИМЕР 11.19 Користејќи го методот на максимална подобност најди оценувач на θ кај рамномерната распределба.

Решение

Имајќи предвид дека густината на рамномерната распределба е

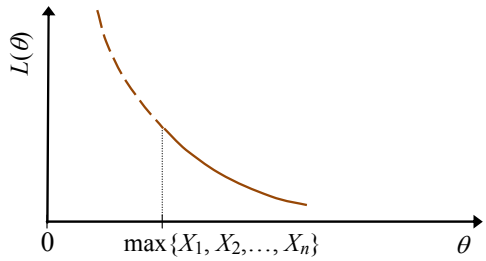
$$f(x;\theta)=\begin{cases}1/\theta, & 0\le x\le\theta\\ 0, & \text{во спротивно}\end{cases}$$

за функцијата на подобност добиваме

$$L(\theta;X_1,X_2,\ldots,X_n)=\left(\frac{1}{\theta}\right)^n,\ 0\le X_i\le\theta \text{ за сите } i.$$

Од условот $0\le X_i\le\theta$ следува дека вредностите на сите примероци X_i мораат да бидат помали или еднакви на θ. Тоа понатаму укажува дека само делот на кривата на десно од $\max\{X_1, X_2, \ldots, X_n\}$ е важечки (види ја сликата подолу што ја прикажува $L(\cdot)$). Оттука следува дека максимумот на $L(\theta; X_1, X_2, \ldots, X_n)$ се добива за

$$\hat{\theta}=\max\{X_1, X_2, \ldots, X_n\}.$$



Ова е подобра оценка од онаа добиена со методот на моменти. Да забележиме дека во овој случај максимумот на функцијата се јавува на границата на функцијата каде што изводот не е 0. ■

Вредноста на оценките добиени со методот на максимална подобност е во добрите особини што тие ги поседуваат кога примерокот е доволно голем. Со сосема благи услови наметнати на густината на распределба на популацијата X, исполнети се две важни карактеристики на ваквите оценки да кога $n \to \infty$,

1) Конзистентност, $E\hat{\theta} \to \theta$, и

2) Асимптотска ефикасност, $D\hat{\theta} \to \left(nE\left(\dfrac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right)^{-1}$.

Додатно, распределбата на $\hat{\theta}$ е и асимптотски нормална и инваријатна во смисла што ако $\hat{\theta}$ е максимално подобна оценка на θ, тогаш $g(\hat{\theta})$ е максимално подобна оценка на $g(\theta)$, за произволна дифернцијабилна еден-еден функција $g(\cdot)$. Од друга страна, дури и за умерено големи примероци, дисперзиите на оценките добиени со методот на максимална подобност може да бидат поголеми од оние добиени со методот на моменти. Ова укажува дека иако методот на максимална подобност има речиси оптимални асимптотски особини, тој може да биде доста несоодветен, т.е. може да дава слаби оценувачи во случаите кога примероците се со помала големина. Се разбира, големината на примерокот е еден од најважните елементи во статистиката, и неговото силно влијание на квалитетот на статистичките оценки и тестовите ќе го дискутираме и понатаму.

На крај да напоменеме дека барањето максимум на нелинеарна функција од повеќе променливи (што е пресметковната основа на овој метод) е често пати тежок проблем што бара приближни нумерички постапки за решавање. Оптимизацијата е една цела гранка во применетата математика што се занимава со слични проблеми.

11.3.3. Метод на најмали квадрати*

Концептот на најмали квадрати (least-sqares) е предложен како процедура за апроксимација на функции уште во 1805 година. Идејата е да се апроксимира

непозната функција $y = g(x)$ со функцијата $h(x) = \sum_{i=0}^{k} a_i \phi_i(x)$,

каде што $\phi_0(x)$, $\phi_1(x)$, …, $\phi_k(x)$ се згодно избрани функции,

(на пример: $\phi_0(x) = 1$, $\phi_1(x) = x$, …, $\phi_k(x) = x^k$),

а a_0, a_1, …, a_k така избрани броеви што обезбедуваат максимална блискост на $g(x)$ и $h(x)$ за некој дискретен домен D од $n > k$ точки.

Попрецизно, за даден домен точки $D = \{(x_j, y_j), j = 1, 2, \ldots, n\}$, параметрите a_0, a_1, …, a_k се избираат така да се минимизира целната функција

$$z(a_0, a_1, \ldots, a_k) = \sum_{i=1}^{n} (y_i - h(x_i))^2 = \sum_{i=1}^{n} \left(y_i - \sum_{i=0}^{k} a_i \phi_i(x_i) \right)^2 .$$

Да забележиме дека тука немаме вклучено никакви веројатносни претпоставки.

ПРИМЕР 11.20 Користејќи го методот на најмали квадрати најди ја најдобрата права линија за множество од n точки $(x_j, y_j), j = 1, 2, \ldots, n$.

Решение

Во ваков едноставен случај имаме $k = 1$, $\phi_0(x) = 1$, $\phi_1(x) = x$, а целната функција е од облик

$$z(a_0, a_1) = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 .$$

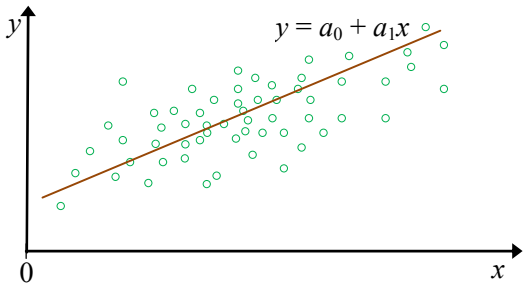
Функцијата $z(\cdot)$ е диференцијабилна, па минимумот го бараме со изедначување на изводот на 0,

$$\frac{\partial z}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0 , \quad \frac{\partial z}{\partial a_1} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) x_i = 0 .$$

Решението на овој систем равенки е

$$a_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \overline{xy}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \quad a_0 = \overline{y} - a_1 \overline{x} , \text{ каде што } \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i , \ \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Сликата подолу прикажува пример на линеарна апроксимација по методот на најмали квадрати.



Геометриски тоа изгледа како повлекување линија, централно низ точките. ■

Како методологија за апроксимација на функци, овој метод бил користен од почетокот на 19 век. Од аспект на условни очекувања тој веќе беше разгледуван во поглавјето за коефициентот на корелација (поглавје 5.5.1) под термин на најдобар линеарен предвидувач (види го примерот 5.6).

Изворно, оправданието за користење на методот на најмали квадрати бил во фактот што кога апроксимативната функција е константа, $h(x) = a_0$, вредноста на a_0 што ја минимизира

$$z(a_0) = \sum_{i=1}^{n} (y_i - a_0)^2 \text{ е едноставно просекот } a_0 = \frac{1}{n}\sum_{i=1}^{n} y_i \,.$$

Во тоа време, просекот се сметал за најдобриот начин да се сумираат информациите содржани во n податоци.

Како да се воведе методот на најмали квадрати во статистичката оценка на непознати параметри?

Ние во претходната глава воведовме концепт на статистички генератор што прави декомпозиција на случајните променливи во облик

$$y_i = E(y_i \mid \mathcal{D}_i) + u_i, \;\; i = 1,2,\ldots,n, \;\; (\text{дисперзијата } Dy_i < \infty),$$

каде што $\mathcal{D}_i$ е условна информација скриена во примерокот. Целта е $\mathcal{D}_i$ да се избере така што деформацијата (disturbance term) u_i да се сведе на минимум, или еквивалентно да се максимизира $E(y_i \mid \mathcal{D}_i)$. Значи ние би сакале

$$E(u_i \mid \mathcal{D}_i) = 0.$$

Врз база на веројатносна претпоставка, компонентата

$$E(y_i \mid \mathcal{D}_i) = g(x_i;\, \theta), \;\; i = 1, 2, \ldots, n,$$

се претставува како параметарска густина на распределба со непознат параметар. Методот на најмали квадрати сугерира минимизација на

$$z(\theta) = \sum_{i=1}^{n} (y_i - g(x_i;\theta))^2$$

по θ, што ја дава бараната оценка.

ПРИМЕР 11.21 Според методот на најмали квадрати, најди оценувач на θ во Бернулиевиот модел $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $0 \le \theta \le 1$, $x = 0, 1$.

Решение

Статистичкиот генератор за Бернулиевиот модел е

$$X_i = E(X_i \mid \mathcal{D}_i) + u_i, \quad i = 1, 2, \ldots, n,$$

каде што $\mathcal{D}_i = \{\Omega, \varnothing\}$ (неинформативно множество), откаде што веднаш следува $E(X_i \mid \mathcal{D}_i) = EX_i = \theta$. Методот на најмали квадрати за оценка на θ, базиран на примерок $X_1, X_2, \ldots, X_n$ треба да ја минимизира

$$z(\theta) = \sum_{i=1}^{n} (X_i - \theta)^2 .$$

Изедначувајќи го првиот извод на 0 добиваме

$$\frac{dz}{d\theta} = -2\sum_{i=1}^{n} (X_i - \theta) = 0 \text{, што дава } \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Овој оценувач се совпаѓа со оценувачите на θ што се добиваат со другите методи. ■

Во случај на нормалниот модел, статистичкиот генератор е од ист облик како кај Бернулиевиот модел

$$X_i = \mu + u_i, \quad i = 1, 2, \ldots, n,$$

така што методот на најмали квадрати повторно дава $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Да забележиме дека методот на најмали квадрати не сугерира оценка за σ^2.

Генерално, методот на најмали квадрати априори не обезбедува оценки со така добри особини, како методите на максимална подобност и моменти. Покажано е дека оценките со овој метод ги имаат особините на конзистентност и асимптотска нормалност.

ЗАДАЧИ

1. За да се испита евентуалниот број на деца за упис во едно училиште земен е примерок на бројот на деца на 100 семејства што живеат во близина. Резултатите се дадени во следната табела:

Број на деца	0	1	2	3	4	5	6	7
Семејства	21	24	30	16	4	4	0	1

   Оцени го просекот и дисперзијата на податоците од примерокот.

2. Објасни накусо што се прави кога се составува точкаст оценувач на непознат параметар. Зошто оценувачот е случајна променлива?

3. Нека имаме три оценувачи $\hat{\theta}_1$, $\hat{\theta}_2$ и $\hat{\theta}_3$ на непознат параметар θ. Знаеме дека $E\hat{\theta}_1 = E\hat{\theta}_2 = \theta$, а $E\hat{\theta}_3 \neq \theta$ и $D\hat{\theta}_1 = 12$, $D\hat{\theta}_2 = 10$ и $E(\hat{\theta}_3 - \theta)^2 = 6$. Спореди ги овие три оценувачи. Кој би го преферирал?

4. Нека $X_1, X_2, \ldots, X_n$ е примерок земен од нормална популација со очекување μ и дисперзија σ^2. Нека X_{min} и X_{max} се најмалиот и најголемиот податок во примерокот.

   а) Дали $(X_{min} + X_{max})/2$ е центриран оценувач на μ и колкава е неговата дисперзија?

   б) Дали овој оценувач е подобар од просекот на примерокот $\overline{X}$?

5. Примерок со 2 податока X_1 и X_2 е земен од популација X со распределба $f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$, $x \geq 0$, каде што θ е непознат параметар. Предложени се два оценувачи на θ, $\hat{\theta}_1 = (X_1 + X_2)/2$ и $\hat{\theta}_2 = \frac{4}{\pi}\sqrt{X_1 X_2}$. Кој од оценувачите е подобра во однос на центрираност и помала дисперзија?

6. Објасни го концептот на доволен оценувач.

7. Дали статистиката $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ за оценка на μ во нормалниот модел е доволна статистика?

8. Објасни ја идејата позади минимална доволна статистика и во каква врска е таа со центрираните оценки.

9. Геометриската средина $\sqrt[n]{X_1 X_2 \cdots X_n}$ се предлага како оценувач на медијаната на логнормално распределена случајна променлива X. Дали тој е центриран? Дали е центриран кога $n \to \infty$?

10. Од n_1 случајно избрани мажи постари од 18 години, X_1 се пушачи, додека од n_2 случајно избрани жени постари од 18 години, X_2 се пушачи. Нека p_1 и p_2 се веројатностите дека случајно избран маж и жена се пушачи.

    а) Покажи дека $(X_1/n_1) - (X_2/n_2)$ е центриран оценувач на $p_1 - p_2$;

    б) Која е стандардната грешка на оценувачот во а);

    в) Како би се користеле добиените вредности x_1 и x_2 (за X_1 и X_2) за оценка на стандардната грешка;

    г) Ако $n_1 = n_2 = 200$, $x_1 = 68$, $x_2 = 52$, пресметај ја оценката за $p_1 - p_2$ и стандардната грешка на оценката;

11. Нека $X_1, X_2, \ldots, X_n$ е примерок земен од распределбата на Реили (Rayleigh)

$$f(x;\theta) = \frac{x}{\theta} e^{-x^2/(2\theta)}, \quad x > 0.$$

    а) Покажано е дека $EX^2 = 2\theta$. Користејќи го ова состави центриран оценувач за θ, базиран на $\sum_{j=1}^{n} X_j^2$;

    б) Оцени го θ од следните податоци за стресот на перките на турбина под специфични услови: 16.88, 10.23, 4.59, 6.66, 13.68, 14.23, 19.87, 9.40, 6.51 и 10.95.

12. Во следните 4 распределби, користејќи го неравенството на Крамер-Рао, определи ја долната граница на дисперзијата на оценувачите на непознатиот параметар θ.

    а) $f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, x \geq 1;$ б) $f(x;\theta) = \theta \cdot x^{\theta-1}, 0 \leq x \leq 1, \theta > 0;$

    в) $f(x;\theta) = \theta^x (1-\theta)^{1-x}, x = 0, 1;$ г) $f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}, x = 0, 1, 2, \ldots .$

13. Нека $X_1, X_2, \ldots, X_n$ е примерок земен од популација X со гама распределба $f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$, за која знаеме дека $EX = \alpha\beta$, а $DX = \alpha\beta^2$ (поглавје 5.3). Најди оценувачи за α и β со методот на моменти.

14. Се тестираат компоненти на една електронска компанија на доверливост. Нека p и $1 - p$ се веројатностите една компонента да биде "успешна" или

"неуспешна". Ако X е бројот на испитани компоненти до првата "неуспешна", X има геометриска распределба $f(k; p) = (1 - p)p^{k-1}$, $k = 1, 2, \ldots$. Ако $X_1, X_2, \ldots, X_n$ е примерок од компонентите, определи го

а) Оценувачот на p со максимална подобност;

б) Оценувачот на p со максимална подобност на $p(X > 9)$. Забележи дека $p(X > 9) = \sum_{k=1}^{9}(1-p)p^{k-1}$.

15. Нека X има поместена експоненцијална распределба $f(x; a) = e^{\alpha - x}$, $x \geq a$. Врз база на примерок со големина n определи ги оценувачите со максимална подобност и моменти за параметарот a.

16. Нека примерокот $X_1, X_2, \ldots, X_n$ е земен од популација со поместена експоненцијална распределба со густина

$$f(x;\lambda,a) = \begin{cases} \lambda e^{-\lambda(x-a)}, & x \geq a \\ 0, & \text{во спротивно} \end{cases}.$$

Определи го оценувачот со максимална подобност за λ и пресметај ја оценката за примерокот: 3.11, 0.64, 2.55, 2.20, 5.44, 3.42, 10.39, 8.93, 17.82 и 1.30.

17. Нека $X_1, X_2, \ldots, X_n$ е примерок од обопштена негативна биномна распределба со параметри r и p (види поглавје 4.1 и 5.3). Состави точкасти оценувачи за овие параметри користејќи го методот на моменти.

18. Користи ја негативната биномна распределба како модел за бројот на голови по натпревар во хокејската (NHL) лига на САД. Следните податоци се земени од 420 натпревари:

Голови	0	1	2	3	4	5	6	7	8	9	10
Натпревари	29	71	82	89	65	45	24	7	4	1	3

Пресметај ги оценките на r и p според формулите од претходната задача.

19. Екологист избира n дисјунктни региони $R_1, R_2, \ldots, R_n$ во кои го испитува (брои) бројот на различни растенија. Бројот на настани што се случуваат во дводимензионални области добро се моделира со Пуасонова распределба. Попрецизно, бројот на настани што се случуваат во регионот R со плоштина a_R има Пуасонова распределба со параметар $\lambda \cdot a_R$, т.е.

$f(x;\lambda) = \frac{(\lambda \cdot a_R)^x}{x!} e^{-\lambda \cdot a_R}$, каде што λ е очекуваниот број настани по единечна плоштина. Најди оценувач на λ по методот на максимална подобност.

20. Нека примерокот $X_1, X_2, \ldots, X_n$ е земен од популација со $\chi 2$ распределба со непознат параметар a и густина од облик

$$f(x;a) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}, \quad x > 0$$ (од гама распределба за $\alpha = p$ и $\beta = 1/a$).

а) Најди ја оценувач на a по методот на максимална подобност;

б) Дали добиениот оценувач е центриран? Ако не е, обиди се да го "поправиш" така тој да стане центриран.

12

Интервални оценки

Точкастите оценки не се доволно информативни бидејќи се сведуваат на обичен број и не даваат информација за прецизноста и доверливоста на оценката. На пример, нека сме оцениле дека $\mu =$ "просечната потрошувачка на гориво на едно возило" е $\bar{x} = 6.4$ литри (на 100 километри). Поради варијабилноста на примерокот, практично никогаш нема да се добие $\mu = \bar{x}$, а самата точкаста оценка $\bar{x}$ не кажува колку е таа блиска до μ. Дали просекот μ е меѓу 6.2 и 7.2 или пак е поверојатно да биде меѓу 6 и 6.8? Токму ваков интервал, во кој со висока веројатност се наоѓа непознатиот параметар се нарекува *интервал на доверба* или *интервална оценка*. Може да се смета за изненадување дека определувањето на вакви интервали е доста лесно, и дека тоа се прави со истите податоци што се користат и за точкастите оценки. Со интервалните оценки се обидуваме да извлечеме повеќе (информации) од примерокот отколку со точкастите оценки.

Дефиниција 12.1 Нека $L_1 = L_1(X_1, X_2, \ldots, X_n)$ и $L_2 = L_2(X_1, X_2, \ldots, X_n)$ се две статистики од примерокот $X_1, X_2, \ldots, X_n$ земен од популацијата X со густина на распределба $f(x; \theta)$, каде што θ е непознат параметар. Нека $L_1 < L_2$ со веројатност 1. Интервалот (L_1, L_2) се нарекува $100(1 - \alpha)\%$-ен интервален оценувач на θ ако $p(L_1 < \theta < L_2) = 1 - \alpha$.

За α обично се земаат мали вредности 0.1, 0.05, 0.01 или дури 0.001 што даваат високи веројатности на доверба $1 - \alpha$ од 0.9, 0.95, 0.99, или 0.999. Да забележиме дека:

1) Границите на интервалот L_1 и L_2 се функции од примерокот, така што за различни реализации на примерокот интервалните оценки варираат во позиција и ширина;

2) За даден примерок, постојат многу парови статистики L_1 и L_2 што даваат доверба $1 - \alpha$. Во многу случаи, симетричните интервали околу θ имаат најдобар однос на доверба и ширина на интервалот;

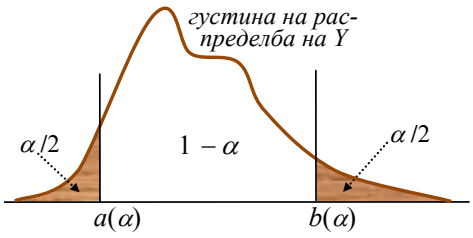
3) За дадена доверба, јасно е дека најдобар интервал е најтесниот, а бидејќи ширината на интервалот $L = L_2 - L_1$ е случајна променлива, логично би можело да се бара "минималната очекувана ширина" како оптимум. Проблемот е што таков минимум не мора да постои за сите вредности на θ.

Интервалите на доверба се конструираат така што се наоѓа погодна случајна променлива $Y = h(X_1, X_2, \ldots, X_n; \theta)$ што е функција и од параметарот θ и од примерокот, а чијашто распределба е позната и не зависи од θ, ниту од други непознати параметри. Сега, поради познатата распределба на $h(\cdot)$, лесно е да се најде интервал таков што

$$p(a < Y < b) = 1 - \alpha, \quad \text{т.е.}$$

$$p(a < h(X_1, X_2, \ldots, X_n; \theta) < b) = 1 - \alpha,$$

каде што a и b не зависат од θ (види слика).



Значи a и b се точки од кои налево и надесно плоштините под густината на распределбата се $\alpha/2$. Од неравенството

$$a < h(X_1, X_2, \ldots, X_n; \theta) < b$$

со обични алгебарски манипулации се добива $L_1(X_1, X_2, \ldots, X_n; a) < \theta < L_2(X_1, X_2, \ldots, X_n; b)$, од што го добиваме интервалниот оценувач

$$p(L_1(X_1, X_2, \ldots, X_n; a) < \theta < L_2(X_1, X_2, \ldots, X_n; b)) = 1 - \alpha.$$

На пример, нека Y има нормална распределба (или приближно нормална) и нека $\hat{\theta}$ е центриран оценувач на θ. Ако се има на располагање приближната стандардна девијација $\sigma_{\hat{\theta}}$ на $\hat{\theta}$, веднаш може да дефинираме случајната променлива $Z = (\hat{\theta} - \theta) / \sigma_{\hat{\theta}}$ што ќе има стандардна нормална распределба $Z(0,1)$. Оттука праволиниски имаме

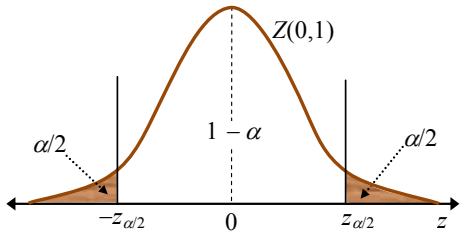
$p\left(-z_{\alpha/2} < \dfrac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) = 1-\alpha$, што понатаму директно дава

$p(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1-\alpha$ – интервален оценувач за θ. Тука главниот проблем е определувањето на $\sigma_{\hat{\theta}}$. Да забележиме дека за поголеми n, секогаш можеме $\sigma_{\hat{\theta}}$ да го замениме со соодветната точкаст оценувач $S_{\hat{\theta}}$.

12.1. Интервални оценки за просекот

Нека $X_1, X_2, \ldots, X_n$ е примерок земен од популација X со нормална распределба, со непознато μ и познато σ. Тогаш (точкастиот оценувач на μ) $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ има нормална распределба $Z(\mu, \dfrac{\sigma}{\sqrt{n}})$. Ние можеме да ја "стандардизираме" $\overline{X}$ со вадење на просекот и делење со стандардната девијација, добивајќи ја случајната променлива

$Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ што има стандардна $Z(0,1)$ распределба.



Сега може да ставиме

$$p\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

каде што $-z_{\alpha/2}$ и $z_{\alpha/2}$ се точки од кои налево и надесно, плоштината под густината на стандардната нормална распределба е $\alpha/2$. Решавајќи ја горната неравенка по μ, го добиваме интервалниот оценувач

$$p\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha\ .$$

Се разбира, за конкретни вредности на примерокот $x_1, x_2, \ldots, x_n$, точкастиот оценувач $\overline{X}$ се заменува со оценката $\bar{x}$.

ПРИМЕР 12.1 Испитувањето на брзината на трансакциски одзив на еден компјутерски систем е нормално распределена случајна променлива со стандардна девијација од 25 милисекунди. По воведување на нова верзија на оперативен

систем, пожелно е повторно да се оцени просечниот одзив μ во "новиот" систем. Земен е примерок од 28 трансакции при што е измерено просечно време на одзив од 118.6 милисекунди. Под претпоставка дека стандардната девијација повторно е $\sigma = 25$ милисекунди, определи 95% интервал на доверба за просекот на времето на одзив. Колкав примерок треба да се земе за ширината на интервалот да биде најмногу 10 милисекунди?

Решение

Имајќи предвид дека $z_{\alpha/2} = z_{0.5/2} = z_{0.25} = 1.96$, добиваме

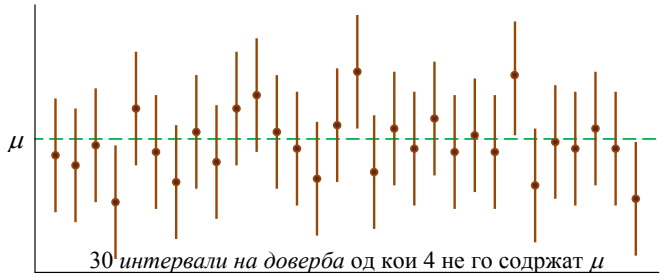
$$p\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = p\left(118.6 - 1.96\frac{25}{\sqrt{28}} < \mu < 118.6 + 1.96\frac{25}{\sqrt{28}}\right)$$

$$= p(109.34 < \mu < 127.86) = p(\mu \in [109.34, 127.86]) = 0.95.$$

Значи со 95% шанси, просечното време на одзив е меѓу 109.34 и 127.86 милисекунди.

Од барањето ширината на интервалот да биде најмногу 10 имаме дека $2{\cdot}1.96{\cdot}25 / \sqrt{n} \le 10$, што дава неравенка по n, т.е. $\sqrt{n} \ge 2{\cdot}1.96{\cdot}25/10 = 9.80 \Rightarrow n \ge 96.04$, т.е. $n = 97$. ■

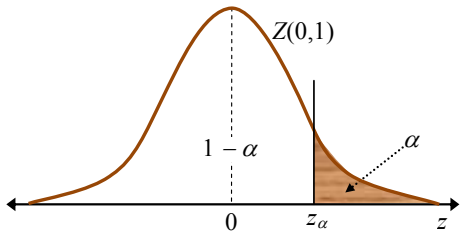
Да забележиме дека интерпретацијата на интервалната оценка како веројатност со која просекот μ припаѓа на даден интервал *не е најпрецизна*. Имено, за секој примерок интервалот е различен, бидејќи нормално, за секој примерок се добива различна вредност за $\bar{x}$. Така, точната интерпретација е дека просекот припаѓа во $100(1 - \alpha)\%$ од генерираните интервали. На пример, на следната слика 26/30 = 86.7% од интервалите го содржат μ.



И покрај ваквата "непрецизност", за поедноставно ние и понатаму слободно ќе користиме термини од облик "веројатност параметарот да е во дадениот интервал". Алтернативна конструкција и интерпретација на интервалите на доверба е преку Баесовиот пристап што тука нема да го разгледуваме.

Ширината на интервалната оценка е $w = 2z_{\alpha/2}\sigma/\sqrt{n}$ што решено по n дава $n = (2z_{\alpha/2}\sigma/w)^2$. Јасно е дека со зголемување на довербата $1 - \alpha$ (намалување на α), ширината на интервалот расте (при фиксно n и σ). Важи и обратното, ако дозволиме растење на ширината на интервалот, довербата расте. Оваа "трговија" може да се наруши само со зголемување на примерокот n. Имено, единствен начин истовремено да се добие потесен интервал и повисока доверба е да се зголеми примерокот. Очигледно е кога $n \to \infty$, ширината на интервалот се стреми кон 0, $w \to 0$.

Понекогаш не е потребен двостран интервал, туку само едностран интервал на доверба. На пример, потребна е долната граница на животниот век или горната граница на времето на реакција на некоја компонента. Во таков случај, горната граница е ∞ или долната граница е $-\infty$. Додатно, $z_{\alpha/2}$ се заменува со z_α (види слика).



Ставаме $p\left(\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha$,

што дава $p\left(\mu < \overline{X} + z_\alpha \dfrac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$.

Сосема идентично може да се добие едностраниот интервал оддолу $p\left(\mu > \overline{X} - z_\alpha \dfrac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$. На идентичен начин може да се добијат едностранн интервали на доверба и за други параметри и соодветни распределби.

Претпоставката за нормалност на популацијата е често пати разумна појдовна точка кај статистичките оценки. Од друга страна, ако вредноста на μ е непозната, вообичаено не е многу логично σ да биде позната. Вредноста на просекот нормално претходи на раштрканоста на податоците околу него, дадена со дисперзијата. Во случаи кога примерокот е доволно голем (обично $n \geq 30$), претпоставката за нормална распределба не е потребна (поради централната гранична теорема), ниту дисперзијата да е позната (S сосема добро ја заменува σ).

ПРИМЕР 12.2 Во една статија објавени се резултати од студија за загадувањето на рибите со жива во езерата на Флорида, САД. Испитани се примероци на риби од 53 езера во Флорида, при што се добиени следните концетрации на жива во мискулите изразени во *ppm*:

1.230	1.330	0.040	0.044	1.200	0.270	0.490	0.190	0.830	0.810	0.710
0.500	0.490	1.160	0.160	0.270	0.050	0.150	0.190	0.770	1.080	0.980
0.630	0.560	0.410	0.730	0.590	0.340	0.340	0.840	0.500	0.340	0.280
0.340	0.750	0.870	0.560	0.170	0.180	0.190	0.040	0.490	1.100	
0.100	0.210	0.860	0.520	0.650	0.270	0.940	0.400	0.430	0.250.	

Определи 99% интервал на доверба за просекот на концетрацијата на жива во рибите. Колкав примерок треба да се земе за ширината на интервалот да биде најмногу 10% (од просекот)?

Решение

Од податоците лесно се добива дека $\bar{x} = 0.5250$ и $s = 0.3486$. Имајќи предвид дека $z_{\alpha/2} = z_{0.005} = 2.5758$, бараната интервална оценка е

$$p\left(0.5250 - 2.5758\frac{0.3486}{\sqrt{53}} < \mu < 0.5250 + 2.5758\frac{0.3486}{\sqrt{53}}\right) =$$

$$= p(0.4017 < \mu < 0.6483) =$$

$$= p(\mu \in (109.34, 127.86)) = 0.99.$$

За ширина од 10%, $n \geq (2 \cdot 2.5758 \cdot 0.3486/(0.0525/2))^2 \approx 706$. ■

Кога примерокот е мал, еден начин да се продолжи е да се направи претпоставка за формата на распределбата на популацијата, и врз база на таквата претпоставка да се определи инервалната оценка. На пример, интервалната оценка може да се направи врз база на гама или веибул распределба на сосема идентичен начин. Сепак, поради големата распространетост, ние тука ќе го разгледаме случајот на мал примерок со нормална распределба со непозната σ. Кај случајот со доволно голем примерок n, случајната променлива $Z = (\bar{X} - \mu)/(S/\sqrt{n})$ има стандардна нормална распределба. Но во ситуација на мало n, S повеќе не е добра апроксимација на σ. Тоа значи дека S ќе се разгледува како случајна променлива, што е корен на збир на квадрати на случајни променливи со нормална распределба, т.е. корен од $\chi 2$ распределба со $n - 1$ степени на слобода (по соодветна нормализација). Еден степен на слобода се губи од условот $\sum(\bar{X} - X_i) = 0$. Како што знаеме од поглавјето 4.2, случајната променливата $T = Z\sqrt{n}/\sqrt{\chi^2} = \frac{X-\mu}{\sigma/\sqrt{n}}\sqrt{n-1}/\sqrt{(n-1)S^2/\sigma^2} =$

$= \frac{X-\mu}{S}\sqrt{n}$ сега ќе има студентова распределба со $n - 1$ степени на слобода. Оттука веднаш следува

$p\left(-t_{\alpha/2} < \dfrac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$, што дава интервален оценувач

$$p\left(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha .$$

ПРИМЕР 12.3 Бројот на жртви при евакуација од пожари во 14 хотели низ САД биле: 5, 36, 5, 8, 10, 4, 7, 8, 5, 9, 4, 0, 16, 0. Јасно е дека бројот на жртви зависи од многу фактори меѓу кои секако е и големината на хотелот. На пример, 36 жртви имало во огромниот хотел MGM во Лас Вегас. Под претпоставка дека бројот на жртви има приближно нормална распределба, најди 98% и 99% интервал на доверба за просечниот број жртви. Колкава е довербата за интервал со ширина 6?.

Решение

Од податоците добиваме $\bar{x} = 117/14 = 8.36$ и $s = 8.94$. Имајќи предвид дека $t_{\alpha/2} = t_{0.01} = 2.65$ за 13 степени на слобода, бараната интервална оценка е

$$p\left(8.36 - 2.65\frac{8.94}{\sqrt{14}} < \mu < 8.36 + 2.65\frac{8.94}{\sqrt{14}}\right) = p(2.03 < \mu < 14.69) = 0.98.$$

За доверба 99%, $t_{\alpha/2} = t_{0.005} = 3.012$, па интервалот е поширок

$$p\left(8.36 - 3.012\frac{8.94}{\sqrt{14}} < \mu < 8.36 + 3.012\frac{8.94}{\sqrt{14}}\right) = p(1.16 < \mu < 15.56) = 0.99.$$

За интервал со ширина до 6, треба $t_{\alpha/2}\dfrac{8.94}{\sqrt{14}} \le 3$, т.е. $t_{\alpha/2} \le 3\dfrac{\sqrt{14}}{8.94} = 1.2555$.

Оттука следува дека $\alpha/2 \ge 0.1$, т.е. $\alpha \ge 0.2$, што дава доверба $1 - \alpha \le 0.8 = 80\%$.

Да забележиме дека тука не може да се зголемува примерокот (се разбира нема да подметнуваме пожари и броиме жртви). ■

12.2. Интервал на предвидување

Во многу апликации, потребно е да се предвиди вредноста на случајната променлива што таа ќе ја добие во иднина. Нека $X_1, X_2, \ldots, X_n$ е примерок земен од популацијата X со нормална распределба. Целта е да се предвиди вредноста X_{n+1}, т.е. една следна вредност на примерокот. Точкастиот оценувач на X_{n+1} е $\overline{X}$, од што следува дека грешката на предвидувањето е $\overline{X} - X_{n+1}$. Очекувана вредност на грешката е $E(\overline{X} - X_{n+1}) = E\,\overline{X} - EX_{n+1} = \mu - \mu = 0$. Бидејќи X_{n+1} е независна од $X_1, X_2, \ldots,$

X_n, таа е независна и од $\overline{X}$, па дисперзијата на грешката на предвидувањето е

$$D(\overline{X} - X_{n+1}) = D\overline{X} + DX_{n+1} = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right).$$

Грешката на предвидувањето $\overline{X} - X_{n+1}$, како линеарна комбинација на независни случајни променливи со нормална распределба има нормална распределба, па случајната променлива

$$Z = \frac{(\overline{X} - X_{n+1}) - 0}{\sqrt{\sigma^2(1+1/n)}} = \frac{\overline{X} - X_{n+1}}{\sigma\sqrt{1+1/n}}$$ има $Z(0,1)$ распределба.

Ако σ се замени со S, за мал примерок случајната променлива

$$T = \frac{\overline{X} - X_{n+1}}{S\sqrt{1+1/n}}$$ добива приближно студентова распределба со $n - 1$

степени на слобода. Оттука на стандарден начин се добива интервалот на предвидувањето

$$p\left(\overline{X} - t_{\alpha/2}S\sqrt{1+\frac{1}{n}} < \mu < \overline{X} + t_{\alpha/2}S\sqrt{1+\frac{1}{n}}\right) = 1 - \alpha\,.$$

ПРИМЕР 12.4 Количеството маснотија во примерок од 10 сендвичи со вршла се измерени на: 25.2, 21.3, 22.8, 17.0, 29.8, 21.0, 25.5, 16.0, 20.9 и 19.5 грама. Под претпоставка дека содржината на маснотиите е со приближно нормална распределба, најди 95% интервал на доверба за просечната содржина на маснотии во сендвичите, како и интервал на предвидување на количеството маснотии во следниот сендвич.

Решение

Од податоците се добива дека $\bar{x} = 21.90$ и $s = 4.134$. Имајќи предвид дека $t_{\alpha/2} = t_{0.025} = 2.262$ за 9 степени на слобода, бараната интервална оценка е

$$p\left(21.9 - 2.262\frac{4.134}{\sqrt{10}} < \mu < 21.9 + 2.262\frac{4.134}{\sqrt{10}}\right) = p(18.94 < \mu < 24.86) = 0.95.$$

Ако сега земеме еден сендвич за јадење, количеството маснотии y што ќе го изедеме може да се процени со

$$p\left(y \in 21.9 \pm 2.262 \cdot 4.134\sqrt{1+\frac{1}{10}}\right) = p(12.09 < y < 31.71) = 0.95.$$

Очигледно овој интервал е многу поширок (повеќе од 3 пати) од интервалот за просекот. Зошто е тоа така? Прво да забележиме дека грешката кај предвидувањето $\overline{X} - X_{n+1}$ е разлика меѓу две случајни променливи, додека кај интервалот на доверба грешката $\overline{X} - \mu$ е разлика меѓу случајна променлива и фиксна, но непозната вредност. Јасно е дека варијабилноста во првиот случај е поголема. Кога n расте ($n \to \infty$), интервалот на доверба се стеснува во една вредност μ, и тогаш интервалот на предвидување очигледно се сведува на ($\mu - z_{\alpha/2}\sigma$, $\mu + z_{\alpha/2}\sigma$). ■

Интервалите на доверба за μ во случај на мал примерок се базирани на студентовата распределба и се покажува дека тие се робусни при мали или дури умерени отстапувања од нормалната распределба. Сепак, ако n е многу мало и распределбата на популацијата е значително "не-нормална", вистинската интервална оценка може да биде многу различна од онаа што е добиена со студентовата распределба. На пример, добиен 95% интервал може објективно да биде 86% интервал, што е доста "незгодно" кога се донесува одлука врз база на оценката. Ситуацијата е уште полоша ако се работи за интервали на предвидување коишто се цврсто врзани за нормалната распределба.

Постојат одредени алтернативни постапки за добивање на интервали на доверба при значителни отстапувања од нормалната распределба. Одлична референца за таквите случаи е [Gerald, Meeker 1991].

12.3. Интервални оценки за пропорцијата

Нека p означува пропорција на "поволни случаи" во популацијата, т.е. релативен број објекти со определено својство. Се зема примерок со големина n, и кога n е мало во споредба со големината на популацијата случајната променлива X = "број на поволни случаи во примерокот" има биномна распределба дадена со законот $p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ со $EX = n{\cdot}p$ и $\sigma_X = \sqrt{n \cdot p(1-p)}$. Уште повеќе, за доволно големо n ($n{\cdot}p \geq 10$ и $n{\cdot}(1{-}p) \geq 10$), двете случајни променливи X и $\hat{P} = X/n$ имаат приближно нормална распределба. Ако се има предвид дека точкастата оценка $\hat{P} = X/n$ на p е центрирана и дека дисперзијата на биномната распределба е $p(1 - p)$, нејзината стандардна девијација е $\sigma_{\hat{P}} = \sqrt{p(1-p)/n}$. Значи случајната променлива

$Z = \dfrac{\hat{P} - p}{\sqrt{p(1-p)/n}}$ има приближно $Z(0,1)$ распределба, од каде што веднаш следува дека

$p(-z_{\alpha/2} < \dfrac{\hat{P} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}) = 1 - \alpha$. Оттука, решавајќи го квадратното неравенството по p, добиваме комплициран интервал на доверба

$p \in \dfrac{\hat{P} + z^2_{\alpha/2}/2n}{1 + z^2_{\alpha/2}/n} \pm z_{\alpha/2} \dfrac{\sqrt{\hat{P}(1-\hat{P})/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}$. Во пракса, сметаме дека за доволно големо n, $z^2_{\alpha/2}/n$ е занемарливо, па конечно имаме

$$p(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}) = 1 - \alpha .$$

ПРИМЕР 12.5 Астронаутите често искусуваат моменти на дезориентираност за време на нивното движење низ летало без гравитација. Како компензација, членовите на екипажот во голема мера се зависни од визуелните информации. Емпириско истражување било спроведено со цел да се утврди ефектот од употребата на светли бои како помош за ориентација. Деведесет студенти, лежејќи на грб во темница, биле дезориентирани (со поставување на ротаричка платформа). Над нив бил поставен диск кој ротира со помала брзина од онаа на платформата и го зазема целото видно поле. Половина од дискот била обоена со посветла боја од останатата половина. Студентите имале за задача да кажат "СТОП" во моментот кога веруваат дека се во вистинска позиција - бојата на дискот во тој момент се бележела. Од 90 студенти, 58 ја одбрале посветлата боја на дискот.

Користејќи ги овие информации одреди ја вистинската пропорција на субјекти кои ја употребиле светлата боја како ориентир. Конструирај 95% интервал на доверба.

Решение

$\hat{p} = \dfrac{m}{n} = \dfrac{58}{90} = 0.64$, $z_{\alpha/2} = z_{0.025} = 1.96$, па

$$p\left(0.64 - 1.96\sqrt{\frac{0.64 \cdot 0.36}{90}} < p < 0.64 + 1.96\sqrt{\frac{0.64 \cdot 0.36}{90}}\right) = p(0.541 < p < 0.739) =$$

0.95, при што интервалот е валиден бидејќи не содржи 0 или 1.

Бидејќи $p = 0.64$ е поголемо од $\alpha/2 = 0.025$, може да се заклучи дека мнозинството на студенти ќе ја одберат посветлата боја како знак дека се во вистинска позиција. ■

Едностраните интервални оценувачи се добиваат праволиниски, како и во случај на очекувањето,

$$p\left(p > \hat{P} - z_\alpha\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = 1-\alpha \quad \text{и} \quad p\left(p < \hat{P} + z_\alpha\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = 1-\alpha\,.$$

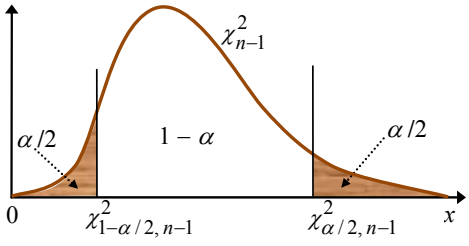
12.4. Интервални оценки за дисперзијата

Иако вообичаено заклучоците во врска со дисперзијата и стандардната девијација на популацијата се помалку интересни од просекот или пропорцијата, тие во многу ситуации се покажуваат како не помалку важни. Типични примери се ситуациите кога треба да се оценат варијациите во крајниот производ, т.е. дали е тој во рамките на стандардите.

Нека $X_1, X_2, \ldots, X_n$ е примерок земен од популацијата X со распределба $Z(\mu, \sigma^2)$. Тогаш случајната променлива

$\dfrac{(n-1)S^2}{\sigma^2}$ има χ^2 распределба со $n-1$ степени на слобода.

Оваа случајната променлива е функција од параметарот σ^2 и од примерокот, има позната распределба и не зависи од други непознати параметри. Значи таа е погодна за креирање на интервална оценка за σ^2. Од друга страна, χ^2 е несиметрична распределба, па определување на точките за област со плоштина $1-\alpha$, ($\alpha/2$ на левата и десната опашка), не е тривијално. Сепак, за χ^2 распределба, може да се определат овие две точки како функции од α, и тоа се $\chi^2_{1-\alpha/2}$ и $\chi^2_{\alpha/2}$ (види ја сликата).



Сега може да ставиме

$$p\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$$

каде што $\chi^2_{1-\alpha/2}$ и $\chi^2_{\alpha/2}$ се точките од кои налево и надесно плоштината под густината на χ^2 распределбата е $\alpha/2$. Оттука, решавајќи ги неравенките по σ^2 се добива интервалниот оценувач

$p\left(\dfrac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$. За оценка на стандардната девијација σ, треба само да се коренуваат двете страни на неравенството.

ПРИМЕР 12.6 Бил спроведен експеримент за испитување на прецизноста на уред за мерење на нивото на јод присутно во супстанци по извесен период на континуирано мешање. Податоците прикажани во табелата претставуваат 10 мери на концентрација на јод во еден ист примерок на супстанца.

Обид	1	2	3	4	5	6	7	8	9	10
Концетрат	5.507	5.506	5.500	5.497	5.506	5.527	5.504	5.490	5.500	5.497

Дисперзијата на популацијата σ^2 ја мери варијабилноста т.е. прецизноста на уредот. Користејќи ги овие податоци најди интервал за σ^2 со 95% сигурност.

Решение

Од податоците лесно се добива дека $\bar{x} = 5.5034$ и $s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2 =$ 0.00009649, т.е. $s = 0.009823$. Понатаму имаме дека

$\chi^2_{0.025} = 19.0228$ и $\chi^2_{1-0.025} = \chi^2_{0.975} = 2.7$ за 9 степени на слобода.

Интервалот на доверба е

$$p\left(\frac{9 \cdot 0.00009649}{19.0228} < \sigma^2 < \frac{9 \cdot 0.00009649}{2.7}\right) = p(0.0000457 < \sigma^2 < 0.0003216) = 0.95.$$

Значи со 95% сигурност може да тврдиме дека варијабилноста на мерењата на концентрацијата на јод во еден ист примерок се движи во интервалот (0.0000457, 0.0003216) што одговара на прецизноста на инструментот. ■

Едностраните интервали на доверба за σ^2 се добиваат стандардно

$$p\left(\sigma^2 > \frac{(n-1)S^2}{\chi^2_\alpha}\right) = 1-\alpha \text{ и } p\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha}}\right) = 1-\alpha .$$

ПРИМЕР 12.7 Автоматска машина полни шишиња со течен детерџент. Земен е примерок од 20 шишиња при што е пресметана дисперзија во полнењата од 0.0153 течни унци. Ако дисперзијата е голема, шишињата ќе имаат премногу или премалку течен детергент што повлекува рекалибрација на машината. Под претпоставка дека волуменот е приближно нормално распределен, состави 95% едностран интервал за горната граница на варијациите во волумените.

Решение

Од $\chi^2_{19,\,0.95} = 10.117$ следува дека $p\left(\sigma^2 < \frac{19 \cdot 0.0153}{10.117}\right) = p(\sigma^2 < 0.0287) =$ 0.95. Така со 95% сигурност можеме да тврдиме дека варијациите во волуменот на полнењето се помали од $\sqrt{0.0287} = 0.17$ течни унци. ■

Речиси сè што некогаш би можело да ни затреба во врска со статистичките интервали може да се најде во одличната книга [Gerald, Meeker 1991].

ЗАДАЧИ

1. За нормална популација со позната дисперзија σ^2, најди ги довербите на интервалите:
   а) $\bar{x} - 2.14\sigma / \sqrt{n} < \mu < \bar{x} - 2.14\sigma / \sqrt{n}$?
   б) $\bar{x} - 2.49\sigma / \sqrt{n} < \mu < \bar{x} - 2.49\sigma / \sqrt{n}$?
   в) $\bar{x} - 1.85\sigma / \sqrt{n} < \mu < \bar{x} - 1.85\sigma / \sqrt{n}$?

2. Студирана е чистотата на екстракт од некој хемиски процес. Од претходни испитувања познато е дека чистотата на екстрактот е нормално распределена со $\sigma = 3$. Примерок од 5 екстракти е испитуван при што се измерени следните чистоти: 91.6, 88.75, 90.8, 89.95 и 91.3. Најди 95% интервал на доверба за просекот на чистотата на екстрактот.

3. Производител на прстени за клипови на автомобилски мотори прави прстени со дијаметри што се нормално распределени со $\sigma = 0.001$ милиметри. Земен е примерок од 15 прстени, при што е пресметан просечен дијаметар од 74.036 милиметри.
   а) Конструирај 99% интервал на доверба за просекот на дијаметрите на прстените;
   б) Конструирај 95% едностран оддолу интервал за просекот на дијаметарот на прстените.

4. Определи ги вредностите за $t_{\alpha,n-1}$ потребни за конструкција на следните еднострани интервали на доверба:
   а) Ниво на доверба = 95%, степени на слобода = 14;
   б) Ниво на доверба = 99%, степени на слобода = 19;
   в) Ниво на доверба = 99.9%, степени на слобода = 24.

5. Од страна на контролата на квалитет мерена е дебелината на ѕидовите на 25 стаклени 2-литерски шишиња, при што е добиен просек од 4.05 милиме-

три со стандардна девијација од 0.08 милиметри. Најди 95% едностран оддолу интервал на доверба за просечната дебелина на шишињата.

6. Познат бренд на диетален маргарин бил анализиран за оценка на нивото на полинезаситени маснотии што ги содржи (во проценти). Примерокот од 6 кутии резултирал со следните податоци: 16.8, 17.2, 17.4, 16.9, 16.5 и 17.1.
   а) Дали има докази за претпоставка дека нивото на полинезаситени маснотии е со нормална распределба?
   б) Најди 99% интервал на доверба на просекот на нивото на полинезаситени маснотии.

7. Министерството за транспорт сака да испита колкава пропорција од луѓето би се согласиле со зголемување на лимитот на брзината на автопатите од 65 на 75 милји на час. Колку луѓе треба да се анкетираат за со 99% сигурност пропорцијата на примерокот да е во маргини 0.05 од вистинската пропорција?

8. Треба да се спроведе студија за процентот на домаќинства што поседуваат најмалку 2 телевизора. Колкав треба да биде примерокот ако сакаме со 99% сигурност грешката во проценката да биде помала од 0.017?

9. Разгледај ја повторно дебелината на ѕидовите на 25 стаклени 2-литерски шишиња од задача 5. Состави 90% интервал на предвидување за дебелината на ѕидот на следната шише што ќе се испитува.

10. За задача 6 состави 99% интервал на предвидување за количеството на полинезаситени маснотии во следната паковка на маргарин што ќе се испитува. Спореди ја ширината на интервалот на предвидување со 99%-иот интервал на доверба од задача 6.

11. Контрола на квалитетот во производство на конзерви во една фабрика со примерок од 10 конзерви утврдила просечна содржина (волумен) од 7.98 унци со стандардна девијација од 0.04 унци. Волуменот е важен, но не помалку е важна варијацијата во волумените. Состави 90% интервал на доверба за варијациите на волуменот на конзервите.

12. Направени се испитувања на цврстината на подлогите од 18% никел – ниско јаглероден челик (во KSI - килофунта по квадратен инч): 69.5, 71.9, 72.6, 73.1, 73.3, 73.5, 75.5, 75.7, 75.8, 76.1, 76.2, 76.2, 77.0, 77.9, 78.1, 79.6, 79.7, 79.9, 80.1, 82.2, 83.7, 93.7. Состави 99% интервал на доверба за стандардната девијација на распределбата на цврстината. Дали интервалот е во ред, без разлика на распределбата?

13
Тестирање хипотези

За природата на некоја појава може да се направат многу хипотези: H_0, H_1, …, H_k. Од различни причини, за нас од посебен интерес е една од нив, да речеме H_0, и неа ќе ја нарекуваме нулта хипотеза, а останатаите ќе ги разгледуваме како една алтернативна хипотеза H_A. Генерално, хипотеза може да биде тврдење за вредноста на некој параметар (карактеристика на популацијата или на распределбата), тврдење за односите меѓу параметри или дури тврдење за обликот на целата распределба.

За да одлучиме која хипотеза да прифатиме, земаме примерок X_1, X_2, …, X_n и формираме статистика $h(X_1, X_2, …, X_n)$. Просторот на примерокот V го делиме на две дисјунктни множестава A и B = V − A. Ако вредноста на статистиката $q = h(x_1, x_2, …, x_n) \in$ A ја прифаќаме H_0, а во спротивно, ако $q \in$ B ја прифаќаме H_A. Множеството B обично се нарекува *критичен домен*. Идеално би било $p(q \in B \mid H_0) = 0$ (никогаш не се отфрла H_0 кога таа е точна), и $p(q \in A \mid H_A) = 0$ (никогаш не се отфрла H_A кога таа е точна). Сепак, таквата идеална поделба на просторот на примерокот не е можна. Затоа избираме мал број $\alpha > 0$ и B така што:

$p(q \in B \mid H_0) \leq \alpha$, каде што α се нарекува *ниво на значајност* или грешка од тип 1, и ја дава веројатноста на отфрлање на H_0 кога таа е точна (вообичаено се зема $\alpha = 0.05$, 0.01 или 0.001); и

$p(q \in A \mid H_A) = \beta$, каде што β се нарекува грешка од тип 2, и ја дава веројатноста на прифаќање на H_0 кога таа не е точна. Вредноста $1 - \beta = p(q \in B \mid H_A)$ се нарекува *јачина на тестот* и ја дава веројатноста на отфрлање на H_0 кога таа не е точна.

Додека вредноста за α вообичаено се задава однапред, за β нема една вредност, туку по една за секоја вредност на статистиката кога H_A е точна.

13.1. Параметарски тестови

Како што веќе видовме, параметрите во распределбите може да се оценуваат со точкасти или интервални оценки. Од друга страна, често пати наместо оценка, треба да се донесе одлука кое од две контрадикторни тврдења за параметарот θ е точно. Наједноставни хипотезите од таков тип се:

H_0: $\theta = \theta_0$, H_A: $\theta = \theta_1 < \theta_0$ или
H_0: $\theta = \theta_0$, H_A: $\theta = \theta_1 > \theta_0$ или
H_0: $\theta = \theta_0$, H_A: $\theta = \theta_1 \neq \theta_0$.

Нека тестираме хипотеза во врска со параметарот θ од распределбата $f(x;\ \theta)$. Независниот и еднакво распределен примерок е случаен вектор $(X_1, X_2, \ldots, X_n)$ со густина на распределба

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta)\ .$$

Денеска прифатен пристап за тестови на параметрите се базира на Нојман-Пирсонов-иот (Neyman–Pearson) метод.

Теорема 13.1 (Нојман-Пирсон)**.** Ако постои област B во $\mathbb{R}^n$ и број c таков што

$$\left(\frac{\prod_{i=1}^{n} f(x_i, \theta_0)}{\prod_{i=1}^{n} f(x_i, \theta_1)} \right) \begin{matrix} < c \\ \geq c \end{matrix} \text{ кога } (x_1, x_2, \ldots, x_n) \begin{matrix} \in \mathrm{B} \\ \notin \mathrm{B} \end{matrix},$$

тогаш множеството B е најдобриот критичен домен за отфрлање на H_0.

Доказ: Може да се најде, на пример во [Трпеновски 1981]**.** ■

Оваа теорема гарантира постоење на оптимален тест и дава постапка за негово изведување само во случај на едноставни хипотези, како што се дадените погоре. Статистиката за тестот е функција од горниот количник

$$q = h\left(\frac{\prod_{i=1}^{n} f(x_i, \theta_0)}{\prod_{i=1}^{n} f(x_i, \theta_1)} \right),$$ а обликот на $h(\cdot)$ зависи од случај до случај.

ПРИМЕР 13.1 Под претпоставка за нормална распределба на популацијата, определи ја статистиката за тестот

H_0: $\mu = \mu_0$, наспроти H_A: $\mu = \mu_1 < \mu_0$, ако σ е познато.

Решение

Во овој случај $f(x_1, x_2, \ldots, x_n; \mu) = \dfrac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i - \mu)^2}$, што дава

$$\left(\frac{e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i - \mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i - \mu_1)^2}} \right) < c \quad \underset{\text{и средување}}{\overset{\text{со логаритмирање}}{\Rightarrow}} \quad \frac{1}{n}\sum_1^n x_i < \overbrace{\frac{2\sigma^2 \ln c + n(\mu_0^2 - \mu_1^2)}{2n(\mu_0 - \mu_1)}}^{\text{НОВО } c}$$

и сега

$$p(\bar{x} < c / H_0) = \alpha \;\Rightarrow\; p\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < \overbrace{\frac{c - \mu_0}{\sigma/\sqrt{n}}}^{z_\alpha} / H_0 \right) = \alpha$$

. Значи ако $\overbrace{\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}^{\text{статистика}} < z_\alpha$,

ја отфрламе H_0 со ниво на значајност α. ■

Сепак, овој пристап обезбедува наоѓање оптимален тест (статистика) само во вакви едноставни случаи. Генерално, статистиката q треба да биде избрана така што,

а) $p(q \in \text{B} \mid H_0 \text{ е точна}) \le \alpha$; и

б) $p(q \in \text{A} \mid H_\text{A} \text{ е точна}) = \beta(\theta)$ е минимална.

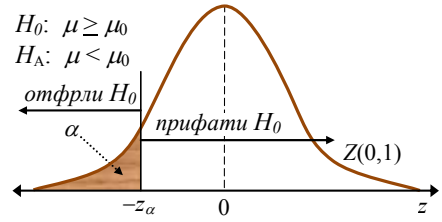
Секој параметарски тест може да се опише низ следните чекори:

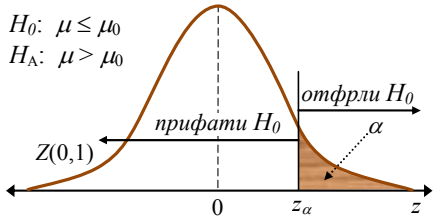
1. Идентификувај ги параметрите што се од интерес;
2. Определи ја нултата H_0, и алтернативната хипотеза H_A;
3. Избери ниво на значајност α и според него областите на отфрлање и порифаќање на H_0;
4. Определи ја статистиката за тестот, евентуалните непознати параметри и пресметај ја вредноста на статистиката;
5. Според областа во која припаѓа вредноста на статистиката, отфрли ја или прифати ја H_0 и интепретирај ја одлуката во светло на конкретниот проблем;
6. За подетален увид во ситуацијата, пресметај ја β (и јачината на тестот $1 - \beta$) и евентуално P-вредноста на тестот и реинтерпретирај го резултатот во светло на нивните вредности.

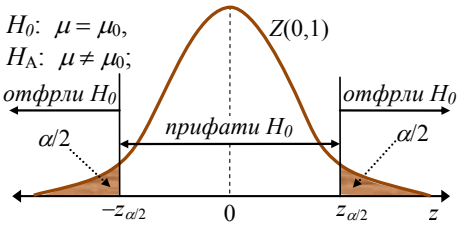
Забележи дека чекорите 1-3 може да се комплетираат пред обезбедувањето на примерокот. *P*-вредноста на тестот ќе ја дискутираме понатаму.

13.1.1. Тестови за просекот

Како и кај оценките на параметрите, најпрво ќе ги разгледаме хипотезите во врска со просекот μ на популацијата. Основна претпоставка за валидност на ваков тест е популацијата да има нормална распределба или примерокот да биде доволно голем, $n \geq 30$, со што според централната гранична теорема може да сметаме дека неговата сума има приближно нормална распределба. Тестовите на просекот μ на популацијата може да земат една од следните три форми:







Во сите 3 случаи, вредноста на статистика е $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ за познато σ, и $z = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$ за непознато σ (за $n \geq 30$, s е добра оценка за σ).

ПРИМЕР 13.2 Производител на распрснувачки систем за заштита од пожар тврди дека температурата на активирање на системот е 130°F (фаренхајтови). Примерок од 9 системи е тестиран и добиено е просечна температура на активација од 131.08°F. Ако распределбата на температурата на активирање е нормална со стандардна девијација од 1.5°F, тестирај дали податоците го потврдуваат тврдењето на производителот со ниво на значајност $\alpha = 0.01$.

Решение

Тестираме H_0: $\mu = 130°$, наспроти H_A: $\mu \neq 130°$.

За вредноста на статистиката имаме $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{131.08 - 130}{1.4 / \sqrt{9}} = 2.16$, а од таблицата за нормална распределба имаме $z_{0.005} = 2.96$. Поради тоа што $z =$

2.16 ∈ [–2.96, 2.96] ја прифаќаме H_0, т.е. заклучуваме дека податоците не нудат доволно докази за отфрлање на тврдењето на производителот. ■

Како да се најде јачината на тестот (или грешката од тип 2)? На пример ако тестираме на алтернативен помал просек би имале дека

$$1-\beta = p(q \in \text{B} \mid H_\text{A}) = p\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha \mid H_\text{A}\right). \text{ Но } \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \text{ нема } Z(0,1)$$

распределба кога H_A е точна (просекот не е μ_0 туку е μ_1), па додаваме соодветен собирок од двете страни на неравенството и добиваме

$$p\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} \mid H_\text{A}\right) = p(z < -z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}) \cdot$$

Така може да се дојде до изразите за β за различните случаи:

$$H_\text{A}:\ p > p_0 \ \Rightarrow\ \beta = \Phi\left(z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

$$H_\text{A}:\ p < p_0 \Rightarrow\ \beta = 1-\Phi\left(-z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

$$H_\text{A}:\ p \neq p_0 \Rightarrow \beta = \Phi\left(z_{\alpha/2} + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

каде што $\Phi(\cdot)$ функцијата на стандардната нормална распределба.

Големината на примерокот n, за која тестот со ниво на значајност α има грешка од тип 2 дадена со β, приближно е,

$$\text{за } H_\text{A}:\ \mu > \mu_0 \text{ и } H_\text{A}:\ \mu < \mu_0 \ \Rightarrow\ n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0-\mu_1}\right)^2$$

$$\text{за } H_\text{A}:\ \mu \neq \mu_0 \ \Rightarrow\ n = \left(\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0-\mu_1}\right)^2.$$

ПРИМЕР 13.3 Службата за одржување на патишта треба да поправи делница од 60 километри. Слојот на асфалт што треба да се стави зависи од бројот на тешки камиони што поминуваат по патот. Државниот извештај тврди дека бројот на тешки камиони по час е 72. Од друга страна, службата има индиции дека овој број може да е поголем. По 50 часовно испитување на сообраќајот (случајно избрани часови во тек на еден месец) добиен е просек од 74.1 тешки камиони по час со стандардна девијација од s = 13.3. Тестирај дали добиените податоци го потврдуваат државниот извештај за $\alpha = 0.1$.

Решение

Тестираме H_0: $\mu = 72$, наспроти H_A: $\mu > 72$.

За $n = 50 \geq 30$ користиме нормална распределба. За статистиката имаме $z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{74.1 - 72}{13.3/\sqrt{50}} = 1.1648$, а од таблицата за нормална распределба читаме $z_{0.1} = 1.28$. Поради $z = 1.1648 < 1.28 = z_{0.1}$ ја прифаќаме H_0 т.е. државниот извештај за бројот на тешки камиони по час.

Нека H_0 не е точна и нека бројот на тешки камиони по час е 78. Колкава е веројатноста дека нашиот тест тоа нема да го детектира (се бара β)?

Имаме дека $1 - \beta = p(h \in \text{B} / H_A) = p\left(\dfrac{\bar{x} - 72}{13.3/\sqrt{50}} > 1.28 / H_A\right) =$

$p\left(\dfrac{\bar{x} - 78}{13.3/\sqrt{50}} > 1.28 + \dfrac{72 - 78}{13.3/\sqrt{50}} / H_A\right) = p(z > -1.91) = 0.9719$. Значи веројатноста да се прифати H_0 кога $\mu = 78$, е само $\beta = 1 - 0.9719 = 0.0281$.

Ако пак бројот на тешки камиони по час е 74, за β добиваме, $1 - \beta =$

$p\left(z > 1.28 + \dfrac{72 - 74}{13.3/\sqrt{50}} / H_A\right) = p(z > 0.22) = 0.4129$, што е слаба јачина, т.е. висока грешка од тип 2, $\beta = 0.5871$. За оваа грешка да ја доведеме до $\beta = 0.1$ би требало да го зголемиме примерокот на $n = \left(\dfrac{13.3(1.28 + 1.28)}{72 - 74}\right)^2 \approx 290$. ■

Кога примерокот е мал ($n < 30$) тестот не може да се прави без претпоставка за (приближно) нормална распределба. Дури и под таква претпоставка, при непознато σ, оценката на σ со s повеќе не е добра и S мора да се разгледува како случајна променлива. Тогаш, кога H_0 е точна ($\mu = \mu_0$), случајната променлива

$T = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}}$ има студентова распределба (однос $Z(0.1)\sqrt{n} / \sqrt{\chi^2}$) со $n - 1$ степени на слобода. Тоа овозможува тестот да остане ист, само што нормалната распределба се заменува со студентова.

ПРИМЕР 13.4 Коските на животните имаат тенденција да бидат со ист однос должина/ширина за едно животно со приближно нормална распределба. Археолозите ископале 20 коски со просечен однос должина/ширина од 9.15 и стандардна девијација од 1.16. Постои претпоставка дека тие се од животно за кое се знае дека односот должина/ширина е 8.5. Дали е тоа така? Користи $\alpha = 0.01$.

Решение

Тестираме H_0: $\mu = 8.5$, наспроти H_A: $\mu \neq 8.5$.

За $n = 20 < 30$ користиме студентова распределба со 19 степени на слобода. За статистиката имаме $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{9.15 - 8.5}{1.16/\sqrt{20}} = 2.506$, а од таблицата за студентова распределба читаме $t_{0.005} = 2.861$. Поради $t = 2.506 < 2.861 = t_{0.005}$ ја прифаќаме H_0 т.е. односот должина/ширина на ископаните коски значајно не се разликува од 8.5.

Ако би тестирале со ниво на значајност $\alpha = 0.1$ би добиле $t = 2.506 > 1.729 = t_{0.05}$ и хипотезата H_0 би била отфрлена.

Ако би тестирале H_0: $\mu = 8.5$, наспроти H_A: $\mu > 8.5$, би добиле исти резултати, но нешто "понаклонети" кон отфрлање на H_0 бидејќи $t_{0.01} = 2.539$, $t_{0.1} = 1.328$ за иста статистика. Јачината на тестот за $\alpha = 0.01$ би била

$$1 - \beta = p\left(t > 2.539 + \frac{8.5 - 9.15}{1.16/\sqrt{20}} / H_A \right) = p(t > 0.033) = 0.4870,$$ а за $\alpha = 0.1$

$$1 - \beta = p\left(t > 1.328 + \frac{8.5 - 9.15}{1.16/\sqrt{20}} / H_A \right) = p(t > -1.178) = 0.6267,$$ што значи дека тестот во кој се отфрла H_0 е појак.

Зголемувањето на примерокот води до пораст на t и зголемени шанси на за отфрлање на H_0. Но тука големината на примерокот не може да се зголемува (постојат само 20 коски). Од самите тестови, нивната јачина и големината на примерокот, сепак би се одлучиле да ја отфрлиме H_0. ■

Некои поважни заклучоци во врска со параметрите α и β се сумирани во следните точки:

а) Големината на критичниот регион В (грешката од тип 1) може секогаш да се редуцира со зголемено α;

б) Грешките од тип 1 и 2, т.е. α и β се зависни. Намалувањето на едната води до зголемувањето на другата, под услов да не се менува големината на примерокот;

в) Зголемувањето на примерокот генерално ги намалува и α и β;

г) Кога H_0 се отфрла, расте β бидејќи се зголемува разликата меѓу вредноста и хипотетичката вредност на параметарот.

Ако се има предвид дека погрешното отфрлање на H_0 е под директна контрола (со зададената веројатност α), отфрлањето на H_0 е силен

заклучок. Од друга страна, веројатноста на грешка од втор тип β зависи од двете, вистинската вредност на параметарот и големината на примерокот, па прифаќањето на H_0 може да се смета за релативно "слаб" заклучок, освен ако β е прифатливо мала. Не-отфрлањето (ова е можеби поадекватен термин од "прифаќање") на H_0 повлекува дека немаме доволно докази за нејзино отфрлање и така да направиме "силен" заклучок. *Отфрлањето на H_0 е како осуда на криминалец во судски процес, а прифаќањето како немање доволно докази за да се осуди.* Значи прифаќањето H_0 не значи дека со висока веројатност таа е точна, туку само дека нема доволно докази, т.е. треба дополнителни докази таа евентуално да се отфрли. Од тие причини, во понатамошниот текст најчесто ќе го користиме терминот "H_0 не се отфрла" наместо "H_0 се прифаќа".

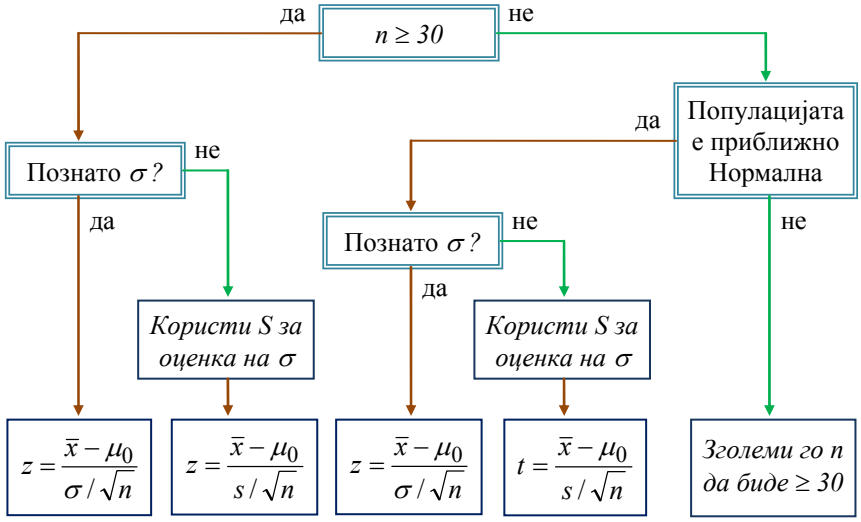
Јачината на тестот $1 - \beta$ (веројатност за коректно отфрлање на H_0) е многу информативна и концизна мера на чуствителноста на тестот (способност да ги детектира разликите). Во случај на слаба јачина, единствено што останува е да се зголеми α или најдобро, големината на примерокот n.

Како да се одлучи што ќе оди како нулта, а што како алтернативна хипотеза? Ова прашање некому на прв поглед може да му изгледа неважно, но не е така.

Тестирањето хипотези е сличен концепт на судски процес, каде што се тестира H_0: *Обвинетиот е невин*, наспроти H_A: *Обвинетиот е виновен*. Затоа обично теоријата што сакаме да ја поддржиме би требало да оди како алтернативна хипотеза. Од тој аспект, истражувачките хипотези би требало да одат како алтернативни, така што нивната вистинитост да произлезе од податоците што ќе ја оспорат нултата хипотеза. Тврдењата на производителите за своите производи обично треба да се сомничат и како такви да бидат зададени како нулти хипотези. Донесувањата на одлуки може да одат и како нулти и како алтернативни хипотези во зависност од секоја конкретна ситуација.

Следниот дијаграм ги сумира можностите за изборот на статистиката за тестирање на просекот на популацијата μ. Како што се гледа, важни одлуки се носат врз база на големината на примерокот. Да се потсетиме дека "магичната" бројка $n = 30$ нема некоја математичка поддршка, туку е искуствена и има мислења дека треба да е поголема (види поглавје 7.4).

Очигледно е дека треба да се биде крајно внимателен со тестирањето кога примерокот е мал, $n < 30$. Во таков случај, за доверлив тест е неопходно популацијата да има (приближно) нормална распределба.



13.1.2. *P*-вредност на тестовите

Прифаќањето или отфрлањето на хипотезите само според нивото на значајност не ни дава идеја за доверливоста на одлуката, т.е. дали статистиката од тестот е само малку или е длабоко во регионот на прифаќање или отфрлање. *P*-вредноста на тестот е алтернативен пристап за доаѓање до заклучок за хипотезите. Таа може да се дефинира како *најмалото ниво на значајност што би водело до отфрлање на хипотезата за дадените податоци*. На *P*-вредноста може да се гледа како на неформална мера за аргументот против (нултата) хипотеза. Еднаш кога *P*-вредноста е позната, ние веднаш можеме да оцениме дали има смисла да ја отфрлиме хипотезата, без формално задавање на нивото на значајност.

Секогаш треба да се имаат предвид следните факти:

а) *P*-вредноста е веројатност што се пресметува под претпоставка дека H_0 е точна;

б) *P*-вредноста не е веројатност дека H_0 е точна, ниту пак веројатност на грешка; *P*-вредноста = *плоштината под густината на распределба во областа на отфрлање на* H_0 *определена од вредноста на статистиката* (наместо од α);

в) *P*-вредноста не може директно да се исчита од таблиците (може да се определи опсегот), туку за нејзино добивање потребно е користење на некој софтвер (на пример Excel).

ПРИМЕР 13.5 Дождовницата во градовите може да биде контаминирана од многу извори, вклучувајќи метали од фрлени батерии. Примерок од 51 панасоникови AAA батерии е испитуван на содржина на цинк, при што е утврден просек од 2.06 грама со стандардна девијација од 0.141 грама. Дали овие податоци се доволен доказ дека содржината на цинк во батериите надминува 2 грама?

Решение

Тестираме H_0: $\mu = 2.0$, наспроти H_A: $\mu > 2.0$ грама.

Примерокот е доволно голем, па статистиката е $z = \dfrac{2.06 - 2.0}{0.141/\sqrt{61}} = 3.04$.

Кои вредности на z се контрадикторни со H_0? Очигледно колку што вредноста на просекот $\bar{x}$ надминува 2.06, толку се "оддалечуваме" од H_0. Вредностите на $\bar{x}$ што надминуваат 2.06 соодвествуваат на вредностите на z што надминуват 3.04. Така P-вредноста на тестот е

P-вредност = $p(z > 3.04 \mid H_0)$ = плоштината под густината надесно од 3.04

$= 1 - \Phi(3.04) = 0.0012.$

Кои P-вредности обезбедуваат доволно докази против H_0?

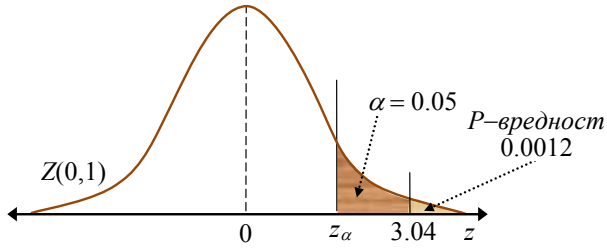
Кога P-вредноста = 0.0012 значи дека само 0.12% од сите можни вредности на статистиката се контрадикторни на H_0 најмалку онолку, колку и нашата статистика. Така, примерокот силно сугерира отфрлање на H_0.

Кога P-вредноста би била, на пример 0.2, тоа би значело дека 20% од сите можни вредности на статистиката се контрадикторни на H_0 најмалку онолку, колку и статистиката од нашиот примерок. Тогаш примерокот не би бил во значителна контрадикција со H_0, па тогаш не би ја отфрлиле. ■

Генерално, колку помала P-вредност на тестот, толку посилен доказ против H_0. Значи H_0 треба да биде отфрлена кога P-вредноста е доволно мала. Но колку е тоа "доволно мала"? Некој "логично" би можел повторно да избере ниво на значајност α (пожелна грешка од тип 1) и тогаш во ваквиот "хибриден" пристап да заклучи

ако P-вредноста $\le \alpha$ *отфрли ја* H_0;
ако P-вредноста $> \alpha$ *не ја отфрлај* H_0.

Ова правило дава идентична област на отфрлање на H_0 како и "стандардното" тестирање, па практично не нуди нешто ново. Во претходниот пример, за сите стандардни вредности $\alpha = 0.1$, 0.05 или 0.01 имаме P-вредноста = 0.0012 < α и секагаш ја отфрламе H_0 (види слика подолу).



Тестирањето хипотези само со користење на *P*-вредност историски претходи на тестирањето со пристапот на Нојман-Пирсон (алтернативна хипотеза и ниво на значајност α). Нејзин творец е Фишер (Ronald Aylmer Fisher, 1890 – 1962) и оригинално тој предложил грубо упатство, за какви *P*-вредности да се отфрла H_0:

P-вредност	**Интерпретација**
$P > 0.10$	податоците нудат силна подршка за H_0
$0.05 < P < 0.10$	податоците нудат некаква подршка за H_0
$0.02 < P < 0.05$	податоците не нудат подршка за H_0
$P < 0.01$	податоците нудат силна подршка за отфрлање на H_0

Критиките на овој пристап се во произволната интерпретација на *P*-вредноста при донесувањето на одлуката. Од друга страна, нивото на значајност α исто така се задава произволно освен барањето вредноста да биде "мала", а ние ги користиме "разумните" вредности 0.1, 0.05 или 0.01. *P*-вредноста понекогаш ја нарекуваат забележано (observed) ниво на значајност наспроти α што е зададено (predefined) ниво на значајност. Модерните книги за статистика ги мешаат и двата пристапи користејќи го како α (Нојман-Пирсоновиот пристап) така и *P*-вредноста (Фишеровиот пристап) заради добивање на подоверливи резултати од тестирањето.

Според некои експерти, Фишер и Нојман би се превртеле во гробовите кога би можеле да видат како таков монструозно "хибриден пристап" се користи во литературата за тестирање на хипотези. Ваквиот присилен брак, на според нив непремостиво различни пристапи има свое оправдание и се разбира и ние го користиме во оваа книга. Сепак, секогаш треба да се има предвид дека тие се концепциски многу различни. Имено, нивото на значајност α е особина на самиот тест, додека *P*-вредноста е мера поврзана директно со разгледуваните податоци.

ПРИМЕР 13.6 Ефикасноста на горивото (миљи по галон - mpg) варира од возило до возило за ист производител. Нека μ биде вистинскиот просек на ефи-

касноста на горивото на 4 различни возила од ист производител за кои е добиено 20.830, 22.232, 20.276 и 17.718. Дали овие податоци се доволен доказ дека ефикасноста на горивото кај овој производител надминува 20, претпоставувајќи негова нормална распределеност?

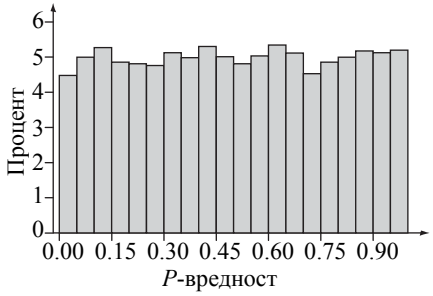
Решение

Тестираме H_0: $\mu = 20$, наспроти H_A: $\mu > 20$ mpg.

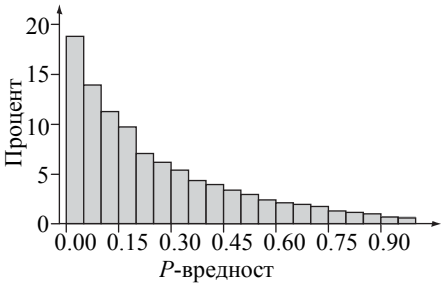
Поради малиот примерок работиме со студентова распределба со 3 степени на слобода. Од податоците имаме $\bar{x} = 20.264$, $s = 1.8864$ и вредност на статистиката

$$t = \frac{20.264 - 20}{1.8864/\sqrt{4}} = 0.2799 < \begin{cases} 4.5407 = t_{0.01} \\ 2.3534 = t_{0.05} \\ 1.6377 = t_{0.1} \end{cases}$$, што значи дека H_0 не се отфрла за ниедно "логично" α. P-вредноста е плоштината под t-густината надесно од 0.2799 и изнесува 0.3938. Се разбира, вака високата P-вредност оди силно во прилог на H_0.

На сликата подолу а) е прикажан хистограм на 10000 P-вредности добиени со симулација, при која се генерираат 10000 пати по 4 вредности (примерок) со нормална распределба со очекување $\mu = 20$ и стандардна девијација $\sigma = 2$. Околу 4.5% од сите P-вредности се меѓу 0 и 0.05, што значи дека H_0 би била отфрлена (за $\alpha = 0.05$) само во 4.5% од 10000 тестови. Како што се гледа од хистограмот, распределбата на P-вредностите е приближно рамномерна. Всушност може да се докаже дека во случај кога H_0 е точна, распределбата на P-вредностите е рамномерна во [0, 1].



а) P-вредности кога H_0 е точна



б) P-вредности кога H_0 не е точна

Од друга страна, нека H_0 не е точна и нека $\mu = 21$. Повторно правиме 10000 симулации на 4 случајни вредности со нормална распределба со очекување $\mu = 21$ и стандардна девијација $\sigma = 2$. Со пресметување на P-вредностите за вредности од истата статистика $t = (\bar{x} - 20)/(s/\sqrt{4})$ добиен е хистограмот б). Гледаме дека сега распределбата е далеку од рамномерна и бројот на P-вредностите постојана се намалува кога нивната вредност се зголемува од 0 до

1. Ако генерирањето на 4-случајни вредности би било од нормална распределба со очекување $\mu = 22$ (H_0 е уште понеточна), опаѓањето на дијаграмот б) би било уште пострмно. ■

Секоја статистика е случајна променлива па и P-вредноста е непрекината случајна променлива што зема вредности во интервалот [0, 1] (е веројатност). Овој пример јасно покажува дека колку повеќе вредноста на параметарот е подалеку од тврдењето на H_0, толку повеќе P-вредности ќе бидат концетрирани околу 0-та зголемувајќи ги шансите H_0 да биде коректно отфрлена што одговара на помало β (подобра јачина на тестот).

13.1.3. Тестови за пропорцијата

Во многу проблеми се користи случајна променлива X со биномна распределба $p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, каде што параметарот p е пропорција (веројатност). На пример, X може да е број на дефектни производи во некој процес на производство. Хипотезите за тестирањето на пропорцијата p во популациите ги опфаќа 3-те стандардни случаи:

H_0: $p \geq p_0$, наспроти H_A: $p < p_0$;

H_0: $p \leq p_0$, наспроти H_A: $p > p_0$;

H_0: $p = p_0$, наспроти H_A: $p \neq p_0$.

За добивање на статистиката да забележиме дека за доволно големо n ($n \cdot p \geq 10$ и $n \cdot (1-p) \geq 10$), двете случајни променливи X и $\hat{P} = X/n$ имаат приближно нормална распределба. Ако се има предвид дека точкастиот оценувач $\hat{P} = X/n$ на p е центриран и дека дисперзијата на биномната распределба е $p(1-p)$, тогаш нејзината стандардна девијација е $\sigma_{\hat{P}} = \sqrt{p(1-p)/n}$. Кога H_0 е точна, имаме $E\hat{P} = p_0$ и $\sigma_{\hat{P}} = \sqrt{p_0(1-p_0)/n}$, па $\sigma_{\hat{P}}$ не вклучува непознати параметри. Оттука, кога H_0 е точна, случајната променлива

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \text{ има приближно } Z(0,1) \text{ распределба.}$$

Сега, на пример ако алтернативната хипотеза е H_A: $p > p_0$, имаме дека

p(грешка од тип 1) = p(H_0 е отфрлена кога е точна) =

$= p(Z \geq z_\alpha$ кога Z има приближно $Z(0,1)$ распределба$) \approx \alpha$.

Слично се добива и за другите 2 случаи. Значи тестот за пропорција во популацијата е

нулта хипотеза: H_0: $p = p_0$; статистика: $Z = \dfrac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Алнернативана хипотеза	*Критичен регион* (*отфрламе* H_0)
H_A: $p < p_0$	$z \le z_\alpha$
H_A: $p > p_0$	$z \ge z_\alpha$
H_A: $p \ne p_0$	$z \le -z_{\alpha/2}$ или $z \ge z_{\alpha/2}$

при што n треба да е доволно големо, т.е. $n{\cdot}p \ge 10$ и $n{\cdot}(1-p) \ge 10$.

ПРИМЕР 13.7 Тапите од плута кај шишињата вино се подложни на деградација што води до намалување на квалитетот на виното. Во еден членак за тестирање на Шардоне е публиковано дека 16 од 91 шише имале некаква контаминација од тапата. Дали е тоа силен доказ да се заклучи дека 15% од таквите шишиња се контаминирани од тапата? Користи $\alpha = 0.1$.

Решение

Нека p = пропорција на контаминирани шишиња Шардоне.

Тестираме H_0: $p = 0.15$, наспроти H_A: $p > 0.15$.

Бидејќи $n{\cdot}p_0 = 91{\cdot}0.15 \ge 10$ и $n{\cdot}(1-p_0) = 91{\cdot}0.85 \ge 10$ користиме стандарден z тест. Од $\hat{p} = 16/91 = 0.1758$, за вредноста на статистиката добиваме

$$z = \frac{0.1758 - 0.15}{\sqrt{0.15 \cdot 0.85/91}} = 0.6898 < 1.28 = z_{0.1},$$ и така H_0 не се отфрла. ■

Како да се определи грешката од тип 2, т.е. β? Кога H_0 не е точна, $p = p_1$, но сепак Z останува да има нормална распределба со просек и дисперзија,

$$EZ = \frac{p_1 - p_0}{\sqrt{p_0(1-p_0)/n}}, \quad DZ = \frac{p_1(1-p_1)}{p_0(1-p_0)}.$$ Отука е можно да се дојде до целосните изрази за β за различните случаи:

$$H_A: p > p_0 \Rightarrow \beta = \Phi\left(\frac{p_0 - p_1 + z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right)$$

$$H_A: p < p_0 \Rightarrow \beta = 1 - \Phi\left(\frac{p_0 - p_1 - z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right)$$

$$H_A:\ p \neq p_0 \ \Rightarrow \beta = \Phi\left(\frac{p_0 - p_1 + z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right) - \Phi\left(\frac{p_0 - p_1 - z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right)$$

каде што $\Phi(\cdot)$ функцијата на стандардната нормална распределба.

Големината на примерокот n за која тестот со ниво на значајност α има грешка од тип 2 (β) приближно е:

$$\text{за } H_A:\ p > p_0 \text{ и } H_A:\ p < p_0 \ \Rightarrow n = \left(\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}}{p_1 - p_0}\right)^2$$

$$\text{за } H_A:\ p \neq p_0 \ \Rightarrow n = \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}}{p_1 - p_0}\right)^2.$$

ПРИМЕР 13.8 Брзата пошта тврди дека најмалку 90% од сите пратки донесени пред 9 часот во градот, ќе бидат до пладне испорачани до примачот. Нека p биде пропорцијата од таквите пратки и нека тестираме H_0: $p = 0.9$, наспроти H_A: $p < 0.9$. Ако само 80% од пратките се примаат до пладне, која е веројатноста дека за $\alpha = 0.01$ тест базиран на $n = 255$ пратки ќе го детектира тоа отстапување од H_0? Колкав треба да биде примерокот да се осигураме дека $\beta = 0.01$?

Решение

За $\alpha = 0.01$, $p_0 = 0.9$, $p_1 = 0.9$, $n = 225$ имаме дека

$$\beta = 1 - \Phi\left(\frac{0.9 - 0.8 - 2.33\sqrt{0.9(1-0.9)/255}}{\sqrt{0.8(1-0.8)/255}}\right) = 1 - \Phi(2.00) = 0.0228.$$

Значи веројатноста дека H_0 ќе биде отфрлена кога $p = 0.8$ (јачина на тестот) е 0.9772, па околу 98% од примероците ќе резултираат во коректно отфрлање на H_0.

$$\text{Од } z_\alpha = z_\beta = z_{0.01} = 2.33 \ \Rightarrow n = \left(\frac{2.33\sqrt{0.9\cdot 0.1} + 2.33\sqrt{0.8\cdot 0.2}}{0.8 - 0.9}\right)^2 \approx 266. \ \blacksquare$$

Во случај на мало n, не користиме апроксимација со нормална распределба, туку директно работиме со Биномна распределба. На пример, ако тестираме H_0: $p = p_0$, наспроти H_A: $p < p_0$, H_0 би ја отфрлиле ако $X < c$, каде што c е критичната вредност што треба да се најде од α. Кога H_0 е точна, X има биномна распределба со параметар p_0, па $\alpha = p(X \leq c \mid H_0)$

$= Bin(c; p_0, n)$ = вредност на функцијата на распределба за c. Тоа значи дека за дадено α, треба да се најде најголемото c такво што $B(c; p_0, n) < \alpha$. Од друга страна, $\beta = p(X > c \mid H_A: p = p_1) = 1 - B(c; p_1, n)$.

ПРИМЕР 13.9 Производител на пластични производи има развиено нов тип корпи за ѓубре, што планира да ги продава со 6 годинишна гаранција. За да види дали тоа е економски исплатливо, примерок од 20 корпи е подложен на забрзано користење за да се симулира 6 годишно користење. Гаранцискиот период ќе се промени само ако помалку од 90% од корпите го "преживеат" гаранцискиот период. Ако p е пропорцијата на "преживеани" корпи, тестирај ја хипотезата за исплатливоста на 6-годишниот гаранциски период со ниво на значајност $\alpha = 0.05$.

Решение

Тестираме H_0: $p = 0.9$, наспроти H_A: $p < 0.9$.

Бидејќи мора $B(c; 0.9, 20) < 0.05$, најголемото c за кое ова е исполнето е $c = 15$, т.е. $B(15; 0.9, 20) = 0.043$ (веќе $B(16; 0.9, 20) = 0.133$). Значи критичниот регион е $X \le 15$, па ако е, на пример, $X = 14$, ја отфрламе H_0 што повлекува промена на гаранцискиот период.

Да ја пресметаме β за $p = 0.8$ (ако пропорцијата на "преживеани" корпи е 0.8, која е веројатноста дека тестот тоа нема да го детектира).

$$\beta = p(\text{прифатена } H_0 \mid X \sim Bin(0.8, 20)) = p(X \ge 16 \mid X \sim Bin(0.8, 20)) =$$
$$= 1 - B(15; 0.8, 20) = 1 - 0.370 = 0.630.$$

Значи има 63% шанси да се даде 6-годишна гаранција кога пропорцијата на корпи со животен век > 6 години е само 80%. Високата грешка од тип 2 (слабата јачина) на тестот произлегува од малиот примерок, како и блискоста на пропорциите 0.8 и 0.9. ■

13.1.4. Тестови за дисперзијата

Понекогаш се јавува потреба од тестирање хипотеза за дисперзијата или стандардната девијација. Тестирањето може да се направи исто како за просекот и пропорцијата:

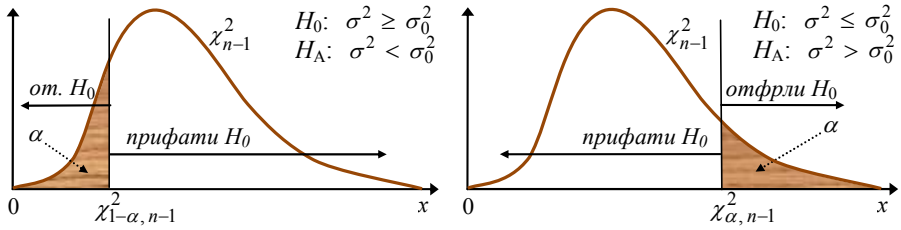
H_0: $\sigma^2 \ge \sigma_0^2$, наспроти H_A: $\sigma^2 < \sigma_0^2$,

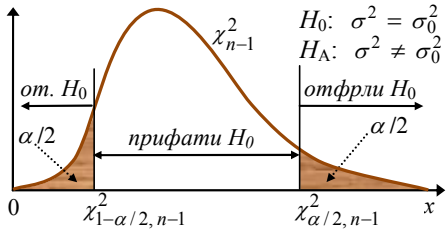
H_0: $\sigma^2 \le \sigma_0^2$, наспроти H_A: $\sigma^2 > \sigma_0^2$,

H_0: $\sigma^2 = \sigma_0^2$, наспроти H_A: $\sigma^2 \ne \sigma_0^2$.

Под претпоставка дека популацијата има приближно нормална распределба и ако хипотезата H_0 е точна, статистиката

$\chi_0^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$ има приближно хи-квадрат распределба со $n-1$ степени на слобода. Критичните области на трите теста се дадени на следните слики:





Значи критичните региони се:

за H_A: $\sigma^2 < \sigma_0^2$ е $\chi_0^2 < \chi^2_{1-\alpha,n-1}$,

за H_A: $\sigma^2 > \sigma_0^2$ е $\chi_0^2 > \chi^2_{\alpha,n-1}$ и

за H_A: $\sigma^2 \neq \sigma_0^2$ е $\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$ или $\chi_0^2 > \chi^2_{\alpha/2,n-1}$.

ПРИМЕР 13.10 Машина за автоматско полнење шишиња со течен детергент треба да работи во предвидена спецификација – наполнетиот волумен на детергент по шише да варира најмногу 0.01 унца^2. Земен е примерок од 20 шишиња од кои со мерење е добиена дисперзија од $s^2 = 0.0153$ унци^2. Дали производителот има проблем со преголема варијација во содржината на шишињата. Користи ниво на значајност $\alpha = 0.05$.

Решение

Тестираме H_0: $\sigma^2 \leq \sigma_0^2$, наспроти H_A: $\sigma^2 > \sigma_0^2$.

Од статистиката добиваме $\chi_0^2 = \dfrac{19 \cdot 0.0153}{0.01} = 29.07$. Од таблица за хи-квадрат распределба читаме $\chi^2_{0.05,19} = 30.1435$. И сега поради $\chi_0^2 = 29.07 < 30.14 = \chi^2_{0.05,19}$ нема доволно докази за отфрлање на H_0.

P-вредноста е 0.0649, што е во согласност со одлуката на тестот. ■

Нека H_A е точна, т.е. нека $\sigma^2 = \sigma_1^2 > \sigma_0^2$. За јачината на тестот добиваме

$$1-\beta = p(\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha,n-1} \mid H_A) = p(\frac{(n-1)S^2}{\sigma_1^2} > \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{\alpha,n-1}) = (\chi^2 > \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{\alpha,n-1})$$

Ако процесот на полнење од претходниот пример варира за 25%, имаме дека $\sigma_1 = 0.125$ ($\sigma_1 = 1.25\sigma_0 = 1.25\sqrt{0.01}$).

Веројатноста дека нашиот тест тоа ќе го детектира е приближно

$$1-\beta = p(\chi^2 > \frac{0.01}{0.125^2}30.14) = p(\chi^2 > 19.29) = 0.4384$$

Значи имаме околу 43.84% шанси дека H_0 ќе биде отфрлена ако вистинската дисперзија е $0.125^2 = 0.0156$. Значи ако вистинската дисперзија е 0.0156, грешката од тип 2 е $\beta = 0.5616$. На пример, ако би сакале да ја симнеме грешката β на $\beta \approx 0.2$, примерокот би требало да се зголеми на околу $n = 61$.

13.1.5. Статистичка наспроти практична значајност на тестовите

Како што веќе беше дискутирано, методологијата на класичното тестирање на хипотезите преку нивото на значајност α користи релативно малку информации скриени во податоците. На пример, кога ја отфрламе H_0 со ниво на значајност $\alpha = 0.05$, секако би биле многу помирни со донесената одлука ако вредноста на статистиката значително ја надмине 5%-ната критична вредност отколку ако е одвај малку над неа. Ова е точно тоа што *P*-вредноста го нуди како алтернатива, давајќи ја значајноста без наметнување конкретна граница, што овозможува секој да донесе свој заклучок според тоа колку е статистиката е "длабоко во периферијата" на распределбата релевантна за тестот.

Сепак, дури и со обезбедена *P*-вредност, се јавуваат сериозни тешкотии со интерпретација на нејзината вредност и донесувањето одлука. Мала *P*-вредност, што обично е силна индикација за отфрлање на H_0, може да биде резултат на голем примерок во комбинација со оддалечување од H_0 коешто има мала практична значајност. Во многу практични ситуации, само големо оддалечување од H_0 е вредно да се детектира, додека малите оддалечувања од H_0 имаат мала практична вредност.

Да претпоставиме дека тестираме H_0: $\mu = 100$, наспроти H_A: $\mu > 100$ каде што μ е просек на популација со нормална распределба со $\sigma = 10$. Нека вистинската вредност на просекот е $\mu = 101$ и нека тоа не биде сериозно отстапување од H_0 во смисла што неотфрлањето на H_0 кога $\mu = 101$ е релативно "ефтина" грешка. За разумно голем примерок n, ова μ води до вредност на $\bar{x}$ блиска до 101 па ние не би сакале овој примерок силно да се "согласува" со отфрлањето на H_0. Следната табела ги дава

P-вредностите кога $\bar{x} = 101$ како и веројатноста на прифаќање на H_0 за ниво на значајност 0.01 кога $\mu = 101$ (β):

n	*P*-вредност	β за $\mu = 101$ и $\alpha = 0.01$
25	0.3085	0.9664
100	0.1587	0.9082
400	0.0228	0.6203
900	0.0013	0.2514
1600	0.0000335	0.0475
2500	0.000000297	0.0038
10000	$7.69 \cdot 10^{-24}$	0.0000

Втората колона во табелата покажува дека дури и за умерено големи примероци, *P*-вредностите за $\bar{x} = 101$ силно сугерираат отфрлање на H_0, додека вредноста на $\bar{x}$ навистина малку, во многу ситуации практично безначајно (1%), се разликува од вистинската вредност на $\mu = 100$. Значи во многу практични ситуации би требало $\bar{x} = 101$ да води кон прифаќање на H_0, и тоа би било така за помали примероци, да речеме $n \leq 250$. Третата колона покажува дека и за практично мала разлика меѓу вистинското $\mu = 100$ и $\bar{x}$, за фиксно ниво на значајност, големите примероци речиси секогаш водат до отфрлање на H_0. Значи дека мора да се биде крајно внимателен при интерпретација на доказите кога примерокот е голем, бидејќи тогаш секое мало отстапување од H_0 речиси сигурно ќе биде детектирано од страна на тестот, иако таквото отстапување има мало практично значење.

13.2. Хи-квадрат тестови

Хи-квадрат ($\chi 2$) тестот се користи да мери колку добро претпоставената распределба се согласува со податоците добиени од случајно избран, независен примерок добиен со повторување на некој експеримент. Разгледуваме експеримент со k исходи $j = 1, \ldots, k$, каде што исходот j се случува со веројатност p_j. Веројатностите p_j не се оценуваат од податоците, туку се сметаат за познати. По n независни повторувања на експериментот, нека N_j е бројот на случувања на настанот j. Класична мера за отстапување на случајна променлива N_j од нејзиното очекување $n \cdot p_j$ е тежинската сума на квадрати на разликите

$$\sum_{j=1}^{k} w_j (N_j - n \cdot p_j)^2 .$$

Како да се изберат w_1, w_2, ..., w_k? Најдобро би било тоа да се направи така што распределбата да биде колку е можно поедноставна, а тоа се постигнува со ставање $w_j = (n \cdot p_j)^{-1}$. За доволно големо n (тука се користи централната гранична теорема), случајната променлива

$$D = \sum_{j=1}^{k} \frac{(N_j - n \cdot p_j)^2}{n \cdot p_j}$$ како сума на квадрати на $Z(0,1)$ случајни променливи има хи-квадрат распределба со $k - 1$ степени на слобода. Еден степен на слобода е изгубен поради $\sum_{j=1}^{k} N_j = n$. На крај останува само да се исчита веројатноста $p(\chi^2_{k-1} > d)$, каде што d е вредност на D, и тоа ја дава веројатноста на "согласност" на примерокот со веројатносниот модел даден со веројатностите p_j. Јасно е дека поголема вредност за d, значи полошо согласување на податоците со моделот.

Интересно е дека распределбата на D не зависи ниту од распределбата на примерокот, ниту од веројатносниот модел. Од друга страна, таа е доста чувствителна на бројот на разгледуваните исходи k.

ПРИМЕР 13.11 Во еден класичен експеримент испитувани се обликот и бојата на грашокот добиен со вкрстување на видови. Примерок од 556 зрна грашок е испитан при што: 315 биле тркалезно-жолти, 108 тркалезно-зелени, 101 наборано-жолти и 32 наборано-зелени. Според одредена теорија (претходни сознанија) фреквенцијата на видовите зрна е со однос 9 : 3 : 3 : 1. Дали според хи-квадрат тестот експериментот дава резултати во согласност со теоријата?

Решение

Со директна пресметка за d добиваме

$$d = \frac{(315 - 556 \cdot 9/16)^2}{556 \cdot 9/16} + \frac{(108 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(101 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(32 - 556 \cdot 1/16)^2}{556 \cdot 1/16}$$

$= 0.4700$. Таблицата за хи-квадрат распределба дава $p(\chi^2_3 > 0.470) = 0.925$. Значи согласноста со теоријата е одлична. Многу луѓе го сметаат овој резултат за сомнително добар. ■

13.2.1. Согласност на податоците со распределбата

Најчестата примена на хи-квадрат тестот е за тестирање на хипотезата дали претпоставена распределба за некој експеримент (случајна променлива) е "добра", т.е. во колкава мера таа се согласува со податоците од земен случаен примерок. Општиот облик на d е

$$d = \sum_{\substack{\text{по сите}\\ \text{полиња}}} \frac{(\text{вредности од примерок} - \text{очекувани вредности})^2}{\text{очекувани вредности}}.$$

Во случај на комплетно претпоставена распределба (познати очекувани вредности), постапката би можела да се сумира во следните 5 чекори:

1) Просторот на примерокот се дели на k дисјунктни интервали A_j, $j = 1, \ldots, k$. Нека N_j е бројот на податоци во интервалот A_j, и притоа ако некој A_j содржи < 5 примероци спој го со A_{j-1} или A_{j+1} за секој интервал да има ≥ 5 податоци;

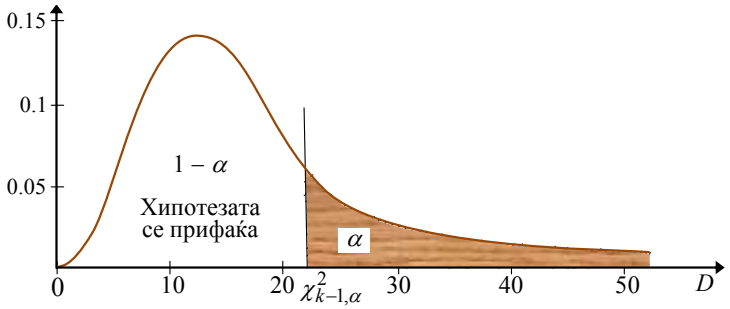
2) Пресметај ги (теоретските) веројатности $p(A_j) = p_j$, $j = 1, \ldots, k$ според претпоставената распределба (хипотезата).

3) Пресметај го d;

4) Избери вредност за α - јачина на тестот и најди ја вредноста $\chi^2_{k-1,\alpha}$ (вообичаено од таблицата на χ^2 распределба). Во пракса, за α се земаат вредности 0.001, 0.01 или 0.05 и притоа вредностите $0.01 \le \alpha \le 0.05$ (меѓу 1% и 5%) се сметаат за *скоро значајни*, $0.001 \le \alpha \le 0.01$ (меѓу 0.1% и 1%) за *значајни* и $\alpha \le 0.001$ (под 1%) за *високо значајни*.

5) Хипотезата (претпоставката за распределбата) се отфрла ако $d > \chi^2_{k-1,\alpha}$, во спротивно таа се прифаќа.

На сл. 13.1 е прикажана соодветната хи-квадрат распределбата.



Слика 13.1 *Хи-квадрат распределба со k – 1 степени на слобода*

Како што се гледа на сликата, областите на прифаќање и отфрлање на хипотезата се аналогни со истите области од параметарските тестови.

ПРИМЕР 13.12 Примерок од 300 сијалици се тестирани на трајност (во денови) и резултатите се дадени во следната табела:

Трајност, t	Број
$t < 100$	121
$100 \le t < 200$	78
$200 \le t < 300$	43
$300 \le t$	58

Претпоставка е дека трајноста на сијалиците има експоненцијална распределба со очекувано траење (просек) од $1/\lambda = 200$ денови, т.е. $\lambda = 0.005$, па

$$f(t) = 0.005e^{-0.005t}, \text{ за } t \ge 0.$$

Тестирај ја хипотезата со значајност $\alpha = 5\%$.

Решение

Областите A_j, $j = 1, 2, 3, 4$ се дадени во левата колона на табелата. Веројатностите $p(A_j)$ се

$$p_1 = p(A_1) = \int_0^{100} 0.005e^{-0.005t}dt = 1 - e^{-0.5} = 0.39;$$

$$p_1 = p(A_2) = \int_{100}^{200} 0.005e^{-0.005t}dt = 1 - e^{-1} - 0.39 = 0.24;$$

и на ист начин $p_3 = p(A_3) = 0.15$ и $p_4 = p(A_4) = 0.22$. За d добиваме

$$d = \frac{(121 - 300 \cdot 0.39)^2}{300 \cdot 0.39} + \frac{(78 - 300 \cdot 0.24)^2}{300 \cdot 0.24} + \frac{(43 - 300 \cdot 0.15)^2}{300 \cdot 0.15} + \frac{(58 - 300 \cdot 0.22)^2}{300 \cdot 0.22} = 1.7$$

и сега од таблицата за χ^2 распределба добиваме $\chi^2_{0.05,3} = 7.815$. На крај, имајќи предвид дека $d = 1.7 < 7.815 = \chi^2_{0.05,3}$, заклучуваме дека со ниво на значајност од 5% (веројатност на грешка), се прифаќа дека примерокот ја има претпоставената експоненцијална распределба. ■

Многу почеста ситуација е кога параметрите на претпоставената распределба се непознати и треба да се оценат од податоците. Тогаш во изразот за случајната променлива D, треба p_j да се заменат со соодветните оценки $\hat{p}_j$

$$D = \sum_{j=1}^{k} \frac{(N_j - n \cdot \hat{p}_j)^2}{n \cdot \hat{p}_j},$$

бидејќи распределбата е со проценетите параметри, па и пресметаните веројатности од неа се исто така само проценети. Притоа се менува и распределбата на D, но за среќа таа останува да биде хи-квадрат но сега со $k - r - 1$ (наместо $k - 1$) степени на слобода, каде што r е бројот на параметрите во распределбата што треба да се оценат.

ПРИМЕР 13.13 Направено е испитување на нивото на холестерол во крвта на примерок од 49 луѓе (во централна америка) во 1964 година. Резултатите се во mg/L (за претварање во европски mmol/L делиме со 38.67) се:

204	108	140	152	158	129	175	146	157	174	192	194	144	152	135	223	145
231	115	131	129	142	114	173	226	155	166	220	180	172	143	148	171	143
124	158	144	108	189	136	136	197	131	95	139	181	165	142	162		

Дали е веројатно дека нивото на холестеролот за оваа популација е нормално распределено ако се знае од претходни испитувања дека очекувањето μ и стандардната девијација σ се 150 и 30 соодветно?

Решение

За поедноставно, земаме 7 еднаковеројатни области $A_j = (a_j, b_j), j = 1, ..., 7$ коишто за стандардната нормална распределба се: $(-\infty, -1.07)$, $(-1.07, -0.57)$, $(-0.57, -0.18)$, $(-0.18, 0.18)$, $(0.18, 0.57)$, $(0.57, 1.07)$ и $(1.07, \infty)$. За $\mu = 150$ и $\sigma = 30$, овие области се трансформираат во: ($x_j = 30a_j+150$, $y_j = 30b_j+150$): $(-\infty, 117.9)$, $(117.9, 132.9)$, $(132.9, 144.6)$, $(144.6, 155.4)$, $(155.4, 167.1)$, $(167.1, 182.1)$ и $(182.1, \infty)$.

Од податоците се добиваат точкастите оценки на просекот и стандардната девијација, $\hat{\mu} = 157.02$, $\hat{\sigma} = 31.42$. Податоците и веројатностите $\hat{p}_2$ во секој интервал се:

Област	Број	$\hat{p}_j$
$(-\infty, 117.9)$	5	0.116
$(117.9, 132.9)$	5	0.115
$(132.9, 144.6)$	11	0.124
$(144.6, 155.4)$	6	0.125
$(155.4, 167.1)$	6	0.145
$(167.1, 182.1)$	7	0.163
$(182.1, \infty)$	9	0.212

Веројатноста, на пример во интервалот (117.9, 132.9) се добива од $\hat{p}_2 = p(117.9 < X < 132.9) = p(-1.25 < z < -0.77) = 0.115$.

Понатаму $d = \dfrac{(5 - 49 \cdot 0.116)^2}{49 \cdot 0.116} + \ \ldots \ = 4.60.$

Од $d = 4.60 < 9.448 = \chi^2_{0.05,4} = \chi^2_{0.05,k-r-1}$, заклучуваме дека нормалната распределба добро ги опишува податоците.

(Ова е построго барање од $\chi^2_{0.05,6} = 12.592$).

P-вредноста на тестот во овој пример е 0.331, што апсолутно оди во прилог на прифаќање на хипотезата. ■

13.2.2. Независност во табели

Во многу ситуации, примерокот може да се класифицира според 2 критерума. Притоа од интерес е да се дознае дали овие 2 методи на класификација се статистички независни. Нека првиот метод на класификација има r нивоа, а вториот c нивоа. Нека O_{ij} биде фреквенцијата за нивото i од првата и нивото j од втората класификација. Овие податоци нормално се сместени во табела што обично се нарекува $r \times c$ случајна табела.

Редици \ Колони	1	. . .	c
1	O_{11}	. . .	O_{1c}
⋮	⋮	⋮	⋮
r	O_{r1}	. . .	O_{rc}

Ние би сакале да ја тестираме хипотезата дека класификациите редица-колона се независни. Се разбира за ова нема точна процедура, но тоа може да се направи приближно ако n е доволно големо.

Нека p_{ij} е веројатноста дека случајно избран податок паѓа во полето ij, кога двете класификации се независни. Тогаш $p_{ij}= u_i\, v_j$, каде што u_i е веројатноста дека случајно избран податок е во редицата на класата i, а v_j е веројатноста дека случајно избран податок е во колоната на класата j. Оценките на u_i и v_j се

$$\hat{u}_i = \sum_{j=1}^{c} O_{ij} \quad \text{и} \quad \hat{v}_i = \sum_{i=1}^{r} O_{ij}\,,$$

па очекуваната фреквенција во секое поле е

$$E_{ij} = n\hat{u}_i\hat{v}_j = \frac{1}{n}\sum_{j=1}^{c} O_{ij}\sum_{i=1}^{r} O_{ij}\,.$$

За доволно големо n, случајната променлива

$$D = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

има приближно хи-квадрат распределба со $(r-1)(c-1)$ степени на слобода, ако хипотезата за независност е точна. Хипотезата ја отфрламе кога вредноста d за D е $d > \chi^2_{\alpha,(r-1)(c-1)}$.

ПРИМЕР 13.14 Направена е студија за врската меѓу опременоста на бензиските станици и цената на горивото. Испитани се 441 бензиски станици. Податоците се дадени во табелата:

Опременост \ Цени	Високи	Средни	Ниски	
Субстандардна	24	15	17	56
Стандардна	52	73	80	205
Модерна	58	86	36	180
	134	174	133	

Дали со ниво на значајност од 0.01, добиените податоци сугерираат дека условите на бензиските станици и цената на горивото се независни? Образложи го заклучокот.

Решение

Очекуваната ценовна политика е пресметана со (сума редица)×(сума колона)/(број на податоци) и е дадена во следната табела:

Опременост	*Цени* Високи	Средни	Ниски
Субстандардна	17.02	22.10	16.89
Стандардна	62.29	80.88	61.83
Модерна	54.69	71.02	54.29

На пример, очекувањето $M_{1,2} = 174 \cdot 56/441 = 22.0952$. Понатаму

$$d = (24-17.02)^2/17.02 + \ldots + (36-54.29)^2/54.29 = 22.47$$

И сега поради $d = 22.47 > 13.227 = \chi^2_{0.01,4} = \chi^2_{0.01,(3-1)(3-1)}$, хипотезата за независност се отфрла. Заклучуваме дека опременоста не ни дава знаење за ценовната политика на бензиските станици, и обратно. ■

Понекогаш наместо тестирање на хипотеза, брза информација за степенот на зависност може да се добие преку *коефициентот на зависност* (Contingency coefficient). Пирсоноиот коефициентот на зависност се пресметува едноставно со

$\sqrt{\dfrac{d}{d+r\cdot c}}$, каде што d е вредноста на статистиката од хи-квадрат тестот на хипотезата за независност на класификацијата редица-колона во табелата, а r и c се бројот на редици и колони соодветно. Очигледно вредноста на коефициентот на зависност оди од блиску до 0 (нема зависност) накај 1 (максимална зависност).

Поради прецизноста, почесто се користи Крамеровиот (Cramer) коефициентот на зависност

$\sqrt{\dfrac{d}{r\cdot c\cdot \min\{r,c\}}}$, бидејќи именителот е максималната вредност што може d да ја достигне.

Вредноста на коефициентите на зависност во примерот 13.14 се:

$$\text{Пирсоновиот} = \sqrt{\frac{22.47}{22.47+9}} = 0.845, \text{ а Крамеровиот} = \sqrt{\frac{22.47}{9\cdot 3}} = 0.912.$$

Високите вредности на коефициентите сугерираат орфрлање на претпоставката за независност на опременоста и ценовната политика на бензиските станици што е очигледно во согласност со тестот од примерот 13.14.

ЗАДАЧИ

1. Дали следните тврдења се легитимни статистички хипотези:

   а) H: $\sigma > 100$; б) H: $\bar{x} = 24$; в) H: $S < 8$; г) H: $\sigma_1/\sigma_2 < 1$;

   д) H: $\lambda \leq 0.01$, каде што λ е параметарот на експоненцијалната распределба;

2. Во секој од следните случаи одговори дали проблемот на тестирање хипотези е правилно формулиран:

   а) H_0: $\mu = 11.2$, наспроти H_A: $\mu \neq 11.2$;

   б) H_0: $\sigma > 9$, наспроти H_A: $\sigma = 9$;

   в) H_0: $S = 5$, наспроти H_A: $S < 5$;

   г) H_0: $p = 0.25$, наспроти H_A: $p = 0.35$;

   д) H_0: $S^2 = 5.1$, наспроти H_A: $S^2 > 5.1$;

   ѓ) H_0: $\sigma = 3.5$, наспроти H_A: $\sigma < 4.2$.

3. Продавница продава автомобилски гуми од 2-ра класа за кои тврди дека имаат просечен животен век од 30000 километри со стандардна девијација $\sigma = 1500$ километри. Даден е примерок од 16 такви гуми на тестирање, при што е добиен просечен животен век од 30822 километри. Под претпоставка дека животниот век на гумите е со нормална распределба:

   а) Дали може да се заклучи со $\alpha = 0.01$ дека гумите се дури подобри од она што го тврдат во продавницата;

   б) Ако витинскиот животен век на гумите е 31000 километри, колкава е веројатноста дека тестот тоа нема да го открие?

   в) Ако витинскиот животен век на гумите е 31000 километри колкав треба да биде примерокот за грешката β да биде најмногу 0.1?

4. Конструран е нов тип вештачко срце во главно од титаниум и пластика што работи на батерии што треба да се полнат на секои 4 часа. Примерок од 50 батерии е испитуван на должината на траење при што е добиен просек од 4.05 часа. Ако траењето на батериите е со нормална распределба со $\sigma = 0.2$ часа, определи:

   а) Дали може да се заклучи дека просечното траење на батериите надминува 4 часа. Користи $\alpha = 0.05$;

   б) Пресметај ја јачината на тестот кога вистинското траење на батериите е 4.5 часа;

в) Колкав треба да биде примерокот кога вистинското траење на батериите е 4.5 часа ако сакаме јачината на тестот да биде најмалку 0.9?

5. Пожелен процент на SiO_2 во одреден тип на цемент е 5.5%. За ова да се провери во една фабрика за производство на цемент, земен се 16 независни примероци за анализа. Просекот на добиената содржина на SiO_2 бил 5.25% со стандардна девијација од 0.3%. Под претпоставка дека процентот на SiO_2 е нормално распределен, определи:

   а) Дали добиените податоци со ниво на значајност 0.01 индицираат дека просечниот процент на SiO_2 е различен од 5.5?

   б) Ако вистинскиот просек е 5.6, колкава е веројатноста дека ова отстапување од 5.5 тестот нема да го открие?

   в) Колкав треба да биде примерокот n, за грешката од тип 2, β да е ≤ 1?

6. Дадени се податоци за времето на поправка (во минути) на прекини на пругата за конкретна железничка линија: 159, 120, 480, 149, 270, 547, 340, 43, 228, 202, 240 и 218. Ако времето на поправка е со приближно нормална распределба (провери со веројатносен график), определи:

   а) Дали има доволно докази да се тврди дека просечното време на поправка надминува 200 минути со ниво на значајност од 0.05?

   б) Колкава е веројатноста за грешка од тип 2, кога вистинското просечно време на поправка би било 300 минути?

7. Во едно списанието меѓу другото, дадена е телесната температура на 25 жени: 97.8, 97.2, 97.4, 97.6, 97.8, 97.9, 98.0, 98.0, 98.0, 98.1, 98.2, 98.3, 98.3, 98.4, 98.4, 98.4, 98.5, 98.6, 98.6, 98.7, 98.8, 98.8, 98.9, 98.9, и 99.0 (во Фаренхајтови°). (Врската меѓу степените е Целзиусови° = (5/9)(Фаренхајтови° – 32)). Ако телесната температура кај жените е со приближно нормална распределба (провери со веројатносен дијаграм), определи:

   а) Дали од податоците може да се заклучи дека просечната температура е различна од 98.6 (37 целзиусови). Земи $\alpha = 0.05$. Најди ја P-вредноста;

   б) Пресметај ја јачината на тестот ако вистинската просечна температура е 98.0;

   в) Колкав би требало да биде примерокот кога вистинската просечна температура е 98.2 степени, ако би сакале јачината на тестот да биде најмалку 0.90?

8. Содржината на катран во примерок од 30 цигари е измерена на: 1.542, 1.622, 1.440, 1.459, 1.598, 1.585, 1.466, 1.608, 1.533, 1.498, 1.532, 1.546, 1.520, 1.532, 1.600, 1.466, 1.494, 1.478, 1.523, 1.504, 1.499, 1.548, 1.542, 1.397, 1.545, 1.611, 1.626, 1.511, 1.487, 1.558.

а) Може ли да се поддржи тврдењето дека просечната содржина на катран во цигарите надминува 1.5, со ниво на значајност $\alpha = 0.05$. Најди ја P-вредноста на тестот;

б) Пресметај ја β ако вистинската просечна содржина на катран е 1.6;

в) Колкав треба да биде примерокот кога вистинската просечна содржина на катран е 1.6, ако сакаме јачината на тестот да биде најмалку 0.8?

9. Земен примерок од 150 крвни групи од една донација на крв. Се покажало дека 82 од нив се од 0-та крвна група. Дали ова сугерира дека процентот на застапеност на 0-тата крвна група во донацијата се разликува од истата застапеност во популација што се проценува на околу 40%? Тестирај ја хипотезата за $\alpha = 0.01$. Дали заклучокот би се сменил за $\alpha = 0.1$?

10. Заедничка карактеристика на дебелите луѓе е дека нивниот индекс на телесната маса (BMI = тежина/висина^2, изразени во метри и килограми) е најмалку 30. Во еден примерок од вработени жени: 262 имале BMI < 25, 159 имале 25 ≤ BMI < 30 и 120 имале BMI ≥ 30. Дали овој примерок оди во прилог на тврдењето дека 20% од луѓето се дебели?

а) Тестирај ја хипотезата за $\alpha = 0.05$;

б) Објасни ги сценаријата за грешките од тип 1 и тип 2;

в) Колкава е веројатноста (грешка од тип 2) да не може да се заклучи дека повеќе од 20% од популацијата е дебела, кога вистинскиот процент на дебели луѓе е 25%?

11. Производител на интраокуларни леќи има нова машина за која тврди дека прави површински дефекти на не повеќе од 2% од полираните леќи. Во примерок од 250 леќи пронајдени се 6 дефектни.

а) Дали ова е во согласност со тврдењето на производителот? Користи $\alpha = 0.05$?

б) Најди ја P-вредноста на тестот.

12. Примерок од 500 регистрирани гласачи во Феникс е анкетиран за тоа дали би користеле ново еколошко гориво за автомобили за да се намали аеро-загадувањето. Ако повеќе од 315 гласачи се изјаснат позитивно, може да се заклучи дека најмалку 60% од гласачите се за користење на еколошкото гориво.

а) Најди ја веројатноста на грешката од тип 1, ако точно 60% од гласачите се за користење на еколошкото гориво;

б) Колкава е грешката од тип 2, ако 75% од гласачите се изјасниле за користење на новото гориво.

13. Производител на автомобилски гуми го испитува животниот век на гуми со нов тип каучук. За таа цел се направени 16 гуми и тестирани на пат, при што е добиен просечен животен век од 60139.7 километри со стандардна девијација од 3645.94 километри.

    а) Дали може да се заклучи со $\alpha = 0.05$ дека стандардната девијација на животниот век на гумите надминува 200 километри. Направи соодветна претпоставка за распределбата.

    б) Најди ја P-вредноста на тестот.

14. Дали во хи-квадрат тестот е поверојатно една хипотеза да се прифати кога е $\alpha = 0.05$ или $\alpha = 0.01$? Образложи.

15. Провери ја рамномерноста на појавување на цифрите 0 до 9 во децималите на ирационалниот број e. Фреквенцијата на појавување на цифрите во 5000 децимали на бројот e се дадени во табелата:

Цифра	0	1	2	3	4	5	6	7	8	9
Фреквенција	4947	5056	4969	5026	4966	5046	5132	4959	4972	4925

16. Во следната табела се даден бројот на дефектни производи на 5 производствени ленти:

Лента	1	2	3	4	5
Дефектни	15	27	31	19	11

    Тестирај ја хипотезата дека пропорцијата на дефектни производи на сите 5 производствени ленти е еднаква. Користи $\alpha = 0.05$.

17. Во следната табела се дадени броевите на предизвикани сообраќајки од страна на 7842 возачи (во Калифорнија):

Број на сообраќајки	**Број на возачи**
0	5147
1	1859
2	595
3	167
4	54
5	14
> 5	6

    Врз база на овие податоци, тестирај ја хипотезата дека $X =$ "број на несреќи на возач" е случајна променлива со Пуасонова распределба со очекување $\lambda = 0.08$ по година, со ниво на значајност од 1%.

    Упатство: Хипотезата е дека распределбата е

    $$p(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} = \frac{0.48^x e^{-0.48}}{x!}, \quad x = 0, 1, 2, \ldots .$$

    Забележи дека ако бројот за > 5 сообраќајки беше помал од 5, нив ќе ги уфрлевме во > 4 сообраќајки.

18. Во студија за грешки во една електронска компонента, заклучено е дека се можни 4 типа грешки. Компонентата има 2 позиции за монтирање. Резултатите се дадени во следната табела:

Позиција на монтирање	Тип грешка A	B	C	D
1	22	46	18	9
2	4	17	6	12

Дали може да се заклучи дека типот на грешката е независен од позицијата на монтирање, Користи ниво на значајност $\alpha = 0.01$. Пресметај ја и P-вредноста на тестот.

14

Оценки и тестови со два примерока

Во многу ситуации потребно е да се споредат ефектите на две различни активности, како на пример, различен третман на пациенти (плацебо наспроти третман) во медицината или пак различен пристап во производството на најразлични производи во инженерството. Ако групирањето на објектите што се користат во споредбата не е контролирано, статистичката анализа обично се нарекува *набљудувачка*. Тешкотиите со ваквите студии е што иако заклучоците од статистичката анализа може да сугерираат значајни разлики во ефектите од две активности, таа не дава одговор на прашањето дали тоа се должи точно на разликите што се предмет на истражување (видот на третманот или постапка во производството) или на некои други влијанија, коишто можеби се и непознати. Со други зборови, со ваквите студии не можеме да бидеме сигурни дека сме ја нашле причинско - последичната врска.

На пример, бројни студии покажуваат значајно зголемен број на случаи на респираторни проблеми како и канцер на белите дробови кај пушачите. Но како да се покаже дека една битна причината за канцерите е токму пушењето? Многу од испитаниците станале пушачи пред почнувањето на студиите, така што истражувањето е типично набљудувачко. Многу други фактори различни од пушењето може да имаат поважна улога во развојот на канцерот на белите дробови отколку пушењето.

Досега разгледуваните оценки и тестови секогаш се однесуваа на еден просек, една пропорција или една дисперзија. Во оваа глава ќе разгледаме некои методи за наоѓање интервални оценки и тестирање хипотези кога се вклучени два параметра од две различни популации. На пример, да претпоставиме дека треба да се спореди пропорцијата на дефектни производи на две производствени ленти. Ние би можеле да земеме примерок од првата и примерок од втората лента и да ја тестираме хипотеза H_0: $p_1 - p_2 = 0$ ($p_1 = p_2$) наспроти H_A: $p_1 - p_2 > 0$ ($p_1 > p_2$). Некој би можел да состави интервална оценка $p(a < p_1 - p_2 < b) = 1 - \alpha$, и од неа да заклучува за разликите во квалитетот на производството на двете ленти.

14.1. Оценки и тестови за разлика на проседи

Тука ќе ги разгледаме оценките и хипотезите што се однесуваат на разликата на просеците $\mu_1 - \mu_2$ земени од две популации со различни распределби. Нека $X_1, X_2, \ldots, X_m$ е примерок земен од популација X со непознат просек μ_1 и нека $Y_1, Y_2, \ldots, Y_n$ е примерок од популацијата Y со непознат просек μ_2. Се разбира, ако големината на примероците зависи од нас, тогаш може да земеме $m = n$, но во општ случај тие ќе имаат различна големина.

Природна центрирана оценка на разликата на просеците на популациите $\mu_1 - \mu_2$ е $\overline{X} - \overline{Y}$ - разликата на просеците на примерокот. Дисперзијата на $\overline{X} - \overline{Y}$ е

$$D(\overline{X} - \overline{Y}) = D\overline{X} + D\overline{Y} = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \quad \Rightarrow \quad \sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\,.$$

Кога σ_1 или σ_2 се непознати, за добивање на $\sigma_{\overline{X}-\overline{Y}}$ мора да се користат соодветните точкасти оценки s_1 и s_2.

Како и во случаите со еден примерок, најпрво претпоставуваме дека и двете популации се со нормална распределба (реално) и со познати дисперзии (не е реално). Додатно, сметаме дека двата примерока се независни (реално). Кога двете популации се со нормална распределба, $\overline{X}, \overline{Y}$ и $\overline{X} - \overline{Y}$ се исто така со нормална распределба. Тогаш случајната променлива

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$$ има стандардна $Z(0,1)$ распределба.

Сега, интервалот на доверба се добива со решавање на неравенството

$$p\left(-z_{\alpha/2} < \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}} < z_{\alpha/2}\right) = 1-\alpha\ ,$$ по $\mu_1-\mu_2$ што дава

$$p\left(\overline{X}-\overline{Y}-z_{\alpha/2}\sqrt{\sigma_1^2/m+\sigma_2^2/n} < \mu_1-\mu_2 < \overline{X}-\overline{Y}+z_{\alpha/2}\sqrt{\sigma_1^2/m+\sigma_2^2/n}\right) = 1-\alpha\ .$$

За тестирање хипотези, ставаме H_0: $\mu_1 - \mu_2 = \Delta_0$, па статистиката $Z = \dfrac{\overline{X}-\overline{Y}-\Delta_0}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}}$ кога H_0 е точна има стандардна $Z(0,1)$ распределба.

Ако вредноста на статистиката е z, тестирањето оди стандардно:

Алнернативана хипотеза	*Критичен регион* (*отфрламе* H_0)
H_A: $\mu_1 - \mu_2 < \Delta_0$	$z \le -z_\alpha$
H_A: $\mu_1 - \mu_2 > \Delta_0$	$z \ge z_\alpha$
H_A: $\mu_1 - \mu_2 \ne \Delta_0$	$z \le -z_{\alpha/2}$ или $z \ge z_{\alpha/2}$.

P-вредноста на тестот, како и грешката од тип 2, β (т.е. јачината на тестот $1-\beta$) се пресметуваат сосема аналогно на случаите со еден примерок. Истото важи и за едностраните интервални оценки.

Кога примероците се доволно големи, $m, n \ge 30$, поради централната гранична теорема, претпоставката за нормална распределеност на популациите како и претпоставката за познати дисперзии не се потребни. Тогаш, како и кај случаите со еден примерок, едноставно σ_1 и σ_2 се заменуваат со точкастите оценувачи s_1 и s_2.

ПРИМЕР 14.1 Од 215 мажи што дипломирале медицина на Харвард и починале во период 1974-1977 година, 125 работеле полно работно време како лекари, додека останатите 90 имале академска кариера (на универзитети). Просечниот животен век на првата група (лекарите) бил 48.9 години по дипломирањето со стандардна девијација од 14.6, додека на втората група (академци) бил 43.2 години по дипломирањето со стандардна девијација од 14.4. Дали овие податоци сугерираат дека академската кариера кај лекарите го скратува животот. Користи $\alpha = 0.01$.

Решение

Нека μ_1 = "просечен животен век на лекарите по дипломирањето" и μ_2 = "просечен животен век на академците по дипломирањето ". Тогаш тестираме H_0: $\mu_1 - \mu_2 = 0$, наспроти H_A: $\mu_1 - \mu_2 > 0$ (тука $\Delta_0 = 0$). Имајќи предвид дека $z_\alpha = z_{0.01} = 2.326$, поради

$$z = \frac{48.9 - 43.2}{\sqrt{14.6^2/125 + 14.4^2/90}} = 2.8467 \ > \ 2.326$$, ја отфрламе H_0.

P-вредноста на тестот е $1 - \Phi(2.8647) = 0.0022$, што оди во прилог на отфрлањето на H_0 ($P < 0.01$). Значи податоците одат во прилог на заклучокот дека дипломираните на медицина од Харвард што работат лекарска пракса имаат подолг живот од оние што отишле на градење академска кариера. ■

Да забележиме дека во овој пример е направена ретроспективна статистичка студија. "Нормален" пристап би бил да се земе примерок и од него да се поделат докторите на 2 групи според нивниот ангажман по дипломирањето. Дали овој статистички значаен резултат се должи навистина на разликата во медицинската пракса по дипломирањето или пак е резултат на други фактори: како годините на студирање, начинот на исхрана, физичката активност итн., немаме одговор.

Грешката од 2 тип, прифаќањето на H_0 кога е точна H_A, т.е. $\mu_1 - \mu_2 = \Delta_1 > \Delta_0$ се наоѓа стандардно

$$\beta = p\left(\frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} < z_\alpha \mid \mathrm{H_A} \right) = p\left(\frac{\overline{X} - \overline{Y} - \Delta_1}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} < z_\alpha + \frac{\Delta_0 - \Delta_1}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \right)$$

$$= p\left(z < z_\alpha + \frac{\Delta_0 - \Delta_1}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \right) = \Phi\left(z_\alpha + \frac{\Delta_0 - \Delta_1}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \right)$$, што е сосема аналогно на случајот со 1 примерок: H_A: $\mu = \mu_1 > \mu_0 \ \Rightarrow \ \beta = \Phi\left(z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \right)$.

Ако разликата во животниот век на дипломираните медицинари на Харвард меѓу оние што се бавеле со лекарска практика и оние со академска кариера била 3 години, веројатноста дека тестот тоа нема да го открие е

$$\beta(3) \ = \ \Phi\left(2.326 + \frac{0 - 3}{\sqrt{14.6^2/125 + 14.4^2/90}} \right) \ = \ \Phi(0.8277) \ = \ 0.7961 .$$

Оваа висока грешка од тип 2 се должи на ниското $\alpha = 0.01$ и малата разлика од 3 години. На пример, ако разликата е 8 години, веројатноста дека тестот тоа нема да го открие е $\Phi(-1.6694) = 0.0475$.

Потребните големини на примероците m и n може приближно да се најдат од $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} = \frac{(\Delta_1 - \Delta_0)^2}{(z_\alpha + z_\beta)^2}$, каде што α е зададената грешка од тип

1, а β е зададена грешка од тип 2 кога $\mu_1 - \mu_2 = \Delta_1 \neq \Delta_0$. За двостран тест, како и претходно, само α се заменува со $\alpha/2$.

Кога барем еден од примероците е мал и дисперзијата непозната (реална ситуција во пракса), а распределбите на популациите непознати, не ни останува ништо друго освен да ја претпоставиме распределбата на популациите и претпоставката да ја провериме користејќи хистограм или некоја друга приближна постапка. Сепак и во ваквите ситуации, се покажува дека претпоставката за нормална распределба на популациите е типично најразумен избор. Тоа може приближно да се провери со нормален веројатносен дијаграм. Сепак, под претпоставка за нормалност на популациите, а недоволна големина на барем еден примерок, S_1 или S_2 повеќе не се добри апроксимации на σ_1 или σ_2, па мора да се интерпретираат како случајни променливи што повторно не води до статистика што има приближно студентова распределба.

Поточно, нека $X_1, X_2, \ldots, X_m$ е примерок земен од популација X со непознат просек μ_1 и нека $Y_1, Y_2, \ldots, Y_n$ е примерок од популацијата Y со непознат просек μ_2, каде што X и Y се приближно нормално распределени. Ако барем еден, m или n е недоволно големо ($m < 30$ или $n < 30$), може да сметаме дека случајната променлива

$T = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/m + S_2^2/n}}$ има приближно студентова распределба со

$\nu = \dfrac{(S_1^2/m + S_2^2/n)^2}{(S_1^2/m)^2/(m-1) + (S_2^2/n)^2/(n-1)}$ степени на слобода (ν се заокружува на првиот помал цел број).

Сега стандардно го добиваме интервалот на доверба

$p\left(\mu_1 - \mu_2 \in \overline{X} - \overline{Y} \pm t_{\alpha/2,\nu}\sqrt{S_1^2/m + S_2^2/n}\right) = 1 - \alpha$. За тестирање хипотези, ставаме H_0: $\mu_1 - \mu_2 = \Delta_0$, па статистиката

$T = \dfrac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S_1^2/m + S_2^2/n}}$ кога H_0 е точна има студентова распределба.

Ако t е вредност на статистиката, тестирањето оди стандардно:

Алнернативана хипотеза	*Критичен регион* (*отфрламе* H_0)
H_A: $\mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha/2,\nu}$
H_A: $\mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha/2,\nu}$
H_A: $\mu_1 - \mu_2 \neq \Delta_0$	$t \leq -t_{\alpha/2,\nu}$ или $t \geq t_{\alpha/2,\nu}$.

ПРИМЕР 14.2 Анализирани се два катализатора на некој хемиски процес, за да се спореди нивното дејство. Првиот, што тековно се користи, е поскап од вториот. Направен е тест да се испита дали премин на вториот катализатор ќе влијае на просечниот исход на хемискиот процес. Резултатите се дадени во табелата:

Катализатор 1	91.50	94.18	92.18	95.39	91.79	89.07	94.72	89.21
Катализатор 2	89.19	90.95	90.46	93.21	97.19	97.04	91.07	92.75

Дали податоците сугерираат дека двата катализатора имаат во просек исто влијание на хемискиот процес. Користи $\alpha = 0.05$.

Решение

Нека μ_1 = "просечно влијание на катализаторот 1" и μ_2 = " просечно влијание на катализаторот 2". Тестираме хипотеза H_0: $\mu_1 - \mu_2 = 0$, наспроти H_A: $\mu_1 - \mu_2 \neq 0$.

Од податоците пресметуваме: $\bar{x}_1 = 92.255$, $s_1 = 2.39$, $\bar{x}_2 = 92.733$ и s_2 =2.98. За степените на слобода добиваме $\nu = \dfrac{(2.39^2/8 + 2.98^2/8)^2}{(2.39^2/8)^2/7 + (2.98^2/8)^2/7} =$ 13.37 $\approx$ 13. Оттука се добива критичната точка $t_{\alpha/2,\nu} = t_{0.025,13} = -2.16$, и поради

$$t = \frac{92.255 - 92.733}{\sqrt{2.39^2/8 + 2.98^2/8}} = -0.3539 > -2.16,$$ и се разбира $-0.3539 < 2.16$, не ја отфрламе H_0.

P-вредноста на тестот е 0.3645 што оди во прилог на прифаќање на H_0 ($P > 0.025$). Значи податоците одат во прилог на користење на вториот катализатор бидејќи во просек тој исто влијае на исходот на хемискиот процес, а е со помала цена. ■

Во минатото доста популарна била комбинираната (pooled) t процедура во чија основа била додатната претпоставка за еднакви дисперзии $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Тогаш статистиката станува

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/m + 1/n)}}$$ со стандардна Z(0,1) распределба.

За оценка на дисперзијата σ^2 (кога е непозната), на прв поглед би можеле да користиме нешто како $(S_1^2 + S_2^2)/2$, но поради различната големина на примероците, подобро е да се земе комбиниран оценувач

$S_c^2 = \frac{m-1}{m+n-2} S_1^2 + \frac{n-1}{m+n-2} S_2^2$, каде што првиот собирок учествува со $m - 1$, а вториот со $n - 1$ во вкупните $m + n - 2$ степени на слобода. Заменувајќи го σ^2 со S_c^2, наместо Z со стандардна нормална, се добива случајна променлива T со студентова распределба и $m + n - 2$ степени на слобода. Ваквиот комбиниран t тест може да се изведе преку принципот на максимална веројатност и има малку помала β при исто α во однос на стандардната постапка со две дисперзии. Сепак, поновите истражувања покажуваат дека тој не е доволно добар во ситуации кога $\sigma_1^2 \neq \sigma_2^2$. Затоа, ваквата постапка за добивање интервална оценка или тестирање хипотеза треба да се користи само во случаи кога има силни аргументи за приближна еднаквост на дисперзиите на двата примерока.

Тестовите што досега ги разгледавме, вообичаено се нарекуваат *неупарени* (unpaired) тестови. Ова е од причина што податоците се обезбедуваат од 2 независни примерока (и еднакво распределени). Кај упарените тестови, примерокот типично се состои од упарени податоци од примерок што се резултат на 2 пати повторен експеримент. На пример, истите пациенти, материјали или производи се испитуваат пред и по некој третман, обработка или промена. Проблемот кај упарените тестови е што 2-та примерока се зависни, па неупарените тестови во таков случај се несоодветни.

Нека повторно $X_1, X_2, \ldots, X_n$ е примерок земен од популација X со непознат просек μ_1 и нека $Y_1, Y_2, \ldots, Y_n$ е примерок од популацијата Y со непознат просек μ_2, каде што X и Y се приближно нормално распределени. Кај упарените тестови ги имаме разликите $R_j = X_j - Y_j$ што се независни. Случајната променлива $R = X - Y$ има нормална распределба (поради нормалноста на X и Y) со очекување $\mu_R = E(X - Y) = \mu_1 - \mu_2$ и дисперзија $DR = D(X - Y) \neq DX + DY$ (поради зависност на X и Y). Сега веднаш на изведениот примерок $R_1, R_2, \ldots, R_n$ може да се примени студентов тест за еден примерок со $n - 1$ степени на слобода.

За статистика $T = \frac{\overline{R} - \mu_R}{S_R / \sqrt{n}}$, со вредност t тестот е:

Алнернативана хипотеза	*Критичен регион* (*отфрламе* H_0)
H_A: $\mu_R < \Delta_0$	$t \leq -t_{\alpha,n-1}$
H_A: $\mu_R > \Delta_0$	$t \geq t_{\alpha,n-1}$
H_A: $\mu_R \neq \Delta_0$	$t \leq -t_{\alpha/2,n-1}$ или $t \geq t_{\alpha/2,n-1}$.

Упарената интервална оценка се добива веднаш од

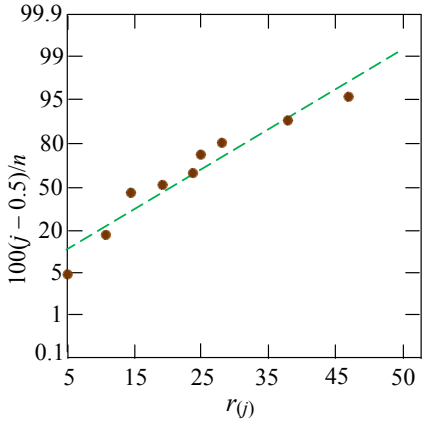
$$p\left(-t_{\alpha/2} \le \frac{\overline{R}-\mu_R}{S_R/\sqrt{n}} \le t_{\alpha/2}\right) = 1-\alpha \quad \Rightarrow \quad p\left(\mu_R \in \overline{R} \pm t_{\alpha/2}\frac{S_R}{\sqrt{n}}\right) = 1-\alpha\,.$$

ПРИМЕР 14.3 Дигиталните медицински слики сместени во соодветна база на поатоци би требало да ја зголемат ефикасноста на нивното користење. Направен е експеримент во кој на 13 лекари, обучени за работа на компјутер, им било мерено времето (секунди) потребно за наоѓање дигитална слика од базата и истиот слајд во картотеката. Резултатите се дадени во табелата:

Слајд	30	35	40	25	20	30	35	62	40	51	25	42	33
Дигитална слика	25	16	15	15	10	20	7	16	15	13	11	19	19

Состави 95% интервал на доверба за разликата на просечно потребното време за наоѓање на слајд и дигитална слика. Тестирај ја (очигледно точната) хипотеза дека дигиталното барање во база е побрзо. Користи $\alpha = 0.05$.

Решение



Приближната линеарност на податоците на веројатносиот дијаграм покажува дека можеме да сметаме на приближно нормална распределба. На дијаграмот има само 9 точки бидејќи има преклопување на точките.

Од вредностите во табелата се добива

$$\bar{r} = 20.5\,, \quad s_R = 11.96\,,$$

а критичната t вредност е $t_{0.025,12} = 2.179$. Сега интервалот на доверба е

$$p(\mu_R \in 20.5 \pm 2.179\frac{11.96}{\sqrt{13}}) = p(\mu_R \in 20.5 \pm 7.2) = p(\mu_R \in (13.3, 27.7)) = 0.95.$$

Нека μ_1 = "просечно време на барање слајд" и μ_2 = "просечен време на барање дигитална слика". Тестираме хипотеза H_0: $\mu_1 - \mu_2 = 0$, наспроти H_A: $\mu_1 - \mu_2 > 0$. Вредноста на статистиката е $t = \dfrac{20.5-0}{11.96/\sqrt{13}} = 6.18$, и поради $t = 6.18 > 1.7823 = t_{0.05,12}$, се разбира ја прифаќаме H_A (што е очекувано).

P-вредност на тестот е 0.0000357 што се разбира, апсолутно сугерира отфрлање на H_0. ■

14.2. Оценки и тестови за разлика на пропорции

Да го разгледаме случајот на разлика на пропорциите p_1 и p_2 на случување на некој настан во две популации X и Y. Веројатностите p_1 и p_2 ги даваат шансите случајно избран објект од популациите да има некоја барана особина. Земаме примерок $X_1, X_2, \ldots, X_m$ и $Y_1, Y_2, \ldots, Y_n$ од популациите, каде што X_k, $Y_k = 0$ (неуспех) или 1 (успех). Ако U и V се број на "успеси" во примероците соодветно, $U = \sum X_k$ и $V = \sum Y_k$ (U и V се независни). Тогаш, кога големината на двата примерока е многу помала од големините на популациите, случајните променливи U и V имаат приближно биномна распределба.

Природен точкаст оценувач за разликата на пропорциите $p_1 - p_2$ е соодветната разлика на пропорциите на примероците $\hat{P}_1 - \hat{P}_2 = U/m - V/n$. Очекувањето и дисперзијата на оценувачот се

$$E(\hat{P}_1 - \hat{P}_2) = \frac{1}{m}EU - \frac{1}{n}EV = p_1 - p_2 \text{ (центрираност)},$$

$$D(\hat{P}_1 - \hat{P}_2) = D\left(\frac{U}{m}\right) + D\left(\frac{V}{n}\right) = \frac{1}{m^2}DU + \frac{1}{n^2}DV = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}.$$

За доволно големи m и n, $\hat{P}_1$ и $\hat{P}_2$ (па и $\hat{P}_1 - \hat{P}_2$) имаат приближно нормална распределба. Оттука случајната променлива

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/m + p_2(1-p_2)/n}}$$ има приближно $Z(0,1)$ распределба.

Да забележиме дека Z не може директно да се користи како статистика бидејќи содржи непознати параметри p_1 и p_2.

За добивање интервална оценка, при доволно големи m и n може да се земе дека $\hat{P}_1(1-\hat{P}_1)/m + \hat{P}_2(1-\hat{P}_2)/n$ е добар оценувач за дисперзијата, па оценката се добива на стандарден начин

$$p(p_1 - p_2 \in \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/m + \hat{p}_2(1-\hat{p}_2)/n}) = 1 - \alpha,$$

каде што $\hat{p}_2$ и $\hat{p}_2$ се конкретни вредности за оценувачите $\hat{P}_1$ и $\hat{P}_2$. Под "доволно големо n" подразбириме $m\hat{p}_1, m(1-\hat{p}_1), n\hat{p}_2$ и $n(1-\hat{p}_2)$ да се најмалку 10. Во спротивно оценката е несигурна.

Кај тестирањето хипотези ситуацијата е нешто посложена. Затоа најпрво за H_0: $p_1 - p_2 = \Delta_0$ го разгледуваме случајот што најчесто се јавu-

ва во пракса $\Delta_0 = 0$, т.е. H_0: $p_1 = p_2 = p$. Значи кога H_0 е точна имаме една пропорција, па примероците може да ги разгледуваме како еден заеднички примерок со големина $m + n$ од една популација. Тогаш природен оценувач на p е

$$\hat{P} = \frac{U+V}{m+n} = \frac{m}{m+n}\hat{P}_1 + \frac{n}{m+n}\hat{P}_2$$ - тежински просек на $\hat{P}_1$ и $\hat{P}_2$.

Сега случајната променлива

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)(1/m + 1/n)}}$$ има приближно $Z(0,1)$ распределба. Се разбира, ова Z не може да се користи како статистика бидејќи p е непознато. Но во ваков случај (со еден параметар p), веќе знаеме (од случајот со еден примерок) дека при замена на p со $\hat{p}$, распределбата останува приближно $Z(0,1)$.

Ако вредноста на статистиката е z, тестирањето на хипотезата H_0: $p_1 - p_2 = 0$, наспроти алтернативите се прави стандардно:

Алнернативана хипотеза	*Критичен регион* (*отфрламе* H_0)
H_A: $p_1 - p_2 < \Delta_0$	$z \le -z_\alpha$
H_A: $p_1 - p_2 > \Delta_0$	$z \ge z_\alpha$
H_A: $p_1 - p_2 \ne \Delta_0$	$z \le -z_{\alpha/2}$ или $z \ge z_{\alpha/2}$.

ПРИМЕР 14.4 Во една студија е испитувано дејството на регуларното земање аспирин кај болните од канцер на дебелото црево. Од 549 пациенти што земале аспирин, кај 81 настапила смрт како последица на канцерот. Од друга страна, од 730 пациенти што не земале аспирин, 141 починале како последица на канцерот. Дали податоците сугерираат дека редовното користење аспирин ги намалува смртните случаи како последица на канцерот. Користи $\alpha = 0.05$.

Решение

Нека p_1 = "пропорцијата на починати што користеле аспирин" и p_2 = " пропорцијата на починати што не користеле аспирин ". Тестираме хипотеза H_0: $p_1 - p_2 = 0$, наспроти H_A: $p_1 - p_2 < 0$, бидејќи би сакале да знаеме дали аспиринот ги намалува смртните случаи од дадениот тип канцер (има индиција во податоците бидејќи ($\hat{p}_1 < \hat{p}_2$).

Од податоците ги пресметуваме точкастите оценки $\hat{p}_1 = 81/549 = 0.1475$, $\hat{p}_2 = 141/730 = 0.1932$ и $\hat{p} = (81+141)/(549+730) = 0.1736$. Тестот е прифатлив бидејќи сите $m\hat{p}_1$, $m(1-\hat{p}_1)$, $n\hat{p}_2$ и $n(1-\hat{p}_2)$ се најмалку 10 (на пример $m\hat{p}_1 = 549 \cdot 0.1475 \approx 81$).

Критичната точка е $z_\alpha = z_{0.05} = -1.645$, и поради

$$z = \frac{0.1475 - 0.1932}{\sqrt{0.1736 \cdot 0.8264(1/549 + 1/730)}} = -2.14 \ < -1.645$$ ја отфрламе H_0.

P-вредноста на тестот е $\Phi(-2.14) = 0.0162$ што е помало од 0.05 што исто така сугерира отфрлање на H_0 со ниво на значајност $\alpha = 0.05$. Но некој што бара поцврст доказ би можел да користи ниво на значајност $\alpha = 0.01$, и тогаш H_0 не би била отфрлена. Во секој случај резултатот е "неубедлив" и би требало да се направат дополнителни испитувања. ■

Грешката од тип 2, прифаќање на H_0 кога е точна H_A, е понезгоден проблем бидејќи тогаш повеќе не важи $p_1 = p_2 = p$. Значи кога H_0 не е точна мора да се користи

$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$, па β повеќе не е функција само од $p_1 - p_2$. Вредностите на β за различните алтернативните хипотези се:

Алнернативана хипотеза	*Вредност на* β
H_A: $p_1 - p_2 < 0$	$1 - \Phi\left(\frac{-z_\alpha\sqrt{\bar{p}\cdot(1-\bar{p})(1/m+1/n)} - (p_1-p_2)}{\sigma}\right)$
H_A: $p_1 - p_2 > 0$	$\Phi\left(\frac{z_\alpha\sqrt{\bar{p}\cdot(1-\bar{p})(1/m+1/n)} - (p_1-p_2)}{\sigma}\right)$
H_A: $p_1 - p_2 \neq 0$	$\Phi\left(\frac{z_{\alpha/2}\sqrt{\bar{p}\cdot(1-\bar{p})(1/m+1/n)} - (p_1-p_2)}{\sigma}\right) -$
каде што	$- \Phi\left(\frac{-z_{\alpha/2}\sqrt{\bar{p}\cdot(1-\bar{p})(1/m+1/n)} - (p_1-p_2)}{\sigma}\right).$
$\bar{p} = (m \cdot p_1 + n \cdot p_2)/(m+n) \approx \hat{p}$	

За примерот 14.4 имаме $\sigma = \sqrt{0.1475 \cdot 0.8525/549 + 0.1932 \cdot 0.8068/730)} =$ 0.021, па кога $p_1 = 0.1475$, $p_2 = 0.1932$ ја добиваме грешката од тип 2,

$$\beta \approx 1 - \Phi\left(\frac{-1.645\sqrt{0.1736 \cdot 0.8264(1/549 + 1/730)} - (0.1474 - 0.1932)}{0.021}\right) =$$

$$= 1 - \Phi(0.505) = 1 - 0.6932 = 0.3068.$$

Во случај на ниво на значајност $\alpha = 0.01$, имаме дека $z_\alpha = -2.326$, па за грешката од тип 2 добиваме $\beta = 1 - \Phi(-0.189) = 1 - 0.425 = 0.575$. Како што може да се очекува, за пониско α добиваме повисока грешка

β. Да забележиме дека овие резултати се доверливи само кога m и n се доволно големи.

За зададени p_1 и p_2 со $p_1 - p_2 = d$, може да се определи големината на примерокот така што да се добие саканото β. За тоа да се направи, на пример за H_A: $p_1 - p_2 < 0$, треба само да се изедначи z_β со аргументот на $\Phi(\cdot)$ (за H_A: $p_1 - p_2 > 0$ изедначуваме $-z_\beta$). Равенките се со 2 непознати m и n, но ако ставиме $m = n$, добиваме

$$n = \frac{\left(z_\alpha \sqrt{(p_1 + p_2)(2 - p_1 - p_2)/2} + z_\beta \sqrt{p_1(1-p_1) + p_2(1-p_2)}\right)^2}{d^2}.$$

За двостран тест, само го заменуваме α со $\alpha/2$.

ПРИМЕР 14.5 Во 1954 год. била тестирана вакцина за детска парализа. Поради објективност, ниту децата ниту администраторите на вакцината не знаеле кој прима вистинска, а кој плацебо (лажна) вакцина. Нека p_1 и p_2 се пропорциите на децата што ќе добијат парализа од плацебо и вистинската група соодветно. Треба да се тестира H_0: $p_1 - p_2 = 0$, наспроти H_A: $p_1 - p_2 > 0$ (инцидентите на парализа треба да се поголеми кај плацебото). Познато е дека случаите на парализа кај децата се $p_1 = 0.0003$ (30 во 100 000), додека со вакцинација би требало барем двојно да се намалат, т.е. $p_2 = 0.00015$. Користејќи ниво на значајност $\alpha = 0.05$, и разумно барање грешката од тип 2 да биде (најмногу) $\beta = 0.1$, определи ја потребната големина на примерокот.

Решение

Претпоставувајќи еднаква големина n на двата примерока (плацебото и вистинската група) добиваме

$$n = \frac{(1.645\sqrt{0.00045 \cdot 1.99955/2} + 1.28\sqrt{0.0003 \cdot 0.9997 + 0.00015 \cdot 0.99985})^2}{(0.0003 - 0.00015)^2} = 171000 \cdot$$

Тестирањето било спроведено на

$m = 201229$ плацебо примерок, со 110 случаи на детска парализа, и

$n = 200745$ примерок со вистинска вакцина, со 33 случаи на детска парализа.

Податоците даваат оценки $\hat{p}_1 = 110/201229 = 0.00054664$, $\hat{p}_2 = 33/200745 = 0.00016439$ и $\hat{p} = 143/401974 = 0.00035574$.

Статистиката е $z = \dfrac{0.00054664 - 0.00016439}{\sqrt{0.00035574 \cdot 0.99964426(1/201229 + 1/200745)}} = 6.4258,$

што е висока вредност и апсолутно сугерира отфрлање на H_0. P-вредноста е практично 0. ■

Понекогаш има потреба да се изведува статистички заклучок за $p_1 - p_2$ кога барем еден од примероците е мал, и нема никаква можност тој да се зголеми. Поуздан метод за такви случаи едноставно нема. Во таква ситуација, најчесто се препорачува и користи таканаречениот Фишер-Ирвин-ов (Fisher-Irwin) тест базиран на хипергеометриската распределба.

Нека X и Y се број на успеси (поволни случаи) во 2-а случајни примерока со големини m и n, и пропорции на успех p_1 и p_2 соодветно. Тестираме хипотеза H_0: $p_1 = p_2$, наспроти алтернатива H_A: $p_1 > p_2$. Значи големи вредности на X одат во прилог на H_A, додека умерените вредности ја подржуваат H_0. Бидејќи комбинираниот $m + n$ примерок содржи вкупно $X + Y = U$ успеси, бројот на начини на избор на X успеси за првиот примерок, оставајќи $U - X$ успеси за вториот е

$\binom{U}{X}\binom{m+n-U}{m-X}$ (од $m + n$ објекти од кои U се успешни бираме m во кои X се од успешните). Сега веројатноста X да прима различни вредности x е

$$p(X = x \mid U \text{ успеси во } m+n \text{ обиди}) = \frac{\binom{U}{x}\binom{m+n-U}{m-x}}{\binom{m+n}{m}} \text{ кога } H_0 \text{ е точна.}$$

Ова е хипергеометриска распределба.

За ова да се искористи за тестирање хипотези, треба да ги сумираме сите веројатности "надесно" од вредноста на X, т.е. за сите вредности на $X \geq$ бројот на успеси од податоците. Оваа веројатност е точно P-вредност на тестот. H_0 се отфрла кога P-вредноста е доволно мала. Сосема аналогно се добиваат P-вредностите за тестот H_A: $p_1 < p_2$ и двостраниот тест H_A: $p_1 \neq p_2$.

ПРИМЕР 14.6 Во производството на одреден тип електронски плочки бројот на несоодветни е релативно голем, и на примерок од 10 плочки само 4 се беспрекорни. По одредена анализа, направени се промени така што на следниот примерок од 10 плочки дури 8 биле соодветни. Дали со ниво на значајност $\alpha = 0.1$ може да се смета дека промените го подобриле производството?

Решение

Ова е типичен проблем од биномен тип, што е погоден за Фишеровиот тест. Тука маме $m = n = 10$, вредноста за X е $x = 8$, а $u = 8 + 4 = 12$. Вредностите $X \geq 8$ се 8, 9 и 10. Соодветните веројатности се:

$$p(X = 8 \mid 12 \text{ успеси}) = \binom{12}{8}\binom{8}{2} / \binom{20}{10} = 0.0750,$$

$$p(X = 9 \mid 12 \text{ успеси}) = \binom{12}{9}\binom{8}{1} / \binom{20}{10} = 0.0095 \text{ и}$$

$$p(X = 10 \mid 12 \text{ успеси}) = \binom{12}{10}\binom{8}{0} / \binom{20}{10} = 0.0003.$$

Оттука P-вредноста на тестот е 0.0750 + 0.0095 + 0.0003 = 0.0848, што значи дека за $\alpha = 0.1$ нултата хипотеза се отфрла. Заклучуваме дека промените го подобриле ороизводството. ■

Поради потребата од фиксирана вредност за $X + Y$ (во нашиот пример не е), за Фишеровиот тест и сличните постапки има доста контраверзни мислења во литературата. Од друга страна, за ситуациите со мал примерок, не постојат алтернативни постапки што би задоволувале.

14.3. Оценки и тестови за разлика на дисперзии

Иако во пракса поретки во однос на просекот и пропорцијата, проблемите на тестирање на разликите во дисперзиите играат важна улога во многу ситуации. За овој тип тестови претпоставуваме дека распределбата на двете популации е приближно нормална.

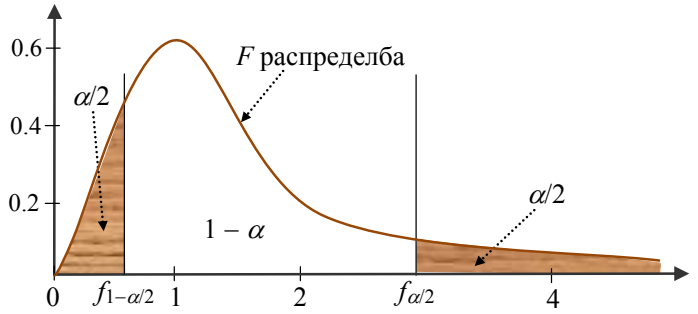
За ваков тест најдобро одговара Фишеровата распределба што веќе ја дискутиравме во поглавјето 4.2. Имено, ако случајната променлива X има има $\chi 2$ распределба со n_1 степени на слобода, а Y има $\chi 2$ распределба со n_2 степени на слобода и ако X и Y се независни, тогаш случајната променлива F дефинирана со

$F = \dfrac{X / n_1}{Y / n_2}$, има Фишерова распределба со (n_1, n_2) степени на слобода.

За добивање на интервална оценка треба да забележиме дека случајната променлива $(n_1 - 1)S_1^2 / \sigma_1^2$ има $\chi 2$ распределба со $n_1 - 1$ степени на слобода и соодветно $(n_2 - 1)S_2^2 / \sigma_2^2$ исто има $\chi 2$ распределба со $n_2 - 1$ степени на слобода. Тогаш статистиката

$F = \dfrac{S_2^2 / \sigma_2^2}{S_1^2 / \sigma_1^2}$, има Фишерова распределба со $(n_1 - 1, n_2 - 1)$ степени на слобода. Интервалниот оценувач го добиваме со решавање на нера-

венството $p\left(f_{1-\alpha/2} < \dfrac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} < f_{\alpha/2}\right) = 1-\alpha$ по σ_1^2/σ_2^2, што го дава интервалот на доверба $p\left(f_{1-\alpha/2}\dfrac{S_1^2}{S_2^2} < \dfrac{\sigma_1^2}{\sigma_2^2} < f_{\alpha/2}\dfrac{S_1^2}{S_2^2}\right) = 1-\alpha$. Забележи дека интервалот е несиметричен поради несиметричноста на Фишеровата распределба (види слика).



При пресметките со Фишеровата распределба, од голема помош е равенството $f_{1-\alpha,(n_1,n_2)} = 1/f_{\alpha,(n_2,n_1)}$.

Тестирањето на разликите на дисперзиите стандардно е од облик

H_0: $\sigma_1^2 = \sigma_2^2$, наспроти H_A: $\sigma_1^2 \neq \sigma_2^2$.

Тогаш, кога H_0 е точна ($\sigma_1^2 = \sigma_2^2$), статистиката $F = S_1^2/S_2^2$ има Фишерова распределба со ($n_1 - 1$, $n_2 - 1$) степени на слобода. Ако вредноста на статистиката е f, тестирањето на хипотезата H_0: $\sigma_1^2 = \sigma_2^2$, наспроти алтернативите се прави стандардно:

Алнернативана хипотеза	*Критичен регион* (*отфрламе* H_0)
H_A: $\sigma_1^2 < \sigma_2^2$	$f \le f_{1-\alpha}$
H_A: $\sigma_1^2 > \sigma_2^2$	$f \ge f_{\alpha}$
H_A: $\sigma_1^2 \neq \sigma_2^2$	$f \le f_{1-\alpha/2}$ или $f \ge f_{\alpha/2}$.

ПРИМЕР 14.7 Во еден прозводствен погон се испитуват два пристапа кон склопување на производи на една производствена лента. И двата пристапа даваат ист број склопени производи на ден. Сепак, поради подобра контрола на процесот на производство, се преферира пристапот што дава помали варијации во бројот на производи по ден. Резултатите од испитувањето се

пристап 1: $n_1 = 21$ ден, $s_1^2 = 1.432$,

пристап 2: $n_2 = 25$ дена, $s_2^2 = 1.708$.

Конструирај 95% интервал на доверба за односот на дисперзиите. Тестирај ја соодветната хипотеза.

Решение

За Фишеровата распределба имаме дека $F_{1-\alpha/2,\ (24,20)} = F_{0.975,\ (24,20)} = 0.430 = 1/F_{0.025,\ (20,\ 24)}$, додека $F_{\alpha/2,\ (24,20)} = F_{0.025,\ (20,\ 24)} = 2.41$.

Интервалот на доверба е

$$p\left(\frac{1.432}{1.708}0.43 < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1.432}{1.708}2.41\right) = p(0.361 < \frac{\sigma_1^2}{\sigma_2^2} < 2.021) = 0.95.$$

Овој интервал ја вклучува 1-цата (длабоко внатре), па со 95% сигурност немаме одлука за тоа која дисперзија е поголема.

Сега да ја тестираме хипотезата H_0: $\sigma_1^2 = \sigma_2^2$, наспроти H_A: $\sigma_1^2 < \sigma_2^2$.

Од статистиката добиваме $f = s_1^2 / s_2^2 = 1.432/1.708 = 0.838$. За критичната точка имаме $F_{1-\alpha,\ (20,\ 24)} = F_{0.95,\ (20,24)} = 0.480$. И сега поради $f = 0.838 > 0.480 = F_{0.95,\ (20,24)}$ не ја отфрламе H_0, т.е нема доволно докази за дека вториот пристап има поголеми варијации од првиот. Во прилог на овој заклучок води и релативно високата P-вредност на тестот од 0.347. ■

Нека H_A е точна, т.е. нека $\sigma_1^2 < \sigma_2^2$, тогаш S_1^2 / S_2^2 нема Фишерова распределба. Во тој случај, Фишерова распределба ќе има случајната променлива $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$. Така, за јачината на тестот добиваме

$$1-\beta = p\left(\frac{S_1^2}{S_2^2} < f_{1-\alpha,(n_1,n_2)} \mid H_A\right) = p\left(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < \frac{\sigma_2^2}{\sigma_1^2} f_{1-\alpha,(n_1,n_2)}\right) =$$

$$= p(f < \frac{\sigma_2^2}{\sigma_1^2} f_{1-\alpha,(n_1,n_2)}).$$

Се разбира, односот σ_2^2/σ_1^2 е непознат. Кога би го знаеле, беспредметни би станале и интервалните оценки и тестирањето хипотези за врската меѓу двете дисперзии. Затоа β може само приближно да се процени. За примерот 14.7 грубо да земаме $\sigma_2^2/\sigma_1^2 \approx 2/(0.361+2.021) \approx 0.84$ (средина на интервалната оценка). Тогаш $1-\beta \approx p(f < 0.84 \cdot f_{0.95,\ (20,24)}) = p(f < 0.4032) = 0.022$, што е многу слаба јачина, т.е. многу е висока грешката од тип 2 ($\beta = 0.978$ да не биде отфрлена H_0 кога е точна H_A).

14.4. Преглед на поважните статистички тестови

Вистинскиот избор на статистички тест за анализа на податоците е секогаш малку незгодна работа. Најпрво треба да се избере меѓу две фамилии тестови, параметарски и непараметарски.

Во оваа книга, ние бевме критички ориентирана кон непараметарските статистички модели и сите модели како и сите оценки и тестови беа параметарски. Најголемиот дел од овие тестови користеа претпоставка за нормална распределба на податоците. Поважните параметарски тестови се наведени во првата колона на следната табела:

ЦЕЛ	*Податоци (од нормална популација)*	*Ранг, скор (ненормална популација)*	*Биномни (два возможни исходи)*
Опис на група	Очекување $\bar{x}$, девијација $\sqrt{s}$	Медијана, процентили	Пропорција
Спореди група со хипотетичка вредност	z, t или χ^2 тест	Wilcoxon-ов тест*	Хи-квадрат или Биномен тест
Споредба на 2 групи (упарени или неупарени)	z, t или Fisher-ов тест	Mann-Whitney-ов, Wilcoxon-ов тест*	Хи-квадрат (голем примерок) или Fisher-ов тест
Споредба на ≥3 групи (упарени или неупарени)	Анализа на дисперзијата – ANOVA тестови*	Kruskal-Wallis-ов Friedman-ов тест*	Хи-квадрат, Cochrane Q тест*
Тип на распределба	Хи-квадрат		Хи-квадрат
Квантитативна асоцијација меѓу 2 променливи	Пирсонова корелација	Сперманова корелација	Коефициенти на зависност (contingency)
Предвидување вредености од други вредности	Линеарна (или нелинеарна) регресија	Непараметарска регресија*	Логистична (logistic) регресија*

Тестовите што не прават претпоставка за обликот на распределбата на популацијата се третираат како непараметарски. Практично сите непараметарски тестови ги рангираат податоците (на пример во растечки редослед) и потоа ги анализираат ранговите. Овие тестови (distribution-free) како Wilcoxon-ов, Mann-Whitney-ов или Kruskal-Wallis-ов, се дадени во втората колона на табелата. Третата колона во табелата се однесува на тестови на пропорциите на популациите во табели на зависност (contingency tables) разгледувани во поглавјето 13.2.2. Овие тестови во главно се од хи-квадрат тип.

Да повториме дека иако досегашните дискусии беа критички ориентирани кон непараметарските модели (во овој случај тестови), тие сепак имаат важна улога во ситуациите кога:

1) Податоците се рангирани или се скор, а популацијата е јасно ненормална. На пример, успехот на студентите од една група, Apgar скор на новородени бебиња, критики на филмови, рангирање ресторани итн.;

2) Некои вредности од податоците се "надвор од скалата", преголеми или премали. Дури и кога популацијата е нормална, тешко е да се анализираат такви податоци со параметарски тестови поради постоење на непознати вредности. Кај непараметарските тестови важен е релативниот ранг, а не конкретните вредности, и тогаш ваквите проблеми исчезнуваат.

Кога се работи за избор меѓу параметарски и непараметарски тестови важно е да се имаат предвид следните забелешки:

а) Кога примерокот е *голем* (на пример ≥ 100), многу е полесно да се провери нормалноста на распределбата, дали со веројатносен дијаграм или тест (хи-квадрат и други). Но во таков случај, едноставно се повикуваме на централната гранична теорема што овозможува параметарските тестови да работат добро. Потребната големина на примерокот зависи од распределбата на популацијата, но во најголем број случаи веројатно сме сигурни со 30-тина податоци. За голем примерок и непараметарските тестови работат добро, но малку се поинфериорни од параметарските;

б) Кога примерокот е *мал* (на пример < 20), тешко е да се провери нормалноста на распределбата, па параметарските тестови се "ризични". Од друга страна, непараметарските тестови на мал примерок немаат јачина и секако не се доволно добри. Малиот примерок секако не соочува со дилеми што понекогаш водат до бирање на помало зло. Во таков случај, некој би избрал параметарски тест бидејќи не знае дали нормалноста на популацијата е нарушена, а друг би избрал непараметарски бидејќи не е сигурен дека популацијата е со нормална распределба. Ако податоците се доста раштркани, се зголемуваат шансите за нормална распределба. Тука претходните податоци (а не само примерокот од тековниот експеримент) може да помогнат да се оцени дали популацијата има нормална (или блиска до неа) распределба. Во случај на мал примерок генерално важи дека параметарските тестови се непоуздани, а непараметарските се слаби.

Кога се споредуваат 2 или повеќе примероци треба да се одлучи дали (ако може) да се оди на упарен или неупарен тест. За 3 и повеќе примероци наместо упарен се користи терминот повторувачки примерок. Упарените тестови се користат само ако податоците се повторувачки мерења на некој објект пред и по интервенција или се однапред упарени податоци - тесно поврзани и од ист тип.

На пример, да претпоставиме дека се испитува ефикасноста на 2 лека за намалување крвен притисок. Да речеме дека 10 пациенти се избрани за третман со првиот и други 10 за третман со вториот лек. Знаејќи дека притисокот зависи од возраста и тежината, некој би сакал да направи упарување на пациентите така што да има 10 пара пациенти со приближно иста возраст и тежина по пар. Секој лек ќе се даде на различните пациенти од секој пар. Без вакво упарување тестот е сомнителен бидејќи може помладите или полесните пациенти да се поподложни на намалување на притисокот, па поради нееднаквоста на групите резултатите нема да бидат објективни. Цената што се плаќа со упарувањето е намалувањето на степените на слобода, со што се намалува јачината на тестот. Да забележиме дека при фиксни останата параметри во тестот, поголемиот број на степени на слобода води кон помало β (грешка од тип 2). Од друга страна, во нашиот пример на тестирање лек за намалување притисок, дисперзијата на упарениот тест ќе биде доста помала од неупарениот поради високата корелација меѓу упарените податоци. Ова превагнува во полза на упарен тест.

Генерално, ако постои голема хетерогеност меѓу податоците и висока корелација (големо позитивно ρ), тогаш губењето на степените на слобода ќе биде компензирано со поголемата прецизност (намалена дисперзија) кај упарениот тест, и тој треба да се преферира.

Кога се разгледуваат табели со 2 редици и 2 колони со биномни податоци може да се работи со Фишеров или хи-квадрат тест. Предност на Фишеровиот тест е што директно ја дава P-вредноста, но е покомплициран за пресметка и има одредени проблеми од теоретска природа. Но во ситуација кога бројот на податоци во табелата е мал (кој било < 6), хи-квадрат тестот е несоодветен. За поголем број податоци, може да се користат и двата теста.

Кога се работи за одлука дали да се користи регресија или едноставно корелација, одлуката зависи од тоа дали податоците за случајните променливи X и Y ги разгледуваме како рамноправни, дали се користи функција од некоја од нив и се разбира кои дополнителни информации се потребни. Понекогаш е корисно да се пресмета и корелација и регресија. Кога се испитува линеарна зависност меѓу податоците, најчесто се

користи Пирсоновата (параметарска) корелација. Некои автори невнимателно наведуваат дека Пирсоновата корелација се користи кога податоците се нормално распределени. Тоа не е така. Во случај на заедничка нормална распределба, Пирсоновата корелација комплетно ја опишува врската меѓу случајните променливи. Инаку за случајни променливи со конечни дисперзии и коваријација, таа секогаш постои.

Спирмановата корелација треба да се преферира кога во податоците имаме натрапници (outliers), т.е. нетипични податоци или кога двете групи податоци се неприлагодени (skewed). Нетипичните податоци се податоци што со вредноста не припаѓаат на променливата (грешки од мерење, пресметка итн.). Спирмановата корелација работи со рангови, па таквите податоци не се битен проблем. Неприлагодените податоци се такви што не може згодно да се споредуваат. На пример, корелацијата меѓу големината на мозокот и тежината на животните не е згодно да се пресметува со стандардна метрика (разлики и збирови). Кај Спирмановата корелација сѐ се претвора во ранг, па автоматски податоците се од ист тип.

Регресијата вообичаено се користи кога едната случајна променлива произлегува од другата или секогаш кога на некоја од случајните променливи и правиме трансформација. Регресијата не е симетрична како корелацијата.

ЗАДАЧИ

1. Машински инженер сака да спореди цврстина на профил од челик и профил од одредена легура. Еднаков број профили n од двата типа е тестиран. Секој профил е поставуван во хоризонтална положба процврстен за краевите, и на средината е применета сила од 2500 lb (либри) при што е мерена деформацијата. Од претходни искуства, инжинерите претпоставуваат стандардна девијација на деформацијата и кај двата типа профили од 0.05 инча. Бидејќи профилите од легура се поскапи, инжинерите би тестирале со ниво на значајност 0.01 дали профилите од легура имаат помала просечна деформација. Колкав примерок n треба да се испита за грешката од тип 2 да е 0.05 кога вистинската просечна деформација за 0.04 инча го фаворизира профилот од легура?

2. Анорексија (AN) е психичка состојба што води до големо намалување на телесната тежина кај жените што се плашат да не станат дебели. Со помош

на магнетна резонанца испитувани се некои карактеристики на ткивата на AN примерок од жени што ја повратиле старата телесна тежина и тие се споредувани со ткивата на контролната група (здрави лица). Во следната табела се дадени сумарни податоци за интермускулното мрсно ткиво (IAT/kg):

Група	Примерок	Просек $\bar{x}$	Ст. девијација s
AN	16	0.52	0.26
Контролна	8	0.35	0.15

Под претпоставка дека примероците се земени од нормална распределба

а) Состави 99% интервал на доверба за просекот на IAT кај AN;

б) Состави 99% интервал на доверба за разликата на вистинските просеци на IAT кај AN и контројната популација. Што може да се заклучи од интервалот?

3. Антипсихотични лекови нашироко се користат во случаи на шизофренија и биполарно нарушување. Во соодветно списание дадени се резултати од испитувањето на нивното дејство врз метаболизмот на краток временски период.

   а) Примерок од 41 лице што земале арипипразол имале просечна промена во вкупниот холестерол од 3.75 mg/dL, со стандардна грешка од 3.878. Состави 95% интервал на доверба за вистинската промена во нивото на холестеролот.

   б) Во истиот чланак дадени се резултати од испитување на примерок од 36 лица што земале кветипин кај кои е регистрирана промена на холестеролот од 9.05 mg/dL со стандардна грешка од 4.256. Правејќи потребна претпоставка за распределбата, тестирај дали вистинскиот просечен холестерол е зголемен? Дали избраното ниво на значајност влијае на резултатот? Најди ја P-вредноста на тестот.

4. Две различни формули на моторно гориво било тестирани за да им се определи октанскиот број. Варијациите во октанскиот број за формулата 1 е $\sigma_1^2 = 1.5$, а за формулата 2, $\sigma_2^2 = 1.2$. Два примерока со големини $n_1 = 15$ и $n_2 = 20$ се тестирани при што се добиени просечни октански броеви $\bar{x}_1 = 89.6$ и $\bar{x}_2 = 92.5$. Под претпоставка за нормалност

   а) Состави 95% двостран интервал на доверба за разликата на просеците на октанските броеви;

   б) Ако формулата 2 дава повисок октански број, производителот секако би сакал тоа да го знае. Тестирај соодветна хипотеза со $\alpha = 0.05$. Најди ја P-вредноста на тестот.

5. Да претпоставиме дека сакаме да тестираме хипотеза H_0: $\mu_1 = \mu_2$, наспроти H_A: $\mu_1 \neq \mu_2$ и планираме да користиме примероци со еднаква големина. Претпоставуваме дека двете популации се нормални со иста, но непозната дисперзија. Ако користиме $\alpha = 0.05$ и вистинскиот просек $\mu_1 = \mu_2 + \sigma$, колкави треба да се примероците за јачината на тестот да биде најмалку 0.90?

6. Мерена е просечната далечина на патување на топче за голф со удирање на топчињата од страна на "механички голфер". Испитани се десет случајно избрани топчиња од два производителa при што по удар се добиени следните далечини (во метри)

   Производител 1: 275, 286, 287, 271, 283, 271, 279, 275, 263, 267;
   Производител 2: 258, 244, 260, 265, 273, 281, 271, 270, 263, 268.

   а) Дали може да сметаме дека растојанијата се приближно нормално распределени? Дали е оправдана претпоставката за еднакви дисперзии?

   б) Тестирај ја хипотезата дека топчињата од двата призводителa имаат иста просечна должина на патување. Користи $\alpha = 0.05$. Пресметај ја P-вредноста на тестот;

   в) Која е јачината на тестот, за да може да се детектира разлика во просечните должини на патување од 5 метри?

   г) Колкав примерок е потребен за да се детектира разлика во просечните должини на патување од 3 метри, со јачина на тестот од приближно 0.75?

   д) Состави 95% двострaн интервал на доверба за разликата во просечните должини на патување на топчињата од двата производителa.

7. Петнаесет мажи меѓу 35 и 50 годишна возраст учествувале во студија за оценка на влијанието на исхраната и вежбањето во намалување на нивото на холестерол во крвта. Вкупниот холестерол е мерен пред и по 3 месеци од практикувањето диета и аеробични вежби. Податоците се дадени во табелата во *mg*/*dl* (за претворање во *mmol*/*l* подели со 39):

Пред	265	240	258	295	251	245	287	314	260	279	283	240	238	225	247
По	229	231	227	240	238	241	234	256	247	239	246	218	219	226	233

   Дали овие податоци го поткрепуваат тврдењето дека дието со малку мрснотии и вежбањето го намалуваат (во просек) нивото на холестеролот во крвта? Користи $\alpha = 0.05$.

8. Десет лица учествувале во програма за намалување на тежината преку промена на исхраната. Нивната тежина (во фунти) пред и по програмата се дадена во следната табела:

Пред	195	213	247	201	187	210	215	246	294	310
По	187	195	221	190	175	197	199	221	278	285

а) Дали податоците покажуваат дека оваа програма е ефикасна во симнување на тежината? Користи $\alpha = 0.05$.

б) Дали има елементи за подршка на тврдењето дека оваа програма резултира во намалување на просечната тежина од најмалку 10 фунти со ниво на значајност $\alpha = 0.05$.

9. Дали некој кој ги менува брендовите поради финансиски причини има помалку шанси да остане лојален од оној кој тоа го прави без финансиски поттик. Нека p_1 и p_2 се пропорциите на оние што го менуваат брендот поради и без финансиски причини соодветно. Направена е анкета која покажала дека од 200 испитаници 30 го смениле брендот од финансиски причини, додека од 600 испитаници 180 го смениле од други причини. Тестирај хипотеза за еднаквост на овие пропорции со ниво на значајност $\alpha = 0.01$.

10. Две машини се користат за формирање пластични делови. Делот е дефектен ако е деформиран. Два примерока од по 300 дела се земени од двете машини при што 15 дефектни биле од првата, а 8 дефектни од втората.

а) Дали е разумно да се заклучи дека двете машини произведуваат ист процент на дефектни делови?

б) Ако претпоставиме дека пропорциите на дефектни делови се $p_1 = 0.05$ и $p_2 = 0.01$, која е јачината на тестот? Колкав треба да биде примерокот за да се открие оваа разлика со веројатност ≥ 0.9. Користи $\alpha = 0.05$.

11. Од примерок од 500 граѓани во некој град, 385 се изјасниле за зголемување на максималната брзина на автопатите на 75 миљи на час, додека од друг примерок од 400 граѓани, 267 се изјасниле во прилог на истото.

а) Дали има разлика во подршката за зголемување на брзината во двете анкети? Користи $\alpha = 0.05$. Која е P-вредноста на тестот?

б) Состави 95% интервал на доверба за разликата на двете пропорции и интерпретирај го добиениот резултат.

12. Олестра е замена за масло одобрена од FDA (Food and Drug Administration) во САД за користење во храната. Поради ретки инциденти на гастро-проблеми направено е испитување со плацебо контролиран експеримент за споредба на чипс со олестра и стандарден чипс. Од 529 испитаници од плацебо групата, 17.6% осетиле гастро-проблеми, додека од 563 испитаници што земале чипс со олестра, 15.8% осетиле гастро-проблеми.

а) Тестирај хипотеза со 5% значајност за тоа дали е различен бројот на гастро-проблемите кај чипсот со и без олестра;

б) Ако вистинските проценти на гастро-проблеми се 15 и 20% соодветно, колкав треба да биде примерокот за таквата разлика да се детектира со веројатност од 0.90?

13. Во медицинските истражувања односот p_1/p_2 често пати е поважен од разликата $p_1 - p_2$ (третманот 1 да е подобар од третманот 2). Нека m и n се големи и $\hat{\theta} = \hat{P}_1 / \hat{P}_2$.Тогаш $\ln(\hat{\theta})$ има приближно нормална распределба со приближно очекување $\ln(\theta)$ и стандардна девијација $((m-x)/mx + (n-y)/ny)^{1/2}$.

    а) Под овие услови состави интервална оценка за θ;

    б) Се смета дека земањето аспирин го намалува ризикот од инфаркт. Од плацебо група на 11034 испитаници, кај 189 бил регистриран инфаркт по фиксен број години. Во групата од 11037 испитаници што земале аспирин, овој број бил само 104. Состави 95% интервал на доверба за односот на бројот на инфаркти кај двете групи. Што овој интервал сугерира за ефикасноста на третманот со аспирин?

14. За Фишеровата распределба најди ги вредностите:
    а) $f_{0.25,(5,10)}$; б) $f_{0.10,(24,9)}$; в) $f_{0.05,(8,15)}$; г) $f_{0.75,(5,10)}$.

15. Две хемиски компании испорачуваат сиров материјал при што е важна концетрацијата на еден елемент во материјалот. Просечната содржина на елементот кај двете компании е приближно ист, но варијациите во концетрациите се различни. Стандардната девијација на концетрацијата на елементот во примерок $n_1 = 10$, земен од компанијата 1 е $s_1 = 4.7$ грама по литар, додека примерок $n_2 = 16$, земен од компанијата 2 е $s_2 = 5.8$ грама по литар. Дали овие податоци се доволен индикатор дека варијациите на елементот во сировиот материја кај двете компании се разликуваат? Користи $\alpha = 0.05$.

16. Направена е студија да се испита дали мажите и жените се разликуваат во варијациите на времето потребно за спојување на компоненти според дадена скица. Примерок од 25 мажи и 21 жена биле тестирани, при што за првите стандардната девијација при спојување била 0.98 минути за мажите и 1.02 минути за жените.

    а) Дали овоие податоци сугерираат разлика во варијациите на времето на спојување на компонентите меѓу мажите и жените? Користи $\alpha = 0.02$;

    б) Најди 98% интервал на доверба за односот на дисперзиите и интерпретирај го добиениот резултат.

15

Регресиона анализа

Многу проблеми во инженерството вклучуваат испитување на непозната зависност меѓу две или повеќе променливи. Типична ситуација е кога случајна променлива Y е функција од една или повеќе независни (детерминистички) променливи $x_1, x_2, \ldots, x_m$. На пример, јачината на струјата Y зависи од напонот (x_1) и отпорот (x_2); периодата на нишалото Y зависи од должината (x_1); цената на некретнина Y зависи од локацијата (x_1), староста (x_2) и квадратурата (x_3); резултатот од некој хемиски процес Y зависи од температурата (x_1), притисокот (x_2) и содржина на супстанца 1 (x_3) и супстанца 2 (x_4) итн.

Ние веќе накусо го дискутиравме генералниот статистички модел (поглавје 9.2) каде што може да го вклучиме овој проблем. Имено, таму веќе воведовме статистички генератор на случајна променлива Y со

$$Y = E(Y \mid \mathcal{D}) + \varepsilon, \;\; EY^2 < \infty$$

каде што $\mathcal{D}$ е условната информација што потекнува од случајниот примерок X, обично $\mathcal{D} = \mathcal{D}(x_1, x_2, \ldots, x_n)$, а ε е несистематска грешка т.е. деформација (disturbance term). Деформацијата го одразува фактот дека вредностите на Y што се добиваат од примерокот не може перфектно да се сложуваат со моделот. Со други зборови, ε е случајна променлива што ги вклучува ефектите од сите немоделирани извори на варијабилност што влијаат на моделот.

За испитување на зависноста меѓу променливите обично се користи *регресиониот модел* каде што $\mathcal{D}$ вклучува зависносни информации

во облик, $\mathcal{D} = (X = x_k)$ (за еднодимензионална зависност) што води до општ облик на статистички генератор

$$Y = E(Y \mid X = x_k) + \varepsilon,\ k = 1, 2,\ \ldots, n.$$

Најчесто користен е нормалниот линеарен регресионен модел. Во едноставниот еднодимензионален случај

$$E(Y \mid x) = \beta_0 + \beta_1 x,\ \text{т.е. } Y = \beta_0 + \beta_1 x + \varepsilon.$$

ε има нормална распределба $Z(0, \sigma^2)$ (очекуваната грешка треба да е 0).

Во секоја реална ситуација, вообичаено е параметрите β_0, β_1 и σ^2 да бидат непознати и да мора да бидат определени од примерокот.

15.1. Проста линеарна регресија

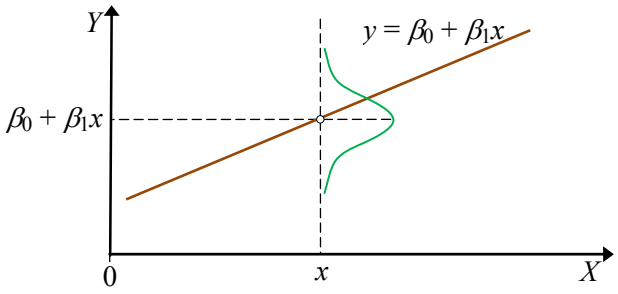
За анализа на линеарен регресионен модел, ќе претпоставиме фиксно x и ќе ја испитуваме вредноста на случајната променлива Y. Јасно е дека случајната променлива ε од десната страна на моделот ги определува особините на Y. Очекувањето и дисперзијата на Y се

$$EY = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x \ \text{ и}$$

$$DY = D(\beta_0 + \beta_1 x + \varepsilon) = D(\beta_0 + \beta_1 x) + D(\varepsilon) = 0 + \sigma^2 = \sigma^2.$$

Значи во случај на нормална линеарна регресија, распределбата на Y е

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\beta_0-\beta_1 y)^2}{2\sigma^2}}.$$ На сл. 15.1 е прикажана распределбата на Y конкретна вредност на x.



Слика 15.1 *Распределба на Y за фиксно x*

Дисперзијата σ^2 ја определува варијабилноста на вредностите на случајната променлива Y. Мало σ^2 значи дека вредностите на Y паѓаат

блиску до правата линија. Бидејќи σ^2 е константна, варијабилноста на Y за секоја вредност x е иста.

Да претпоставиме дека располагаме со примерок од n парови вредности $\{(x_j, y_j), j = 1, 2, \ldots, n\}$. Моделот се трансформира во

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j,\ j = 1, 2, \ldots, n.$$

Оценките за β_0 и β_1 може да се добијат со методот на најмали квадрати (види поглавје 11.3.3 и пример 11.20). Тоа значи дека β_0 и β_1 се избираат така да се минимизира функцијата

$L(\beta_0, \beta_1) = \sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_j)^2$, што ги дава оценките

$\hat{\beta}_1 = \dfrac{\sum_{j=1}^{n} x_j y_j - n\overline{xy}}{\sum_{j=1}^{n} (x_j - \overline{x})^2}$, $\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$, каде што $\overline{x} = \dfrac{1}{n}\sum_{1}^{n} x_i$, $\overline{y} = \dfrac{1}{n}\sum_{1}^{n} y_i$.

Разликата меѓу податокот и соодветната оценка $y_j - \hat{y}_j$, вообичаено се нарекува остаток (residual). Да забележиме дека соодветните оценувачи се добиваат кога y_j се разгледуваат како случајни променливи од примерокот Y_j (независни).

ПРИМЕР 15.1 Во следната табела е дадена чистотата на кислородот произведен со процес на дестилација при што првата редица го дава процентот на хидрокарбон во дестилациониот уред (вкупно 20 податоци):

Хидрокарбон %	0.99	1.02	1.15	1.29	1.46	1.36	0.87	1.23	1.55	1.40
Чистота %	90.01	89.05	91.43	93.74	93.74	94.45	87.59	91.77	99.42	93.65
Хидрокарбон %	1.19	1.15	0.98	1.01	1.11	1.20	1.26	1.32	1.43	0.95
Чистота %	93.54	92.52	90.56	89.54	89.85	90.39	93.25	93.41	94.98	87.33

Состави линеарен регресионен модел за овие податоци, т.е. за зависноста на чистотата на кислородот од количеството хидрокарбон.

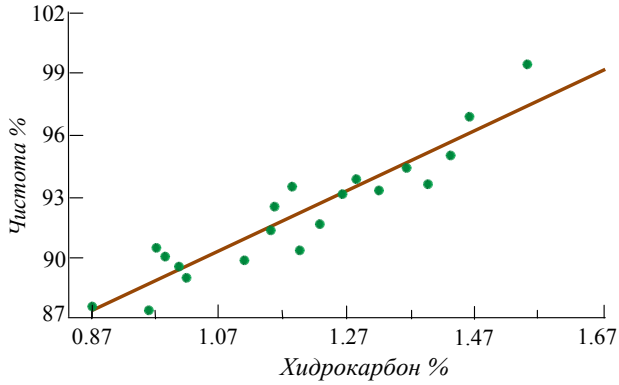
Решение

Од 20-те податоци во првата редица (за хидрокарбон) имаме $\overline{x} = 1.1960$, а од втората (чистота) $\overline{y} = 92.1605$. Понатаму добиваме

$$\hat{\beta}_1 = \frac{2214.6566 - 20 \cdot 1.1960 \cdot 92.1605}{0.68088} = 14.9475 \text{ и}$$

$$\hat{\beta}_0 = 92.1605 - 14.9475 \cdot 1.1960 = 74.2833 .$$

Значи моделот е $y = 74.2833 + 14.947x$, и тој е прикажан на следната слика



Користејќи го моделот ние сега можеме да ја предвидиме чистотата на кислород од 89.23% за ниво на хидрокарбон од 1.00% или обратно, да пресметаме дека за добивање чистота од 90% потребно е ниво на хидрокарбон од (90 – 74.2833)/14.947 = 1.05%. ■

Тука е важно да нагласиме дека регресионата врска е валидна само за вредности на x во рангот на податоците. Екстраполацијата на резултатите надвор од овој ранг (во примерот 15.1 рангот е [0.87, 1.55]) е сомнителна и генерално неточна. Колку сме "подалеку" од рангот на податоците, екстраполацијата станува сè понеточна.

Исто така би нагласиле дека линеарната регресија претпоставува однапред линеарна зависност меѓу Y и x. Ако оваа зависност е нелинеарна, линеарната регресија дава бесмислени резултати, дури и кога правата линија обезбедува добро сложување со податоците.

Третиот непознат параметар на моделот σ^2 (покрај β_0 и β_1) е дисперзијата на грешката ε, и таа може лесно да се оцени од податоците. Од $\hat{\varepsilon} = \sum_1^n (y_i - \hat{y}_i)^2$ добиваме очекување $E\hat{\varepsilon} = (n-2)\sigma^2$ (y_j се вредности на случајните променливи Y_j), па центрираната оценка на σ^2 би била $\hat{\sigma}^2 = \sum_1^n (y_i - \hat{y}_i)^2 / (n-2)$. Во примерот 15.1, добиваме $\hat{\sigma}^2 = 1.18$.

Пред да ги најдеме очекувањата и дисперзиите на точкастите оценувачи $\hat{\beta}_1$ и $\hat{\beta}_0$ ќе воведеме ознаки

$$S_{xx} = \sum_{j=1}^n (x_j - \bar{x})^2 = \sum_{j=1}^n x_j^2 - \frac{1}{n}\left(\sum_{j=1}^n x_j\right)^2 \text{ и } S_{xy} = \sum_{j=1}^n x_j y_j .$$

Ако $\hat{\beta}_1$ го напишеме во облик

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n}(y_j - \bar{y})(x_j - \bar{x})}{S_{xx}} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})y_j}{S_{xx}} = \sum_{j=1}^{n} c_j y_j ,$$

се гледа дека $\hat{\beta}_1$ е линеарна комбинација (c_j се константи) на независни и нормално распределени случајни променливи Y_j. Оттука следува дека

$E\hat{\beta}_1 = \mu_{\hat{\beta}_1} = \beta_1$, т.е. $\hat{\beta}_1$ е центрирана оценка на β_1.

Дисперзијата е $D\hat{\beta}_1 = \sigma^2 / S_{xx}$ каде што $\sigma^2 = DY = D(\varepsilon)$.

За $\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$ исто така се добива дека е центрирана оценка со дисперзија $D\hat{\beta}_0 = \sigma^2 / (1/n + \bar{x}^2 / S_{xx})$. Оценувачот е повторно со нормална распределба, како линеарна комбинација на независни и нормално распределени случајни променливи Y_j

Особините на параметрите на регресијата се потребни за составување интервали на доверба и тестирање хипотези за нивната вредност.

15.2. Интервални оценки за параметрите

Како што веќе видовме, случајните променливи Y_j имаат распределба $Z(\beta_0 + \beta_1 x, \sigma^2)$. $\hat{\beta}_1$ и $\hat{\beta}_0$ имаат исто така нормални распределби $Z(\beta_1, \sigma^2 / S_{xx})$ и $Z(\beta_0, \sigma^2 (1/n + \bar{x}^2 / S_{xx}))$ соодветно. Од друга страна, случајната променлива $(n-2)\ \hat{\sigma}^2/\sigma^2$ има χ^2 распределба со $(n-2)$ степени на слобода и притоа $\hat{\beta}_1$ и $\hat{\beta}_2$ се независни од $\hat{\sigma}^2$. Тогаш статистиките

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \quad \text{и} \quad (\hat{\beta}_0 - \beta_0) / \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

имаат студентова распределба со $(n-2)$ степени на слобода. Имениелите на горните 2 статистики се нарекуваат *стандардни грешки* за β_1 и β_0 соодветно. Ова води до следните интервални оценувачи за β_1 и β_0:

$$p\left(\hat{\beta}_1 - t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \right) = 1 - \alpha \quad \text{и}$$

$$p\left(\hat{\beta}_0 - t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \right) = 1 - \alpha .$$

ПРИМЕР 15.2 Најди 95% интервал на доверба за β_1 и β_0 од примерот 15.1.

Решение

Да се потсетиме дека $\hat{\beta}_1 = 14.9475,\ S_{xx} = 0.68088,\ \hat{\sigma}^2 = 1.18$, па оценката е

$$p\left(\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right) =$$

$$= p\left(14.9475 - 2.101\sqrt{\frac{1.18}{0.68088}} < \beta_1 < 14.9475 + 2.101\sqrt{\frac{1.18}{0.68088}}\right) =$$

$$= p\left(\beta_1 \in (12.197,\ 17.697)\right) = 0.95$$

Знаеме дека $\hat{\beta}_0 = 74.2833$, па интервалната оценка за β_0 е

$$p\left(\beta_0 \in \left(74.2833 \pm 2.101\sqrt{1.18\left(\frac{1}{20} + \frac{1.43}{0.68088}\right)}\right)\right) =$$

$$= p\left(\beta_0 \in (70.6479,\ 77.9187)\right) = 0.95\ .\ \blacksquare$$

15.2.1. Интервална оценка на регресионата линија

Интервална оценка може да се контруира и околу очекуваната вредност на Y за конкретна вредност на x, да речеме x_0. Значи интервалот на доверба е околу $E(Y \mid x_0) = \mu_{Y|x_0}$ и често се нарекува интервална оценка околу регресионата линија. Од $E(Y \mid x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$, за точкаста оценка $\hat{\mu}_{Y|x_0}$ на просекот на Y во точката x_0 може да се земе $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$, што е центрирана оценка на $\mu_{Y|x_0}$ (поради центрираноста на $\hat{\beta}_0$ и $\hat{\beta}_1$). Дисперзијата на $\hat{\mu}_{Y|x_0}$ е

$$D\hat{\mu}_{Y|x_0} = \sigma^2\left(1/n + (x_0 - \bar{x})^2 / S_{xx}\right).$$

Сега ако се има предвид дека $\hat{\mu}_{Y|x_0}$ е со нормална распределба (поради нормалноста на $\hat{\beta}_0$ и $\hat{\beta}_1$), и го земеме $\hat{\sigma}^2$ како оценувач за σ^2, лесно е да се покаже дека статистиката

$$T = (\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}) / \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

има студентова распределба со $n - 2$ степени на слобода. Оттука (со решавање на двостраното неравенство $-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}$ по $\mu_{Y|x_0}$) следува

$$p\left(\mu_{Y|x_0} \in \left(\hat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right)\right) = 1-\alpha\,.$$

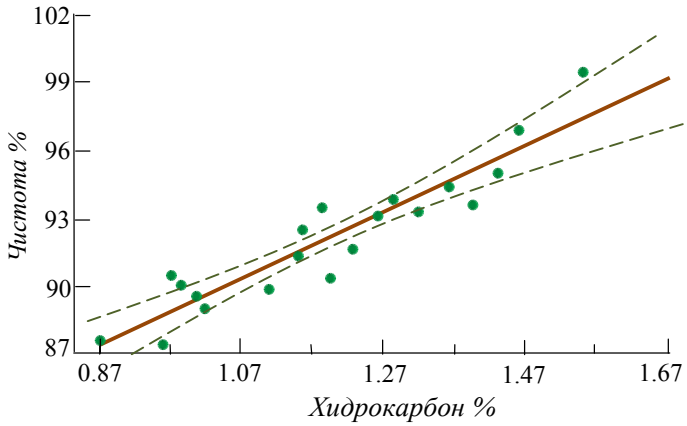
ПРИМЕР 15.3 Најди 95% интервал на доверба за просекот на чистотата на кислородот ($\hat{\mu}_{Y|x_0}$) (види пример 15.1) за ниво на хидрокарбон од 1.00% (x_0 = 1.00%).

Решение

Да се потсетиме дека $\hat{\beta}_1 = 14.9475$, $\hat{\beta}_0 = 74.2833$, па $\hat{\mu}_{Y|x_0}$ = 74.2833 + 14.9475·1.00 = 89.23. Имајќи предвид дека $n = 20$, $S_{xx} = 0.68088$, $\hat{\sigma}^2 = 1.18$, $\bar{x}$ = 1.1960, бараната интервална оценка е

$$p\left(\mu_{Y|x_0} \in \left(89.23 \pm 2.101\sqrt{1.18\left(\frac{1}{20}+\frac{(1-1.1960)^2}{0.68088}\right)}\right)\right) =$$

$$= p\left(\mu_{Y|x_0} \in (88.48,\ 89.98)\right) = 0.95 \text{ (види слика).}$$



Сликата ја прикажува регресионата линија и 95%-ен интервал на доверба за $\mu_{Y|x_0}$ околу линијата. Таа се добива од интервалите на доверба за многу вредности x_0. Се разбира, ширината на интервалите се зголемува со зголемување на $|x_0 - \bar{x}|$. ■

15.2.2. Интервална оценка за нови податоци

Важна примена на регресионите модели е предвидувањето вредности на Y, за нови вредности на x. Ако x_0 е вредноста за која сакаме предвидување, точкастиот оценувач $\hat{Y}_0$ за предвидената вредност е

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 .$$

За конструкција на интервален оценувач за Y_0, треба да забележиме дека новиот податок е независен од податоците користени за креирање на регресиониот модел. Интервалите на доверба на регресионата линија околу $\mu_{Y|x_0}$ не се соодветни бидејќи тие се однесуваат на $x = x_0$, што е параметар на популацијата, а не иден - нов податок.

На пример, да претпоставиме дека зависноста на возраста на детето и бројот на зборови што ги користи може да се опише со проста линеарна регресија. Тогаш, интервалната оценка за $\mu_{Y|x_0}$ обезбедува информација за просечната големина на речникот на сите 6-годишни деца ($x_0 = 6$), т.е. интервалот се однесува на параметар чијашто вредност е фиксна, но непозната. Од друга страна, идната вредност на Y не е параметар туку случајна променлива.

Ако Y_0 е идна вредност за новиот податок $x = x_0$, а $\hat{Y}_0$ е точкаст оценувач за Y_0, тогаш грешката на предвидувањето е $Y_0 - \hat{Y}_0$ и е со нормална распределба со очекување 0 и дисперзија

$$D(Y_0 - \hat{Y}_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right), \text{ бидејќи } Y_0 \text{ и } \hat{Y}_0 \text{ се независни.}$$

Ако ја земеме $\hat{\sigma}^2$ како оценувач за σ^2, лесно може да се покаже дека статистиката

$$T = (Y_0 - \hat{Y}_0) / \sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

има студентова распределба со $n - 2$ степени на слобода. Оттука (со решавање на неравенството $-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}$ по Y_0) веднаш се добива интервалната оценка за предвидувањето

$$p\left(Y_0 \in \left(\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)\right) = 1 - \alpha \text{, за } \hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 .$$

Забележи дека интервалот на доверба е најтесен за $x_0 = \bar{x}$, и се проширува со растењето на $|x_0 - \bar{x}|$. Ако ги споредиме интервалната оценка на регресионата линија (од претходното поглавје) со интервалната оценка за нови податоци, веднаш се забележува дека втората е секогаш поширока $(1+1/n+(x_0-\bar{x})^2/S_{xx}$ наспроти $1/n+(x_0-\bar{x})^2/S_{xx})$. Интуитивно, тоа доаѓа оттаму, што таа ги вклучува и двете, грешката од моделот и грешката придружена на новиот податок.

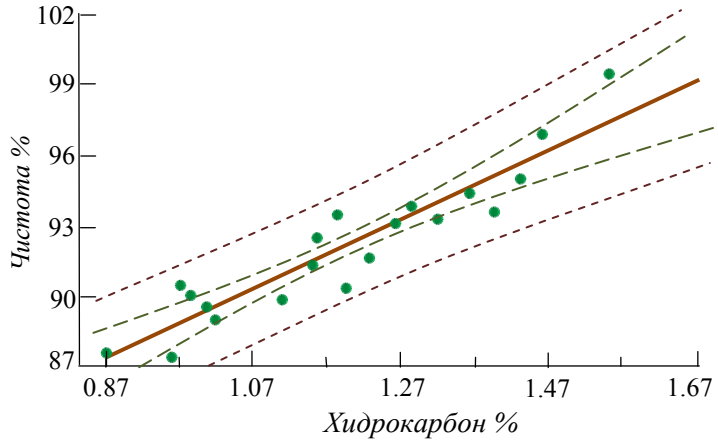
ПРИМЕР 15.4 Најди 95% интервал на доверба за просекот на чистотата на кислородот (Y_0) (види пример 15.1) за нов податок на ниво на хидрокарбон од 1.00% ($x_0 = 1.00\%$) .

Решение

Од $\hat{\beta}_1 = 14.9475$ и $\hat{\beta}_0 = 74.2833$, добиваме $\hat{Y}_0 = 74.2833 + 14.9475 \cdot 1.00 = 89.23$. Сега ги имаме сите потребни вредности ($n = 20$, $S_{xx} = 0.68088$, $\hat{\sigma}^2 = 1.18$, $\bar{x} = 1.1960$), па бараната интервална оценка е

$$p\left(Y_0 \in \left(89.23 \pm 2.101\sqrt{1.18\left(1+\frac{1}{20}+\frac{(1-1.1960)^2}{0.68088}\right)}\right)\right) =$$

$$= p\big(Y_0 \in (86.83,\ 91.63)\big) = 0.95 \text{ (види слика).}$$



Сликата ги прикажува регресионата линија и 95%-ните интервали на доверба за $\mu_{Y|x_0}$ и за нови податоци Y_0 околу линијата. Надворешните линии се однесуваат на интервалите на доверба за новите податоци и тие очигледно се доста поширoки. ■

15.3. Тестирање хипотези за простата линеарна регресија

Адекватноста на линеарниот регресионен модел може да се проверува со тестирање хипотези за параметрите. И тука работиме под претпоставка дека грешката ε е со нормална распределба $Z(0, \sigma^2)$.

Најпрво да ги разгледаме тестовите за падот (slope) β_1 и пресекот (intercept) β_0 на регресионата линија. Соодветни хипотези се:

H_0: $\beta_1 = \beta_T$ наспроти H_A: $\beta_1 \neq \beta_T$, и

H_0: $\beta_0 = \beta_T$ наспроти H_A: $\beta_0 \neq \beta_T$.

Тука, како и во некои претходни случаи, користиме исти ознаки за статистиките T и тестовите β_T, што не е некој проблем бидејќи секогаш од контекстот знаеме за што се работи.

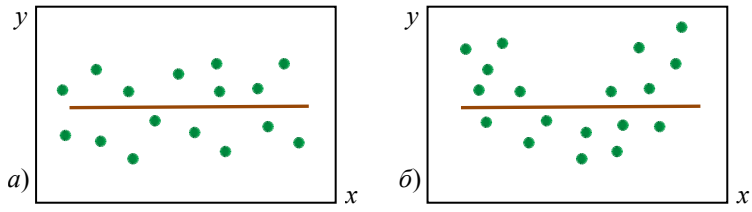
Претходно веќе видовме дека статистиките

$$T = \frac{\hat{\beta}_1 - \beta_T}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \quad \text{и} \quad T = (\hat{\beta}_0 - \beta_T) / \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

имаат студентова распределба со $(n - 2)$ степени на слобода.

Имајќи ја вредноста t за статистиката T, и во двата случаја H_0 се отфрла со ниво на значајност α кога вредноста на статистиката е $|t| > t_{\alpha/2,n-2}$.

Важен специјален случај на тестирање хипотези е H_0: $\beta_1 = 0$ наспроти H_A: $\beta_1 \neq 0$. Овој тест се однесува на *значајноста на регресијата*. Имено, неотфрлањето на H_0 е еквивалентно со заклучокот дека нема линеарна зависност меѓу x и Y. Ова повлекува едно од двете, или x има мало влијание на варијациите на Y, па најдобрата оценка за Y е $\bar{y}$, или вистинката зависност меѓу x и Y е нелинеарна (види сл. 15.2).



Слика 15.2 *Хипотезата* H_0: $\beta_1 = 0$ *не се отфрла*

Отфрлањето на H_0: $\beta_1 = 0$ може да значи дека праволинискиот модел на сл. 15.2 *а*) е соодветен или иако имаме линеарен ефект од x, подобар би можел да биде некој полиномен модел со степени на x, сл. 15.2 *б*).

ПРИМЕР 15.5 Тестирај ја хипотезата H_0: $\beta_1 = 0$ наспроти H_A: $\beta_1 \neq 0$ како и H_0: $\beta_0 = 0$ наспроти H_A: $\beta_0 \neq 0$ за просекот на чистотата на кислородот (пример 15.1) со ниво на значајност $\alpha = 0.01$.

Решение

Од $\hat{\beta}_1 = 14.9475$, $n = 20$, $S_{xx} = 0.68088$ и $\hat{\sigma}^2 = 1.18$, добиваме вредност за статистиката

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{14.9475}{\sqrt{1.18 / 0.68088}} = 11.35$$

Сега, бидејќи $t = 11.35 > 2.88 = t_{0.005,18} = t_{\alpha/2,n-2}$, ја отфрламе H_0 со ниво на значајност $\alpha = 0.01$. P-вредноста на тестот е 0.00000000123 (многу мала, добиена од Excel) е исто во прилог на отфрлање на H_0. Во случај на тестот за β_0 за статистиката се добива $t = 46.62$, па повторно се отфрла H_0 со ниво на значајност $\alpha = 0.01$.

Значи регресионата линија има и пад и пресек (со y оската), т.е не е дегенерирана. ■

15.4. Соодветност на простата линеарна регресија

Линеарните регресионите модели се базираат на некои претпоставки. Оценките на параметрите на моделот (β_0, β_1 и σ^2) се под претпоставка дека грешките ε_j се некорелирани случајни променливи со очекување 0 и фиксна дисперзија. Интервалните оценки и тестовите бараат грешките ε_j да бидат со нормална распределба. Се разбира, се чини најважно е испитуваниот феномен да се "однесува" на линеарен начин.

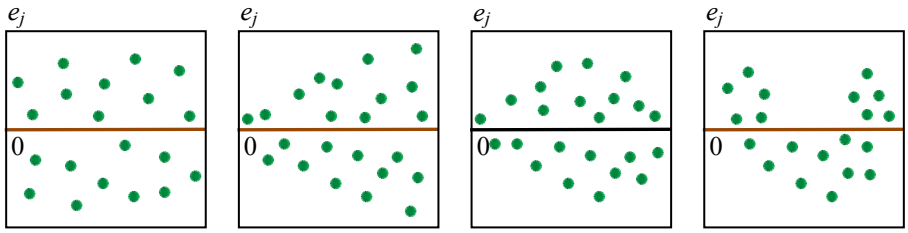
15.4.1. Анализа на остатоците

За проверка на приближната "нормалност" на грешките ε_j, може да се користат остатоците $e_j = y_j - \hat{y}_j$, $j = 1, 2, \ldots, n$.

За таа цел може да се искористат хистограмите или нормалните веројатносни дијаграми. Од друга страна, бројот на податоци често па-

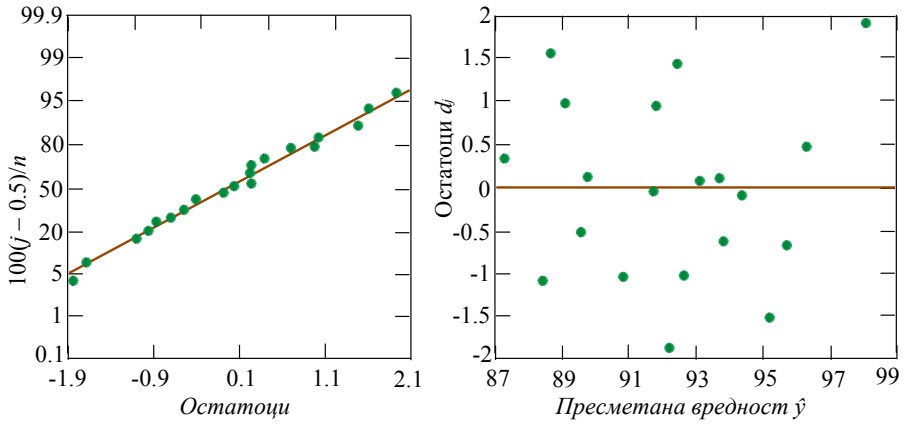
ти е мал за добивање на релевантен хистограм, но веројатносниот дијаграм е прифатлива опција. Остатоците може да се стандардизираат ставајќи $d_j = e_j / \sqrt{\hat{\sigma}^2}$, $j = 1, 2, \ldots, n$ и сега ако грешките се нормално распределени, приближно 95% од вредностите d_j треба да се во интервалот (–2, 2). Податоците што се надвор од интервалот укажуваат на присуство на натрапници (outliers), т.е. нетипични податоци во однос на мнозинството. Иако постојат различни методи за отфрлање на натрапниците, ваквите податоци понекогаш даваат важни информации за необичното однесување на податоците и не треба автоматски да се отфрлаат.

Често пати од помош се дијаграмите од тип: остатоци - временска секвенца (ако податоците се такви), остатоци - $\hat{y}_j$ или остатоци - x_j. Овие дијаграми обично имаат еден од следните 4 форми:



Идеална ситуација е кога податоците се приближно распоредени како на првата слика. Останатите три слики сугерираат постоење на аномалии во однос на претпоставките. На пример, сликите 2 и 3 сугерираат променлива дисперзија (се зголемува или менува), додека сликата 4 сугерира дека линерниот модел е веројатно несоодветен.

На следните слики е прикажан веројатносниот дијаграм и дијаграмот остатоци - $\hat{y}$ за регресиониот модел од примерот 15.1.



Како што се гледа и од двете слики, тие не индицираат никакви несоодветности на регресиониот модел. Нормалниот веројатносен дијаграм оди во прилог на нормална распределба, а остатоците d_i се очигледно во интервалот (–2, 2).

15.4.2. Коефициент на детерминираност

Да се потсетиме дека ние веќе ја разгледувавме можната линеарна зависност меѓу две случајни променливи преку коефициентот на корелација ρ (види поглавје 5.5.1). Вредноста

$$R^2 = 1 - \frac{SS_E}{SS_T}, \text{ за } SS_E = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \text{ и } SS_T = \sum_{j=1}^{n} (y_j - \bar{y}_j)^2$$

се нарекува *коефициент на детерминираност* и често се користи за оценка на соодветноста на регресиониот модел. Ако дополнително ја воведеме *регресионата сума на квадрати* со

$SS_R = \sum_{j=1}^{n} (\hat{y}_j - \bar{y}_j)^2$, имаме дека важи

$$\sum_{j=1}^{n} (y_j - \bar{y}_j)^2 = \sum_{j=1}^{n} (\hat{y}_j - \bar{y}_j)^2 + \sum_{j=1}^{n} (y_j - \hat{y}_j)^2, \text{ т.е.}$$

$SS_T = SS_R + SS_E$.

Сега коефициентот на детерминираност може да се напише во друга форма, како

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T}.$$

Ако X и Y се две случајни променливи, R^2 е едноствно квадрат на коефициентот на корелација, $R^2 = (\rho_{X,Y})^2$. Оттука следува дека секогаш важи $0 \le R^2 \le 1$. Обично R^2 го интерпретираме како степен на варијабилност на податоците во однос на регресиониот модел. За примерот 15.1 (чистота на кислородот) имаме $R^2 = 1 - SS_E/SS_T = 1 - 21.25/173.38 = 0.877$, што значи дека моделот зема предвид 87.7% од варијабилноста на податоците.

Генерално, коефициент на детерминираност не го "мери" падот на регресионата линија. Поголеми вредности на R^2 не повлекуваат пострмен пад. Исто така, поголеми вредности на R^2 не повлекуваат соодветност на моделот, т.е. R^2 може да биде големо, а линеарниот регресионен модел сепак да биде несоодветен. И на крај, големо R^2 не повлекува подобро предвидување на идните податоци.

15.5. Општа линеарна регресија*

Во општ случај, случајната променлива Y може да зависи од m независни променливи, што води до модел од облик

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon,$$

што се нарекува повеќедимензионален (multiple) линеарен регресионен модел. Параметрите β_j, $j = 0, 1, \ldots, m$ се регресиони коефициенти што ја даваат очекуваната промена на одговорот Y при промена на регресионите променливи x_j. Геометриски овој модел опишува хипер-рамнина во m-димензионалниот простор $\mathbb{R}^m$.

Линеарната регресија често се користи за апроксимација на функции. Тоа значи дека вистинската функцопнална зависност на Y од $x_1, x_2, \ldots, x_m$ е непозната и под одредени услови линеарниот регресионен модел дава соодветна апроксимација.

Тука е важно да нагласиме дека многу други модели со наизглед покомплексна структура може да бидат анализирани преку линеарната регресија. На пример, секој полиномен модел како што е

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon,$$

може да се стандардизира со ставање $x_1 = x$, $x_2 = x^2$ и $x_3 = x^3$, и така да се добие линеарен регресионен модел со 3 променливи. Слично, моделот од втор ред

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon,$$

станува стандарден линеарен со $x_3 = x_1^2$, $x_4 = x_2^2$ и $x_5 = x_1 x_2$.

Да претпоставиме дека имаме $n > m$ податоци каде што x_{ij} ќе означува i-ти податок за променливата x_j. Секој податок $(x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$, $i = 0, 1, \ldots, n$, го задоволува регресиониот модел, т.е.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i, \quad i = 0, 1, \ldots, n.$$

Ова многу позгодно се претставува во матрична форма

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{каде што}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \quad \text{и} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}.$$

Целта е да се минимизира грешката $\boldsymbol{\varepsilon}$ по $\boldsymbol{\beta}$, т.е. да се минимизира

$$L(\beta_0,\beta_1,\ldots,\beta_m)=\sum_{j=1}^{n}\varepsilon_j^2=\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\ =$$
$$=\ \mathbf{y}^{\mathrm{T}}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$$
$$\mathrm{d}L/\mathrm{d}\boldsymbol{\beta}=0-(\mathbf{y}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}-\mathbf{X}^{\mathrm{T}}\mathbf{y}+(\mathbf{X}^{\mathrm{T}}\mathbf{X}+(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}})\boldsymbol{\beta}=-2\mathbf{X}^{\mathrm{T}}\mathbf{y}+2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$$

(тука се користени 3 правила за диференцирање на матрично-векторски изрази: $\mathrm{d}\mathbf{z}\mathbf{x}/\mathrm{d}\mathbf{x}=\mathbf{z}^{\mathrm{T}}$, $\mathrm{d}\mathbf{x}^{\mathrm{T}}\mathbf{z}/\mathrm{d}\mathbf{x}=\mathbf{z}$ и $\mathrm{d}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}/\mathrm{d}\mathbf{x}=(\mathbf{A}+\mathbf{A}^{\mathrm{T}})\mathbf{x}$).

Изедначувајќи го $\mathrm{d}L/\mathrm{d}\boldsymbol{\beta}=\mathbf{0}$, добиваме

$-2\mathbf{X}^{\mathrm{T}}\mathbf{y}\ +2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}=\mathbf{0}$, т.е. $\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^{\mathrm{T}}\mathbf{y}$ што дава $\hat{\boldsymbol{\beta}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Да забележиме дека $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ е секогаш несингуларна матрица бидејќи во спротивно, би имале две пропорционални редици или колони во матрицата што би повлекувало 2 исти податока или исти вредности за 2-е променливи, што треба да се елиминира од моделот.

Се разбира, ваквите матрични пресметки (се бара и инверзна матрица) вообичаено се прават на компјутер.

ПРИМЕР 15.6 Да ги разгледаме повторно податоците од примерот 10.7 за полупроводниците. Секој полупроводник од земен примерок од 25 полупроводници е поврзан на рамка со жица. Трите променливи во табелата се: влечна сила (силата потребна да се скрши обвивката), должина на жицата и висината на пакувањето на полупроводникот.

Влечна сила	*Должина на жица*	*Висина на пакување*	24.35	9	100			
9.95	2	50	27.50	8	300	10.30	1	585
24.45	8	110	17.08	4	412	34.93	10	540
31.75	11	120	37.00	11	400	46.59	15	250
35.00	10	550	41.95	12	500	44.88	15	290
25.02	8	295	11.66	2	360	54.12	16	510
16.86	4	200	21.65	4	205	56.63	17	590
14.38	2	375	17.89	4	400	22.13	6	100
9.60	2	52	69.00	20	600	21.15	5	400

Состави линеарен регресионен модел за влечната сила потребна да се скрши обвивката на полупрводникот како функција од должината на жицата и висината на пакувањето.

Решение

Моделот е од облик $Y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$, каде што Y е потребната влечна сила. Матриците се

$$\mathbf{y}=\begin{bmatrix}9.95\\24.45\\\vdots\\21.15\end{bmatrix},\quad \mathbf{X}=\begin{bmatrix}1&2&50\\1&8&110\\\vdots&\vdots&\vdots\\1&5&400\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix}.$$ Понатаму добиваме

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}=\begin{bmatrix}1&1&\dots&1\\2&8&\dots&5\\50&110&\cdots&400\end{bmatrix}\begin{bmatrix}1&2&50\\1&8&110\\\vdots&\vdots&\vdots\\1&5&400\end{bmatrix}=\begin{bmatrix}25&206&8294\\206&2396&77177\\9294&77177&3531848\end{bmatrix}$$ и

$$\mathbf{X}^{\mathrm{T}}\mathbf{y}=\begin{bmatrix}1&1&\dots&1\\2&8&\dots&5\\50&110&\cdots&400\end{bmatrix}\begin{bmatrix}9.95\\24.45\\\vdots\\21.15\end{bmatrix}=\begin{bmatrix}725.82\\8008.37\\274811.31\end{bmatrix},$$ па оттука следува

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}=\begin{bmatrix}25&206&8294\\206&2396&77177\\9294&77177&3531848\end{bmatrix}^{-1}\begin{bmatrix}725.82\\8008.37\\274811.31\end{bmatrix}=$$

$$=\begin{bmatrix}0.214653&-0.007491&-0.000340\\-0.007491&0.001671&-0.000019\\-0.00340&-0.000019&0.0000015\end{bmatrix}\begin{bmatrix}725.82\\8008.37\\274811.31\end{bmatrix}=\begin{bmatrix}2.26379\\2.74427\\0.01253\end{bmatrix}.$$

Значи бараниот регресионен модел е $y = 2.26379 + 2.74427x_1 + 0.01253x_2$.

Сега моделот може да се користи за наоѓање различни вредности на y за дадени вредности на x_1 и x_2. На пример, за 5-тиот податок (8, 295), се добива

$\hat{y} = 2.26379 + 2.74427\cdot 8 + 0.01253\cdot 295 = 27.9143.$

Остатокот е $e_5 = y - \hat{y} = 25.02 - 27.9143 = -2.8943$. ■

Исто како и кај едноставниот линеарен регресионен модел, важна е оценката на σ^2 (m + 2 - от параметар), т.е. дисперзијата на грешката ε. Центрирана оценка на σ^2 е повторно сума на квадратите на остатоците поделена со $n - m - 1$ (сега $m + 1$ е број на параметри што кај простата регресија беше 2). Значи $\hat{\sigma}^2 = \sum_1^n (y_i - \hat{y}_i)^2 / (n - m - 1)$.

Во примерот 15.6, добиваме $\hat{\sigma}^2 = 115.2/22 = 5.2364$.

Очекувањата и дисперзиите на точкастите оценки $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m$, т.е. на точкастата оценка $\hat{\boldsymbol{\beta}}$ се наоѓаат лесно под претпоставка грешките $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ да се со очекување 0 и дисперзија σ^2. За очекувањето имаме

$$E\hat{\boldsymbol{\beta}} = E((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}) = E((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) =$$
$$= E((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon})) = E\boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}E\boldsymbol{\varepsilon} = \boldsymbol{\beta}$$

т.е. $\hat{\boldsymbol{\beta}}$ е центрирана оценка на $\boldsymbol{\beta}$.

Дисперзиите на $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ се дијагоналните елементи на дисперзијата на $\hat{\boldsymbol{\beta}}$, т.е. на матрицата на коваријантност $\Sigma = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Недијагоналните елементи на матрицата Σ ($m \times m$ матрица) се коваријациите меѓу $\hat{\beta}_i$ и $\hat{\beta}_j$. Корените на дијагоналните елементи $\sqrt{\sigma^2\Sigma_{ii}}$ се стандардните девијации, но ако дисперзијата σ^2 се замени со оценката $\hat{\sigma}^2$, елементите $\sqrt{\hat{\sigma}^2\Sigma_{ii}}$ се нарекуваат стандардни грешки на $\hat{\beta}_i$ со ознака $se(\hat{\beta}_i)$ и се користат за проценка на точноста на регресионите коефициенти.

Во нашиот пример, стандардните грешки се $se(\hat{\beta}_0) = 1.060$, $se(\hat{\beta}_1) = 0.09352$ и $se(\hat{\beta}_2) = 1.060$.

Очекувањето и дисперзијата на параметрите на линеарната регресија ќе бидат потребни за составување интервали на доверба и тестирање хипотези за нивната вредност.

15.6. Интервални оценки на параметрите на линеарната регресија*

Како и во други слични ситуации, и кај општата линеарна регресија е корисно да се состават инрервали на доверба за регресионите коефициенти. И тука претпоставуваме дека грешките ε_j се независни со нормална распределба $Z(0, \sigma^2)$. Тоа значи дека случајните променливи Y_i се независни и нормално распределени со очекување $\beta_0 + \sum_{j=1}^{m}\beta_j x_{ij}$ и дисперзија σ^2. Бидејќи оценката $\hat{\boldsymbol{\beta}}$ е линеарна комбинација од податоците, таа има исто така нормална распределба со очекување $\boldsymbol{\beta}$ и матрица на коваријантност $\Sigma = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Тогаш секоја од m-те статистики

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2\Sigma_{jj}}},\ j = 0, 1, \ldots, m,$$

каде што Σ_{jj} е j-тиот дијагонален елемент на матрицата на коваријантност, има студентова распределба со n

– m – 1 степени на слобода. Ова води до следните интервални оценувачи на β_j,

$$p\left(\hat{\beta}_j - t_{\alpha/2,n-m-1}\sqrt{\hat{\sigma}^2\Sigma_{jj}} < \beta_j < \hat{\beta}_j + t_{\alpha/2,n-m-1}\sqrt{\hat{\sigma}^2\Sigma_{jj}}\right) = 1-\alpha\ .$$

ПРИМЕР 15.7 Состави 95% интервал на доверба за коефициентот β_1 во моделот за влечната сила од примерот 15.6.

Решение

Точкастата оценка за β_1 е $\hat{\beta}_1$ = 2.7443, а соодветниот дијагонален елемент на матрицата на коваријантност е Σ_{11} = 0.001671. Оценката $\hat{\sigma}^2$ = 5.2352, а $t_{0.025,22}$ = 2.074. Така ја добиваме интервалната оценка

$$p\left(\beta_j \in 2.7443 \pm 2.074\sqrt{5.2352 \cdot 0.001617}\right) =$$

$$= p\left(\beta_j \in (2.5503, 2.9383)\right) = 0.95\ . \blacksquare$$

15.6.1. Интервална оценка за регресионата линија*

Покрај регресионите коефициенти, може да се состави интервална оценка за просечната вредност на регресијата (mean response) во конкретна точка, на пример $x_{01}, x_{02}, \ldots, x_{0m}$. За оценка на просечната вредност ставаме $\mathbf{x}_0 = [1\ x_{01}\ x_{02}\ \ldots\ x_{0m}]^{\mathrm{T}}$ и добиваме

$$E(Y \mid \mathbf{x_0}) = \mu_{Y|\mathbf{x}_0} = \mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}\ ,\ \text{што се оценува со}\ \hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}\ .$$

$\hat{\mu}_{Y|\mathbf{x}_0}$ е центрирана оценка со дисперзија $D\hat{\mu}_{Y|\mathbf{x}_0} = \sigma^2\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0$. Интервалната оценка се конструира од статистиката

$$T = \frac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{\sqrt{\hat{\sigma}^2\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0}}$$

што има студентова распределба со $n - m - 1$ степени на слобода. Така се добива

$$p\left(\mu_{Y|\mathbf{x}_0} \in \hat{\mu}_{Y|\mathbf{x}_0} \pm t_{\alpha/2,n-m-1}\sqrt{\hat{\sigma}^2\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0}\right) = 1-\alpha\ .$$

ПРИМЕР 15.8 Состави 95% интервал на доверба за влечната сила за должина на жицата x_1 = 8 и висината на пакувањето x_2 = 275 од примерот 15.6.

Решение

Имамр дека $\mathbf{x}_0 = [1\ 8\ 275]^{\mathrm{T}}$ па точкастата оценка на просечно очекуваната вредност е

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} = [1\ 8\ 275]\begin{bmatrix} 2.2638 \\ 2.7443 \\ 0.0125 \end{bmatrix} = 27.66.$$ За дисперзијата се добива

$$\hat{\sigma}^2\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0 =$$

$$= 5.2352\cdot[1\ 8\ 275]\begin{bmatrix} 0.214653 & -0.007491 & -0.00034 \\ -0.00749 & 0.001671 & -0.000019 \\ -0.00034 & -0.000019 & 0.0000015 \end{bmatrix}\begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix} = 0.2327.$$

Така ја добиваме интервалната оценка

$$p(\mu_{Y|\mathbf{x}_0} \in 27.66 \pm 2.074\sqrt{0.2327}) = p(\mu_{Y|\mathbf{x}_0} \in (26.66, 28.66)) = 0.95.$$ ■

15.6.2. Интервална оценка за нови податоци*

Линеарниот регресионен модел може да се користи за предвидување на вредности на Y за нови вредности на $\mathbf{x}$. Ако $\mathbf{x}_0 = [1\ x_{01}\ x_{02}\ \ldots\ x_{0m}]^{\mathrm{T}}$ е нов податок, тогаш точкастата оценка за овој податок се добива од

$$\hat{Y}_0 = \mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}.$$

Слично како кај простата линеарна регресија, интервалниот оценувач за нови податоци се добива со едноставна интервенција во дисперзијата, што наместо $\sigma^2\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0$ станува $\sigma^2(1+\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0)$. Тоа води до интервален оценувач

$$p\left(Y_0 \in \hat{Y}_0 \pm t_{\alpha/2,n-m-1}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0)}\right) = 1-\alpha.$$

За новите податоци (предвидувања) важи слична дискусија како и за простата линеарна регресија. Имено, веднаш се забележува дека овој интервал е поширок од интервалите на регресионата линија (дисперзијата е поголема). Тоа доаѓа оттаму што предвидувањето покрај стандардната грешка на оценката на просекот во дадената точка $\mathbf{x}_0$, вклучува и грешка од варијабилноста на случајната променлива Y за истата вредност $\mathbf{x} = \mathbf{x}_0$.

ПРИМЕР 15.9 Користејќи ги податоците од примерот 15.6, состави 95% интервал на доверба за влечната сила кога должина на жицата $x_1 = 8$ и висината на пакувањето $x_2 = 275$ се разгледуваат како нов податок.

Решение

Имаме дека $\mathbf{x}_0 = [1\ 8\ 275]^{\mathrm{T}}$, па точкастата оценка на просечно очекуваната вредност е $\hat{y}_0 = \mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} = 27.66$.

Од $\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0 =$ од претходниот пример $= 0.2327/5.2352 = 0.04445$, за дисперзијата се добива $\hat{\sigma}^2(1+\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0) = 5.4679$. Оттука интервалот на предвидување е

$$p(Y_0 \in 27.66 \pm 2.074\sqrt{5.4679}) = p(\mu_{Y|\mathbf{x}_0} \in (22.81, 32.51)) = 0.95\,.$$

Очигледно интервалот е доста поширок од интервалот добиен со интервална оценка на регресионата линија. ■

15.7. Тестирање хипотези за општата линеарна регресија*

Тестирањето хипотези кај линеарниот регресионен модел е корисно од аспект на проверка на соодветноста на параметрите. Тестирањето претпоставува дека грешките ε_j се независни и со нормална распределба $Z(0, \sigma^2)$.

Најпрво ќе го разгледаме тестот за значајноста на регресијата

H_0: $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ наспроти H_{A}: $\beta_j \neq 0$, за барем едно j.

Отфрлањето на H_0 значи дека најмалку еден од регресионите променливи $x_1, x_2, \ldots, x_m$ влијае (придонесува) значајно во моделот. Кога H_0 е точна, SS_R/σ^2 е случајна променлива со хи-квадрат распределба и m степени на слобода, додека SS_E/σ^2 е случајна променлива со хи-квадрат распределба, но со $n - m - 1$ степени на слобода. Оттука е јасно дека соодветно формиран количник на овие случајни променливи може да даде Фишерова распределба. Значи тест статистиката е

$$F = \frac{SS_R/m}{SS_E/(n-m-1)},$$

и сега хипотезата H_0 се отфрла со ниво на значајност α кога вредноста f на статистиката е поголема од критичната вредност $f_{\alpha,(m,n-m-1)}$.

SS_E се добива од $SS_E = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

(другите два собирока се $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}\mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}^{\mathrm{T}}\mathbf{y})^{\mathrm{T}}\hat{\boldsymbol{\beta}} = 0$)

За наоѓање SS_R може најпрво да најдеме дека

$$SS_T = \sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2, \text{ и сега}$$

$$SS_R = SS_T - SS_E = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2.$$

ПРИМЕР 15.10 Тестирај ја значајноста на регресијата од примерот 15.6 за ниво на значајност $\alpha = 0.05$.

Решение

Налпрво ги пресметуваме:

$$SS_T = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 = 27\,177.951 - (725.82)^2/25 = 6105.9447,$$

$$SS_R = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 = 27\,062.7775 - (725.82)^2/25 = 5990.7712,$$

и сега $SS_E = SS_T - SS_R = 115.1735$.

Вредноста на статистиката е $f = \dfrac{5990.7712/2}{115.1735/(25-2-1)} = \dfrac{2995.3856}{5.2352} = 572.17$.

Сега, бидејќи $f = 572.17 > 3.44 = f_{0.05,(2,22)}$ (3.44 е од десна страна на распределбата) ја отфрламе H_0 и заклучуваме дека влечната сила е линеарно зависна од должина на жицата или висината на пакувањето или од двете. Се разбира и многу малата P-вредност на тестот (2.0285×10^{-21}) го потврдува тоа. Ова веднаш не повлекува дека моделот е соодветен. Потребни се понатамошни тестирања пред тој да може да се користи во праксата. ■

За оценка на соодветноста на регресиониот модел може да се искористи и коефициент на детерминираност R^2. Во нашиот пример, имаме $R^2 = SS_R/SS_T = 5990.7712/6105.9447 = 0.9811$. Тоа значи дека моделот зема предвид 98.11% од варијабилноста на податоците. Од друга страна, треба да се иа предвад дека R^2 е проблематичен кротериум за мера на квалитетот на согласување на моделот со податоците бидејќи тој секогаш расте со додавањето на нови податоци. Од тие причини, често се преферира подесената R^2 статистика

$$R_a^2 = 1 - \frac{SS_E/(n-m-1)}{SS_T/(n-1)}.$$

Сега R_a^2 се зголемува со додавање на нов податок само ако новиот податок го намалува квадратот на просечната грешка. Во нашиот пример $R_a^2 = 0.979$. Во основа R_a^2 го пенализира додавањето на нови податоци и како такво е чувар на моделот од пресогласување (overfitting) со расположивите податоци. Имено, недоволното согласување како и преголемото согласување на моделот со податоците не е добро и води кон поголеми грешки при предвидувањата. Фигуративно тоа може да се изрази, на пример, преку машина што треба да препознае дали еден објект е дрво врз база на претходни податоци за дрвата. Машината пресогласена со податоците од претходните дрва кога ќе види ново дрво ќе заклучи дека тоа не е дрво бидејќи нема ист број лисја со веќе виденитe. Од друга страна машина со мало согласување со претходните податоци може да заклучи дека сѐ што е зелено е дрво.

Често пати е потребно тестирање хипотеза за конкретен регресионен корфициент. На пример, моделот може да стане подобар со бришење или додавање на некој коефициент. Хипотезите за индивидуален регресионен коефициент β_j се

H_0: $\beta_j = 0$ наспроти H_A: $\beta_j \neq 0$.

Отфрлањето на H_0 значи дека регресорот x_j треба да остане во моделот. Статистиката за тестирање е со студентова распределба

$T = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \Sigma_{jj}}}$, каде што Σ_{jj} е j-тиот дијагонален елемент на матрицата на коваријантност $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Хипотезата H_0: $\beta_j = 0$ се отфрла кога вредноста на статистиката $|t| > t_{\alpha/2,n-m-1}$. Ваквиот тест се нарекува маргинален бидејќи регресониот коефициент β_j зависи од сите регресиони променливи.

ПРИМЕР 15.11 Тестирај го регресониот коефициент β_2 (висината на пакувањето) во моделот од примерот 15.6 за ниво на значајност $\alpha = 0.05$.

Решение

Вториот дијагонален елемент Σ_{22} во матрицата $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ е 0.0000015, па за статистиката добиваме

$$t = \frac{0.01253}{\sqrt{5.2352 \cdot 0.0000015}} = 4.4767 .$$

Сега бидејќи $t = 4.4767 > 2.074 = t_{0.025,22}$ ја отфрламе H_0 и заклучуваме дека висината на пакувањето е битна променлива во моделот. P-вредноста на тестот за $t = 4.4767$ е 0.0002, што тоа го потврдува. ■

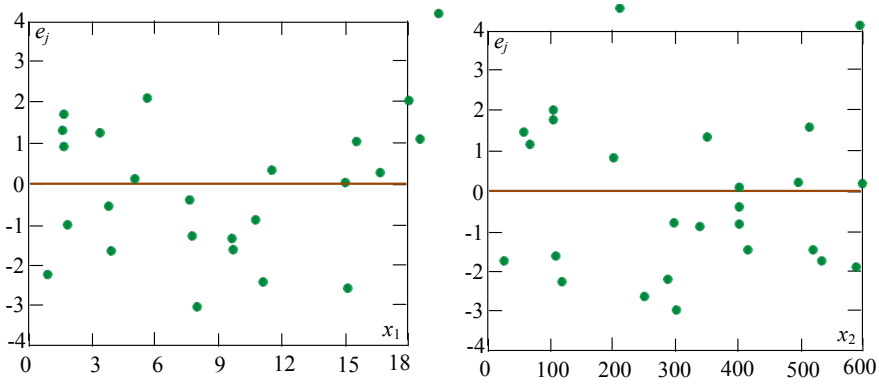
Постојат и други пристапи за тестирање на регресионите коефициенти индивидуално [Montgomery 2003], што тука не ги разгледуваме.

15.8. Соодветност на општата линеарна регресија*

Слично како и кај простата линеарна регресија, има различни пристапи за приближна проверка дали податоците се погодни за линеарна регресиона анализа.

И тука како и кај простата линеарна регресија за проверка на приближната "нормалност" на грешките ε_j, може да се користат остатоците $e_j = y_j - \hat{y}_j$, $j = 1, 2, \ldots, n$.

За таа цел може да се искористат хистограмите или нормалните веројатносни дијаграми. Од друга страна бројот на податоци често пати е мал за добивање на релевантен хистограм, но веројатносниот дијаграм е прифатлива опција. Остатоците може да се стандардизираат ставајќи $d_j = e_j / \sqrt{\hat{\sigma}^2}$, $j = 1, 2, \ldots, n$ и сега ако грешките се нормално распределени, приближно 95% од вредностите d_j треба да се во интервалот (–2, 2). Податоците што се надвор од интервалот укажуваат на присуство на натрапници (outliers), т.е. нетипични податоци во однос на мнозинството. На следните слики е прикажани дијаграмите на остатоци - $\hat{y}$ за променливите x_1 и x_2 од регресиониот модел од примерот 15.6.



Како што се гледа и од двете слики, врските меѓу влечната сила и должината на жицата, како и меѓу влечната сила и висината на пакувањето не изгледаат линеари поради многуте натрапници. Тоа индицира

дека во моделот недостасува некоја променлива (на пример x_1^2) што би требала да се вклучи во моделот.

Тука само ќе напоменеме дека покрај стандардизираните остатоци, во литературата и компјутерските програми од оваа област се користат и други скалирани остатоци со цел полесно да се забележат невообичаените податоци. Еден од популарните такви пристапи се студентизираните остатоци од облик

$$s_j = e_j / \sqrt{\hat{\sigma}^2 (1 - H_{jj})}\ , j = 1, 2, \ldots, n$$

каде што H_{jj} е j-тиот дијагонален елемент на $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\ \mathbf{X}^{\mathrm{T}}$. Да забележиме дека студентизираните остатоци се секогаш поголеми од стандардизираните остатоци.

Кај општата линеарна регресија се случува некои податоци да се натрапници, или позитивно гледано *влијателни*. Обично овие податоци се далеку од другите и имаат поголемо влијание на моделот. Се разбира, би било згодно овие влијателни податоци да се откријат, и ако се само "погрешни" да се елиминираат од моделот, аке не се, тогаш сознанието за нив може да има битно значење при користењето на моделот. Една таква интересна мера се базира на растојанието меѓу обичната оценка $\hat{\boldsymbol{\beta}}$ базирана на сите n податоци и оценката $\hat{\boldsymbol{\beta}}_{(i)}$ добиена со исклучување на i-тиот податок (повторно работејќи по најмали квадрати). Оваа мера [Cook 1982] е дефинирана со

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2},\ i = 1, 2, \ldots, n.$$

Сега ако i-тиот податок е влијателен, тогаш $\hat{\boldsymbol{\beta}}_{(i)}$ ќе се разликува значително од $\hat{\boldsymbol{\beta}}$, па големо D_i значи дека податокот е влијателен. Растојанијата D_i се пресметуваат преку

$$D_i = \frac{s_i}{m+1} \frac{H_{ii}}{1 - H_{ii}},\ i = 1, 2, \ldots, n,$$

каде што s_i се студентизираните остатоци, а H_{ii} е i-тиот дијагонален елемент од матрицата $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$.

На пример, растојанието D_i на првиот податок од примерот 15.6 е

$$D_1 = \frac{s_i}{m+1} \frac{H_{ii}}{1 - H_{ii}} = \frac{\left(e_1 / \sqrt{\hat{\sigma}^2 (1 - H_{jj})} \right)^2}{m+1} \frac{H_{ii}}{1 - H_{ii}} =$$

$$= \frac{(1.57/\sqrt{5.2352(1-0.1573)})^2}{3}\frac{0.1573}{1-0.1573} = 0.035.$$

За $D_i < 1$, податокот не се смета за влијателен.

Бидејќи остатоците имаат иста логика и кај простата и кај општата линеарна регресија, методите што ги дискутиравме кај простата линеарна регресија се применливи и тука, како на пример коефициентот на детерминираност R^2, веројатносните дијаграми итн.

15.9. Сведување на линеарна регресија*

Во некои случаи може да процениме дека простиот линеарен регресионен модел $Y = \beta_0 + \beta_1 x + \varepsilon$ е несоодветен поради линеарноста. Потребата од нелинеарност може да се забележи на дијаграмот со податци $\{(x_j, y_j), j = 1, 2, \ldots, n\}$, или од претходното искуство со таквиот тип податоци. Тука е битно да се нагласи дека линеарноста се однесува на регресионите коефициенти β_0 и β_1, а не на регресионата променлива x.

Така на пример моделот $Y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$ е линеарен иако регресионата променлива е нелинеарна. Тука $1/x$ едноставно го заменуваме со друга променлива z и моделот станува станува стандарден линеарен.

Во некои ситуации, нелинеарната зависност може да се претстави линеарно, користејќи соодветни трансформации. Таквите нелинеарни модели се нарекуваат внатрешно (intrinsically) линеарни.

На пример, таков модел е нелинеарниот експоненцијален модел

$$Y = \beta_0 e^{\beta_1 x}\varepsilon .$$

Оваа функција е внатрешно линеарна и може да се трансформира во права линија со логаритмирање

$$\ln Y = \ln\beta_0 + \beta_1 x + \ln\varepsilon .$$

И тука се бара несистематската грешка $\ln\varepsilon$ да биде нормално распределена со очекување 0 и дисперзија σ^2.

Понекогаш може да се искористат повеќе трансформации за моделот, за тој да се линеарисира. На пример, во моделот

$$Y = \frac{1}{e^{\beta_0+\beta_1 x+\varepsilon}},$$

со ставање $Z = 1/Y$ и логаритмирање се добива

$\ln Z = \beta_0 + \beta_1 x + \varepsilon$.

Од друга страна, моделот $Y = \beta_0 \beta_1^x + \varepsilon$ е суштински нелинеарен и нема смена што би го направила да биде линеарен.

Слична дискусија важи и за општиот линеарен регресионен модел

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon.$$

Ваквиот модел вклучува многу нелинеарни модели по променливите, а типичен пример се полиномните модели, како на пример следниот од 4-ред со една променлива

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \varepsilon,$$

или следниот полиномен модел од 3-ред со 2 променливи

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \\ + \beta_{111} x_1^3 + \beta_{222} x_2^3 + \beta_{122} x_1^2 x_2 + \beta_{122} x_1 x_2^2 + \beta_{123} x_1 x_2 x_3 + \varepsilon.$$

Во општ случај, линеарниот регресионен модел може да се напише во облик

$$Y = \beta_0 + \beta_1 h_1 + \beta_2 h_2 + \ldots + \beta_m h_m + \varepsilon$$

каде што $h_i = h_i(x_1, x_2, \ldots, x_m)$ се произволни функции од регресионите променливи. Сите овие модели се линеарни бидејќи линеарноста се однесува на регресионите коефициенти $\beta_0, \beta_1, \ldots, \beta_m$.

ПРИМЕР 15.12 Популацијата на еден град, во просек се зголемува (y) во зависност од бројот на поминати години од 1970-а (x). Според моделот

$$EY = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Користејќи ги податоците од следната табела, определи ги оценките на β_0, β_1 и β_2 по методот на најмали квадрати.

x	0	1	2	3	4	5
y(%)	1.03	1.32	1.57	1.75	1.83	2.33

Решение

Ставаме $x_1 = x$ и $x_2 = x^2$ и добиваме линеарен регресионен модел. Од

$$\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 4 & 9 & 16 & 25 \end{bmatrix} \text{ и } \boldsymbol{y}^{\mathrm{T}} = \begin{bmatrix} 1.03 & 1.32 & 1.57 & 1.75 & 1.83 & 2.33 \end{bmatrix} \text{ следува}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix}^{-1} \begin{bmatrix} 9.83 \\ 28.68 \\ 110.88 \end{bmatrix} = \begin{bmatrix} 1.07 \\ 0.20 \\ 0.01 \end{bmatrix}$$

Значи регресиониот модел е $y = 1.07 + 0.20x + 0.01x^2$. ■

Општиот облик на регресионен модел е $Y = f(\mathbf{x}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$, каде што $\mathbf{x}^{\mathrm{T}} = [1\ x_1\ x_2\ \ldots\ x_m]$. Кај линеарниот модел имаме просто $f(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. Кај нелинеарниот модел барем еден од изводите на $f(\mathbf{x}, \boldsymbol{\beta})$ по регресионите коефициенти $\beta_0, \beta_1, \ldots, \beta_m$ зависи од барем еден од овие параметри. И кај нелинеарните регресиони модели би можело коефициентите да се бараат по методот на најмали квадрати, т.е. да се минимизира

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{n} \left(y_i - f(\mathbf{x}_i, \boldsymbol{\beta})\right)^2 ,$$

но познато е дека барање екстрем на нелинеарни функции од повеќе променливи вообичаено бара користење на приближни методи.

ЗАДАЧИ

1. Специјален случај на линеарна регресија е кога $\beta_0 = 0$, т.е. $Y = \beta x + \varepsilon$.
   а) Определи ја оценката $\hat{\beta}$ на регресиониот коефициент β по методот на најмали квадрати;
   б) Најди го очекувањето и дисперзијата на $\hat{\beta}$;
   в) Најди центрирана оценка за σ^2 (дисперзијата на Y).

2. Да претпоставиме дека врската меѓу постојан притисок x и време на откажување Y на изложената компонента е опишано преку проста линеарна регресија $y = 65 - 1.2x$, со $\sigma = 8$. Најди ја веројатноста $Y > 50$, кога $x = 20$. Ако Y_1 и Y_2 се регресиони вредности за податоците x_1 и x_2, колкава е веројатноста $Y_1 > Y_2$?

3. Под претпоставка дека Y (линеарен регресионен модел) е со нормална распределба, определи оценувач на дисперзијата σ^2 на Y користејќи го методот на максимална подобност.

4. Во следната табела се дадени податоци за потрошувачката на бензин MPG (на автопат во милји по галон) и запремината на моторите (во кубни инчи) за 20 модели на автомобили во САД:

Модел	MPG	Волумен in^3	Модел	MPG	Волумен in^3
Acura Legend	30	97	Ford Taurus	27	153
BMW 735i	19	209	Ford Tempo	33	90
Buick Regal	29	173	Honda Accord	30	119
Chevrolet Cavalier	32	121	Mazda RX-7	23	80
Chevrolet Celebrity	30	151	Mercedes 260E	24	159
Chrysler Conquest	24	156	Mercury Tracer	29	97
Dodge Aries	30	135	Nissan Maxima	26	181
Dodge Dynasty	28	181	Oldsmobile Cutlass	29	173
Ford Escort	31	114	Plymouth Laser	37	122
Ford Mustang	25	302	Pontiac Grand Prix	29	173

а) Состави линеарен регресионен модел за потрошувачката (Y) според запремината на моторот (x);

б) Оцени ја потрошувачката на бензин за автомобил со мотор од 150 in^3;

в) Оцени ја потрошувачката на бензин на Ford Escort (114 in^3) и соодветното отстапување.

5. Има индиции дека количество пареа што месечно се користи во една хемиска фабрика е во некаква релација со просечната околна температура (во F°) за тој месец. Во следната табела се дадени податоците за претходната година:

Месец	Темпер.	Потр./1000
Јануар	21	185.79
Февруар	24	214.47
Март	32	288.03
Април	47	424.84
Мај	50	454.58
Јуни	59	539.03
Јули	68	621.55
Август	74	675.06
Септември	62	562.03
Октомври	50	452.93
Ноември	41	369.95
Декември	30	273.98

а) Под претпоставка дека простиот линеарен регресионен модел е соодветен за оваа зависност, состави го моделот користејќи го методот на најмали квадрати;

б) Која е проценката за користење на пареата кога температурата е 55° F;

в) Да претпоставиме дека просечната месечна температура е 47° F. Најди ја вредноста на потрошувачката на пареа според моделот како и соодветното отстапување.

6. Во студиите на испитување на транспортот, се претпоставува дека бучавоста на камионите (Y) е линеарно поврзана со логаритамот на брзината (x). Обезбедени се податоци, дадени во табелата:

x	20	30	40	50	60	70	80	90	100	x - брзина во km на час
y	55	63	68	70	72	78	74	76	79	y - децибели

Под претпоставка дека $Y = \beta_0 + \beta_1 \cdot \log_{10} x + \varepsilon$, определи го регресиониот модел (параметрите β_0 и β_1).

7. Бетон направен од рамномерна мешавина на паста цемент-вода се користи во некои области со обилни врнежи, поради одличните дренажни особини. Во следната табела се дадени податоци за тоа како порозноста (y) во % е во врска со тежината по единица (x):

x	99.0	101.1	102.7	103.0	105.4	107.0	108.7	110.8
y	26.8	27.9	27.0	25.2	22.8	21.5	20.9	19.6

x	112.1	112.4	113.6	113.8	115.1	115.4	120.0
y	17.1	18.9	16.0	16.7	13.0	13.6	10.8

а) Состави регресиона линија за овие податоци;

б) Интерпретирај го падот (slope) на линијата;

в) Што се случува ако моделот се користи за предвидување на порозноста за тежина 135?

г) Пресметај ги отстапувањата на првите две вредности;

д) Пресметај ја точкастата оценка на σ.

8. Густината на популацијата во Јапонија носи проблеми со недостаток на ресурси. Еден од многуте проблеми е отстранувањето на отпадоците (сметот). Обид овој проблем да се ублажи е дизајн на нова компресиона машина за обработка на отпадниот талог. Најважниот параметар на машината е односот на влажноста на компресираните топчиња (y во %) и степенот на филтрирање x даден во килограми/метар/час (kg-DS/m/hr). Следната табела дава дел од овие податоци:

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2

x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Состави едноставен линеарен регресионен модел за овие податоци.

9. Варијациите во квалитетот на малтерот употребен во градбата влијае не само на структурниот и акустички дизајн на објектот, туку и на дизајнот на греењето, вентилацијата и ладењето. Во следната табела се дадени податоци за густината на малтерот (y) во либри/фити^3 во зависност од процентот на содржина на воздух (x):

x	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0
y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1

x	15.3	16.2	17.8	18.7	19.7	20.6	25.0
y	111.3	107.2	108.9	107.8	111.0	106.2	105.0

а) Состави едноставен линеарен регресионен модел за податоците;

б) Состави 95%-ен интервал на доверба за β_1.

10. Корозијата на челичните шипки е најголемиот проблем за трајноста на металните конструкции. Карбонизацијата е хемиски процес што води до намалување на pH вредноста и почеток на корозијата. Во следната табела се дадени мерењата на длабочината на карбонизацијата (x) во милиметри и цврстината (y) во мегапаскали на една конструкција:

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4

x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8

а) Состави прост линеарен регресионен модел за податоците;

б) Состави 95%-ен интервал на доверба за просечната цврстина на шипките со длабочина на карбонацијата од 45 милиметри;

в) Состави 95%-ен интервал на доверба за цврстина на шипка со измерена длабочина на карбонацијата од 45 милиметри.

11. Разгледај ги следните 4 интервали во врска со податоците од задача 7:

а) 95%-ен интервал за просекот на порозноста за единешна тежина 110;

б) 95%-ен интервал на предвидување за просекот на порозноста за единечна тежина 110;

в) 95%-ен интервал за просекот на порозноста за единешна тежина 115;

г) 95%-ен интервал на предводување за просекот на порозноста за единечна тежина 115.

Без пресметка на овие интервали, спореди ги нивните широчини.

12. Разгледај ги повторно податоците од задача 8 за степенот на филтрација и влажноста на топчињата од машината за компресија.

а) Состави 90%-ен интервал на доверба за вистинскиот процент на содржина на влажност кога степенот на филтрација е 125.

б) Предвиди ја вредноста на степенот на влажност за степенот на филтрација 125 користејќи 90%-ен интервал. Каков е овој интервал во споредба со интервалот од а);

в) Какви се интервалите од а) и б) во споредба со соодветните интервали за степен на филтрација од 115?

г) Тестирај хипотеза H_0: $\hat{\beta}_0 + \hat{\beta}_1 \cdot 125 = 80$ наспроти H_A: $\hat{\beta}_0 + \hat{\beta}_1 \cdot 125 < 80$ со ниво на значајност 0.01.

13. За податоците на потрошувачката на бензин според кубикажата на автомобилот (задача 4):

а) Оцени ги стандардните грешки на падот и пресекот (β_1 и β_0);

б) Тестирај ја значајноста на регресијата за $\alpha = 0.01$. Најди ја P-вредноста на тестот;

в) Тестирај H_0: $\beta_1 = -0.05$ наспроти H_1: $\beta_1 < -0.05$ за $\alpha = 0.01$. Најди ја P-вредноста на тестот;

г) Тестирај H_0: $\beta_0 = 0$ наспроти H_0: $\beta_0 \neq 0$ за $\alpha = 0.01$. Најди ја P-вредноста на тестот.

14. За податоците на потрошувачката на бензин според кубикажата на автомобилот (задача 4), најди ја вкупната варијабилност на потрошувачката на бензин придружена на кубикажата на моторот (коефициентот на детерминираност). Состави нормален веројатносен дијаграм и процени дали претпоставката за нормалност е задоволена.

15. Просечната месечна потрошувачка на струја (y) на еден производствен погон е линеарно зависна од просечната температура на околината (x_1) и бројот на работни деиови во месецот (x_2). Во следната табела се дадени податоци за 12 месеци (1 година):

x_1	20	26	41	55	60	67	75	79	70	55	45	33
x_2	23	21	24	25	24	26	25	25	24	25	25	23
y	210	206	260	244	271	285	270	265	234	241	258	230

Состави линеарен регресионен модел користејќи најмали квадрати.

16. Направена е студија за врската на триењето (y) со вискозноста на маслото (x_1) и количината на полнење (x_2). Податоците се дадени во табелата:

x_1	1.6	15.5	22.0	43.0	33.0	40.0
x_2	851	816	1058	1201	1357	1115
y	293	230	172	91	113	125

а) Состави линеарен регресионен модел за овие податоци;

б) Оцени ја σ^2 и начајноста на регресијата за $\alpha = 0.01$. Најди ја P-вредноста на тестот;

в) Пресметај ја вредноста на y за $x_1 = 25$ и $x_2 = 1000$;

г) Состави линеарен регресионен модел за овие податоци што го вклучува и мешаниот член x_1x_2;

д) Пресметај ја вредноста на y за $x_1 = 25$ и $x_2 = 1000$ во новиот модел.

17. Разгледај го регресиониот модел за потрошувачката на струја од задача 15.

а) Тестирај ја значајноста на моделот за $\alpha = 0.05$ и најди ја P-вредноста.

б) Тестирај ги коефициентите на регресијата индивидуално за $\alpha = 0.05$.

18. Разгледај го повторно регресиониот модел за триењето во зависност од вискозноста на маслото и полнењето од задачата 16. Преработи го моделот со вклучување на додатен коефициент на интеракција x_{12} од задача 16 г).

а) Тестирај ја значајноста на моделот за $\alpha = 0.05$;

б) Дали додатниот регресионен коефициент е битен за регресиониот модел?

19. Состави 90% интервал на доверба за просечната месечна потрошувачка на струја од задача 15, за просечната температура на околината $x_1 = 28$ и бројот на работни денови во месецот $x_2 = 22$. Состави интервал на доверба за истото како нов податок.

20. Разгледај ја повторно студијата за врската на триењето (y) со вискозноста на маслото (x_1) и количината на полнење (x_2) од задача 16.

а) Најди го коефициентот на детерминираност R^2 на овој модел;

б) Најди го коефициентот на детерминираност R^2 на моделот со додатен член x_1x_2. Дали ова повлекува подобрување на моделот?

Додаток А

Комбинаторика

Потребата од комбинирање на броеви, букви или други објекти се јавува во различни области на науката, но и во секојдневниот живот. На пример, во процесот на производство треба да се распореди работата меѓу луѓето и машините, во училиштата да се направи распоред на часовите, во хемијата да се испита распоредот на атомите во молекулите, во практично сите научни области да се процени веројатноста на појавување на некој настан, во игрите на среќа треба да се направат скратени системи за спортски прогнози или лото итн.

Комбинаториката е дисциплина што се занимава со структури дефинирани на конечни множества. На неа може да се гледа како на методологија на подредување на елементите на конечните множества во структури, според оредени правила. Овие структури ќе ги нарекуваме комбинаторни објекти.

На пример, ако сакаме да ги најдеме можните начини на конструкција на сите четирицифрени броеви составени од цифрите 1, 2, 3, 4, 5, 6 и 7 со или без повторување на цифрите, добиваме комбинаторни објекти наречени варијации што соодветно можат да бидат со и без повторување. Ако сакаме да ги најдеме сите броеви што можат да се конструираат од цифрите 2, 7, 8 и 9 добиваме комбинаторни објекти наречени пермутации, додека ако меѓу цифрите има еднакви, на пример, 2, 2, 6, 6, 8 добиваме пермутации со повторување. Ако сакаме да ги најдеме сите можности за избор на 7 броја од 39 различни броја, добиваме комбинаторни објекти наречени комбинации, додека ако меѓу броевите има еднакви, комбинациите се со повторување.

Комбинаториката како дисциплина е многу повеќе широка, отколку длабока. Таа има допирни точки со речиси сите математички дисциплини, појавувајќи се и на места каде што не ја очекуваме. Нејзината фокусираност е повеќе на техниките отколку на резултатите.

Трите основни комбинаторни проблеми се:

1) Проблем на *постоење*: дали некој комбинаторен објект постои, т.е. дали некое подредување е возможно?
2) Проблем на *пребројување*: на колку различни начини комбинаторниот објект може да биде конструиран?
3) Проблем на *конструкција*: како да се состави алгоритам за конструкција (генерирање) на комбинаторни објекти од некој тип?

Покрај проблемот на пребројување, тука големо внимание ќе посветиме на конструкцијата на комбинаторните објекти од даден тип, т.е. на алгоритмите за нивно генерирање, што е веројатно најважен проблем во практиката. Ефективните алгоритми за конструкција на комбинаторните објекти не се тривијални и бараат познавање на некои важни алгоритмски техники. Генерирањето на комбинаторните објекти обично се прави во три чекори: избор на почетниот објект, трансформација на тековниот објект во следен и проверка на условот за крај:

дефинирај прв објект
while *не е исполнет условот за крај* **do** { **write** *објект*
транформирај во нов објект

Дефинирањето на комбинаторните објекти ќе го направиме со терминологија од формалните јазици (букви, азбука, зборови итн.). Ова се чини е поприроден пристап од користење на терминологијата на множества и колекции објекти, што ја нема таквата едноставност и елеганција.

Нека е дадено множеството $A = \{a_1, a_2, ..., a_n\}$. Без губење на општоста, наместо директно разгледување на комбинаторните објекти на множеството A, може да се разгледуваат комбинаторните објекти на множеството $I_n = \{1, 2, ..., n\}$. Множеството A, односно I_n ќе го нарекуваме *азбука*, а неговите елементи *букви*.

Сите подредени, непразни и конечни низи на елементи од множеството I_n ќе ги нарекуваме зборови на I_n, а нивното множество ќе го означиме со I_n^+.

На пример, за $I_8 = \{1, 2, ..., 8\}$, низите 221134, 87654322, 22222, 88 се зборови над I_8 и припаѓаат на I_8^+.

Должина на зборот z, е природниот број $|z|$ еднаков на бројот на букви во z. Така, $|221134| = 6$ и поопшто, ако $z = z_{i_1} z_{i_2} \ldots z_{i_k}$, $|z| = k$.

Опсег на зборот z, е природниот број $||z||$ еднаков на бројот на различните букви во z. На пример, $||221134|| = 4$.

Јасно е дека секогаш важи $||z|| \leq |z|$.

Градба на зборот z, што ќе ја означуваме со $G(z)$, е векторот $(g_1, g_2, \ldots, g_n)$ каде што елементите g_i, $i = 1, 2, \ldots, n$ се природни броеви (заедно со 0) и $g_i = m$ акко буквата i се појавува точно m пати во z. На пример, $G(221134) = (2, 2, 1, 1, 0, 0, 0, 0)$.

Од дефиницијата на градбата на збор непосредно следува дека, $|z| = g_1 + g_2 + \ldots + g_n$, а $||z||$ = бројот на ненулти елементи во $(g_1, g_2, \ldots, g_n)$.

На пример, за $I_5 = \{1, 2, 3, 4, 5\}$ и $z = 2313511$ имаме, $|z| = 7$, $||z|| = 4$ и $G(z) = (3, 1, 2, 0, 1)$.

Сите комбинаторни објекти разгледувани во овој текст ќе бидат дефинирани како зборови над азбуката I_n.

А.1. Варијации

Варијација без повторување, или едноставно *варијација* од ред k е секој збор $z \in I_n$ за кој важи $|z| = ||z|| = k$, т.е. секој збор z со k различни букви. На пример, за азбуката $I_3 = \{1, 2, 3\}$ варијациите без повторување од ред 2 се: 12, 13, 21, 23, 31 и 32.

Типичен случај на јавување на варијациите без повторување е кога ги бараме можностите за извлекување на k елементи од множество со n елементи при што редоследот е важен. На пример, ако од шпил карти влечеме 3 една по една, за првата карта имаме 52 можности, па за секоја извлечена прва карта имаме 51 можности за извлекување на втората и за секоја двојка претходно извлечени карти имаме 50 можности да ја извлечеме третата. Така, вкупниот број на можности е $52 \cdot 51 \cdot 50$. Втор типичен случај кога се јавуваат варијациите без повторување е кога ги разгледуваме можностите за разместување на n елементи на k места. На пример, ако разместуваме на 3 места карти од еден шпил, на првото место можеме да ставиме една од 52-те карти, па за секоја така ставена карта можеме една од 51-те карти да ја ставиме на второто место и за секоја двојка претходно ставени карти можеме една до останатите 50 карти да ја ставиме на третото место. Така, вкупниот број на можни разместувања е $52 \cdot 51 \cdot 50$.

Бројот на варијациите без повторување од ред k може да се одреди вака: Имаме n можности да ја избереме првата буква од I_n. Потоа, за секој вака избрана буква имаме $n-1$ можности да ја избереме втората, па за секоја избрана двојка букви имаме $n-2$ можности да ја избереме третата, итн. Заклучно со k-тата буква бројот на различни избори на k-те букви изнесува $n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$, па постојат точно толку зборови со должина k во кои буквите не се повторуваат.

Редоследот во кој се одредуваат варијациите може да биде различен, а од посебен интерес е таканаречениот лексикографски редослед. При овој редослед, првата буква од претходната варијација одејќи од лево на десно што е различна од соодветната буква во следната варијација е помала од неа. Со други зборови, варијациите се генерираат во растечки редослед. Следниот алгоритамот ги генерира сите варијации без повторување (од секој ред $k \le n$) на азбуката I_n, при што варијациите за фиксно k се во лексикографски редослед.

```
procedure VAR(n)
  v_k ← k ← i ← 1
  while i ≠ 0 do
    write v_1 v_2 … v_k
    if k < n
      then
        l = најмалиот број што не е еден од v_1, v_2, …, v_k
        k ← k + 1
        v_k ← l
      else
        // Најди го најдесното место каде што v_i > v_{i+1}
        i ← n − 1
        while v_i > v_{i+1} do i ← i − 1
        // v_j е најмалиот поголем елемент надесно од v_i
        j ← n
        while v_i > v_j do j ← j − 1
        v_i ← v_j
        k ← i
  return
```

Генерирањето започнува со варијација од една буква и се проширува со нови варијации добиени со додавање на по една буква на почетната. Проширувањето се прави со додавање на најмалиот број што не се наоѓа во тековната варијација $v_1v_2...v_к$. На пример, ако $n = 5$, и тековната варијација е 135, тогаш $l = 2$, или ако тековната варијација е 34, тогаш $l = 1$. Променливата k во секој момент го дава бројот на букви во тековната варијација. Кога k станува еднакво на n, буквите по i-тата едноставно се забораваат ($k \leftarrow i$) и потоа повторно алгоритамот ја проширува варијацијата додека тоа е можно ($k < n$).

Варијација со повторување од ред k е секој збор z со должина k, $|z| = k$. На пример, варијации со повторување од ред 2 на азбуката $I_3 = \{1, 2, 3\}$ се: 11, 12, 13, 21, 22, 23, 31, 32 и 33.

Типичен случај на јавување на варијациите со повторување е кога ги бараме можностите за извлекување на k елементи од множество со n елементи со враќање, при што редоследот е важен. На пример, ако од шпил карти влечеме 3, за првата имаме 52 можности, па за секоја извлечена прва карта имаме повторно 52 можности за извлекување на втората (картата е вратена во шпилот) и за секоја двојка претходно извлечени карти имаме 52 можности да ја извлечеме третата. Така, вкупниот број на можности е $52 \cdot 52 \cdot 52 = 52^3$. Втор општ случај кога се јавуваат варијации со повторување е кога ги разгледуваме можностите за разместување на n елементи на k места со повторување.

Бројот на варијациите со повторување од ред k на азбука од n букви може да се одреди на следниов начин. Имаме n можности да ја избереме првата буква од I_n. Потоа за секоја вака избран прва буква имаме повторно n можности да ја избереме втората буква од I_n, па за секоја вака избрана двојка букви имаме n можности да ја избереме третата и така натаму. Заклучно со k-тата буква бројот на различни избори на k-те букви изнесува $n \cdot n \cdot \ldots \cdot n = n^k$, па постојат точно толку зборови со должина k во кои буквите може да се повторуваат.

Следниот алгоритам ги генерира варијациите со повторување од ред k во лексикографски редослед.

```
procedure VARP(n, k)
  i ← 1
  v_1 ← 0
  while i > 0 do { v_i ← v_i + 1
                   if v_i > n then i ← i − 1
                                  else if i = k then write v_1 v_2 ... v_k
                                                else { i ← i + 1
                                                       v_i ← 0
return
```

А.2. Пермутации

Нека е даден зборот $z \in I_n^+$. *Пермутација* на z е секој збор z' со иста градба, $G(z) = G(z')$. Да забележиме дека зборот z е една пермутација на себеси. Ако $|z| = ||z||$ (во z нема повторување на букви) пермутациите се

без повторување и тогаш едноставно ги викаме *пермутации*, а во спротивно, кога $||z|| < |z|$ (во z има повторување на букви) пермутациите се *со повторување*. На пример, пермутациите без повторување на зборот z = 132 над азбуката $I_3 = \{1,2,3\}$ се: 123, 132, 213, 231, 312 и 321.

Пермутациите се јавуваат во случаи кога треба n елементи да разместиме на n места, т.е. да направиме подредувања на n елементи. На пример, сакаме 5 работни задачи да ги доделиме на 5 работника или да ги напишеме сите зборови со буквите: м, и, л, а, н (не е битно дали зборот има значење). И во двата случаи сите можности се определени со пермутациите од 5 елементи и нивниот број е 120.

Имајќи предвид дека пермутациите на збор z за кој важи $|z| = ||z|| = n$ се всушност варијации без повторување од ред n, нивниот број може да се добие кога во изразот $n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$ ставиме $k = n$. Така, веднаш се добива дека бројот на пермутации на збор со должина n е $n\cdot(n-1)\cdot\ldots\cdot2\cdot1 = n!$. Генерирањето на пермутации во лексикографски редослед може да се направи со следниот алгоритам.

```
procedure PERM(n)
 for j ← 0 to n do p_j ← j
 i ← 1
 while i ≠ 0 do
        write p_1 p_2 ... p_n
        // Најди го најдесното место каде што p_i > p_{i+1}
        i ← n − 1
        while p_i > p_{i+1} do i ← i − 1
        // Најди p_j, најмалиот поголем елемент на десно од p_i
        j ← n
        while p_i > p_j do j ← j − 1
        // Промени места на p_i и p_j и преуреди ги p_{i+1}, ..., p_n
        p_i ↔ p_j
        d ← i + 1
        g ← n
        while g > d do
               p_g ↔ p_d
               g ← g − 1
               d ← d + 1
return
```

Генерирањето на пермутациите започнува со пермутацијата 12 ...n. Преминот од пермутацијата $p_1p_2...p_n$ кон следната се прави со нејзино разгледување оддесно налево и наоѓање на првата позиција каде што $p_i < p_{i+1}$. Потоа надесно од p_i се пронаоѓа најмалата буква p_j поголема од p_i и нивните места се менуваат. На крај, буквите $p_{i+1}p_{i+2}...p_n$ што се во опа-

ѓачки редолед се преуредуваат во растечки редослед. На пример, за n = 8 и пермутацијата 26587431 имаме дека $p_i = 5$ и $p_j = 7$. Откако нивните места се променат и преуредат во растечки редослед буквите надесно од i-тата, се добива новата пермутација 26713458.

Постојат значително поефикасни пристапи за генерирање на пермутациите од погоре опишаниот. Така, еден од најефикасните алгоритми е генерирањето на пермутациите со минимални измени. Кај тој пристап, при премин од претходна кон следна пермутација само две соседни букви треба да си ги сменат местата. Тука главен проблем е како да се направи добар редослед на генерирање при кој ќе се добијат сите пермутации само еднаш.

Рекурзивна верзија на ваков алгоритам не е многу тешко да се конструира. За n = 1, пермутацијата 1 е добра. Да претпоставиме дека имаме низа од $(n-1)!$ пермутации на зборот $12...n-1$ каде што секоја пермутација се разликува од претходната само со промена на две соседни букви. Тогаш секоја од добиените $(n-1)!$ пермутации ја прошируваме со уфрлување на буквата n на секое од n-те возможни места меѓу останатите букви.

На пример, за n = 4, редоследот на така добиените пермутации ќе биде следниот:

$$\begin{array}{cccccc}
1234 & 4132 & 3124 & 4321 & 2314 & 4213 \\
1243 & 1432 & 3142 & 3421 & 2341 & 2413 \\
1423 & 1342 & 3412 & 3241 & 2431 & 2143 \\
4123 & 1324 & 4312 & 3214 & 4231 & 2134 \\
\underbrace{123} & \underbrace{132} & \underbrace{312} & \underbrace{321} & \underbrace{231} & \underbrace{213} \\
\multicolumn{3}{c}{\underbrace{\qquad\qquad\qquad}_{12}} & \multicolumn{3}{c}{\underbrace{\qquad\qquad\qquad}_{21}} \\
\multicolumn{6}{c}{\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{1}}
\end{array}$$

Рекурзивната верзија на овој пристап за генерирање пермутации е релативно едноставна, но и помалку ефикасна. Ние тука ќе ја дадеме итеративната верзија што е значително посложена. За нејзина ефективна реализација се потребни три вектори што ќе ги чуваат: тековната пермутација $p_1p_2...p_n$, нејзината обратна пермутација $q_1q_2...q_n$[5] и вектор на насоката $(s_1, s_2, ..., s_n)$, каде што секое s_i е -1 (движењето е налево), $+1$ (движењето е надесно) или 0 (нема движење). Буквата се движи налево или надесно сé додека не стигне до поголема буква (затоа на поче-

[5] Ако пермутациите $p_1p_2...p_n$, се разгледуваат како пресликување $f(i) = p_i$, обратната пермутација е обратно пресликување f^{-1}. На пример, обратната пермутација на пермутацијата 4213 е 3241, т.е. $f(1) = 4 \Rightarrow f^{-1}(4) = 1$, значи во обратната пермутација на 4-тата позиција стои 1, $f(2) = 2 \Rightarrow f^{-1}(2) = 2$, $f(3) = 1 \Rightarrow f^{-1}(1) = 3$ и $f(4) = 3 \Rightarrow f^{-1}(3) = 4$.

токот се става $n_0 = n_{n+1} = n+1$). Во тој момент насоката на движење на буквата се менува и се придвижува следната помала од неа буква (ако е можно). Обратната пермутација се користи за лесно пронаоѓање на следната помала буква. На почетокот, s_1 се става 0, за кога m станува рамно на 1 во внатрешниот циклус сè да биде во ред.

Иако алгоритамот за реализацијата на оваа постапка е кус, тој во имплементацијата на деталите не е едноставен. Овој алгоритам е еден од најефективните познати алгоритми за генерирање пермутации.

procedure PERM(n)
  for $i \leftarrow 1$ **to** n **do** $\begin{cases} p_i \leftarrow q_i \leftarrow i \\ s_i \leftarrow -1 \end{cases}$
  $s_1 \leftarrow 0$
  $p_0 \leftarrow p_{n+1} \leftarrow m \leftarrow n+1$
  while $m \neq 1$ **do** $\begin{cases} \textbf{write } p_1 p_2 \dots p_n \\ m \leftarrow n \\ \textbf{while } p_{q_m+s_m} > m \textbf{ do } \begin{cases} s_m \leftarrow -s_m \\ m \leftarrow m-1 \end{cases} \\ p_{q_m} \leftrightarrow p_{q_m+s_m} \\ \text{// во овој момент } p_{q_m+s_m} = m \\ q_{p_{q_m}} \leftrightarrow q_m \end{cases}$
return

Како што веќе дефиниравме на почетокот, пермутација на збор z за кој важи $||z|| < |z|$ (имаме повторувања на букви во z) ја викаме пермутација со повторување. На пример, пермутациите на зборот 1313 над азбуката $I_3 = \{1, 2, 3\}$ се: 1133, 1313, 1331, 3113, 3131 и 3311.

Пермутациите со повторување се јавуваат во случаите кога сакаме да разместиме група од n елементи меѓу кои има еднакви на n места, т.е. да направиме подредувања на n-те елементи. На пример, сакаме 5 работни задачи од кои 3 се од еден тип а другите 2 од од друг тип да ги доделиме на 5 работника или да ги напишеме сите зборови со буквите: м, а, т, е, м, а, т, и, к, а (не е битно значењето). Во првиот случај имаме пермутации со повторување од 5 елементи и бројот на можности е 10, а во вториот случај имаме пермутации со повторување од 10 елементи и број на можности 30240.

Бројот на пермутации со повторување на зборот z, со должина n ($|z| = n$) е помал од бројот на пермутации без повторување на збор со иста должина, бидејќи поради повторувањата на буквите во z, нивното преместување не води секогаш до нов збор. Нека е даден збор z со должина $|z| = n$ и градба $(g_1, g_2, ..., g_n)$. Сега на $g_1!$ начини можеме да ја преместу-

ваме буквата 1 без да се промени зборот z, на $g_2!$ начини можеме да ја преместуваме буквата 2 без да се промени зборот z, итн., на $g_n!$ начини можеме да ја преместуваме буквата n без да се промени зборот z. Така, вкупниот број пермутации со повторување на z, се добива кога пермутациите (без повторување) на збор со иста должина се скратат $g_1!\cdot g_2!\cdots g_n!$ пати и изнесува

$$\frac{n!}{g_1!\,g_2!\cdots g_n!}.$$

Да забележиме дека тука премолчано користиме дека 0! = 1.

За генерирање пермутации со повторување ќе дадеме еден не така ефективен алгоритам што ги пронаоѓа пермутациите на произволен збор z во лексикографски редослед. Зборот z може да биде со повторување на букви $||z|| < |z|$, или без повторување $||z|| = |z|$. Алгоритамот повторно се базира на техниката на пребарување со враќање (backtracking).

```
function PROV(m, p(1:m), gz(1:n))
  ind ← true
  for i ← 1 to m do gp_i ← 0
// Ја пресметуваме градбата на зборот p(1:m) до m симболи
  for i ← 1 to m do gp_{p_i} ← gp_{p_i} + 1
// Дали зборот p(1:m) води до нова пермутација?
  i ← 1
  while gp_i ≤ gz_i and i ≤ m do i ← i + 1
  if i ≤ m then ind ← false
return ind

procedure PERMP(n, z(1:n))
// Ја пресметуваме градбата и описот на зборот z
  for i ← 1 to n do gz_i ← 0
  for i ← 1 to n do gz_{z_i} ← gz_{z_i} + 1
// ops е најголемиот елемент во z
  for i ← 1 to n do if gz_i > 0 then ops ← i
  m ← 1
  p_m ← 0
  while m > 0 do
      while p_m ≤ ops do
          p_m ← p_m + 1
          if p_m ≤ ops and PROV(m, p(1:m), gz(1:n))
              then if m = n then write p_1 p_2 ... p_n
              else
                  m ← m + 1
                  p_m ← 0
      m ← m − 1
end
```

Функцијата PROV(·) ја проверува градбата на до тој момент изградената пермутација $p_{(1:m)}$ на зборот $z_{(1:n)}$. Ако која било буква на градбата $gp_{(1:m)}$ на $p_{(1:m)}$ е поголема од соодветниот елемент на градбата $gz_{(1:n)}$ на почетниот збор $z_{(1:n)}$, конструкцијата на пермутацијата се прекинува и се прави обид со наредна буква. Јасно е дека во таков случај не може да се добие збор со иста градба со почетниот збор.

Иако овој алгоритам работи и со зборови без повторувачки букви (генерира пермутации без повторување), во таквите случаи тој е неефективен. Општоста и едноставноста е платена со неговата недоволна ефективност. Да забележиме дека како влез на алгоритамот може да биде кој било збор со која било должина (на пример, $z_{(1:5)} = 21331$, $z_{(1:10)} = 4412113222$ а излез се сите негови пермутации во лексикографски редослед. Влезот мора да биде таков што n е поголем или рамен (во случај на збор без повторувања) од најголемата вредност на буквите во зборот z, што е секогаш исполнето ако почнеме од 1 и немаме "празнини" во растот на буквите. На пример, алгоритамот нема да работи добро за влез $z_{(1:4)} = 1215$ ($n = 4 < 5 = ops$), но сепак ќе работи за $z_{(1:4)} = 1214$ ($n = 4 = 4 = ops$).

А.3. Подмножества и комбинации

Потребата од анализа на комбинаторните објекти што соодветствуваат на сите подмножества од едно множество е честа во комбинаторните алгоритми. На пример, подмножествата на множеството {♣,♦,♥} се: ∅, {♣}, {♦}, {♥}, {♣,♦}, {♣,♥}, {♦,♥} и {♣,♦,♥}. Подмножествата како комбинаторни објекти е згодно да се дефинираат како варијации со повторување од ред n на азбуката {0, 1}. Елементот i од множеството $I_n = \{1, 2, ..., n\}$ припаѓа на едно негово подмножество акко i-тиот елемент во варијацијата е еднаков на 1. Овие варијации се едноставно низи на нули и единици со должина n. Варијацијата 00...0 соодветствува на ∅ множество, а варијацијата 11...1 на целото множество I_n. На пример, за $I_8 = \{1, 2, ..., 8\}$ и подмножеството {1, 4, 7, 8}, соодветната варијација е 10010011. Бројот на варијациите со повторување од ред n над азбуката {0, 1} е 2^n, па толкав е и бројот на подмножествата на множеството I_n.

Генерирањето на варијациите со повторување над азбуката {0, 1} е дадено во следниот алгоритам. Тој во основа ги генерира сите бинарни низи со должина n.

```
procedure BNIZI(n)
  for i ← 1 to n+1 do b_i ← 0
  while b_{n+1} ≠ 1 do
      write b_n b_{n-1} ... b_1
      i ← 0
      while b_i = 1 do
          b_i ← 0
          i ← i+1
      b_i ← 1
return
```

Преминот од овој алгоритам, кон алгоритам што директно ги дава подмножествата на I_n е тривијален и тоа му го оставаме на читателот.

Многу почесто, наместо разгледување на сите подмножества на едно множество, потребно е разгледување само на комбинаторните објекти што соотвествуваат на подмножествата со фиксна големина.

Комбинација без повторување, или просто *комбинација*, од ред k е секој збор $z = z_1z_2...z_k$ од I_n^+ за кој важи $z_1 < z_2 < ... < z_k$. На пример, комбинациите од ред 2 на азбуката $I_3 = \{1,2,3\}$ се: 12, 13 и 23.

Бројот на комбинациите од ред k на азбука од n букви може да се одреди на следниот начин. Постојат $n\cdot(n-1)\cdot(n-2)\cdot...\cdot(n-k+1)$ начини да се одберат зборови со k различни букви $z_1z_2...z_k$ (види ја дискусијата за варијациите); во сите $k!$ можности за избор на овие букви имаме само една комбинација (една можност да е $z_1 < z_2 < ... < z_k$). Оттука бројот на комбинациите од ред k на азбука од n букви изнесува

$$\frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} .$$

Следниот алгоритам ги генерира комбинациите од ред k над I_n.

```
procedure KOMB(n,k)
  c_0 ← -1
  for i ← 1 to k do c_i ← i
  j ← 1
  while j ≠ 0 do
      write c_1 c_2 ... c_k
      // Најди прв елемент од десно што не е максимален
      j ← k
      while c_j = n-k+j do j ← j-1
      // Зголеми го за 1
      c_j ← c_j + 1
      // Секој наредан е претходен зголемен за 1
      for i ← j+1 to k do c_i ← c_{i-1} + 1
return
```

Генерирањето започнува со комбинацијата 12...k. Следната комбинација се добива од претходната така што тргнувајќи оддесно налево се бара првата буква што не ја достигнала својата максимална вредност. Таа буква се зголемува за 1, а на сите букви надесно од неа се доделуваат нови најмали возможни вредности. На пример, ако тековната комбинација од ред 7 над азбуката I_9 е 1234589, следната ќе биде 1234678 бидејќи 5 е првата буква оддесно што не е максимална. Таа се зголемува за 1 и станува 6, а останатите се 6 + 1 = 7 и 7 + 1 = 8.

Генерирањето комбинации може да се направи со сосема сличен алгоритам на алгоритамот за генерирање на варијациите со повторување од ред k во лексикографски редослед. Таквиот пристап ги генерира комбинациите постапно, вклучувајќи буква по буква во тековната комбинација, што може да биде корисно во случаите кога некои комбинации треба да се отфрлат во корен.

```
procedure KOMB(n,k)
  i ← 1
  c_1 ← 0
  while i > 0 do { c_i ← c_i + 1
                   if c_i > n then i ← i − 1
                              else if i = k then write c_1 c_2 ... c_k
                                            else { i ← i + 1
                                                   c_i ← c_{i−1}
return
```

Единствената, но суштинска промена во однос на алгоритамот за генерирање на варијации е наредбата $c_i \leftarrow c_{i-1}$ во последниот **else** блок. Со тоа се постигнува секоја следна буква да е поголема од претходната во процесот на градење на комбинацијата.

Комбинациите без повторување се веројатно најчесто користените комбинаторни објекти.

Комбинациите со повторување се исто така важни комбинаторни објекти што често се јавуваат во пракса. *Комбинација со повторување* од ред k е секој збор $z = z_1z_2...z_k$ за кој важи $z_1 \le z_2 \le ... \le z_k$. На пример, комбинациите со повторување од ред 2 на азбуката $I_3 = \{1, 2, 3\}$ се: 11, 12, 13, 22, 23 и 33.

Комбинациите со повторување се јавуваат во случаи кога треба да одбереме k елементи од n групи елементи. Вкупниот број елементи не е важен. На пример, ако во цвеќара има 5 видови цвеќе, правење различни букети од 3, 7 или 11 цвета се комбинации со повторување и нивниот број е 21, 330 или 1365 соодветно.

Бројот на комбинациите со повторување од ред k на азбука од n букви може да се одреди на следниот начин: Од дефиницијата следува дека бројот на комбинации со повторување е еднаков на бројот на целобројни решенија на системот неравенки $1 \le z_1 \le z_2 \le ... \le z_k \le n$; Овој систем неравенки е еквивалентен со системот $1 \le z_1 < z_2 + 1 < ... < z_k + k - 1 \le n + k - 1$; Имајќи предвид дека бројот на целобројни решенија на системот неравенки $1 \le b_1 < b_2 < ... < b_к \le n + k - 1$ е еднаков на бројот на комбинации од ред k на азбуката од $n + k - 1$ букви, добиваме дека бројот на комбинации со повторување изнесува:

$$\binom{n+k-1}{k} = \frac{n \cdot (n+1) \cdot \ldots \cdot (n+k-1)}{k!} \ .$$

На пример, бројот на целобројни решенија на системот неравенки $1 \le z_1 \le z_2 \le 3$ е 6 (тие се: 11, 12, 13, 22, 23, 33), а тие решенија се идентични со решенијата на системот неравенки $1 \le z_1 < z_2 + 1 \le 4$. Решенијата на системот неравенки $1 \le b_1 < b_2 \le 4$ се исто така 6 (тие се: 12, 13, 14, 23, 24, 34).

Патот по кој дојдовме до бројот на комбинациите со повторување го покажува и начинот на кој може тие да се генерираат. За тоа се доволни две мали интервенции во алгоритамот за генерирање на комбинации без повторување.

```
procedure KOMBP(n,k)
  c_0 ← −1
  for i ← 1 to k do c_i ← i
  j ← 1
  while j ≠ 0 do
      // Намали ја најдената комбинација за i − 1
      for i ← 1 to k do c_i ← c_i − i + 1
      write c_1 c_2 ... c_k
      j ← k
      // Сега тргнуваме од n + k − 1
      while c_j = n + j − 1 do j ← j − 1
      c_j ← c_j + 1
      for i ← j + 1 to k do c_i ← c_{i−1} + 1
return
```

Двете интервенции се во замената на n со $n + k - 1$ (бараме комбинации над азбука со $n + k - 1$ букви, а не со n) и пред испишувањето на резултатот ги намалуваме буквите за $i - 1$ (за колку се зголемени во азбуката со $n + k - 1$ букви).

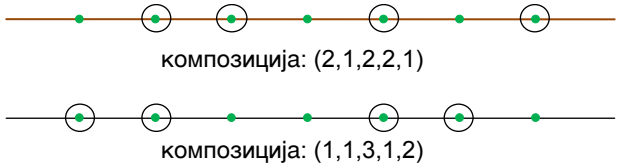
А.4. Композиции и разбивања

Сега ќе го разгледаме проблемот на разбивање на природниот број n во низа од природни броеви (r_1, r_2, ..., r_k) чија сума го дава n, $r_1 + r_2 + \ldots + r_k = n$. Во разбивањето редоследот на елементите не е важен и вообичаено се наметнува ограничување $r_i > 0$. Ако редоследот на броевите r_i е важен, тогаш разбивањето се нарекува *композиција* на n, и тогаш исто така се наметнува $r_i > 0$. Кога k е фиксно, добиваме композиции на n од ред k (од k делови). Во тој случај се дозволува $r_i = 0$ и тогаш добиваме слаби (weak) композиции на n од ред k. Се разбира, возможно е наметнување на различни ограничувања на композицајата и нејзините елементи. На пример, одредување максимална или минимална вредност на елементите r_i, ограничување на повторувањата на поедини елементи итн.

Следниов пример ја илустрира разликата меѓу композициите, композициите од ред k и разбивањата, за $n = 3$.

Композиции се: (3), (1, 2), (2, 1), (1, 1, 1);
Композиции од ред 2 се: (1, 2), (2, 1);
Слаби композиции од ред 2 се: (0, 3), (1, 2), (2, 1), (3, 0);
Разбивања се: (3), (1, 2), (1, 1, 1).

За да го одредиме бројот на композиции од ред k, згодно е бројот n да го претставиме како линија разделена со $n - 1$ точка (составена е од n единечни отсечки). Композициите од ред k се добиваат со различни избори на $k - 1$ точки на линијата и броење на отсечките меѓу избраните точки. На сликата се прикажани две композиции на 8 од ред 5.



Графичко претставување на композициите

Јасно е дека со сите избори на $k - 1$ точки од сите $n - 1$ точки се добиваат сите можни композиции од ред k, така што нивниот број изнесува: $\binom{n-1}{k-1}$.

Бројот на сите композиции на број n лесно се добива ако се забележи дека тој е еднаков на бројот на сите бинарни низи од $n - 1$ елемен-

ти што соотвествува на различните избори на сите $n-1$ точки на линијата. Така добиваме дека овој број е 2^{n-1}. Се разбира, истото може да се добие од

$$\sum_{k=1}^{n}\binom{n-1}{k-1}=2^{n-1}.$$

Бројот на слабите композиции (дозволени 0) може да се добие со дозволено повторување на изборот на n-те точки на линијата, што дава комбинации со повторување

$$\binom{n+k-1}{k-1}.$$

Следниот алгоритам ги генерира композициите на бројот n во редослед што е лексикографски ако композициите се замислат со иста должина со додавање 0 на десно. Како основа е користен алгоритамот за генерирање на бинарни низи со должина n.

```
procedure KOMP(n)
 for i ← 1 to n+1 do b_i ← 0
 while b_{n+1} ≠ 1 and b_n ≠ 1 do
                                  // бинарна низа во композиција
                                  l ← 0
                                  k ← 1
                                  for i ← 1 to n−1 do if b_i = 0 then
                                                                     r_k ← i − l
                                                                     l ← i
                                                                     k ← k + 1
                                  r_k ← n − 1
                                  write (r_k, r_{k−1}, ..., r_1)
                                  // генерирање нова бинарна низа
                                  i ← 0
                                  while b_i = 1 do
                                                   b_i ← 0
                                                   i ← i + 1
                                  b_i ← 1
return
```

Трансформацијата на бинарната низа во композиција се прави со броење на последователните низи од единици при што разликите меѓу позициите на прекин на низите единици ги даваат броевите во композицијата. На пример за $n = 8$ и низата (11000100) се добива композицијата (3, 1, 1, 2, 1) бидејќи низата од единици се прекинува на 3-та позиција ($3 - 0 = 3$), па на 4-та (немаме единица $4 - 3 = 1$), на 5-та ($5 - 4 = 1$), на 7-та ($7 - 5 = 2$) и на 8 ($8 - 7 = 1$). Поедноставниот пристап на просто броење на поднизите од единици и нули дава решение кај кое композициите не се во редослед што асоцира на лексикографски.

За разбивањата на n, каде што редоследот на компонентите не е од интерес постојат повеќе алгоритами. Следниот алгоритам се смета за средно ефикасен и тој ги генерира разбивањата во растечка должина и во лексикографски редослед за секоја фиксна должина.

procedure RAZBIVANJE(n)
$k \leftarrow 1$
$r_1 \leftarrow n$
$r_0 \leftarrow -1$
while $k \le n$ **do**
write $(r_1, r_2, \ldots, r_k)$
$i \leftarrow k-1$
while $r_k - r_i < 2$ **do** $i \leftarrow i-1$
if $i \neq 0$ **then for** $j \leftarrow k-1$ **to** i **by** -1 **do** $r_j \leftarrow r_i + 1$
else **for** $j \leftarrow 1$ **to** k **do** $r_j \leftarrow 1$
$k \leftarrow k+1$
$r_k \leftarrow n - \sum_{j=1}^{k-1} r_j$
return

Во овој алгоритам разбивањето на k компоненти почнува со $r_1 = r_2 = \ldots = r_{k-1} = 1$ и $r_k = n - k + 1$. Понатаму, елементите се разгледуваат од десно налево, застанувајќи на најдесното r_i такво што $r_k - r_i \ge 2$. Сите r_j за $j = i, i + 1, \ldots, k - 1$ се заменуваат со $r_i + 1$, и на крај r_k се заменува со $n - \sum_{j=1}^{k-1} r_j$.

На пример, за $n = 12$, $k = 5$ и разбивање (1, 1, 3, 3, 4), наоѓаме дека 4 е поголем за повеќе од 2 од најдесната единица па следното разбивање ќе биде (1, 2, 2, 2, 5). Кога ниеден од елементите на разбивањето не се разликува од последниот за повеќе од 1, алгоритамот завршува.

Додаток Б

Веројатносни симулации

Веројатносните модели се понекогаш толку комплицирани што алатките од математиката стануваат недоволни за добивање одговори на релевантните прашања. Веројатносните (stochastic) симулации се алтернативен пристап: се генерираат вредности за случајните променливи и уфрлуваат во моделот и така се имитираат (симулираат) излезите за целиот модел.

Веројатносните симулации може да се гледаат како дел од симулациите Монте Карло (МК), поим што нашироко се користи во литературата. Во најширока смисла, МК симулација значи каква било симулација (не мора компјутерска) што користи случајни броеви во симулационата постапка. Терминот МК симулација, иако помалку соодветен од терминот веројатносна симулација, нашироко се користи бидејќи веројатно звучи "повозбудливо".

Идејата за користење на случајност во научни цели е многу постара од компјутерите. На пример, во проблемот на Буфон (пример 2.17), веројатноста p фрлена игла со должина $2l$ да пресече некоја од паралелните линии што се на растојание $2a$ ($l < a$) беше

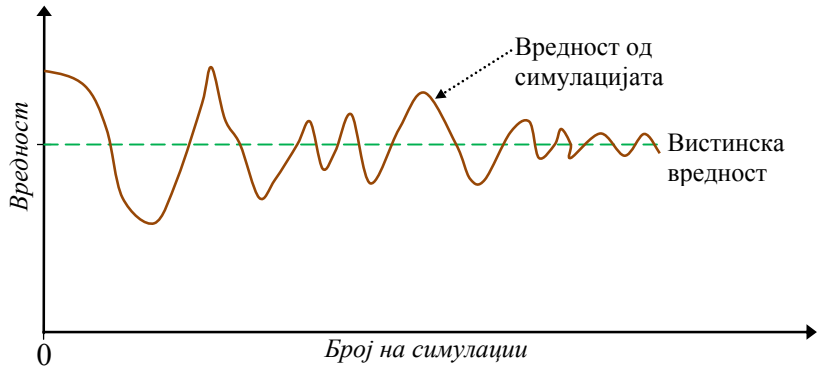
$$p = \frac{2l}{a\pi}, \text{ од каде за } \pi \text{ се добива } \pi = \frac{2l}{ap}.$$

Користејќи го овој резултат, повеќе истражувачи во периодот 1850-1925 година се обидувале приближно да го пресметаат бројот π. Обидите се сумирани во следната табела:

Истражувач	*l/a*	Бр. фрлања	Бр. пресеци	π
Wolf, 1850	0.80	5000	2532	3.1596
Smith, 1855	0.60	3204	1218.5	3.1553
De Morgan, 1860	1.00	600	382.5	3.1370
Fox, 1864	0.75	1030	489	3.1595
Lazzerini, 1901	0.83	3408	1808	3.1415929
Reina, 1925	0.5419	2520	869	3.1795

Очигледно, резултатот од симулациите на Lazzerini е предобар, што изгледа доста сомнително.

Со зголемувањето на бројот на симулациите, вредноста што од нив се добива ќе се приближува кон вистинската вредност приближно на начин како што е прикажано на следниот графикон:



Идеално би било да ги запреме симулациите во моментот кога вредноста од симулацијата ја сече линијата на вистинската вредност. Се разбира, тоа е невозможно да се направи, и обично ние однапред одредуваме голем број на симулации со цел што подобро да се приближиме до вистинската вредност. Ако ги тргнеме сомненијата околу резултатот на Lazzerini, тој веројатно на среќа запрел на број во кој имаме (речиси) пресек на вредноста од симулацијата со вистинската вредност.

МК симулациите нашироко се користат во различни научни области. Во математичката анализа (на пример, решавање интеграли), веројатноста и статистиката (на пример, оценки на веројатности и непознати параметри), физиката (на пример, во нелинеарната динамика или статистичката физика), оптимизацијата (на пример, симулирано подобрување (annealing), генетски алгоритми), инженерството (на пример, пресметка на ризици), општествените науки (на пример, разни квантитативни методи применети на популацијата), економија (на пример, финансиско инженерство), итн.

Методите на користење на случајни броеви за решавање на проблемите е слично како и секоја статистичка анализа само што наместо "реален" примерок, користиме конструиран примерок што се добива од случајните броеви. Секој примерок е во основа со нумерички вредности и се избира на случаен начин, па логично е да може и вештачки да се конструра со помош на случајни броеви.

Б.1. Случајни броеви

Во сржта на секоја симулација се случајните броеви. Оттука е особено важно да се има добар извор на случајни броеви за реализација на симулацијата. Обезбедување на "добра" низа случајни броеви не е едноставна работа како што некој би помислил (види ја дискусијата од поглавјето 8.2.1). Низите случајни броеви, генерално може да се поделат во три широки категории:

1) *Вистински случајни*: нема начин да се предвиди следниот случаен број во низата;
2) *Псевдо-случајни*: низата случајни броеви што е алгоритамски генерирана не е вистински случајна бидејќи таа може да биде комплетно повторена за познати почетни услови користени во алгоритамот;
3) *Квази случајни*: се користат како случајни броеви во некои симулации, но тие воопшто не се случајни броеви.

Поранешните, некомпјутерски методи за генерирање случајни броеви, како на пример користење децимали на ирационалните броеви, во основа даваат псевдо-случајни броеви. За генерирање случаен број од [0, 1] интервал може да се земат 4 по 4 последователни цифри од π,

3.14159265358979323846264338327950288419716939937510582097494459230781640628620899862803482534211706 79…

и тие броеви да се делат со 10000.

Со еволуцијата на технологијата почнале да се користат "машини" што генерираат случајни броеви од тип на лото, рулет итн. Користејќи специјализирана машина, фирмата RAND во 1955 публикувала книга со 1 милион случајни броеви. Се разбира и тие се само псевдо случајни.

Со развојот на компјутерската технологија се појавиле голем број методи што го користат компјутерскиот хардвер. Тие се базираат на читање на вредности од мемориски локации што практично дава "вистински" случајни броеви, иако формално тие се само псевдо-случајни. На пример, програмскиот пакет HAVEGE (HArdware Volatile Entropy

Gathering and Expansion) е софтвер за генерирање непредвидливи (апроксимација на вистински) случајни броеви [http://www.irisa.fr/caps/projects/hipsor/index.php, 2006]. Тој користи комплициран механизам на собирање на десетици илјади битови при повиците до оперативниот систем (системски прекини), користејќи го часовникот на хардверот. Ваквите псевдо-случајни броеви можеме да ги сметаме за непредвидливи. Генерално, непредвидливите случајни броеви се најпогодни за дисциплините како што е, на пример, криптографијата.

За поголемиот број симулации во науката, пожелно е користење на предвидливи случајни броеви со цел да се овозможи повторливост. Тоа значи кога програмата се повикува при исти влезни (почетни) параметри, низата генерирани случајни броеви е секогаш иста. Обично генераторот се иницијализира со почетно зададен цел број што се нарекува семе (seed). За иста вредност на семето се добива иста низа случајни броеви. За семе често се зема вредноста на системскиот часовник или просто се задава рачно. Можноста за повторување се базира на сознанието дека науката треба по дефиниција да изучува повторливи феномени. Квалитетот на генераторот може да се оценува според должината на низата без повторување на самата себе (периода). Периодата треба да е поголема од бројот на потребните случајни броеви за симулацијата. Кај многу денешни генератори, периодата е од ред на 4 бајтен цел број, $2^{32} \approx 4\times10^{9}$, што може да биде недоволно. Важна е и максималната вредност *max* што еден целоброен генератор ја дава (обично $2^{32}-1$). За добивање децимален случаен број r_{float} во [0, 1) интервал од целоброен случаен број r_{int}, едноставно ставаме

$$r_{float} = \frac{r_{int}}{max},$$

а обратно, ако сакаме од децимален случаен број од [0, 1) интервал да добиеме случаен цел број $\le k$, ставаме

$$r_{int} = (\text{int})(r_{float} \cdot k + 1)$$

каде што (int) значи цел дел.

Б.1.1. Генератори на случајни броеви

Постојат многу алгоритми за генерирање случајни броеви. Секој програмски јазик, како Це, Паскал или Јава има стандардни функции, но често и посебни библиотеки со понапредни функции за нивно генерирање. На пример, стандардните функции во Це се од облик

```
void srand(unsigned _seed);
int  rand(void);
```

каде што првата го задава семето, а втората генерира случајни броеви, коишто во Це се цели броеви од 0 до $2^{16} = 32768$.

Тука би нагласиле дека стандардните функции за генерирање случајни броеви (најчесто rand()) имплементирани во програмските јазици не се доволно добри за "сериозни" симулации. Повеќето од нив се базираат на линеарни конгруентални алгоритми кои генерираат случајни броеви според итеративната релација

$$I_{j+1} = a \cdot I_j + c \pmod{m}$$

каде што сите се цели броеви, а "mod m" означува остаток при делење со m. Почетната вредност I_0 е семе. Очигледно е дека периодата е најмногу $m - 1$ како и максималната вредност. На пример, за семе $I_0 = 4$, $a = 7$, $c = 4$ и $m = 2^4 - 1 = 15$ имаме

$$I_1 = a \cdot I_0 + c \pmod{m} = 32 \pmod{15} = 2.$$

Продолжувајќи ја оваа постапка се добива низата

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
I_i	4	2	3	10	14	12	13	5	9	7	8	0	4	2	3	10	14	12	13

од каде се гледа дека $I_{12} = I_0$, т.е. периодата е 12. Ова е доста добро ако се има предвид дека периодата не може да надмине 15. Ако за семе би земале $I_0 = 11$, периодата би била само 3 ($I_3 = I_0$) што е катастрофално. Ова води кон следните заклучоци:

а) Не е добро да се користи генератор каде што периодата зависи од семето;

б) Со внимателност треба да се одбира не само генераторот, туку и неговите константи (произволното менување на константите не гарантира ништо).

За линеарниот конгруентален алгоритам, условите за достигнување максимална периода од $m - 1$ се: c и m треба да се взаемно прости, $a - 1$ да е деливо со сите прости делители на m, и $a - 1$ да е деливо со 4 ако m е деливо со 4. Да забележиме дека $a = 7$, $c = 4$ и $m = 15$ не го задоволува вториот од овие услови (6 и 15 имаат заеднички прост делител 3), па така генераторот не ја достигнува максималната периода.

Во Це, стандардниот генератор е реализиран во слична форма со следниот код:

```
unsigned long seed=1;
void srand(unsigned int seed_set) {
  seed=seed_set;
}

int rand(void) {  // Not recommended for anything
  long a=1103515245;
  long c=12345;
  long div=32768;
  seed = seed*a + c;
  return (unsigned int) (seed/65536) % div;
}
```

Во симулационата програма најпрво се повикува функцијата

srand(unsigned int seed_set)

која го поставува семето (инаку останува seed = 1), а потоа се повикува функцијата rand(void) што генерира случаен број (остатокот при делење со div).

Многу често е користен генераторот со минимален стандард [Park, Miller 1988] што ги користи параметрите

$$a = 7^5 = 16\,807,\ c = 0 \quad \text{и} \quad m = 2^{31} - 1 = 2\,147\,483\,647.$$

Ваквиот генератор е релативно брз, во одредена смисла перфектен, но во многу апликации се покажува како лош. Слабостите се во: одредената корелација меѓу две последователни генерирања (по генериран мал број повторно следува мал број), генераторот не може да генерира последователно два исти броја, постои корелација на куси растојанија и др.

На крај би кажале дека постојат многу варијации и подобрувања на линеарниот конгруентален алгоритам. Едно негово обопштување е во насока на нелинеарни конгруентални генератори (на пример, $I_{j+1} = a \cdot I_j^2 + b \cdot I_j + c$) или нивни соодветни комбинации.

Б.1.2. Тестирање на генераторите*

Како би можеле да го испитаме квалитетот на еден генератор на случајни броеви?

Како прво, периодата на генераторот секако треба да биде поголема од потребниот број случајни броеви за конкретната симулација. Пе-

риодата за даден генератор обично се знае или може едноставно програмски да се најде.

Како второ, би можело да се провери дали генераторот дава случајни броеви со рамномерна распределба. На пример, ако генерираме случајни броеви во [0, 1) интервал, би можеле него да го поделиме на одреден број еднакви подинтервали и тогаш бројот на генерираните случајни броеви во секој од подинтервалите би требало да е приближно еднаков. Истиот тест би можел да се направи во повеќе димензии, бидејќи многу генератори тука започнуваат да покажуваат слабости.

На рамномерната распределеност е фокусиран и еден од базичните тестови за генераторите на случајни броеви што ја испитува флуктуацијата на генерираните броеви околу просекот користејќи, на пример, χ^2 тест. Нека рангот на случајните броеви ([0, 1) или 0, 1, …, $2^{32} - 1$) е поделен на m еднакви дисјунктни области. Сега ако генерираме n случајни броеви, очекуваниот број броеви E_i во областа i е $E_i = n/m$. Ако пак бројот на генерираните вредности што паднале во областа i е x_i, тогаш χ^2 статистиката е дадена со

$$\chi^2 = \sum_{i=1}^{m} \frac{(x_i - E_i)^2}{E_i}.$$

За доволно големо m ($m \geq 30$), треба веројатностите $\chi^2 > m - 2/3$ и $\chi^2 < m - 2/3$ да бидат по 0.5 (еднаковеројатни настани). Овој тест би можел да се направи со пресметка на вредностите на χ^2 поголем број пати (се разбира за различни низи случајни броеви) и проверка дали просекот е $m - 2/3$. Исто така, може да се предвидуваат веројатностите за различни точки и провери колку често χ^2 ја надминува точката. На пример, за $m = 50$, χ^2 треба да надмине $1.52m$ најмногу 1% пати.

Сигурно еден од најважните тестови би бил да се провери евентуалната корелација меѓу генерираните броеви, но тоа е тешко да се направи. Во основа, покомплексните корелациони тестови се емпириски и се надвор од интересот на оваа книга. Тука единствено би можеле да го споменеме автокорелациониот тест кој ја испитува корелацијата на "куси растојанија". Ако со X_i ја означиме низата случајни променливи чиишто вредности се дадени со низата генерирани случајни броеви x_i, автокорелационата функција се дефинира со

$$\rho(k) = \frac{K_{X_{i+k}X_i}}{DX_i} = \frac{E(X_{i+k} - \mu)(X_i - \mu)}{\sigma^2} = \frac{E(X_{i+k}X_i) - \mu^2}{EX_i^2 - \mu^2} \approx$$

$$\approx \frac{\frac{1}{n}\sum_{i=1}^{n} x_{i+k}x_i - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

$\rho(k)$ треба да биде 0 кога $k \to \infty$.

Некој тука би поставил прашање за тоа каков или кој генератор на случајни броеви да користи. Се разбира, за поедноставни апликации сосема е доволен стандардниот генератор (обично се нарекува rand()) и неговите варијации, што стандардно го има во програмските јазици. Тој најчесто е имплементиран како Park-Miller-ов минимален генератор. За покомплексни симулации, каде што е потребен голем број случајни броеви, веројатно е подобро да се користи некој попрофесионален генератор, како на пример RAN2 (http://www.nr.com) или Mersenne twister (http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html). Подолу се дадени функциите за генерирање случајни броеви превземени директно од random.h на програмскиот јазик Це.

```
//
// Mersenne Twister random number generator:
//
void mt_init_genrand(unsigned long s);
void mt_init_by_array(unsigned long init_key[], int key_length);
/* Generates 32 bit random */
unsigned long genrand_int32(void);
/* Generates 31 bit random */
long genrand_int31(void);
/* Generates a random number on [0,1]-real-interval */
double genrand_real1(void);
/* generates a random number on [0,1)-real-interval */
double genrand_real2(void);
/* generates a random number on (0,1)-real-interval */
double genrand_real3(void);
/* generates a random number on [0,1) with 53-bit resolution */
double genrand_res53(void);
```

Како што се гледа, првата функција е за задавање на семе, втората за задавање семиња за генерирање во повеќе димензии. Другите 6 функции генерираат случајни броеви. Се разбира, сите овие функции се многу помоќни генератори на случајни броеви од функцијата rand(void) (декларирана е во stdlib.h).

Б.1.3. Генерирање нерамномерни случајни броеви*

Генераторите на случајни броеви стандардно се дизајнираат така да генерираат случајни броеви со рамномерна распределба. Причината за тоа е јасна, случајните броеви со која било друга распределба речиси секогаш се генерираат преку случајни броеви со рамномерна распределба во [0, 1] интервал. Генерално, податоците од експериментите доаѓаат во многу други распределби како што е нормалната, експоненцијалната, гама итн. Постојат два основни метода за генерирање случајни броеви со друга распределба од рамномерната. Првиот од нив е аналитички метод што не секогаш е применлив, а вториот е Вон Ноимановиот (John von Neumann, 1903 – 1957) нумерички метод на отфрлање.

Аналитички метод. Треба да генерираме случајни броеви распределени според некоја произволна функција (густина на распределба) $f(x)$, за која се разбира важи

$$f(x) > 0, \text{ за секој } x \quad \text{и} \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

Соодветната финкција на распределба $F(x)$ (е монотона) има инверзна $F^{-1}(y)$,

$$y = F(x) \quad \Leftrightarrow \quad x = F^{-1}(y).$$

Генерирањето на случаен број r распределен како $f(x)$ може да се направи во два чекора,

1) Генерирај рамномерно распределен случаен број u и

2) Стави $r = F^{-1}(u)$.

За да покажеме дека r е навистина случаен број со бараната распределба, доволно е да покажеме дека распределбата $F_1(x)$ на r е токму $F(x)$. Ќе тргнеме од функцијата на распределба на r

$F_1(x) = p(r < x)$, и сега од 2) следува

$F_1(x) = p(F^{-1}(u) < x)$.

Понатаму, ако ја примениме $F(x)$ (монотона) од двете страни на неравенството добиваме

$$F_1(x) = p(F(F^{-1}(u)) < F(x)) = p(u < F(x)).$$

Имајќи предвид дека u е со рамномерна распределба за која едноставно важи $p(x < b) = b$, добиваме

$$F_1(x) = p(u < F(x)) = F(x).$$

ПРИМЕР Б.1 Генерирај случајни броеви со експоненцијална распределба

$$f(x)=\begin{cases} e^{-x}, \text{ за } x\ge 0 \\ 0, \text{ во спротивно} \end{cases}.$$

Решение

Функцијата на распределба е $F(x)=\int_0^x f(t)dt=\int_0^x e^{-t}dt=1-e^{-x}$. Оттука добиваме

$$y=1-e^{-x} \;\Rightarrow\; x=-\log(1-y), \text{ т.е. } F^{-1}(u)=-\log(1-u).$$

На крај, бидејќи u е случаен број од интервалот [0, 1], таков е и $1-u$, па генерирањето на случајните броеви со експоненцијална распределба може да се редуцира на

$$r=F^{-1}(u)=-\log(u)$$

каде што u е случаен број со рамномерна распределба. ■

Истата идеја може да се примени и за дискретен случај. Имено, нека е даден дискретниот закон на распределба

x_1	x_2	…	x_n	…
p_1	p_2	…	p_n	…

Функцијата на распределба е дискретна (скалеста) со вредности што се добиваат од сумите

$$F_k=\sum_{i=1}^{k} p_i\,,\; k=1, 2, \ldots, n\,,$$

при што дополнително ставаме $F_0=0$.

Генерирањето на случаен број r распределен според горниот закон на распределба, повторно може да се направи во два чекора, сосема аналогни на непрекинатиот случај,

1) Генерирај рамномерно распределен случаен број u во [0, 1] и

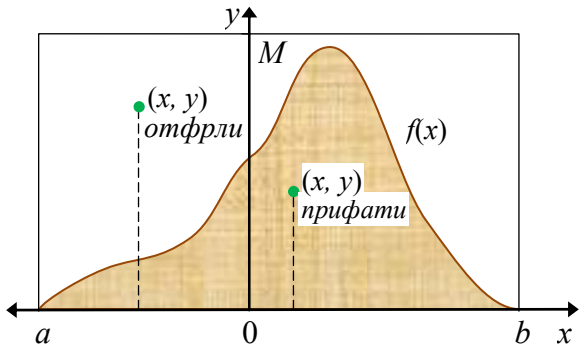
2) Најди го r за кој важи $F_{r-1}<u\le F_r$.

Тука r е цел случаен број со веројатност на генерирање пропорционална на p_r.

Забележи дека во непрекинатиот случај, кога не е возможно експлицитно да се најде инверзната функција $F^{-1}(x)$, една можност е таа да се апроксимира табеларно со дискретна распределба и потоа примени горната метода.

Вон Нойманов метод на отфрлање. Овој метод работи за секоја конечно-вредносна густина на распределба на конечен интервал, независно од тоа дали функцијата може да се интегрира или на соодветната функција на распределба да се најде инверзната. Идејата на овој метод е "помеханичка", т.е. посоодветна за компјутерска реализација.

Нека е дадена густината на распределба $f(x)$ на некој интервал $[a, b]$. Нека M биде избран број таков што $f(x) \le M$ за секој x во интервалот (види сл. Б.1).



Слика Б.1 *Вон Нойманов метод на отфрлање*

Постапката за генерирање случаен број со распределба $f(x)$ со методот на отфрлање е крајно едноставна:

1) Генерирај рамномерно распределен случаен број x во $[a, b]$;

2) Генерирај рамномерно распределен случаен број y во $[0, M]$;

3) Ако $y > f(x)$ отфрли го генерирањето и врати се на 1);

4) Во спротивно, прифати го генерирањето и стави $r = x$.

Така, случајниот број r има густина на распределба $f(x)$. Забележи дека y служи само за контрола дали случајниот број x да се прифати или отфрли. Тука сé изгледа убаво и лесно. Единствен проблем е што многу од генерираните броеви се отфрлаат, т.е. генерирањето оди во празно. Веројатноста генерираниот број да биде прифатен е геометриската веројатност

$$p(\text{"прифати"}) = \frac{\int_a^b f(x)dx}{M(b-a)} = \frac{1}{M(b-a)}.$$

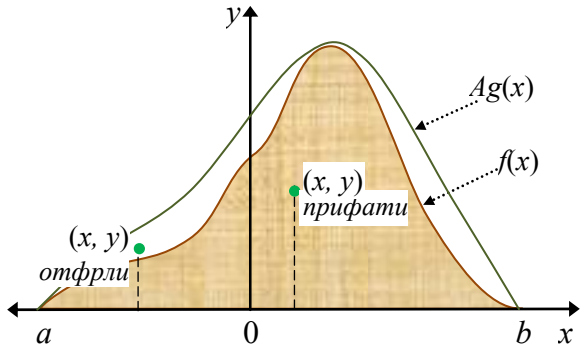
За функција со висок пик и долга ниска опашка, оваа веројатност е мала и бројот на отфрлања станува голем. Сепак, гледано од аспект на пер-

формансите, ваквиот пристап во многу случаи е веројатно поприфатлив, бидејќи не користи ниту функција на распределба ниту нејзина инверзна функција.

За да се намали бројот на отфрлања (промашувања) би можеле наместо во "правоаголник", функцијата $f(x)$ да ја "сместиме" во друга функција што е доволно блиска до $f(x)$, и така да ги минимизираме отфрлањата. Попрецизно, ние би можеле да најдеме функција (густина на распределба) $g(x)$ и константа A таква што да важи

$f(x) \le Ag(x)$, за сите $x \in [a, b]$.

Тука константата A е неопходна за да може $Ag(x)$ да ја надмине $f(x)$, бидејќи и двете се густини на распределба (види сл. Б.2).



Слика Б.2 *Комбиниран метод на генерирање случаен број со распределба $f(x)$*

За сето ова да биде корисно, мора да може да се најде инверзната функција на распределба $G^{-1}(x)$ на густината $g(x)$. Тогаш постапката би се реализира со следните чекори:

1) Генерирај рамномерно распределен случаен број u во [0, 1];
2) Генерирај случаен број x со распределба $g(x)$: $x = G^{-1}(u)$;
3) Генерирај рамномерно распределен случаен број y во $[0, Ag(x)]$;
4) Ако $y > f(x)$ отфрли го генерирањето и врати се на 1);
5) Во спротивно, прифати го генерирањето и стави $r = x$.

Да забележиме дека тука не генерираме случаен број во $[a, b]$, така што границите може да бидат и бесконечни. Тоа е и главната предност на ваквиот покомплициран (комбиниран) метод во однос на едноставниот метод на отфрлања, кој за генерирање на случаен број со дадената густина на распределба бара конечен интервал $[a, b]$.

Б.1.4. Генерирање случајни броеви со нормална распределба*

Нормалната распределба $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, се разбира, е најраспространета распределба и се користи во огромен број апликации при што постојано се јавува потреба од генерирање случајни броеви распределени како $f(x)$. Како што знаеме, $f(x)$ не е интеграбилна, што го прави невозможно праволиниското користење на методите со инверзна функција (аналитички или комбиниран). Дополнително, $f(x)$ е со домен од $-\infty$ до ∞ што исто така е одреден проблем за методот на отфрлање.

За решавање на проблемот со нормалната распределба обично се користи мал трик на премин од една во две димензии. За поедноставно, наместо $Z(\mu, \sigma^2)$ ќе работиме со $Z(0,1)$, т.е.

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.$$

Интегралот на $f(x)$ може да се најде со премин во две димензии

$$\left(\int_{-\infty}^{\infty}e^{-\frac{1}{2}x^2}dx\right)^2=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-\frac{1}{2}x^2}e^{-\frac{1}{2}y^2}dxdy=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(x^2+y^2)}dxdy=$$

и сега со смената во поларни координати $\rho^2=x^2+y^2$, $dxdy=\rho d\rho d\varphi$

$$=\int_0^{\infty}\int_0^{2\pi}e^{-\frac{1}{2}\rho^2}\rho d\rho d\varphi=2\pi\int_0^{\infty}e^{-\frac{1}{2}\rho^2}d\left(\frac{1}{2}\rho^2\right)=2\pi.$$

Методот на Бокс-Милер [Box-Muller 1958] го користи овој трик и тргнува од нормалната распределба во две димензии

$$f(x,y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}=\frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)},$$

што во поларни координати (јакобијанот на трансформацијата е ρ) е

$$g(\rho,\varphi)=f(x,y)\rho=\frac{1}{2\pi}\rho e^{-\frac{1}{2}\rho^2}=\left(\frac{1}{2\pi}\right)\left(\rho e^{-\frac{1}{2}\rho^2}\right)=g_{\varphi}(\varphi)g_{\rho}(\rho).$$

Така, ако генерираме одделно случајни броеви за g_ρ и g_φ ние автоматски генерираме заедничка дводимензионална нормална распределба, т.е. два нормално распределени случајни броја.

Генерирање случајни броеви за g_φ е тривијално бидејќи само треба да се генерира рамномерно распределен случаен број во [0, 1] и помножи со 2π.

Генерирањето случајни броеви за g_ρ може да се направи аналитички со инверзна функција бидејќи соодветната функција на распределба

$$G_\rho(\rho) = \int_0^\infty \rho e^{-\frac{1}{2}\rho^2} d\rho = 1 - e^{-\frac{1}{2}\rho^2},$$ инвертирано дава

$$G_\rho^{-1}(u) = \sqrt{-2\ln(1-u)}.$$

Така може да се генерираат ρ и φ, и од нив x и y според поларната смена $\begin{cases} x = \rho\cos\varphi \\ y = \rho\sin\varphi \end{cases}$. Постапката ја сумираме во следните чекори:

1) Генерирај рамномерно распределен случаен број u_1 во [0, 1];
2) Стави $\varphi = 2\pi u_1$;
3) Генерирај рамномерно распределен случаен број u_2 во [0, 1];
4) Стави $\rho = \sqrt{-2\ln(1-u_2)}$, т.е. $\rho = \sqrt{-2\ln u_2}$;
5) Случајните броеви x и y се добиваат од $\begin{cases} x = \rho\cos\varphi \\ y = \rho\sin\varphi \end{cases}$.

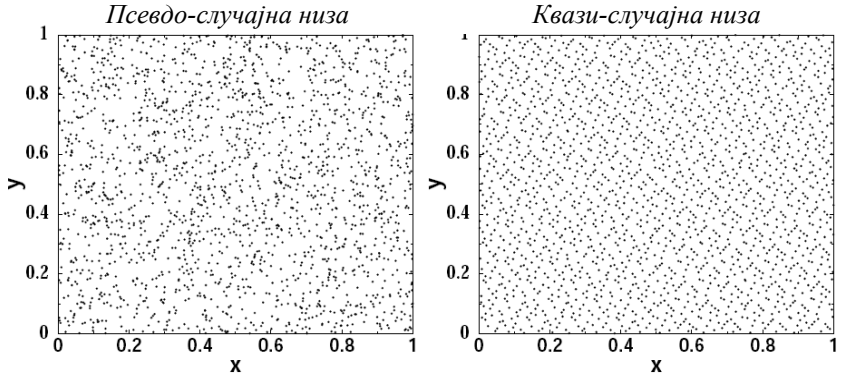
Како што се гледа, постапката генерира два нормално распределени броја, $z_0 = \sqrt{-2\ln u_2}\cos(2\pi u_1)$ и $z_1 = \sqrt{-2\ln u_2}\sin(2\pi u_1)$. Во практична реализација, згодно е чекорите 1) – 5) да се извршуваат во два премина, за генерирање на x, па y. Во првиот премин се зема x и запомнува y, а во вториот премин само се зема y без правење какви било пресметки.

Оваа постапка изгледа максимално ефикасна бидејќи нема отфрлања или некакви проблеми од таков вид. Сепак, од аспект на компјутерска ефикасност, таа не е така добра поради постојаната пресметка на функциите $\sqrt{\ }$, log, sin и cos. Нивната пресметка е многу побавна од пресметките со стандардните математички операции. Оттука повторно изгледа дека постапката со отфрлања е речиси секогаш добра алтернатива, се разбира, кога може директно да се примени.

Б.1.5. Користење на квази-случајни броеви*

Квази-случајните броеви не се случајни броеви. Интуитивно, тие во главно се користат за порамномерно "исполнување на просторот" (во

однос на псевдо-случајните броеви), кога бројот на симулации е лимитиран. Ваквата ситуација е прикажана на сл. Б.3.



Слика Б.3 *Псевдо-случајни и квази-случајни броеви*

Идејата е генерираните броеви да се аранжираат така што тие да бидат што порамномерно распределени без разлика на нивниот број. Покажано е дека конвергенцијата при решавањето интеграли со случајни броеви (Монте Карло) е

а) пропорционална со $1/\sqrt{n}$ за псевдо-случајните броеви и

б) пропорционална со $1/n$ за квази-случајните броеви.

Наједноставен начин да се дојде до квази-случајни броеви е да се генерираат псевдо-случајни броеви по области (stratified sampling). Тоа значи областа на генерирање на случајните броеви да се подели на повеќе подобласти и потоа да се генерира еднаков број случајни броеви во секоја подобласт.

Постојат многу методи на генерирање квази-случајни броеви без потреба од дефинирање подобласти. За разлика од вистинските случајни броеви, ваквите шеми генерираат високо корелирани "случајни" броеви на начин што тие доста рамномерно ја исполнуваат областа на генерирање.

На пример, во еден метод [Richtmeyer 1951], за да се дојде до такво множество вектори во k-димензионален простор, компонентите на векторите

$(x_{i1}, x_{i2}, \ldots, x_{ik})$ се генерерираат според $x_{ij} = i\sqrt{P_j} \pmod 1$,

каде што (mod 1) значи земање само на децималниот дел, а P_j е j–тиот прост број. Овој едноставен концепт има мал недостаток во тоа што мора да се чува листа на првите k прости броеви.

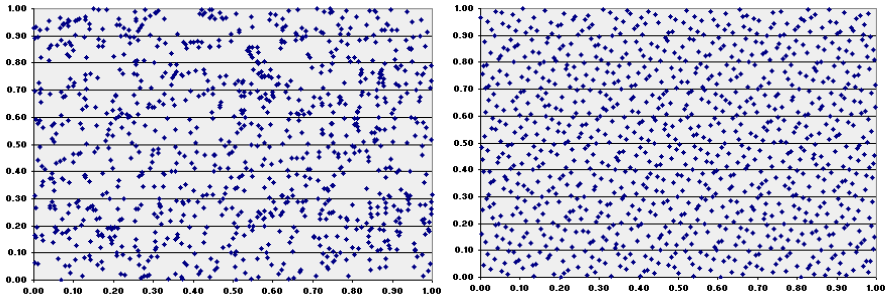
Друг метод (дури од 1935), првично дизајниран за 1-на димензија се базира на следните чекори:

1) Земи цел број и запиши го во некој броен систем (на пример, во бинарен);
2) Сврти го редоследот на цифрите;
3) Стави децимална точка на почеток и интерпретирај го бројот како децимален.

За бинарен систем би имале:

i	i_2	$.i_2$	Децимално
1	1	0.1	0.5
2	10	0.01	0.25
3	11	0.11	0.75
4	100	0.001	0.125
5	101	0.101	0.625
6	110	0.011	0.375

Ова е базична квази-случајна 1-димензионална низа. Таа има различни проширувања за повеќе димензии. На пример Халтоновата низа [Halton 1960], користи броен систем со различна основа за секоја димензија. За првата димензија основата е 2 (како во табелата), за втората основата е 3, за третата основата е 5, итн. Повисока база значи подолг циклус и повеќе пресметки.



Слика Б.4 *Псевдо-случајни и квази-случајни броеви по Халтон* (*бази 2 и 3*)

На крај би го споменале Соболовиот метод [Sobol 1967] за генерирање квази-случајни броеви што за сите димензии користи база 2. Од тој аспект тој е доста ефикасен, но од друга страна тој користи комплицирана постапка на преуредување на броевите по различни димензии. Квази-случајните броеви прикажани на сл. Б.3 се генерирани со Соболовиот метод. Деталите и алгоритмите за Соболовиот метод може да се најдат на многу места, на пример [Galanti, Jung 1997].

Во 40-тите години на минатиот век кога биле дизајнирани првите генератори на случајни броеви, Вон Ноиман искажал мислење дека секој оној што користи аритметички методи за генерирање случајни броеви е на погрешен пат. Денеска, кога генераторите се значително софистицирани и нашироко се користат во речиси сите дисциплини, ова мислење за среќа се покажало како невистинито.

Б.2. Примери на веројатносни симулации

Во ова поглавје ќе разгледаме 7 примери на симулации за пресметка на веројатности. Примерите опфаќаат веројатносни проблеми што вообичаено се решаваат со различни техники: класична веројатност, условна веројатност, Баесова формула, серии независни експерименти, како и геометриска веројатност. Симулациите се дадени преку алгоритамски кодови при што сите користат функција *rand*() за генерирање децимални случајни броеви во [0,1) интервал. Резултатите од работата на алгоритмите (од симулациите) се дадени во табели што содржат пресметани веројатности за различен број симулации. За добивање на овие резултати, алгоритмите се реализирани во Це со користење на Mersenne Twister генераторот на случајни броеви. Користените функции се:

а) за задавање на семето е користен остаток од тековното време

```
seed = (unsigned)time(NULL)%900000000;
mt_init_genrand(seed);
```

б) за генерирање на случаен број од [0, 1) интервал е користена

```
genrand_real2();
```

чиишто декларации се во random.h.

ПРИМЕР Б.2 Колкава е веројатноста во група од k луѓе барем двајца да имаат ист роденден?

Решение

Овој пример веќе е решен (пример 2.12). Нека

$\Omega = \{(x_1, x_2, ..., x_k) \mid x_i = 1, 2, ..., 356\}$, $n = |\Omega| = 356^k$ и сега ако

A = "барем 2-ца имаат ист роденден" тогаш

$\bar{A}$ = "сите имаат различен роденден", и тоа дава

$$p(A) = 1 - p(\bar{A}) = 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - k + 1)}{365^k}.$$

Да се потсетиме дека првата вредност на k за која веројатноста е поголема од 50% е $k = 23$, а веројатноста е 0.5073.

Да ги провериме резултатите со симулации. Алгоритамот е практично праволиниски.

```
read n, k
// ns – број на симулации, pos – број на поволни случаи
ns ← pos ← 0
while ns ≤ n do
    ns ← ns + 1
    // испитуваме k – луѓе
    for j ← 1 to k
        do
            r_j ← [rand() · 365 + 1]  // [...] е цел дел
            i ← 1
            while i ≤ j and r_i ≠ r_j do i ← i + 1
            if i < j then
                pos ← pos + 1
                break // крај на for циклусот
write pos / ns
```

Во овој алгоритам едноставно генерираме k цели случајни броеви r_j од 1 до 365, и потоа проверуваме дали меѓу нив има два исти. За поефикасно, при секое генерирање на случајниот број r_j, веднаш проверуваме дали тој се наоѓа во дотогаш генерираните, и ако е така веднаш одиме на следна симулација (со **break**). Резултатите од симулацијата за 23 луѓе (k = 23) се дадени во табелата (десно). ■

Симулации	*Веројатност*
10^3	0.5010000000
10^4	0.5027000000
10^5	0.5089300000
10^6	0.5077700000
10^8	0.5072939700
10^9	0.5072801230
10^{10}	0.5073039373

ПРИМЕР Б.3 На шаховска табла на случаен начин се стават две дами. Колкава е веројатноста дека тие ќе се напаѓаат?

Решение

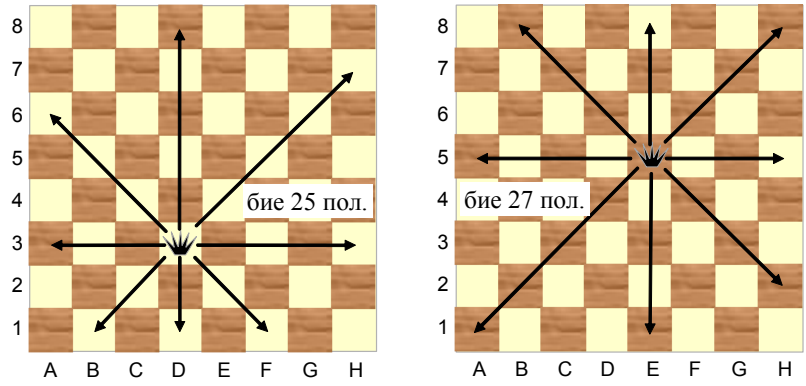
Дама поставена на некое од 28-те рабни полиња бие 21 поле, од некој од 20-те полиња до рабните бие 23 полиња, на следниот "квадрат" од 12 полиња бие 25 полиња и од средните 4 полиња бие 27 полиња (види сл. Б.5). Оттука, веројатноста втората дама да се најде на удар на првата се добива на класичен начин. Сите можни случаи се

$$\Omega = \{(x, y), x, y = 1, 2, \ldots, 8 \text{ и } x \neq y\}, \quad n = |\Omega| = 64 \cdot 63 = 4032,$$

додека поволните случаи се добиваат кога ќе се соберат сите биени полиња од сите позиции на едната дама

$m = 28 \cdot 21 + 20 \cdot 23 + 12 \cdot 25 + 4 \cdot 27 = 1456.$

Значи p("2 дами да се напаѓаат") = m/n = 1456/4032 = 0.3611111111.



Слика Б.5 *Биени полиња од дама на шаховска табла*

Со симулација, решението би можело да се добие со следниот алгоритам, каде што *ns* е бројот на симулации, а *pos* бројот на "позитивните" симулации (дамите се напаѓаат). Позицијата на првата дама е ($r1$, $c1$), а на втората ($r2$, $c2$).

```
read n
ns ← pos ← 0
while ns ≤ n
    do { ns ← ns + 1
         // редица и колона на првата дама
         r1 ← [rand() · 8 + 1] // [...] е цел дел
         c1 ← [rand() · 8 + 1]
         // редица и колона на втората дама
         r2 ← [rand() · 8 + 1]
         c2 ← [rand() · 8 + 1]
         if r1 = r2 and c1 = c2
            then ns ← ns − 1 // се повторува симулацијата
            else if r1 = r2 or c1 = c2 or abs(r1 − r2) = abs(c1 − c2)
                    then pos ← pos + 1 }
write pos / ns
```

Симулации	*Веројатност*
10^3	0.3600000000
10^4	0.3572000000
10^5	0.3623700000
10^6	0.3620470000
10^8	0.3611181400
10^9	0.3611070030
10^{10}	0.3611154798

Како што се гледа од алгоритамот, ако дамите "паднат" во исто поле, симулацијата се повторува. Неиоходни се 4 генерирања на цели случајни броеви од 1 до 8. Дамите се напаѓаат ако се во иста редица, колона

или дијагонала. Резултатите од симулацијата се прикажани во табелата погоре-десно. По 10^{10} симулации, резултатот до 5-тата децимала се совпаѓа со вистинскиот пресметан со класичната веројатност. ■

ПРИМЕР Б.4 Во една кутија има 3 бели и 4 црни топчиња, а во друга 5 бели и 3 црни топчиња. Случајно се одбираат 2 топчиња од првиот сад и 1 од вториот сад и без гледање се стават во трета кутија. Од третата кутија потоа се влече топче. Колкава е веројатноста тоа да е бело?

Решение

B = "извлечено црно топче од 3 кутија", $p(\mathrm{B}) = ?$

A_1 = "третата кутија има 3 бели топчиња"
A_2 = "третата кутија има 2 бели и 1 црно топче"
A_3 = "третата кутија има 1 бели и 3 црни топчиња"
A_4 = "третата кутија има 0 бели топчиња"

$$p(\mathrm{B}) = p(\mathrm{A}_1)p(\mathrm{B} \mid \mathrm{A}_1) + p(\mathrm{A}_2)p(\mathrm{B} \mid \mathrm{A}_2) + p(\mathrm{A}_3)p(\mathrm{B} \mid \mathrm{A}_3) =$$

$$= \frac{3}{\binom{7}{2}}\frac{5}{8}1 + \left(\frac{3\cdot 4}{\binom{7}{2}}\frac{5}{8} + \frac{3}{\binom{7}{2}}\frac{3}{8}\right)\frac{2}{3} + \left(\frac{3\cdot 4}{\binom{7}{2}}\frac{3}{8} + \frac{\binom{4}{2}}{\binom{7}{2}}\frac{5}{8}\right)\frac{1}{3} =$$

$$= \frac{15}{168} + \frac{69}{168}\frac{2}{3} + \frac{66}{168}\frac{1}{3} = \frac{249}{504} = 0.4940476190.$$

Решението со симулација праволиниски ја следи активноста опишана во проблемот.

```
read n
ns ← pos ← 0
a3 ← a2 ← a1 ← b5 ← b4 ← b3 ← b2 ← b1 ← 1   // ai – прва, bi – втора
a7 ← a6 ← a5 ← a4 ← b8 ← b7 ← b6 ← 0   // кутија; 0 – црно, 1 – бело
while ns < n do
    ns ← ns + 1
    // j, k – од прва кутија, l – од втора кутија
    j ← [rand() · 7 + 1]
    k ← [rand() · 7 + 1]
    l ← [rand() · 8 + 1]
    // ставаме во трета кутија
    c1 ← aj; c2 ← ak; c3 ← bl
    m ← [rand() · 3 + 1]
    if cm = 1 then pos ← pos + 1
write pos / ns
```

На случаен начин се земаат две топчиња од првата кутија (индексите j и k) и едно топче од втората кутија (индексот l). Тие се стават во третата кутија (низата c од 3 елементи), од која потоа се влече едно топче (дадено со индексот m). Овој алгоритам убаво го опишува духот на симулациите. Тој едноставно ги извршува сите активности што ние без користење на компјутер би морале рачно да ги извршиме. ■

Симулации	***Веројатност***
10^3	0.5160000000
10^4	0.4976000000
10^5	0.4950000000
10^6	0.4945000000
10^8	0.4940130200
10^9	0.4940440940
10^{10}	0.4940481121

ПРИМЕР Б.5 Во една специјализирана болница се лекуваат болни и тоа: 50% од болест А, 30% од болест Б и 20% од болест В. Шансите за излекување се: 0.7, 0.8 и 0.9 за болестите А, Б и В соодветно. Еден излекуван пациент ја напуштил болницата. Колкава е веројатноста дека тој боледувал од болеста А?

Решение

Ставаме X = "болниот е излекуван",

Y_1 = "пациентот ја има болеста А", $p(Y_1 \mid X) = ?$
Y_2 = " пациентот ја има болеста Б",
Y_3 = " пациентот ја има болеста В",

$p(Y_1) = 0.5, \quad p(X \mid Y_1) = 0.7,$
$p(Y_2) = 0.3, \quad p(X \mid Y_2) = 0.8,$
$p(Y_3) = 0.2, \quad p(X \mid Y_2) = 0.9.$

Со директна примена на формулата на Баес добиваме

$p(Y_1 \mid X) = p(X \mid Y_1)\, p(Y_1) / p(X)$, и имајќи предвид дека

$p(X) = p(Y_1)p(X \mid Y_1) + p(Y_2)p(X \mid Y_2) + p(Y_3)p(X \mid Y_3) =$

$= 0.5 \cdot 0.7 + 0.3 \cdot 0.8 + 0.2 \cdot 0.9 = 0.77$, што дава

$p(Y_1 \mid X) = 0.7 \cdot 0.5 / 0.77 = 0.4545454545.$

Со симулација, решението би можело да се добие со следниот (даден подолу), не така тривијален, алгоритам. И тука *ns* е бројот на симулации, а *pos* бројот на "позитивните" симулации. Алгоритамот работи во 3 гранки, по една (еден **if**) за секоја болест, и во секоја ги "прифаќа" само излечените.

```
read n
ns ← pos ← 0
while ns < n do {
    ns ← ns + 1
    // j – болест, k – излекување
    j ← [rand() · 100 + 1]
    k ← [rand() · 100 + 1]
    // за секоја болест испитуваме излекување
    if j <= 50 then
        if k <= 70 then pos ← pos + 1
                   else ns ← ns – 1 // се повторува сим.
    if j > 50 and j <= 80 then
        if k > 80 then ns ← ns – 1 // се повторува сим.
    if j > 80 then
        if k > 90 then ns ← ns – 1 // се повторува сим.
write pos / ns
```

Главниот трик во овој алгоритам е што ако пациентот со која било болест не е "во излечените", симулацијата се повторува. Бараната веројатност е условна, па симулацијата едноставно ги брои пациентите со болест А (такви се 50%), од сите пациенти коишто се излечени. Ако пациентот не е излечен, симулацијата не важи и оди одново ($ns \leftarrow ns - 1$). За $ns = 10^{10}$, точноста е до 4-тата децимала. ■

Симулации	*Веројатност*
10^3	0.4530000000
10^4	0.4529000000
10^5	0.4543800000
10^6	0.4551240000
10^8	0.4545686900
10^9	0.4545704780
10^{10}	0.4545503702

ПРИМЕР Б.6 Веројатноста за изработка на предмет што задоволува стандарди во еден процес на производство е 0.96. Еден упростен систем за контрола на квалитет дава позитивен резултат со веројатност 0.98 за предметите што го задоволуваат стандардот и 0.05 за предметите што не одговараат на стандардите. Колкава е веројатноста предметот што во оваа контрола е прогласен за стандарден, навистина ги задоволува стандардите?

Решение

Ставаме

B = "предметот ги задоволува стандардите",

A = "позитивен резултат од контролата на квалитет"

$p(\mathrm{B} \mid \mathrm{A}) = ?$

$p(\text{B}) = 0.96,\ p(\text{A} \mid \text{B}) = 0.98,\ p(\text{A} \mid \overline{\text{B}}) = 0.05.$

Од формулата на Баес добиваме

$p(\text{B} \mid \text{A}) = p(\text{A} \mid \text{B})\, p(\text{B}) / p(\text{A})$, и имајќи предвид дека

$p(\text{A}) = p(\text{B})p(\text{A} \mid \text{B}) + p(\overline{\text{B}})p(\text{A} \mid \overline{\text{B}}) = 0.96 \cdot 0.98 + 0.04 \cdot 0.05 = 0.9428,$

добиваме $p(\text{B} \mid \text{A}) = 0.98 \cdot 0.96 / 0.9428 = 0.9978786593.$

Со симулација, решението може да се добие со сосема сличен алгоритам на оној од претходниот пример. Се разбира и тука *ns* е бројот на симулации, а *pos* бројот на "позитивните" симулации.

```
read n
ns ← pos ← 0
while ns < n do
    ns ← ns + 1
    // j – стандарден, k – прогласен стандарден
    j ← [rand() · 100 + 1]
    k ← [rand() · 100 + 1]
    if j <= 96 then
        if k <= 98 then pos ← pos + 1
                   else ns ← ns – 1 // се повторува сим.
    if j > 96 then
        if k > 5 then ns ← ns – 1 // се повторува сим.
write pos / ns
```

Повторно бараната веројатност е условна, па симулацијата едноставно ги брои предметите што го задоволуваат стандардот (такви се 96%), од сите предмети што се прогласени за стандардни од контролата на квалитетот. Резултатите од симулацијата се подобри од оние од претходниот пример бидејќи самиот алгоритам е поедноставен. За $ns = 10^{10}$, точноста оди до 6-тата децимала. ■

Симулации	*Веројатност*
10^3	0.9990000000
10^4	0.9978000000
10^5	0.9981200000
10^6	0.9978170000
10^8	0.9978767100
10^9	0.9978790780
10^{10}	0.9978784782

ПРИМЕР Б.7 Да го разгледај го Бертрандовиот парадокс од примерот 2.18. Направи симулации на трите решенија и провери дали симулациите даваат соодветни резултати.

1) За првото решение "метод на случајни крајни точки" случајно бираме две точки од кружницата $(1, \alpha)$ и $(1, \beta)$ во поларни координати и тогаш должината на тетивата $d = \sqrt{2 - 2\cos(\alpha + \beta)}$.

```
read n
ns ← pos ← 0
while ns < n do
    ns ← ns + 1
    // a, b – случајни броеви
    a ← 2π · rand() // во [0,2π]
    b ← 2π · rand() // во [0,2π]
    d ← √(2 − 2cos(a + b))
    if d > √3 then pos ← pos + 1
write pos / ns
```

Симулации	Веројатност
10^8	0.3332936900
10^9	0.3333099550
10^{10}	0.3333359768

Резултатите од симулацијата (види ја табелата) одговараат на резултатот 1/3 што ова решение го дава. Тука симулацијата можеше да се упрости со земање 0 за едниот агол.

2) За второто решение "метод на случаен радиус" случајно бираме една точка во областа на кружницата (ρ, α) во поларни координати на некој радиус (во декартови е ($\rho\cos\alpha$, $\rho\sin\alpha$), и тогаш должината на тетивата е $d = 2\sqrt{1-\rho^2}$. Симулацијата е праволиниска и ја оставаме на читателот.

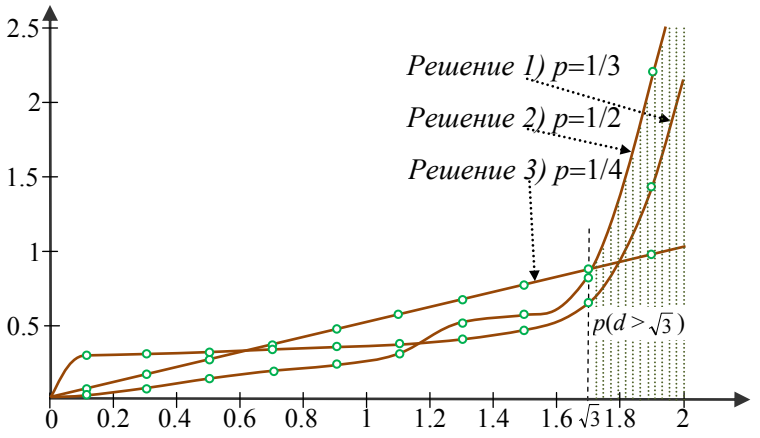
3) За третото решение "метод на случајна средна точка", случајно бираме една точка со декартови координати (x, y) во областа на кружницата, и тогаш должината на тетивата е $d = 2\sqrt{1-(x^2+y^2)}$. Симулацијата е едноставна.

```
read n
ns ← pos ← 0
while ns < n do
    ns ← ns + 1
    // x, y – случајни броеви
    x ← 2 · rand() − 1 // во [−1,1]
    y ← 2 · rand() − 1 // во [−1,1]
    d ← x² + y²
    // d – дали сме во кружницата
    if d ≤ 1 then
        d ← 2√(1 − d)
        if d > √3 then pos ← pos + 1
    else ns ← ns − 1 // симулацијата се повторува
write pos / ns
```

Симулации	Веројатност
10^8	0.2499466900
10^9	0.2499749110
10^{10}	0.2499939488

Резултатите од симулацијата одговараат на резултатот 1/4 што ова решение го дава. Забележи дека тука мора да се направи проверка дали точката е во областа на кружницата.

На сл. Б.6 се дадени приближните густини на распредлба на должините на тетивите за трите методи на избирање на случајна тетива. За таа цел конструиран е хистограм користејќи 10^{10} симулации на случајно генерирање тетиви по трите методи. Што може да се согледа од овие густини?



Слика Б.6 *Густини на распределба на должините на тетивите за трите класични решенија на Бертрановиот парадокс*

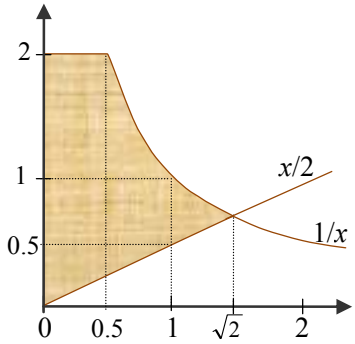
Прво што се забележува е дека распределбата на должините на тетивите во сите методи е различна и оттука веројатностите во врска со должините на тетивите се различни. Како второ, сите густини се растечки одејќи од најкусите кон најдолгите тетиви (максималната можна должина е 2). Тоа значи дека бројот на тетиви расте со приближување на нивните средини кон центарот на кружницата. Ова е најизразено кај второто решение каде што дури 50% од тетивите се со должина поголема од $\sqrt{3}$ (исенчениот дел на сликата). Како трето, решението 2) најсилно го одразува фактот што огромен број тетиви се со средини блиски на центарот на кружницата и во него. Тоа значи дека "густината" на радиусите блиску до центарот е далеку најголема. Интересно е дека кај решението 3) истата густина (на средините на тетивите) се зголемува линеарно. На крај, кога некој "површно" би ја погледнал сл. Б.6 веднаш би можел да заклучи дека најлогична е распределбата на должините на тетивите од решението 3) (линеарно се зголемува со приближувањето кон центарот). Интересно е дека ова решение во сите анализи се смета за погрешно, а главното внимание се посветува на решението 2), коешто има најкомплицирана густина на распределба на должините на тетивите. ■

ПРИМЕР Б.8 Случајно се бираат два позитивни децимални броја помали од 2. Да се определи веројатноста дека нивниот производ е помал од 1, а количникот помал од 2.

Решение

Ако со x и y ги означиме бараните броеви, имаме дека

$\Omega = \{(x, y) \mid x, y \in [0, 2]\}$, $m(\Omega) = 4$, додека



$S = \{(x, y) \mid x \cdot y \le 1, x/y \le 2\}$, и оттука (види ја сликата лево)

$$m(S) = \int_0^{0.5} (2 - x/2)dx + \int_{0.5}^{\sqrt{2}} (1/x - x/2)dx .$$

Значи бараната веројатност $p(S)$ е

$$p(S) = \frac{1}{8}(1 + 3\ln 2) = 0.3849301927.$$

Симулацијата се сведува на решавање определен интеграл.

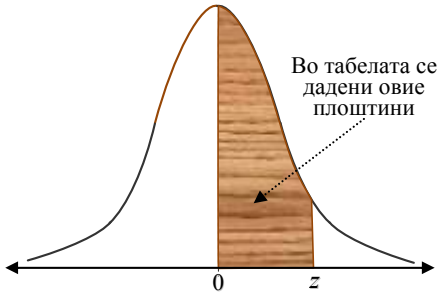
```
read n
ns ← pos ← 0
while ns < n do
    ns ← ns + 1
    // (x, y) – точка во [0,2]×[0,2]
    x ← 2rand()
    y ← 2rand()
    // дали е во областа на функциите
    if x < 0.5 and y > x/2 and y < 2 or
       x > 0.5 and x < √2 and y > x/2 and y < 1/x
       then pos ← pos + 1
write pos / ns
```

Секоја симулација за решавање определен интеграл (пресметување плоштина на област) има слична постапка. Генерираме случајно точки околу областа (кај нас е [0,2]×[0,2]) и само броиме колку од нив се внатре во областа. Инаку областа чијашто плоштина ни треба е дефинирана со функции. Како што гледаме од табелата, по извршени 10^{10} симулации точноста оди до 5-тата децимала. ■

Симулации	*Веројатност*
10^3	0.3830000000
10^4	0.3825000000
10^5	0.3841100000
10^6	0.3842260000
10^8	0.3850217300
10^9	0.3849339560
10^{10}	0.3849357559

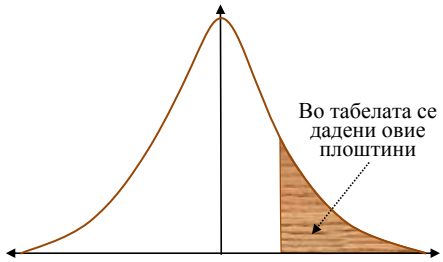
Табели на распределби

Табела на стандардна нормална распределба



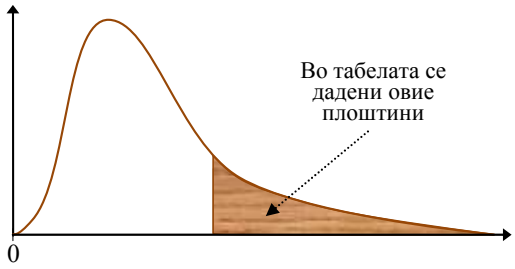
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Табела на студентова распределба



df	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	3.07768	6.31375	12.70620	31.82052	63.65674	127.32134	318.30884	636.61925
2	1.88562	2.91999	4.30265	6.96456	9.92484	14.08905	22.32712	31.59905
3	1.63774	2.35336	3.18245	4.54070	5.84091	7.45332	10.21453	12.92398
4	1.53321	2.13185	2.77645	3.74695	4.60409	5.59757	7.17318	8.61030
5	1.47588	2.01505	2.57058	3.36493	4.03214	4.77334	5.89343	6.86883
6	1.43976	1.94318	2.44691	3.14267	3.70743	4.31683	5.20763	5.95882
7	1.41492	1.89458	2.36462	2.99795	3.49948	4.02934	4.78529	5.40788
8	1.39682	1.85955	2.30600	2.89646	3.35539	3.83252	4.50079	5.04131
9	1.38303	1.83311	2.26216	2.82144	3.24984	3.68966	4.29681	4.78091
10	1.37218	1.81246	2.22814	2.76377	3.16927	3.58141	4.14370	4.58689
11	1.36343	1.79588	2.20099	2.71808	3.10581	3.49661	4.02470	4.43698
12	1.35622	1.78229	2.17881	2.68100	3.05454	3.42844	3.92963	4.31779
13	1.35017	1.77093	2.16037	2.65031	3.01228	3.37247	3.85198	4.22083
14	1.34503	1.76131	2.14479	2.62449	2.97684	3.32570	3.78739	4.14045
15	1.34061	1.75305	2.13145	2.60248	2.94671	3.28604	3.73283	4.07277
16	1.33676	1.74588	2.11991	2.58349	2.92078	3.25199	3.68615	4.01500
17	1.33338	1.73961	2.10982	2.56693	2.89823	3.22245	3.64577	3.96513
18	1.33039	1.73406	2.10092	2.55238	2.87844	3.19657	3.61048	3.92165
19	1.32773	1.72913	2.09302	2.53948	2.86093	3.17372	3.57940	3.88341
20	1.32534	1.72472	2.08596	2.52798	2.84534	3.15340	3.55181	3.84952
21	1.32319	1.72074	2.07961	2.51765	2.83136	3.13521	3.52715	3.81928
22	1.32124	1.71714	2.07387	2.50832	2.81876	3.11882	3.50499	3.79213
23	1.31946	1.71387	2.06866	2.49987	2.80734	3.10400	3.48496	3.76763
24	1.31784	1.71088	2.06390	2.49216	2.79694	3.09051	3.46678	3.74540
25	1.31635	1.70814	2.05954	2.48511	2.78744	3.07820	3.45019	3.72514
26	1.31497	1.70562	2.05553	2.47863	2.77871	3.06691	3.43500	3.70661
27	1.31370	1.70329	2.05183	2.47266	2.77068	3.05652	3.42103	3.68959
28	1.31253	1.70113	2.04841	2.46714	2.76326	3.04693	3.40816	3.67391
29	1.31143	1.69913	2.04523	2.46202	2.75639	3.03805	3.39624	3.65941
30	1.31042	1.69726	2.04227	2.45726	2.75000	3.02980	3.38518	3.64596
40	1.30308	1.68385	2.02108	2.42326	2.70446	2.97117	3.30688	3.55097
50	1.29871	1.67591	2.00856	2.40327	2.67779	2.93696	3.26141	3.49601
60	1.29582	1.67065	2.00030	2.39012	2.66028	2.91455	3.23171	3.46020
70	1.29376	1.66691	1.99444	2.38081	2.64790	2.89873	3.21079	3.43501
80	1.29222	1.66412	1.99006	2.37387	2.63869	2.88697	3.19526	3.41634
90	1.29103	1.66196	1.98667	2.36850	2.63157	2.87788	3.18327	3.40194
100	1.29007	1.66023	1.98397	2.36422	2.62589	2.87065	3.17374	3.39049
500	1.28325	1.64791	1.96472	2.33383	2.58570	2.81955	3.10661	3.31009
1000	1.28240	1.64638	1.96234	2.33008	2.58075	2.81328	3.09840	3.30028
∞	1.28155	1.64485	1.95996	2.32635	2.57583	2.80703	3.09023	3.29053

Табела на хи-квадрат распределба



df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	40.25602	43.77297	46.97924	50.89218	53.67196
40	20.7065	22.1643	24.4331	26.5093	29.0505	51.8050	55.7585	59.3417	63.6907	66.7659
50	27.9907	29.7076	32.3574	34.7642	37.6886	63.1671	67.5048	71.4202	76.1539	79.4900
60	35.5346	37.4848	40.4817	43.1879	46.4589	74.3970	79.0819	83.2976	88.5794	91.9517
70	43.2752	45.4418	48.7576	51.7393	55.3290	85.5271	90.5312	95.0231	100.425	104.2148
80	51.1720	53,5400	57.1532	60.3915	64.2778	96.5782	101.8794	106.6285	112.3287	116.3210
90	59.1963	61.7541	65.6466	69.1260	73.2912	107.5650	113.1452	118.1358	124.1163	128.2989
100	67.3276	70.0648	74.2219	77.9295	82.3581	118.4980	124.3421	129.5611	135.8067	140.1694

Решенија на задачите

Глава 2

1. а) точно; б) точно; в) неточно; г) точно; д) неточно; ѓ) неточно.
2. а) $A\overline{B}\overline{C}$; б) $AB\overline{C}$; в) ABC; г) $A+B+C$; д) $A\overline{B}\overline{C}+\overline{A}B\overline{C}+\overline{A}\overline{B}C$; ѓ) $\overline{A}\overline{B}\overline{C}$; е) $(A+B+C)-ABC$.
3. а) $A+B$; б) $A\overline{B}+\overline{A}B$;
4. а) 0.4; б) $p(\overline{A}B)=p(B)-p(AB)=0.1$;
   в) 0.6; г) $p(\overline{A}+B)=p(\overline{A})+p(B)-p(\overline{A}B)=0.8$.
5. а) 0.9; б) 0.0236; в) 0.0156.
6. 10^{-8}
7. а) $\dfrac{1}{10^3\cdot 26^2}$; б) $\dfrac{1}{3\cdot 10^2\cdot 2\cdot 26}$; в) $\dfrac{1}{10^2\cdot 26}$.
8. 0.36.
9. 5/9.
10. 99/126.
11. Општо, p("има комплетни точно k пара") $= \dfrac{\binom{n}{k}\binom{n-k}{r-k}2^{r-k}}{\binom{2n}{2r}}$.
12. а) 15/16; б) 2/3; в) 1.
13. а) 0.3568; б) 0.4625.
14. 0.8828

    Нека A, B, C се настаните дека вработениот е член на 1-та, 2-та, 3-та комисија соодветно. Тогаш бараната веројатност е $p(A+B+C)=p(A)+p(B)+p(C)-p(A)p(B)-p(A)p(C)-p(B)p(C)+p(A)p(B)p(C)$.

15. $p(\mathrm{A}) + p(\mathrm{B}) + p(\mathrm{C}) - 2p(\mathrm{AB}) - 2p(\mathrm{AC}) - 2p(\mathrm{BC}) + 3p(\mathrm{ABC})$

    p("точно еден од A,B,C") = p(A–AB–AC + B–AB–BC + C–AC–BC) =
    = p(A–AB–AC) + p(B–AB–BC) + p(C–AC–BC) = итн.

16. 0.5436

    Веројатноста фаталниот исход да се случи во n-тиот обид е $\left(\frac{5}{6}\right)^{n-1}\frac{1}{6}$.

    Бараната веројатност е $\sum_{n=0}^{\infty}\left(\frac{5}{6}\right)^{2n}\frac{1}{6}=\frac{1}{6}\sum_{n=0}^{\infty}\left(\frac{25}{36}\right)^{n}=\frac{1}{6}\frac{1}{1-\frac{25}{36}}$.

17. а) 1/3; б) 1/2.

18. а) $\big(1-(1-p(\mathrm{A})(1-p(\mathrm{C}))\big)\big(1-(1-p(\mathrm{B})(1-p(\mathrm{D}))\big) = 0.8788$; б) 0.9434.

19. $1 - 2p^2 - 2p^3 + 5p^4 - 2p^5$

    Веројатноста на една конфигурација на патиштата со блокирани k од нив е $p^k(1-p)^{5-k}$. Треба да се избројат сите конфигурации за кои има (или нема) пат од A до D.

20. 0.2645

    Нека A_1, A_2, A_3, A_4 се настаните дека 5-те карти не содржат каро, херц, треф и пик соодветно. Тогаш бараната веројатност е $1 - p(\mathrm{A}_1 + \mathrm{A}_2 + \mathrm{A}_3 + \mathrm{A}_4) = 1 - \big(p(\mathrm{A}_1) + p(\mathrm{A}_2) + p(\mathrm{A}_3) + p(\mathrm{A}_4) - p(\mathrm{A}_1\mathrm{A}_2) - \ldots - p(\mathrm{A}_3\mathrm{A}_4) + p(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3) + \ldots + p(\mathrm{A}_2\mathrm{A}_3\mathrm{A}_4) - p(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3\mathrm{A}_4)\big)$. Внимателно, настаните не се независни, на пример $p(\mathrm{A}_i\mathrm{A}_j) \neq p(\mathrm{A}_i)p(\mathrm{A}_j)$.

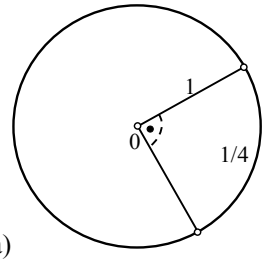
21. 0.1207.

22. $a^2(1 - \pi/4)$

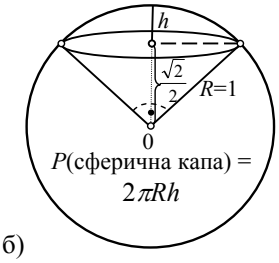
23. а) 1/2; б) да.

24. а) $\frac{(a-2r)^2}{a^2}$; б) $1-\frac{4r^2}{a^2}$.

25. а) 1/4 (слика а)); б) $(1 - \sqrt{2}/2)/2$ (слика б)).

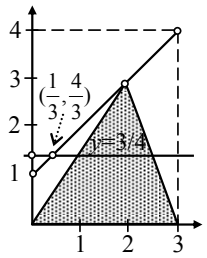


Слика а)



Слика б)

26. 8/9 (види слика)



Плоштината на триаголникот треба да биде

$3 \cdot h/2 > 2 \Rightarrow h > 4/3$.

Правата $y = x + 1$ се сече со $y = 4/3$ во (1/3, 4/3).

Оваа точка ја дели отсечката во однос 1:8.

Поволните случаи се 8 од 9 делови на отсечката.

Глава 3

1. Во случај на дисјунктност: а) точно; б) неточно; в) точно; г) неточно;
   Во случај на независност: г) неточно; а) точно; б) неточно; в) точно;

2. а) 0.1333; б) 0.2000; в) 0.1667

   A = "различен број", P(A) = 30 / 36

   B = "сумата е 6", AB = { (1,5), (5,1), (2,4), (4,2) }
   C = "сумата е 7", AC = { (1,6), (6,1), (2,5), (5,2), (3,4), (4,3) }
   D = "првата коцка е 6-ка", AD = { (6,1), (6,2), (6,3), (6,4), (6,5) }

   а) p(B | A) = (4 / 36) / (30 / 36) = 0.1333 < p(B) = 0.1389
   б) p(C | A) = (6 / 36) / (30 / 36) = 0.2000 > p(C) = 0.1667
   в) p(D | A) = (5 / 36) / (30 / 36) = 0.1667 = p(D) = 0.1667.

3. а) 0.4388/0.6962 = 0.6303; б) 0.1097/0.25 = 04388.

4. 0.6923

   Сите парови прогнози се (Т=точна, П=погрешна): (Т, Т), (Т, П), (П, Т), (П, П), со веројатности 0.72, 0.18, 0.08 и 0.02 соодветно. Бараната веројатност е p((Т, П) | ((Т, П) + (П, Т))) .

5. а) 0.0642; б) 0.7692.

6. а) 0.3500; б) 0.1225; в) 0.6500.

7. а) 0.0080; б) 0.00008; в) 0.9800; г) Да, Не.

8. 2/3.

9. 0.9468

   Ова е проблем на колекција: n објекти случајно распоредуваме на k места. Нека A_i е настанот дека i-тото место останало празно. Тогаш

   $p(A_i) = \left(\frac{k-1}{k}\right)^n$, $p(A_iA_j) = \left(\frac{k-2}{k}\right)^n$, итн. Веројатноста дека барем едно место ќе остане празно е (со принцип на вклучување-исклучување)

$$p\left(\sum_{j=1}^{k} \mathrm{A}_j\right) = k\left(\frac{k-1}{k}\right)^n - \binom{k}{2}\left(\frac{k-2}{k}\right)^n + \ldots + (-1)^k \binom{k}{k-1}\left(\frac{1}{k}\right)^n .$$

За k=4, n=15, $p\left(\sum_{j=1}^{4} \mathrm{A}_j\right) = 4\left(\frac{3}{4}\right)^{15} - 6\left(\frac{1}{2}\right)^{15} + 4\left(\frac{1}{4}\right)^{15} = 0.0532.$

10. а) 0.351; б)) 0.917; в) 0.2500.

11. а) 0.002; б) 0.086; в) 0.4904.

12. а) 0.0225; б) 0.1250.

13. 1/506;

14. а) 0.9596 (види подолу); б) 0.9950

p("ис.0") = 0.8, p("ис.1") = 0.1,
p("пр.1 | ис.0") = 0.01 $\Rightarrow p$("пр.0 | ис.0") = 0.99,
p("пр.0 | ис.1") = 0.05 $\Rightarrow p$("пр.1 | ис.1") = 0.95,
p("ис.0 | пр.0") = p("пр.0 | ис.0")p("ис.0") / p("пр.0"), каде што
p("пр.0") = p("пр.0 | ис.0")p("ис.0") + p("пр.0 | ис.1")p("ис.1")

15. 0.5135

16. а) 0.9847; б) 0.1184.

17. 74, Се решава неравенството по n: $1 - 0.98 = \binom{n}{0} \cdot 0.052^0 \cdot 0.948^n$.

18. а) 0.00003; б) 0.00024; в) 0.00107.

19. а) 0.5177; б) 0.4914

p("барем една 6-ка од 4 фрлања") =

= 1– p("ниедна 6-ка од 4 фрлања") = 1– $\binom{4}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4$

p("барем еднаш 2-е 6-ки од 24 фрлања") =

= 1– p("ниеднаш 2-е 6-ки од 24 фрлања") = $\binom{24}{0}\left(\frac{1}{36}\right)^0\left(\frac{35}{36}\right)^{24}$.

20. 0.0000000017326

Геометриските веројатности стрелецот да погоди во кружните прстени се 1/16, 3/16, 5/16 и 7/16. За стрелецот да собере најмалку 90 поени од 10 пукања постојат 5 можности: 9×10 + 1×6 + 0×4 + 0×1, 9×10 + 0×6 + 1×4 + 0×1, 9×10 + 0×6 + 0×4 + 1×1, 8×10 + 2×6 + 0×3 + 0×1 и 8×10 + 1×6 + 1×4 + 0×1. Веројатностите на овие настани се:

$\frac{10!}{9!\cdot 1!}\left(\frac{1}{16}\right)^9\left(\frac{3}{16}\right)^1$, $\frac{10!}{9!\cdot 1!}\left(\frac{1}{16}\right)^9\left(\frac{5}{16}\right)^1$, $\frac{10!}{9!\cdot 1!}\left(\frac{1}{16}\right)^9\left(\frac{7}{16}\right)^1$, $\frac{10!}{8!\cdot 2!}\left(\frac{1}{16}\right)^8\left(\frac{3}{16}\right)^2$ и

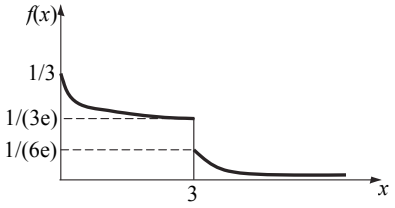
$\frac{10!}{8!\cdot 1!\cdot 1!}\left(\frac{1}{16}\right)^8\left(\frac{3}{16}\right)^1\left(\frac{5}{16}\right)^1$. Нивниот збир го дава решението.

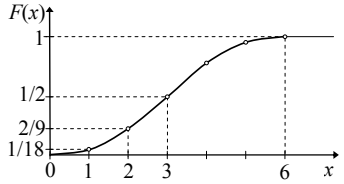
Глава 4

1. а) 0.9939; б) 0.0007; в) 0.3188.
2. $p(X=0)=0.00001$; $p(X=1)=0.00167$; $p(X=2)=0.07663$; $p(X=3)=$ $=0.92169$.
3. а) 0.9; б) 0.7; в) 0.1; г) 0.8.
4. $p(X=n)=(n-1)p^2(1-p)^{n-2}$, $n=2, 3, \ldots$.
5. а) $p(X>2)=1-p(X\le 2)=1-F(2)=1-(1-e^{-2/3})=e^{-2/3}$;

   б) $p(2<X\le 6)=F(6)-F(2)=(1-e^{-2}/2)-(1-e^{-2/3})=e^{-2/3}-e^{-2}/2$;

   в)





6. а) 1/3; б) в) 1/18, 1/2.
7. $p(x>12.6)=\int_{12.6}^{\infty}20e^{-20(x-12.5)}dx=-e^{-20(x-12.5)}\Big/_{12.6}^{\infty}=0.1353$;

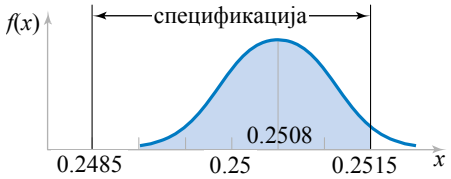
   $p(12.5<x<12.6)=\int_{12.5}^{12.6}20e^{-20(x-12.5)}dx=-e^{-20(x-12.5)}\Big/_{12.5}^{12.6}=0.8647$.
8. 2/3.
9. а) 0.25; б) 0.2140.
10. $3/a$.
11. а) 0.0498; б) 0.8775.
12. а) 0.0025; б) 0.6321.
13. $p(0.2485<X<0.2515)=p(\frac{0.2485-0.2508}{0.0005}<z<\frac{0.2515-0.2508}{0.0005})=$

    $=p(-4.6<z<1.4)=p(z<1.4)-p(z<-4.6)=0.91924$;

Кога процесот би се подобрил така што просекот да биде еднаков на целната вредност 0.25, би имале

$$p(0.2485 < X < 0.2515) = p(\frac{0.2485 - 0.2500}{0.0005} < z < \frac{0.2515 - 0.2500}{0.0005}) =$$

$$= p(-3 < z < 3) = p(z < 3) - p(z < -3) = 0.9973\,.$$



14. а) 0.06681; б) 0.86638; в) 0.000214.

15. а) 0.0062; б) 0.0124; в) 5.33 милиметри.

16. $p(X \le 950) = \sum_{x=0}^{950} \frac{e^{-1000}x^{1000}}{x!} \approx p(z \le \frac{950 - 1000}{\sqrt{1000}}) = p(z \le -1.58) = 0.057\,.$

17. б) 330; в) 0.0089.

18. а) 0.983; б) 0.45.

19. Случајната променлива $T = \frac{X - \mu}{S}\sqrt{n} = -2.9059$ (S е стандардна девијација) има студентова распределба со 20 – 1 = 19 степени на слобода. Од статистички алатки (или таблици) читаме $p(T > -2.9059) = 0.996$.

20. а) 0.275; б) 0.685.

21. а) 0.5; б) 0.35; в) 0.5; г) 0.8; д) 0.30; ѓ) 0.4; е) 0.4;
ж) $p(X = 1 \mid Y = 1, Z = 2) = 0.4$; $p(X = 2 \mid Y = 1, Z = 2) = 0.6$;

22. а) $\{(X, Y, Z), 0 \le X, Y, Z \le 4$ и $X + Y + Z = 4\}$;
б) $p(X = 0 \mid Y = 2) = 0.3333$, $p(X = 1 \mid Y = 2) = 0.5332$, $p(X = 2 \mid Y = 2) = 0.1335$, $p(X = 3 \mid Y = 2) = p(X = 4 \mid Y = 2) = 0$;
в) 0.1758, 0.2198, 0.4761, 0.1335.

23. а)

Y \ X	0	1	2	3	4
0	0.1244	0.2618	0.1964	0.0621	0.0070
1	0.0873	0.1354	0.0666	0.0104	0
2	0.0203	0.0207	0.0050	0	0
3	0.0018	0.0009	0	0	0
4	0.0001	0	0	0	0

б) $f_X(0) = 0.2338$, $f_X(1) = 0.4188$, $f_X(2) = 0.2679$, $f_X(3) = 0.0725$, $f_X(4) = 0.0070$;
в) $f_{Y|3}(0) = 0.857$, $f_{Y|3}(1) = 0.143$; г) не.

24. а) $f_{X,Y}(x, y) > 0$ и $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = \int_0^{\infty}\left(\int_x^{\infty} 6 \cdot 10^{-6} e^{-0.001x - 0.002y} dy\right)dx =$

$$= 6\cdot 10^{-6}\int_0^{\infty}\left(\int_x^{\infty} e^{-0.002y}dy\right)e^{-0.001x}dx = 6\cdot 10^{-6}\int_0^{\infty}\left(\frac{e^{-0.002x}}{0.002}\right)e^{-0.001x}dx =$$

$$= 0.003\int_0^{\infty} e^{-0.003x}dx = 0.003\,\frac{1}{0.003} = 1;$$

б) $p(X<1000, Y<2000) = 6\cdot 10^{-6}\int_0^{1000}\left(\int_x^{2000} e^{-0.002y}dy\right)e^{-0.001x}dx = 0.915;$

в) $p(Y>2000) = 6\cdot 10^{-6}\int_0^{2000}\left(\int_{2000}^{\infty} e^{-0.002y}dy\right)e^{-0.001x}dx +$

$$+ 6\cdot 10^{-6}\int_{2000}^{\infty}\left(\int_x^{\infty} e^{-0.002y}dy\right)e^{-0.001x}dx = 0.05$$ или

$f_Y(y) = 6\cdot 10^{-6}e^{-0.002y}\int_0^y e^{-0.001x}dx = 6\cdot 10^{-6}e^{-0.002y}(1-e^{-0.001y})$, за y > 0

$p(Y>2000) = 6\cdot 10^{-6}\int_{2000}^{\infty} e^{-0.002y}(1-e^{-0.001y})dy = 0.05$

г) $f_X(x) = 6\cdot 10^{-6}\int_x^{\infty} e^{-0.001x-0.002y}dy = 0.003e^{-0.003x}$, за $x>0$, и сега

$$p(Y>2000 \mid X=1500) = \int_{2000}^{\infty}\frac{6\cdot 10^{-6}e^{-0.001x-0.002y}}{0.003e^{-0.003x}}dy = 0.368;$$

д) не.

25. а) 0.0439, 0.0019; б) 0.065.

26. а) 0.032; б) 0.0267.

Глава 5

1. а) $3.91\cdot 10^{-19}$; б) 200; в) $2.56\cdot 10^{-18}$.

2. а) 3000; б) 1431.18.

3. а) 0.3679; б) 50.51.

4. $EX = \int_{12.5}^{\infty} x20e^{-20(x-12.5)}dx = -xe^{-20(x-12.5)} - \frac{1}{20}e^{-20(x-12.5)}\Big|_{12.5}^{\infty} = 12.55;$

$DX = \int_{12.5}^{\infty}(x-12.55)^2 20e^{-20(x-12.5)}dx = 0.0025.$

5. $EX = \int_0^{\infty} x(2e^{-2x})dx = \frac{1}{2}$ минути;

$EX^2 = \int_0^{\infty} x^2(2e^{-2x})dx = \frac{1}{2}, \Rightarrow DX = EX^2 - (EX)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$ минути.

6. а) $EX = 109.39$ микрометри, $DX = 33.19$ микрометри; б) 54.70 евра.

7. $EX = 1.85$ минути, $DX = 0.0408$ минути.

8. $(1 - \bar{u})\lambda,\ 1/\lambda$.

9. Нека X = "тежина на патиките";

   а) $p(X > 13) = p\left(\frac{X-12}{0.5} > \frac{13-12}{0.5}\right) = p\ (z > 2) = 0.228$;

   б) $p\,(z < \frac{13-12}{\sigma}) \geq 0.999 \Rightarrow \frac{13-12}{\sigma} \geq 3.72 \Rightarrow \sigma \leq 0.2688$;

   в) $p\,(z < \frac{13-\mu}{0.5}) \geq 0.999 \Rightarrow \frac{13-\mu}{0.5} \geq 3.72 \Rightarrow \mu \leq 11.14$.

10. а) 500 000; б) 223607.

11. а) 443.11; б) 53650.5; в) 0.2212.

12. $EX = 1.067$, $DX = 0.6146$.

13. а) $EX = 1.2$; б) $E(Y \mid X = 3) = 0.143$.

14. Маргиналната густина на распределба на X е $f_X(x) =$

$$= \int_x^\infty 6 \cdot 10^{-6} e^{-0.001x - 0.002y} dy = 6 \cdot 10^{-6} e^{-0.001x} \left(\frac{e^{-0.002y}}{-0.002} \Big/ {}_x^\infty \right) = 0.003 e^{-0.003x}$$

Условната густина на распределба на $X \mid Y$ е $f_{Y|X}(y) = f_{X,Y}(x, y) / f_X(x) =$

$6 \cdot 10^{-6} e^{(-0.001x - 0.002y)} / 0.003 e^{-0.003x} = 0.002\ e^{-0.002x - 0.002y}$, за $0 < x < y$.

Бараното условно очекување е $E(Y \mid X = 1500) =$

$$\int_{1500}^\infty y(0.002) e^{0.002 \cdot 1500 - 0.002y} dy = 0.002 e^3 \int_{1500}^\infty y e^{-0.002y} dy =$$

$$= 0.002 e^3 \left(\frac{1500}{0.002} e^{-3} + \frac{e^{-3}}{0.002 \cdot 0.002} \right) = 2000 \cdot$$

15. $\int_0^1 \int_0^{1-x} \int_0^{1-x-y} c\,dz\,dy\,dx = c \int_0^1 \int_0^{1-x} (1 - x - y)\,dy\,dx = c \int_0^1 \left(\frac{x^2}{2} - x + \frac{1}{2} \right) dx = \frac{c}{6} \Rightarrow$

$c = 6$, $f_X(x) = \int_0^{1-x} \int_0^{1-x-y} 6\,dy\,dx = 6\left(\frac{x^3}{2} - \frac{x^2}{2} + \frac{x}{2} \right) \Rightarrow$

$EX = 6 \int_0^1 \left(\frac{x^3}{2} - \frac{x^2}{2} + \frac{x}{2} \right) dx = \frac{1}{4}$.

16. Бидејќи коефициентот на корелација $\rho = 0.964549$, одговорот е да.

17. X и Y се независни, па корелацијата е 0.

18. а) $EX = 4.3$, $EY = 82.8$, $DX = 6.21$, $DY = 152.96$, $E(XY) = 382.7 \Rightarrow \rho = 0.8650$;
б) $a = 4.2930$, $b = 64.3401$, $\sigma^2 = 1.5635$;
в) 4.8125.

19. Секоја линеарна комбинација на $X+Y$ и $X-Y$ е линеарна комбинација и на X и Y, па така $(X+Y, X-Y)$ има дводимензионална нормална распределба. Од $K_{X+Y,\,X-Y} = K_{X,X} - K_{X,Y} + K_{X,Y} - K_{Y,Y} = DX - DY = 0$.

20. а) 0.023; б) 4558.

Глава 6

1. Бидејќи е $x > 0$, трансформацијата е еднозначна, т.е. од $y = x^2$ следува $x = \sqrt{y}$, па бараната распределба е $f(y) = p(1-p)^{\sqrt{y}-1}$, $y = 1, 4, 9, \ldots$.

2. Функцијата $y = 2x + 4$ е монотона со инверзна $(y-4)/2$ и јакобијан еднаков на $d(y/2 - 2)/dy = 1/2$. Оттука, распределбата на Y е

$$f(y) = \frac{y-4}{32}, \quad 4 \le y \le 12.$$

3. $$f(y) = \begin{cases} 0, & \text{за } y < -1 \\ (y+1)/9, & \text{за } -1 \le y < 2 \\ (5-y)/9, & \text{за } 2 \le y < 5 \\ 0, & \text{за } y \ge 5 \end{cases}.$$

4. $$f_W(w) = \begin{cases} \dfrac{0.19}{2k\sqrt{w/k}} \left(\dfrac{\sqrt{w/k}}{36.6} \right)^{-7.96} e^{-\left(\frac{\sqrt{w/k}}{36.6} \right)^{-6.96}}, & \text{за } w > 0 \\ 0, & \text{во спротивно} \end{cases}.$$

5. $$f_Y(x) = \begin{cases} \dfrac{2}{\pi\sqrt{1-x^2}}, & \text{за } x \in (0,1) \\ 0, & \text{во спротивно} \end{cases}.$$

6. $$f_Y(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi x}} e^{-x/2}, & \text{за } x > 0 \\ 0, & \text{во спротивно} \end{cases}.$$

7. а) $$f_P(x) = \begin{cases} \dfrac{1}{0.08 r_0 \sqrt{\pi x}}, & \text{за } 4\pi(0.99 r_0)^2 \le x \le 4\pi(1.01 r_0)^2 \\ 0, & \text{во спротивно} \end{cases}$$

б) $f_V(x)=\begin{cases}\dfrac{1}{0.08\pi r_0}\left(\dfrac{3x}{4\pi}\right)^{-2/3}, & \text{за } \dfrac{4}{3}\pi(0.99r_0)^3\le x\le\dfrac{4}{3}\pi(1.01r_0)^3\\ 0, & \text{во спротивно}\end{cases}$

8. Бараните закони на распределба се:

U	11	12	13	14	17	18
$p(U=x_i+y_j)$	0.08	0.32	0.02	0.08	0.10	0.40

V	10	12	16	20	24	32
$p(V=x_i\cdot y_j)$	0.08	0.02	0.10	0.32	0.08	0.40

9. $f(x,y)=\begin{cases}c, & \text{за } x^2+y^2\le r^2\\ 0, & \text{во спротивно}\end{cases}$, и од $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)dxdy=1 \Rightarrow c=\dfrac{1}{r^2\pi}$.

Маргиналните густини се $f_1(x)=f_2(x)=\begin{cases}\dfrac{2\sqrt{r^2-x^2}}{r^2\pi}, & \text{за } |x|\le r\\ 0, & \text{во спротивно}\end{cases}$

Не, поради $f(0,0)=\dfrac{1}{r^2\pi}\ne\dfrac{4}{r^2\pi}=f(0)f(0)$.

10. $f(t)=\begin{cases}(a_1+a_2+\ldots+a_n)e^{-(a_1+a_2+\ldots+a_n)t}, & \text{за } t>0\\ 0, & \text{во спротивно}\end{cases}$

11. $f_Y(y)=\begin{cases}\dfrac{3y^2}{(1+y)^4}, & \text{за } y\ge 0\\ 0, & \text{во спротивно}\end{cases}$

12. Ако означиме $y_1=h_1(x_1,x_2)=x_1+x_2,\ y_2=h_2(x_1,x_2)=x_1-x_2$, бараната заедничка густина на распределба е

$g(y_1,y_2)=f(h_1^{-1}(y_1,y_2),h_2^{-1}(y_1,y_2))|J|$ при што инверзните функции се

$x_1=h_1^{-1}(y_1,y_2)=\dfrac{y_1+y_2}{2}$, $x_2=h_2^{-1}(y_1,y_2)=\dfrac{y_1-y_2}{2}$, а јакобијанот се парцијалните изводи на h_1^{-1} и h_2^{-1}, $J=\begin{vmatrix}1/2 & 1/2\\ 1/2 & -1/2\end{vmatrix}=-\dfrac{1}{2}$.

Поради независноста на X и Y добиваме

$g(y_1,y_2)=f(h_1^{-1}(y_1,y_2))f(h_2^{-1}(y_1,y_2))|J|=\dfrac{1}{4\pi}e^{-(y_1^2+y_2^2)/4}$.

Глава 7

1. $p(|X-1| \le 0.75) \ge 0.41$ според неравенството Чебишев.
   Инаку $p(|X-1| \le 0.75) = 0.75$.

2. а) $p(80 \le X \le 120) \ge 0$; б) $p(80 \le X \le 120) \ge 5/9$ што е значително подобрена граница.

3. Бројот на предмети X што одговара на стандардите има биномна распределба, со параметри $p = 0.9$ и $n = 18000$. Очекувањето кај биномната распределба е $EX = n \cdot p = 0.9 \cdot 18000 = 16200$, а дисперзијата е $DX = n \cdot p(1-p) = 1620$. Сега добиваме

   $$p(|X - EX| \le 200) \ge 1 - DX/200^2, \text{ т.е. } p(|X - 16200| \le 200) \ge 0.955.$$

4. Нека $X_1, X_2, \ldots, X_n$ се случајни променливи (независни) – дијаметри на n дупки. Просечниот дијаметар е $E(\frac{1}{n}\sum_{k=1}^{n} X_k) = \mu$, а стандардната девијација е $\sqrt{D(\frac{1}{n}\sum_{k=1}^{n} X_k)} = \sqrt{0.01^2/n}$. Сега неравенството на Чебишев дава

   $$p(|X - EX| < k\sqrt{0.01^2/n}) \ge 1 - \frac{1}{k^2}, \text{ и за } k=3,\ p(|\overline{X} - \mu| < 3\sqrt{0.01^2/n}) \ge \frac{8}{9}.$$

5. 3/4.

6. $\varphi(t) = e^{5it}$, 5, 0; $\quad \phi(t) = \dfrac{2e^{it}}{it}\left(1 - \dfrac{1}{it}\right) + \dfrac{2}{(it)^2}$, $\dfrac{2}{3}$, $\dfrac{1}{18}$.

7. а) $\varphi(t) = \sin(bt)/bt$; б) $\varphi(t) = 1/(1+t^2)$;

8. Не. Тоа е "коцкарска илузија".

9. Да.

10. Конвергенцијата по веројатност ја повлекува конвергенцијата по распределба бидејќи интуитивно распределбата е составена од "специфични" веројатности. Конвергенцијата "скоро сигурно" ја повлекува конвергенцијата по веројатности бидејќи интуитивно таа е конвергенција на множества – настани што повлекува и конвергенција на соодветните веројатности.

11. Со неравенството на Чебишев се добива $n \ge 50000$ (пример 7.4).
    Користејќи ја централната гранична теорема добиваме

    $$p(|Y_n - p| \le 0.01) = p\left(\left|\frac{1}{n}\sum_{j=1}^{n}(\delta_i - E\delta_i)\right| \le 0.01\right) =$$

    $$= p\left(\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\frac{\delta_i - E\delta_i}{\sigma_\delta}\right)\right| \le \frac{0.01\sqrt{n}}{\sigma_\delta}\right) = p\left(|z| \le \frac{0.01\sqrt{n}}{\sigma_\delta}\right).$$

Имајќи предвид дека $\sigma_\delta \le 1/2$, добиваме $p(|E_n - p| \le 0.01) = p(|z| \le 0.02\sqrt{n}) \ge 0.95$. Тоа конечно дава $0.02\sqrt{n} \approx 1.96 \Rightarrow n \approx 9604$, што е многу подобра проценка од онаа добиена со го неравенството на Чебишев.

12. И двете гранични теореми даваат делумно решение на општиот проблем: "Какво е граничното однесување на $S_n = (X_1 + X_2 + ... + X_n)/n$ кога n тежи кон бескрајност?" за произволни случајни променливи X_1, X_2, ... X_n. Според законот на големите броеви S_n конвергира по веројатност или "скоро сигурно" кон вистинската вредност μ.

    Од друга страна, централната гранична теорема тврди дека распределбата на $\sqrt{n}(S_n - \mu)$ конвергира кон нормална распределба $Z(0, \sigma^2)$, што е по-општо тврдење.

Глава 8

1. Се базира на сомнителен концепт на еднакви шанси што сомнително се објаснува преку физичка симетрија. Многу е ограничен. Што со несиметричните исходи на експериментите?
2. Секоја има свои предности, во зависност од реалната ситуација.
3. Веројатностите зависат од оној што го работи моделот. Не овозможува објективна статистичка анализа.
4. Самиот процес на добивање на експерименталните податоци содржи субјективни елементи.
5. Немаат периода и приближно се рамномерно распределени. Понекогаш се користат (во различни варијанти) како генератор на случајни броеви.
6. Одговорите зависат од филозофското убедување на читателот.
7. Податоците ги имаме групно, па работиме со средините на интервалите:

x	$x \le 100$	$100 < x \le 150$	$150 < x \le 200$	$200 < x \le 250$	$250 < x \le 300$
$F_n(x)$	0	1/40	6/40	15/40	21/40

$300 < x \le 350$	$350 < x \le 400$	$400 < x \le 450$	$450 < x \le 700$	$x > 700$
29/40	35/40	37/40	39/40	1

8. Рени смета дека нема циркуларност во дефинирање на веројатноста преку законот на големите броеви.
9. Поради комплицираност на проблемот и развојот на математичката мисла низ историјата.
10. Точно ... .

Глава 9

1. Примерок со големина n се n случајни променливи $X_1, X_2, \ldots, X_n$ или идентично, случаен вектор $(X_1, X_2, \ldots, X_n)$. Реализација на примерок се n конкретни вредности $x_1, x_2, \ldots, x_n$ на случајните променливи $X_1, X_2, \ldots, X_n$ (x_j е една вредност на случајната променлива X_j), т.е обичен вектор на броеви $(x_1, x_2, \ldots, x_n)$.

2. Распределба на примерок е заедничката распределба на случајни променливи $X_1, X_2, \ldots, X_n$ или идентично, распределба на случајниот вектор $(X_1, X_2, \ldots, X_n)$.

3. Кај набљудуваните податоци немаме никакво влијание.

4. $E(X \mid \{\Omega, \varnothing\}) = EX, \quad E(X \mid \mathcal{F}) = X.$

5. Условните распределби на X, за $Y = -1, 0, 1$ се:

X	-1	1	X	-1	1	X	-1	1
$p(X \mid Y=-1)$	1/3	2/3	$p(X \mid Y=0)$	2/3	1/3	$p(X \mid Y=1)$	1/4	3/4

   Очекувањата се:

   $E(X \mid Y=-1) = (-1)(1/3) + 1(2/3) = 1/3$

   $E(X \mid Y=0) = (-1)(2/3) + 1(1/3) = -1/3$

   $E(X \mid Y=1) = (-1)(1/4) + 1(3/4) = 1/2$ и оттука

   $E(X) = 0.3E(X \mid Y=-1) + 0.3E(X \mid Y=0) + 0.4E(X \mid Y=1) = 0.2.$

   Ова е исто со стандардното очекување $EX = -1(0.4) + 1(0.6) = 0.2$.

6. Спецификацијата е претпоставка за статистичкиот модел. Респецификацијата е промена на/во моделот во светло на анализа на податоците од примерокот.

7. Сè што имаме е примерок и знаење од веројатност и статистика. Ако додатно ја имаме (ја добиеме барем приближно) и распределбата, имаме сè што ни треба.

8. Поради проблеми врзани со претпоставките како за обликот на распределбата така за независноста и еднаквата распределеност. Многу претпоставки што може да ја "нарушат" доверливоста на статистичката анализа и добиените резултати.

Глава 10

1. а) $\bar{x} = 7.184$, $s^2 = 0.0004268$, $s = 0.02066$; б) 7.18, 7.20.

2. а) $\bar{x} = 65.85$, $s = 12.16$, 58; б) Цртеж; в) $\bar{x} = 66.86$, $s = 10.74$, 60.

3. а) Стебло-лисја дијаграм:

Стебло	Лисја	Фрек.
8	3	1
8	4 5 5 5	4
8	6 6 7 7	4
8	8 8 8 9 9 9 9 9 9	9
9	0 0 0 0 0 0 0 1 1 1 1 1 1 1	14
9	2 2 2 2 3	5
9	4 5 5	3

Над 50% од оценките се околу 90 со блага предност на 90 и 91. Распределбата на оценките е доста симетрична.

б) $\bar{x} = 89$, $s = 2.8$, 90;

в) 22/40.

4. а) лисја: една цифра

12 2
12 445
12 6667777
12 889999
13 00011111111
13 222222222233333333333
13 44444444444444444555555555555555555
13 6666666666777777777
13 888888888888999999
14 0000001111
14 2333333
14 444
14 77

Поради симетрија;

б) Блиска до ѕвонест облик (нормален) со очекување ≈ 135 и релативно значајна дисперзија.

5. Стеблата се двоцифрени, а лисјата се едноцифрени:

Американски	*Стебла*	Француски
	8	1
157020153504	9	00645632
9324	10	2563
6306	11	6913
058	12	325528
8	13	7
	14	
	15	8
2	16	

Репрезентативни вредности: малку над 100 за американските и малку над 110 за француските. Нешто поголема варијабилност во траењето на француските филмови.

Повеќе екстремни позитивни прирасти кај американските филмови.

Најдолгиот американски филм е 162 минути, додека соодветниот француски е 158 минути.

6. а) Фреквенциите и релативните фреквенции се:

Класа	Фрек.	Рел. фрек.
0	7	0.117
1	12	0.200
2	13	0.217
3	14	0.233
4	6	0.100
5	3	0.050
6	3	0.050
7	1	0.017
8	1	0.017

б) 0.917, 0.867, 1 – 0.867 = 0.133;

в) Хистограмот е значително позитивно закривен и центриран некаде меѓу 2 и 3. Во 39 од 60-те случаи податоците се во интервалот [1, 3].

7. а) Раштрканоста на податоците е доста нерамномерна и податокот 50 е на граница на класа.

б) Фреквенциите и релативните фреквенции се:

Класа	Фрек.	Рел. фрек.
0 - 49	9	0.18
50 - 99	19	0.38
100 - 149	11	0.22
150 - 199	4	0.08
200 - 299	4	0.08
300 - 399	2	0.04
400 - 499	0	0.00
500 - 599	1	0.02

Централната вредност на податоците е некаде околу 100. Постои голема варијабилност во животниот век, посебно кај поголемите податоци. Последните 2-3 интервали би можеле а се спојат во еден.

в) Фреквенциите и релативните фреквенции се:

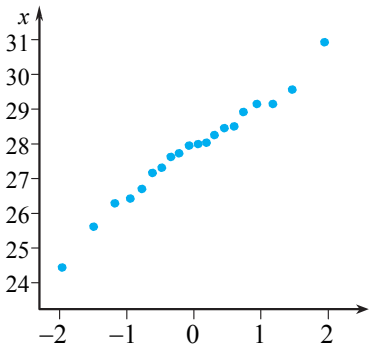
Класа	Фрек.	Рел. фрек.
2.25 - 2.74	2	0.04
2.75 - 3.24	2	0.04
3.25 - 3.74	3	0.06
3.75 - 4.24	8	0.16
4.25 - 4.74	18	0.36
4.75 - 5.24	10	0.20
5.25 - 5.74	4	0.08
5.75 - 6.25	3	0.06

Кај ваквите $\ln(x)$ вредности имаме многу поголема симетрија и помала варијабилност. Исто така има помали празнини низ податоците.

г) 0.38, 0.14.

8. Да.

9.



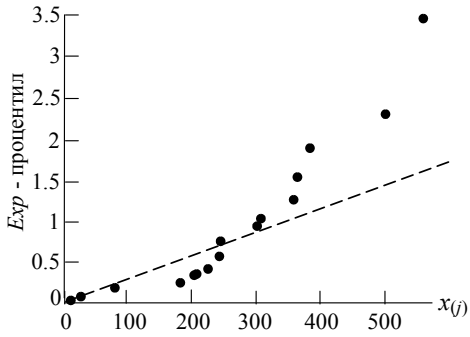
Како што се гледа од сликата, обликот на точките е доста праволиниски што укажува дека има добри шанси распределбата на волтажите на прекин да биде нормална.

10. Да.

11. Податоците со процентили како и соодветните процентили на експоненцијалната распределба ($\lambda e^{-\lambda x}$) за $\lambda = 1$ се дадени во табелата:

$x_{(j)}$	11.6	26.5	82.8	179.7	204.6	212.6	229.9	242.0
$(j - 0.5)/29$	0.031	0.094	0.156	0.219	0.281	0.344	0.406	0.469
exp - процентил	0.031	0.099	0.170	0.247	0.330	0.344	0.421	0.633

$x_{(j)}$	244.8	304.3	307.8	359.5	366.7	379.1	502.5	558.9
$(j - 0.5)/29$	0.531	0.594	0.656	0.719	0.781	0.844	0.906	0.969
exp - процентил	0.757	0.901	1.067	1.269	1.519	1.859	2.364	3.474



Како што се гледа од сликата, обликот на точките не е праволиниски, па не може да се смета дека распределбата е експоненцијална

λ е скалирачки параметар исто како што е σ кај нормалната распределба.

12. а) 5.42, $\sqrt{0.33} = 0.57$, 5.46;

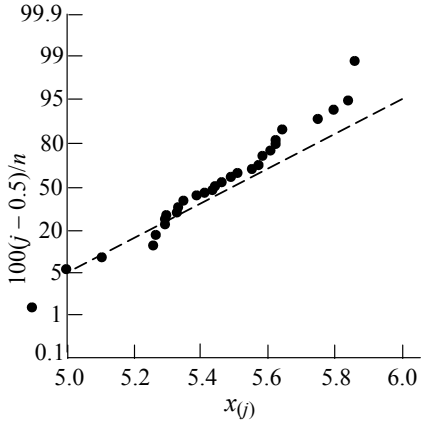
б) Веројатно да, бидејќи податокот 4.07 се чини нереално мал, а тој влегува во просекот но не и во медијаната (денеска знаеме дека вистинската вредноста е приближно 5.52);

в) Табелата на вредностите и соодветните процентили е

$x_{(j)}$	4.07	4.88	5.10	5.26	5.27	5.29	5.29	5.30	5.34	5.34
$(j - 0.5)/29$	0.017	0.052	0.086	0.121	0.155	0.190	0.224	0.259	0.293	0.328

$x_{(j)}$	5.36	5.39	5.42	5.44	5.46	5.47	5.50	5.53	5.55	5.57
$(j - 0.5)/29$	0.362	0.397	0.431	0.466	0.5	0.534	0.569	0.603	0.638	0.672

$x_{(j)}$	5.58	5.61	5.62	5.63	5.65	5.75	5.79	5.85	5.86
$(j - 0.5)/29$	0.707	0.741	0.776	0.810	0.845	0.879	0.914	0.948	0.983



Нормалниот веројатносен дијаграм покажува дека податоците во значителна мера ја следат правата линија. Тоа укажува дека има добри шанси распределбата на мерењата на густината на земјата да е нормална.

Глава 11

1. 1.75, 27.96.
2. Се изведува формула што за кои било вредности на примерокот дава вредност за параметарот. Оценката е случајна променлива бидејќи е функција од примерокот, т.е од случајни променливи.
3. Оценувачот $\hat{\theta}_3$ е најефикасен, додека оценувачот $\hat{\theta}_2$ е најдобро центриран. Изборот меѓу овие две зависи од тоа дали центрираноста е важна за конкретниот проблем.
4. а) Да, $\sigma^2/2$ (грешката е $\sigma/\sqrt{2}$); б) не, $\overline{X}$ има помала дисперзија σ^2/n.
5. $\hat{\theta}_1$ е подобар.
6. Интуитивно, доволните оценувачи се "најпреспективни" бидејќи потенцијално имаат дисперзии што се помали од дисперзиите на сите други (што не зависат од доволна статистика) центрирани оценувачи.
7. $$f(x_1, x_2, \ldots, x_n; \mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-1/(2\sigma^2)\sum(x_k-\mu)^2} =$$
$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-1/(2\sigma^2)\sum(x_k-\overline{x}+\overline{x}-\mu)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}(\sum(x_k-\overline{x})^2+n(\overline{x}-\mu)^2)}$$
Ова може да се факторизира како

$f(x_1,x_2,\ldots,x_n;\mu) = e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-1/(2\sigma^2)\sum(x_k-\bar{x})^2}$, па според теоремата 11.4, $\overline{X}$ е доволна статистика.

8. Го трасира патот кон наоѓање на надобрата центрирана оценка "економизирајќи" го барањето низ можните доволни оценки. За испитување на статистиката на минимална доволност постои релативно лесна постапка што не е случај за "само" доволните статистики.

9. Не е центриран. Кога $n \to \infty$ е центриран.

10. а) Следува од $EX_i = n_i p_i$, за $i = 1, 2$; б) $\sqrt{p_1q_1/n_1 + p_2q_2/n_2}$;
в) Стави $\hat{p}_i = x_i/n_i$ и $\hat{q}_i = 1-\hat{p}_i$ за $i = 1, 2$ во б); г) 0.08, 0.0457.

11. а) $\hat{\theta} = \sum_{j=1}^{n} X_j^2/(2n)$; б) 74.505.

12. а) θ^2/n; б) θ^2/n; в) $\theta(1-\theta)/n$; г) θ/n.

13. $EX^2 = DX - (EX)^2 = \alpha\beta^2(\alpha+1)$ па системот равенки е
$\overline{X} = \alpha\beta$, $(1/n)\sum X_i^2 = \alpha\beta^2(\alpha+1)$.

Решението на системот е $\hat{\alpha} = \dfrac{\overline{X}^2}{\frac{1}{n}\sum X_i^2 - \overline{X}^2}$, $\hat{\beta} = \dfrac{\frac{1}{n}\sum X_i^2 - \overline{X}^2}{\overline{X}}$.

14. а) $1-\frac{1}{\overline{X}}$; б) $1-\sum_{k=1}^{n}\frac{1}{\overline{X}}\left(1-\frac{1}{\overline{X}}\right)^{k-1}$

15. $X_{(1)} = \min(X_i)$, $\overline{X}-1$.

16. $\hat{\lambda} = \dfrac{n}{\sum(X_i - \min(X_i))}$, 0.202.

17. Од $EX = r(1-p)/p$ и $DX = r(1-p)/p^2$ и добиваме дека $EX^2 = DX + (EX)^2 = r(1-p)(\mathrm{r} - rp + 1)/p^2$. Изедначувајќи $EX = \overline{X}$ и $EX^2 = (1/n)\sum X_i^2$ се добиваат оценувачите

$$\hat{p} = \frac{\overline{X}}{\frac{1}{n}\sum X_i^2 - \overline{X}^2} \quad \text{и} \quad \hat{r} = \frac{\overline{X}^2}{\frac{1}{n}\sum X_i^2 - \overline{X}^2 - \overline{X}}.$$

18. Добиваме $\bar{x} = \sum x_i / 420 = (0 \cdot 29 + 1 \cdot 71 + 2 \cdot 82 + \ldots)/420 = 2.98$ и $\sum x_i^2 / 420$ $= (0^2 \cdot 29 + 1^2 \cdot 71 + 2^2 \cdot 82 + \ldots)/420 = 12.40$, и заменуваме во оценувачите

$$\hat{p} = \frac{2.98}{12.40 - 2.98^2} = 0.85 \ , \quad \hat{r} = \frac{2.98^2}{12.40 - 2.98^2 - 2.98} = 16.5 \cdot$$

Се разбира r е секогаш позитивно, додека $\hat{r}$ може да испадне негативно (поради именителот). Ако нешто такво се случи, или негативната биномна распределба е несоодветна или се такви оценките на параметрите.

19. Заедничката распределба за n-те региони е

$$f(x_1, x_2, \ldots, x_n; \lambda) = \frac{(a_{R_1})^{x_1} (a_{R_2})^{x_2} \cdots (a_{R_n})^{x_n} \cdot (\lambda)^{\sum x_i}}{x_1! x_2! \cdots x_n !} e^{-\lambda \cdot \sum a_{R_i}} \text{ , т.е.}$$

$$\ln f(x_1, x_2, \ldots, x_n; \lambda) = \sum x_i \ln(a_{R_i}) + \ln(\lambda) \cdot \sum x_i - \lambda \sum a_{R_i} - \sum \ln(x_i!) \ .$$

Земајќи извод $\frac{\partial}{\partial \lambda} \ln f(x_1, x_2, \ldots, x_n; \lambda)$ и изедначувајќи го на 0 добиваме

$\frac{1}{\lambda} \sum x_i - \sum a_{R_i} = 0$, што дава $\lambda = \frac{\sum x_i}{\sum a_{R_i}}$, т.е. оценувач $\hat{\lambda} = \frac{\sum \xi_i}{\sum a_{R_i}}$.

20. а) $f(x_1, x_2, \ldots, x_n; a) = \frac{a^{np}}{(\Gamma(p))^n} (x_1 x_2 \cdots x_n)^{p-1} e^{-a \cdot \sum x_i}$ што дава

$\ln f(x_1, x_2, \ldots, x_n; a) = np \ln a - n \ln \Gamma(p) + (p-1) \sum \ln(x_i) - a \sum x_i$ и

конечно $\frac{\partial}{\partial a} \ln f(x_1, x_2, \ldots, x_n; \lambda) = \frac{np}{a} - \sum x_i = 0 \quad \Rightarrow \quad \hat{a} = \frac{np}{\sum X_i}$.

б) $Y = \sum X_i$ има $\chi 2$ распределба (сума на независни случајни променливи со $\chi 2$ распределба) со густина $f_Y(x; a) = \frac{a^{np}}{\Gamma(np)} x^{np-1} e^{-ax}$, $x > 0$.

Сега $E\hat{a} = np \int_0^\infty \frac{1}{x} \frac{a^{np}}{\Gamma(np)} x^{np-1} e^{-ax} dx = \frac{npa^{np}}{\Gamma(np)} \int_0^\infty x^{np-2} e^{-ax} dx = \ \ldots$

$= \frac{npa}{np-1} = a + \frac{a}{np-1} \neq a$ (не е центриран). $\hat{a}_1 = \hat{a} - \frac{\hat{a}}{np}$ е центриран.

Глава 12

1. а) 97.93%; б) 99.36%; в) 96.78%.
2. (89.471, 91.489).

3. а) (74.0353, 74.0367); б) (74.0355, ∞).
4. а) $t_{0.05,14} = 1.761$; б) $t_{0.01,19} = 2.359$; в) $t_{0.001,24} = 3.467$.
5. (4.023, ∞).
6. а) Да; б) (16.455, 17.505).
7. 666.
8. 5759.
9. $3.19 < X_{n+1} < 4.19$.
10. $\bar{x} = 16.9833$, $s = 0.3189$, $t_{\alpha/2} = t_{0.005} = 4.0321$
$$\left(16.983 - 4.032 \cdot 0.319\sqrt{1+\frac{1}{6}},\ 16.983 + 4.032 \cdot 0.319\sqrt{1+\frac{1}{6}}\right) = (15.59,\ 18.37);$$
11. $\chi^2_{0.05} = 16.9190$ и $\chi^2_{0.95} = 3.3251$ за 9 степени на слобода.
$p(0.000851 < \sigma^2 < 0.004331) = 0.90$.
12. (3.6, 8.1); не.

Глава 13

1. а) да; б) не; в) не; г) да; д) да.
2. а) да; б) не; в) не; г) не; д) не; ѓ) не.
3. а) Тестираме H_0: $\mu = 30000$, наспроти H_A: $\mu > 30000$ километри.
Статистиката е $z = \dfrac{30822 - 30000}{1500/\sqrt{16}} = 2.192 < 2.33 = z_{0.01}$, H_0 не се отфрла;
б) $\beta = \Phi\left(2.33 + \dfrac{30000 - 31000}{1500/\sqrt{16}}\right) = \Phi(-0.34) = 0.3669$;
в) Имаме $z_\beta = z_{0.1} = 1.28$ па $n = \left(\dfrac{1500(2.33 + 1.28)}{30000 - 31000}\right)^2 = (-5.42)^2 = 29.32$.
4. а) $z = 1.77$, отфрли ја H_0; б) $\beta \approx 1$; в) $n = 35$.
5. а) $z = -3.33 < -2.59$, па H_0 се отфрла; б) 0.1056; в) $n = 217$.
6. а) $\bar{x} = 249.7$, $S = 145.1$. Не, бидејќи $1.19 < 1.796$; б) 0.30.
7. а) $t = -3.48$, па отфрли ја H_0, P-вредноста = 0.002; б) $\beta \approx 1$; в) $n = 35$.
8. а) $t = 3.018$ и H_0 се отфрла, P-вредноста = 0.0038; б) $\beta = 0.8$; в) $n = 38$.

9. $z = 3.67 > 2.58$, па отфрли ја H_0: $p = 0.40$. Не.

10. а) Не, бидејќи $1.28 < 1.645$;

    б) Тип 1: Заклучок дека повеќе од 20% се дебели кога тоа не е така;

    Тип 2: Заклучок дека 20% се дебели кога вистинскиот процент надминува 20%;

    в) 0.121.

11. а) $z = 0.452$, не ја отфрлај H_0; б) P-вредност = 0.67364.

12. а) $\alpha = 0.0853$; б) $\beta \approx 0$.

13. а) $\chi_0^2 = 4984.83$, отфрли ја H_0; б) P-вредност < 0.005.

14. Поголеми шанси за прифаќање има за $\alpha = 0.01$. Областа на прифаќање (под графикот на хи-квадрат функцијата) е поголема.

15. $d = 7.274$, $p(\chi_9^2 > 7.274) = 0.609$ што значи дека има доволна основа за прифаќање на хипотезата. Исто така, со ниво на значајност 0.05 (а тогаш и со 0.01 и 0.001) имаме дека $d = 7.274 < 16.9 = \chi^2_{0.05,9}$, што значи дека хипотезата е поддржана од податоците.

16. Тестираме хипотеза дека $p_1 = \ldots = p_5 = 0.2$ наспроти дека тоа не е така. Од податоците имаме дека $n \cdot p_i = 103 \cdot 0.2 = 20.6$, и за d добиваме

$$d = \frac{(15-20.6)^2}{20.6} + \frac{(27-20.6)^2}{20.6} + \ldots + \frac{(19-20.6)^2}{20.6} = 13.36 \cdot$$

Понатаму поради $d = 13.36 > 9.4877 = \chi^2_{0.05,4} \Rightarrow$ хипотезата се отфрла, т.е. барем една производствена лента има поголем (помал) број дефектни производи.

17. $p_j = p(A_j) = \dfrac{0.48^{j-1} e^{-0.48}}{(j-1)!} =$ {0.6188, 0.2970, 0.0713, 0.0114, 0.0013, 0.0001, 0.0001}. $d = \sum_{j=1}^{7} \dfrac{N_j^2}{n \cdot \hat{p}_j} - n = 8413 - 7842 = 571 > 16.812 = \chi^2_{0.01,6}$ и

хипотезата се отфрла со ниво на значајност од 1%.

18. а) $\chi^2_{0.01,3} = 10.71$, H_0 не се отфрла;

    б) $0.05 < P$- вредност < 0.10.

Глава 14

1. 50.
2. а) 99% интервал: (0.33, 0.71);
   б) 99% интервал: (–0.07, 0.41), па 0 е можна вредност за разликата.
3. а) (–3.85, 11.35);
   б) За ниво $\alpha = 0.05$, холестеролот е зголемен, но не и за $\alpha = 0.01$. Да. P-вредноста = 0.02.
4. а) (–3.684, –2.116);
   б) $z_0 = -7.254$, отфрли ја H_0, P-вредноста ≈ 0.
5. 26.
6. а) да; б) $t_0 = 2.558$, отфрли ја H_0, P-вредноста = 0.020; в) 0.05; г) $n = 51$; д) (1.86, 18.94).
7. Да. $t_0 = 5.465$, отфрли ја H_0.
8. а) Да. $t_0 = 8.387$ и H_0 се отфрла; б) $t_0 = 3.45$, отфрли ја H_0.
9. H_0 се отфрла бидејќи $-4.18 \le -2.33$.
10. а) $z_0 = 1.49$, не се отфрла H_0;
    б) 0.81859, 383.
11. а) $z_0 = 3.42$, па H_0 се отфрла. P-вредноста е 0.00062;
    б) (0.0434, 0.1616).
12. а) $z = 0.80 < 1.96 \Rightarrow$ H_0 не се отфрла; б) $n = 1211$.
13. а) Интервалот за $\ln(\theta)$ е $\ln(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\left(\frac{m-x}{mx} + \frac{n-y}{ny}\right)}$. Сега, интервалот за θ се добива со директно антилогаритмирање.

б) (1.43, 2.31), се покажува дека аспиринот е корисен.

14. а) 1.59; б) 2.28; в) 2.64; г) 0.529.
15. $f = 0.657$, H_0 не се отфрла.
16. а) Тестираме H_0: $\sigma_1^2 = \sigma_2^2$
    наспроти H_A: $\sigma_1^2 \neq \sigma_2^2$, и
    бидејќи $f = 0.961 \in (0.365, 2.859) = (f_{1-\alpha/2,\,(24,20)}, f_{\alpha/2,\,(24,20)})$ $\Rightarrow$ не ја отфрламе H_0 за $\alpha = 0.02$;
    б) (0.3369, 2.640).

Глава 15

1. а) $\hat{\beta} = \dfrac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2}$; б) $M\hat{\beta} = \beta$, $D\hat{\beta} = \dfrac{\sigma^2}{\sum_{j=1}^{n} x_j^2}$;

   в) $\hat{\sigma}^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(y_j - \hat{\beta}x_j)^2$.

2. За $x = 20$, $\hat{y} = 65 - 1.2 \cdot 20 = 41$, па $p\left(Z > \dfrac{50-41}{8}\right) = 0.1292$.

   $Y_1 - Y_2$ има нормална распределба со очекување $E(Y_1 - Y_2) = \beta_1 = -1.2$, и дисперзија $D(Y_1 - Y_2) = \sigma^2 + \sigma^2 = 128$, па бараната веројатност е

   $p(Y_1 - Y_2 > 0) = p\left(Z > \dfrac{0-(-1.2)}{\sqrt{128}}\right) = 0.4562..$

3. $\hat{\sigma}^2 = \dfrac{1}{n}\sum_{j=1}^{n}(y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)^2$, $E\hat{\sigma}^2 = \dfrac{n-2}{n}\sigma^2$, значи не е центрирана.

4. а) $\hat{\beta}_1 = -0.0353971$, $\hat{\beta}_0 = 33.5348$, $\hat{\sigma}^2 = 13.392$; б) 28.226;

   в) 29.4995, 1.50048.

5. а) $\bar{x} = 46.5$, $\bar{y} = 421.85$, $S_{xy} = 265864.63$, $S_{xx} = 3309.0002$,
   $\hat{\beta}_1 = 9.2089$, $\hat{\beta}_0 = -6.364$ (Коефициентот на корелација е $\rho = 0.999933$);

   б) 500.1255;

   в) 426.4543, 1.6143.

6. $14.98 + 32.14 \cdot \log_{10} x$.

7. а) $118.91 - 0.905x$;

   б) Очекуваното опаѓање на порозноста со зголемувањето на единечната тежина е 0.905;

   в) Негативно предвидување, но y не може да биде негативно;

   г) –0.52, 0.49;

   д) 0.938.

8. $\bar{x} = 140.895$, $\bar{y} = 78.74$, $S_{xy} = 776.434$, $S_{xx} = 18921.8295$
   $\hat{\beta}_1 = 0.41$, $\hat{\beta}_0 = 72.96$, па линијата е $y = 72.96 + 0.041x$.

9. а) $\bar{x} = 14.54$, $\bar{y} = 112.9067$, $S_{xx} = 405.836$, $\hat{\beta}_1 = -0.917622$, $\hat{\beta}_0 = 126.248889$;

   б) $S_{\hat{\beta}_1} = \dfrac{S}{\sqrt{S_{xx}}} = \dfrac{2.941}{\sqrt{405.836}} = 0.1460$, $t_{0.025,\,13} = 2.160$, Бараниот интервал е

   $-0.918 \pm (2.160)(0.1460) = -0.918 \pm 0.315 = (-1.233, -0.60340)$.

10. а) $\bar{x} = 36.6111$, $\bar{y} = 16.2889$, $S_{xy} = 9293.95$, $S_{xx} = 4840.7778$

$\hat{\beta}_1 = -0.297561$, $\hat{\beta}_0 = 27.182936$;

б) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 45 = 13.79$

$$S_{\hat{y}} = 2.8640\sqrt{\frac{1}{18} + \frac{(45 - 36.6111)^2}{4840.7778}} = 0.7582\,,\quad t_{0.025,\,16} = 2.120$$

Бараниот интервал е (12.18, 15.40).

в) (7.51, 20.07).

11. а) е потесен од б); в) е потесен од г); а) е потесен од в); б) е потесен од г).

12. а) (77.80, 78.38); б) (76.90, 79.28), ист центар но поширок;
в) пошироки, бидејќи 115 е подалеку од $\bar{x}$; г) $t = -11$, P-вредност = 0.

13. а) 0.3902, 2.5717; б) $t = 8.518$, па H_0 се отфрла, 0.000000084965;
в) $t = 72.5563$, H_0 се отфрла, ≈ 0; г) $t = 5.2774$, H_0 се отфрла, 0.00004596.

14. $R^2 = 20.1121\%$, да.

15. $$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 20 & 26 & \dots & 33 \\ 23 & 21 & \cdots & 23 \end{bmatrix} \begin{bmatrix} 1 & 20 & 23 \\ 1 & 26 & 21 \\ \vdots & \vdots & \vdots \\ 1 & 33 & 23 \end{bmatrix} = \begin{bmatrix} 12 & 626 & 290 \\ 626 & 36\,776 & 15\,336 \\ 290 & 15\,336 & 7028 \end{bmatrix}$$ и

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 20 & 26 & \dots & 33 \\ 23 & 21 & \cdots & 23 \end{bmatrix} \begin{bmatrix} 210 \\ 206 \\ \vdots \\ 230 \end{bmatrix} = \begin{bmatrix} 2974 \\ 159\,011 \\ 72\,166 \end{bmatrix},$$ па оттука следува

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 12 & 626 & 290 \\ 626 & 36\,776 & 15\,336 \\ 290 & 15\,336 & 7028 \end{bmatrix}^{-1} \begin{bmatrix} 2094 \\ 159\,011 \\ 72\,166 \end{bmatrix} = \begin{bmatrix} -33.84 \\ 0.39 \\ 10.80 \end{bmatrix}$$

Значи бараниот регресионен модел е $y = -33.84 + 0.39x_1 + 10.80x_2$.

16. а) $y = 383.80 - 3.6381x_1 - 0.1119x_2$;

б) $\hat{\sigma}^2 = 153.0$, а стандардните грешки се $se(\hat{\beta}_0) = 36.22$, $se(\hat{\beta}_1) = 0.5665$ и $se(\hat{\beta}_2) = 0.04338$;

в) 180.95;

г) $y = 484.0 - 7.656x_1 - 0.222x_2 - 0.0041x_{12}$;

д) −31.3.

17. а) $SS_T = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 = 743\,764 - 2974^2 / 12 = 6707.6667$,

$$SS_R = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 = 740\ 766.93 - 2974^2 / 12 = 3710.5967,$$

и сега $SS_E = SS_T - SS_R = 2997.07.$

Статистиката е $f = \dfrac{3710.5967/2}{2997.07/(12-2-1)} = \dfrac{1855.298}{333.0078} = 5.571\cdot$

Сега бидејќи $f = 5.571 > 4.2565 = f_{0.05,2,9}$ ја отфрламе H_0;

P-вредност = 0.0266;

б) Имаме $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 12 & 626 & 290 \\ 626 & 36\,776 & 15\,336 \\ 290 & 15\,336 & 7028 \end{bmatrix}^{-1} = \begin{bmatrix} 51.16998 & 0.10536 & -2.3414 \\ 0.10536 & 0.0005 & -0.0055 \\ -2.3414 & -0.0055 & 0.1087 \end{bmatrix}\cdot$

$$\hat{\sigma}^2 = \sum_1^n (y_i - \hat{y}_i)^2 / (n-m-1) = 1880.4916 / 9 = 208.9435.$$

Статистиките се: за β_0: $t = \dfrac{-33.84}{\sqrt{208.9435 \cdot 51.16998}} = -0.3273$;

за β_1: $t = \dfrac{0.39}{\sqrt{208.9435 \cdot 0.0005}} = 1.2066$; за β_2: $t = \dfrac{10.80}{\sqrt{208.9435 \cdot 0.1087}} = 2.2662$;

а критичната точка е $t_{0.025,9} = 2.2622$. За β_2 ја отфрламе H_0, додека за β_0 и β_1 би ја прифатиле.

18. а) $f = 67.92$, reject H_0; б) не, не се отфрла H_0.

19. Точкастата оценка е $\hat{\mu}_{Y|\mathbf{x}_0} = -33.84 + 0.39x_1 + 10.80x_2 = 214.68$

$$\hat{\sigma}^2 \mathbf{x}_0^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0 =$$

$$= 222.7495 \cdot [1\ 28\ 22] \begin{bmatrix} 51.16998 & 0.10536 & -2.3414 \\ 0.10536 & 0.0005 & -0.0055 \\ -2.3414 & -0.0055 & 0.1087 \end{bmatrix} \begin{bmatrix} 1 \\ 28 \\ 22 \end{bmatrix} = 61.332$$

$$p\left(\mu_{Y|\mathbf{x}_0} \in 214.68 \pm 1.8331\sqrt{61.332}\right) = p\left(\mu_{Y|\mathbf{x}_0} \in (200.324, 229.036)\right) = 0.90;$$

$\hat{\sigma}^2(1 + \mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0) = 284.0815$, па предвидувањето е

$$p\left(Y_0 \in 214.68 \pm 1.8331\sqrt{284.0815}\right) = p\left(Y_0 \in (183.784, 245.576)\right) = 0.90.$$

20. а) 0.985; б) 0.99, не.

Литература

Biswas S., *Topics in Statistical Methodology*, Wiley, New Delhi, 1991.

Box G.E.P., Muller M.E., A Note on the Generation of Random Normal Deviates, *The Annals of Mathematical Statistics*, Vol. 29, No. 2, 1958, 610–611.

Cook R. D., Weisberg S., *Residuals and Influence in Regression*, Chapman & Hall, New York, 1982.

DeGroot H.M., *Probability and Statistics*, Second Ed., Addison-Wesley, Reading, 1989.

Dekking F.M., Kraaikamp C., Lopuhaä H.P., Meester L.E., *A Modern Introduction to Probability and Statistics*, Springer-Verlag, London, 2005.

Devore J.L., *Probability and Statistics for Engineering and the Sciences*, Eighth Edition, Brooks/Cole, Boston, 2012.

Fisher, R. A., *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh, 1956.

Galanti S., Jung A., Low-Discrepancy Sequences: Monte Carlo Simulation of Option Prices, *Journal of Derivatives*, 1997, 63-83.

Gerald H., Meeker W., *Statistical Intervals,* Wiley, New York, 1991.

Hacking I., *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*, Cambridge University Press, Cambridge, 1975.

Halton J., Algorithm 247: Radical-inverse quasi-random point sequence, *ACM*, 1964, 701-702.

Härdle W., *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.

Jaynes E.T., The Well-Posed Problem, *Foundations of Physics 3*, 1973, 477–493.

Jaynes E.T., *Probability Theory: The Logic of Science*, Cambridge University Press, New York, 2003.

Li M., Vitanyi P., *An Introduction to Kolmogorov Complexity and Its Applications*, 3nd Edition, Springer Verlag, New York, 2008.

Mendenhal W., Sincich T., *Statistics for Engineering and the Sciences*, Maxwel Macmillan Int. Ed., New York, 1992.

Montgomery D.C., Runger G.C., *Applied Statistics and Probability for Engineers*, Third Edition, John Wiley & Sons, Inc., New York, 2003.

Motulsky H., *Intuitive Biostatistics*, Oxford University Press, Oxford 1995.

Park S.K., Miller K.W., Random Number Generators: Good Ones Are Hard To Find, *Communications of the ACM 31*(*10*), 1988, 1192–1201.

Poirier D.J. *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Cambridge, MA, 1995.

Quenouille M.H., Notes on bias in estimation, *Biometrika*, 43, 1956, 353–360.

Rao, C. R., Sufficient statistics and minimum variance estimates, *Proceedings of the Cambridge Philosophical Society*, 45, 1949, 218–231.

Renyi A., *Probability Theory*, North-Holland, Amsterdam, 1970.

Richtmyer R.D., The evaluation of definite integrals, and quasi-Monte Carlo method based on the properties of algebraic numbers, *Report LA-1342*, Los Alamos Scientific Laboratory, NM, 1951.

Rodgers J.L., Nicewander W.A., Thirteen Ways to Look at the Correlation Coefficient, *The American Statistician 42*, 1988, 59–66.

Schervish M. J., *Theory of Statistics*, Springer-Verlag, New York, 1995.

Scott D.W., *Multivariate Density Estimation*,Wiley, New York, 1992.

Silverman B.W., *Density Estimation for Statistics and Data Analysis*, Chapman andHall, London, 1986.

Sobol I.M., Distribution of Points in a Cube and Approximate Evaluation of Integrals, *USSR Comput. Math. Phys. 7*, 1967, 86–112.

Soong T.T., *Fundamentals of Probability and Statistics for Engineers*, John Wiley & Sons Ltd, Hoboken, NJ 07030, 2004.

Spanos A., *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, New York, 1999.

Tijms H., *Understanding Probability: Chance Rules in Everyday Life*, Sec.Ed., Cambridge University Press, New York, 2007.

Trpenovski B., *Verojatnost i statistika*, Univerzitet "Kiril i Metodij", Skopje, 1981.

Индекс

Е

Ж

З

И

J

К

Л

Љ

М

Н

О

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

519.2:62(075.8)

ЧАКМАКОВ, Душан
Веројатност и статистика за инженери [Електронски извор] : [учебник] / Душан Чакмаков. - Скопје : Универзитет "Св. Кирил и Методиј", Машински факултет, Оддел за математика и информатика, 2015

Начин на пристап (URL): http://www.ukim.edu.mk/e-izdavastvo - Текст во ПДФ формат, содржи 512 стр. - Наслов преземен од екранот. - Опис на изворот на ден 31.08.2015. - Библиографија: стр.[491]-492. - Регистар

ISBN 978-9989-43-376-4

а) Теорија на веројатност - Инженерство - Високошколски учебници
б) Математичка статистика - Инженерство - Високошколски учебници
COBISS.MK-ID 99201290

Книгата „Веројатност и статистика за инженери“ содржи значително повеќе материјал од еден почетен курс по веројатност и статистика, прилагоден за инженерите. Изложувањето и содржината на материјалот се такви што книгата има многу поширока примена вклучувајќи ја медицината, економијата, природно-математичките науки, информатиката... Книгата во основа е еден модерен курс по веројатност и статистика што комплетно ја покрива наставната програма по соодветните предмети на Машинскиот факултет во Скопје, како на додипломските така и на постдипломските студии. Исто така, таа делумно или комплетно ја покрива наставната програма по многуте варијации на предметот веројатност и статистика на другите факултети на Универзитетот „Св. Кирил и Методиј“, но и на факултетите на другите универзитети во државата..

ISBN 978-9989-43-376-4