Душан Чакмаков

Статистика

Интерна скрипта, Машински факултет
Скопје, 2014

Содржина

1

Вовед

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 непознати. 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

Од веројатност до статистика

Статистиката во некоја смисла е обратна од теоријата на веројатност. Во теоријата на веројатност, врз база на веројатносниот модел (Ω, Φ, p) со зададени веројатности на елементарните настани, проблемот е да се пресметаат веројатностите на произволните настани од Φ. Во статистиката, врз база на емпириските податоци од кои може да се проценат веројатностите на некои настани, проблемот е да се дефинира веројатносниот модел (Ω, Φ, 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.

2.1. 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.

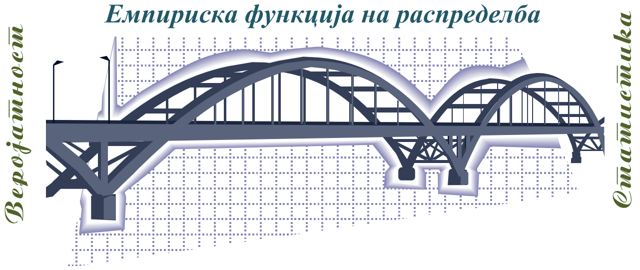
Популацијата може да смета за случајна променлива. Тогаш, анализата се сведува на определување на распределба на соодветната случајна променлива, а понекогаш само некои нејзини бројни карактеристики како: просек, дисперзија, моменти итн. Ако X е популацијата од која сме земале примерок со вредности x_1, x_2, …, 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 2.1 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 карактерстки 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, …, 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 емпирискита функција на распределба да биде доволно добра апроксимација на $F(x)$. Следната теорема го обезбедува многу посилно ова барање.

Теорема 2.1 (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

$$p\left(\lim_{n\to\infty}\left(\sup_{-\infty<x<\infty}\left|F_n(x)-F(x)\right|\right)=0\right)=1\,.\ ■$$

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 ≤ 1/2 станува ≤ 1.

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.

2.2. 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. 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,

$$\Phi = \{f(x\,;\,\Theta) \mid x \in \mathrm{R}\}.$$

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.

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.

2.3. 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: $\Phi = \{f(x;\Theta) \mid x \in \mathrm{R}\}$,
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,

$$f(x_1,x_2,\ldots,x_n;\Theta)=\prod_{k=1}^{n} f_k(x_k;\boldsymbol{\theta}_k)\text{, за сите } (x_1, x_2,\ldots, x_n) \in \mathrm{R}^n;$$

идентична распределба,

$$f_k(x_k;\ \boldsymbol{\theta}_k) = f(x_k;\ \boldsymbol{\theta})\text{, за сите } k = 1, 2, ..., n.$$

Оттука, заедничката распределба едноставно се редуцира на производ на идентичните маргинални распределби

$$f(x_1,x_2,\ldots,x_n;\Theta)=\prod_{k=1}^{n} f(x_k;\boldsymbol{\theta})\text{, за сите } (x_1, x_2,\ldots, x_n) \in \mathrm{R}^n.$$

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 2.2 Да се разгледа случајот кога распределбата на примерокот, т.е. на случајниот вектор $(X_1, X_2, \ldots, X_n)$ е нормална

$$f(x_1,x_2,\ldots,x_n;\Theta)=Z\left(\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix},\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}\end{bmatrix}\right)$$

каде што $\mu_i = EX_i$ се очекувањата, а $\sigma_{ij} = K_{X_i,X_j} = K_{X_j,X_i}$ се коваријациите 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 коваријациите.

Ако се наметне условот за независност, коваријациите 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

$$\theta_k = \{\mu_k, \sigma_{kk}, k = 1, 2, \ldots, n\}$$

чиј број расте со зголемување на примерокот.

Сега, ако дополнително го примениме условот за идентична распределеност на случајните променливи од примерокот:

$$\mu_1 = \mu_2 = \ldots = \mu_n = \mu, \quad \text{и} \quad \sigma_{11} = \sigma_{22} = \ldots = \sigma_{nn} = \sigma, \ \text{т.е. } \theta = \{\mu, \sigma\}$$

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

$$\Theta = \{\mu_i, \sigma_{ij}, i,j=1, 2, \ldots, n\} \to \theta_k = \{\mu_k, \sigma_{kk}, k=1, 2, \ldots, n\} \to \theta = \{\mu, \sigma\},$$

а ова понатаму води до едноставен нормален модел,

1) Веројатностен модел

$$\{f(x; \Theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mid x \in \mathrm{R}, \Theta = \{\mu, \sigma\} \in \mathrm{R} \times \mathrm{R}^+\},$$

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.

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.

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,
   $\Phi = \{f(x\,;\Theta) \mid x \in \mathrm{R}\}$,
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

$$(X_1, X_2, \ldots, X_n) : \Omega \rightarrow \mathrm{V} \subseteq \mathrm{R}^n,$$

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 → лош излез (*garbage in → 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 термин примерок, а од контекстот ќе биде јасно дали се работи за случаен вектор или за обични податоци.

ПРИМЕР 2.3 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 придружена на различните вредности за $\theta \in \Theta$ да бидат вистински параметри на моделот во светло на поедина реализација на примерокот x_1, $x_2, \ldots, x_n$,

$$L(\theta; x_1, x_2, \ldots, x_n) : \theta \to [0, \infty).$$

ПРИМЕР 2.4 Во Бернулиевиот модел

1) Веројатносен модел, $\Phi = \{f(x; \Theta) = \theta^x(1 - \theta)^{1-x} \mid 0 \le \theta \le 1, x = 0, 1\}$,

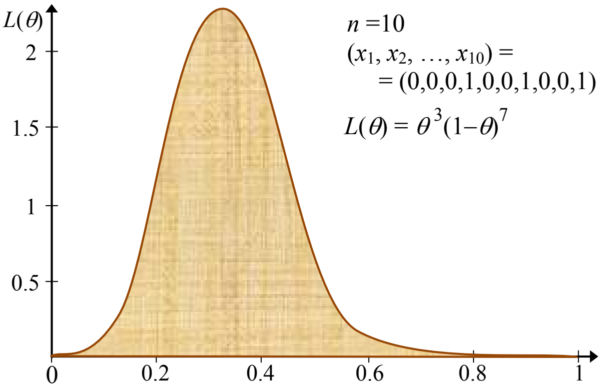
2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \to \{0, 1\}^n$,

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.

2.4. 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 $(X_1, X_2, \ldots, X_n)$.

2.4.1. 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 $(X_1, X_2, \ldots, X_n)$, 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 $V \subseteq R^n$ во множеството параметри Θ,

$$h(\cdot): V \rightarrow \Theta.$$

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 2.5 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, $(X_1, X_2, \ldots, X_n): \Omega \rightarrow \{0, 1\}^n$,

бидејќи знаеме дека $\theta = EX$ кога X има Бернулиева распределба, за оценувач $\hat{\theta}$ на θ е природно да се земе

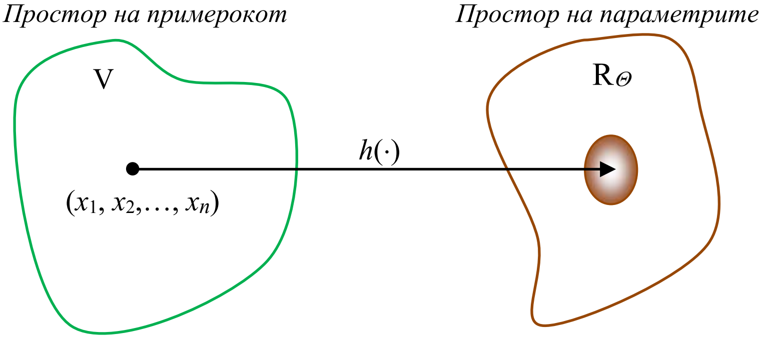
$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} X_k \ . \ \blacksquare$$

$\hat{\theta}$ 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 возможно 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 $h(\cdot)$: V → Θ 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 R_θ 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.

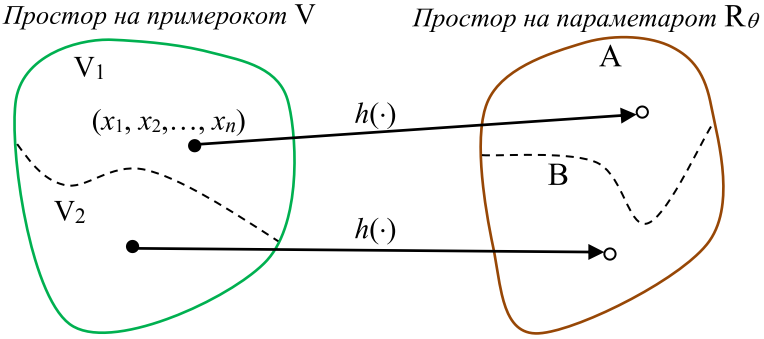
2.4.2. Тестирање хипотези

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 R или 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 B = R_θ / 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 R_θ, т.е. $V_1 = h^{-1}(A)$ и $V_2 = h^{-1}(B)$. Тука главен проблем е определувањето на подмножествата A и B како и функцијата $h(\cdot)$. Се разбира, како што понатаму 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.

2.4.3. 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 2.6 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 R_θ, и потоа од R_θ во просторот на предвидувања којшто е дел од просторот на примерокот, да речеме V_p, т.е.

$$q(h(\cdot)) : \mathrm{V} \to \mathrm{R}_\theta \to \mathrm{V}_p.$$

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 $\hat{\theta}$. ■

2.5. 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.

2.5.1. 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 истражување 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 лабораторискиот 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 2.7 Нашироко 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.

2.5.2. 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 вeредности на набљудуваните променливи. Ова е спротивно во однос на експерименталните и збирните податоци 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 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, термините *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 секогаш 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.

Оценките на параметрите во моделот, креирањето интервали на доверба и тестирањата хипотези ќе бидат предмет на изучување во следните глави.

ЗАДАЧИ

1. 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	425-474	475-524	525-574	575-624	625-674	675-725
Бр.тел.	0	1	5	9	6	8	6	2	2	0	0	1	0

   Состави емпириска функција на распределба за овие податоци.

2. 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?

3. Објасни од веројатносен аспект што значи примерок, а што е реализација на примерокот.

4. Објасни го поимот "распределба на примерок".

5. Која е разликата меѓу експерименталните и набљудуваните податоци од аспект на статистичката анализа?

6. Зошто распределбата на примерокот е суштински концепт во статистичката анализа?

7. Зошто е практично многу тешко да се најде распределбата на примерокот?

3

Описна статистика

Сумарното прикажување на податоците од примерокот е важен чекор во секоја статистичка анализа бидејќи нѐ 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.

3.1. 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.

3.1.1. 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 3.1 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). ■

3.1.2. 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 3.2 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.

Меѓуквартилниот 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. ■

3.1.3. 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.

Стандардизиранaта 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 3.3 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 $z_{70} = (615 - 490.80)/54.74 = 2.27$.

$\overline{x} \pm k \cdot s$	**Интервал**	**% во интервалот**
Во $\overline{x} \pm s$	$490.80 \pm 54.74 = [436.06, 545.54]$	48/70 = 68.57%
Во $\overline{x} \pm 2s$	$490.80 \pm 2{\cdot}54.74 = [381.32, 600.28]$	68/70 = 97.14%
Во $\overline{x} \pm 3s$	$490.80 \pm 3{\cdot}54.74 = [326.58, 655.02]$	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

$$\overline{x} = \frac{\sum_{j=1}^{n} f_j M_j}{n}$$, каде што M_j е средината на класата j.

Дисперзијата би се пресметувала соодветно со

$$s^2 = \frac{\sum_{j=1}^{n} f_j (M_j - \overline{x})^2}{n-1}.$$

ПРИМЕР 3.4 Да претпоставиме дека податоците од примерокот на цени за закуп на еднособни 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. ■

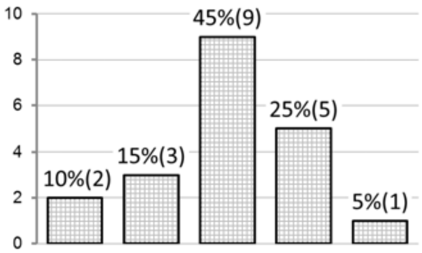
3.2. Визуелно претставување на податоци

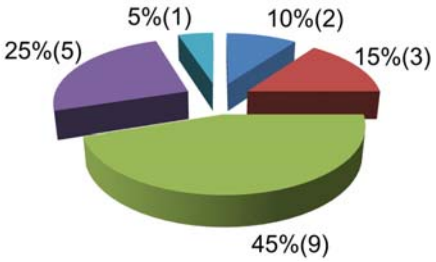
За визуелно претставување на податоците вообичаено се користат дијаграми со барови или пити во 2 или 3 димензии кои обично ги изразуваат процентуалните (или други) односи извлечени од податоците.

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, 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.

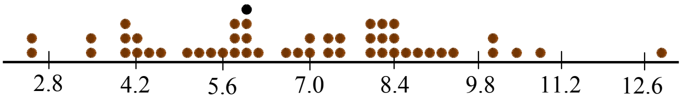
ПРИМЕР 3.6 Во следната табела се дадени податоци за процент од даноците што оди на високото образование за поедини државо во САД (по азбучен редослед на државите):

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.

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.

3.3. 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, … во секоја од класите k_1, k_2, k_3, …, а потоа најди ги релативните фреквенции $f_1/n, f_2/n, f_3/n$, … ;

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 ставаат 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 3.9 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.

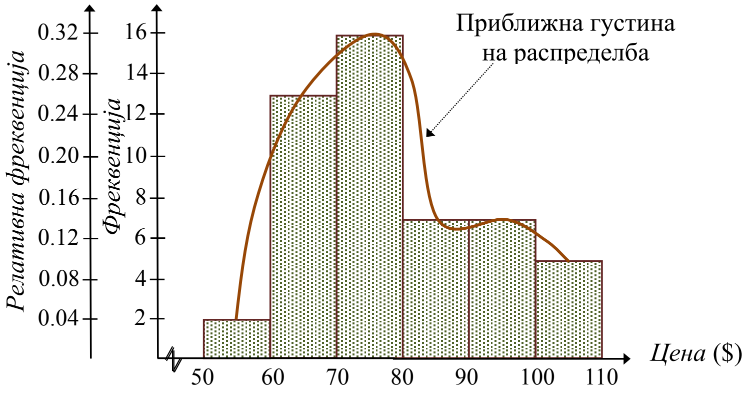
Решение

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

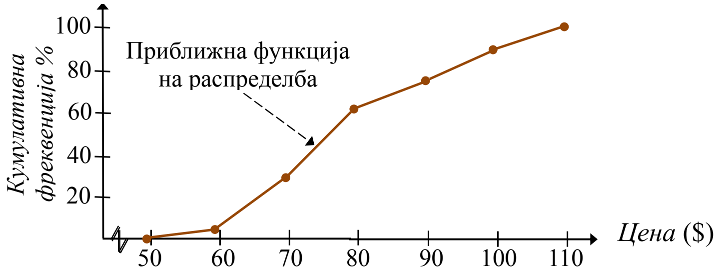
Хистограмот, преку фреквенции е даден на следната слика



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 3.10 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:

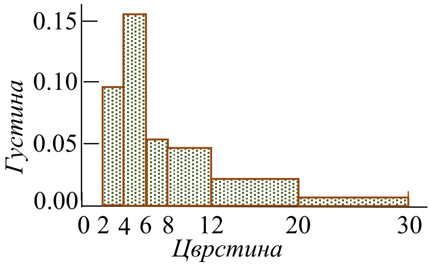
11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

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.

3.4. 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 зборови, $b(p)$ 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 рашireноста 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

$$\textit{проценил на } Z(\mu, \sigma) = \mu + \sigma \cdot \textit{проценил на } Z(0,1).$$

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 3.1 Цртежот на n-те точки

$$(j\text{-та најмала вредност на примерокот, } 100(j - 0.5)/n\text{-ти процентил})$$

во дводимензионален координатен систем се нарекува нормален веројатносен дијаграм.

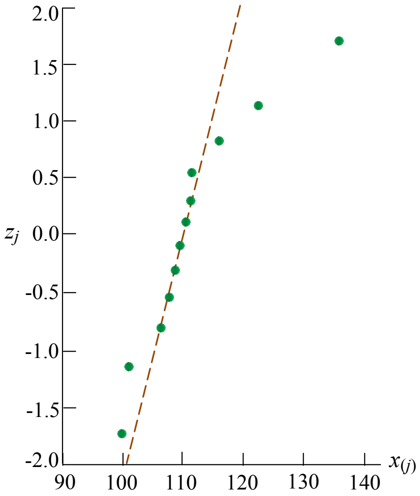
Ако точките од нормалниот веројатносен дијаграм се приближно вдолж права линија, тогаш има индиции дека распределбата на популацијата е приближно нормална. Ако отстапувањето од права линија е значително, може да сметаме дека податоците се од некоја друга распределба. Оценката дали податоците се сложуваат 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 3.11 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 придружениот 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. ■

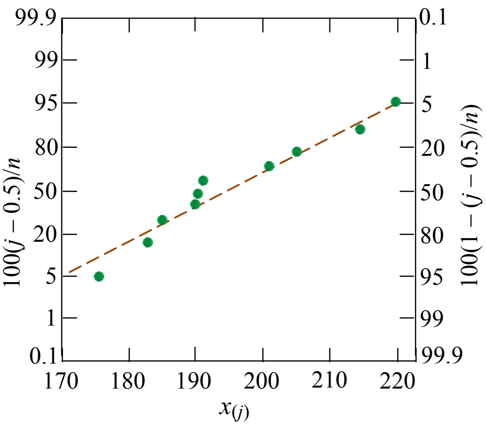
Алтернативен (поедноставен) начин за конструкција на нормален веројатносен дијаграм е z-процентилите на вертикалната оска да се заменат со нелинеарно претставени веројатности $(j-0.5)/n$. Скалингот на оската се прави таков што точките повторно паѓаат на права линија кога распределбата е нормална. На 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 $(x_{(j)}, (j-0.5)/n)$, $j = 1, 2, \ldots, n$. Ако претпоставената нормална распределба адекватно ги опишува податоците, точките ќе бидат приближно вдолж права линија.

ПРИМЕР 3.12 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 придружениот 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;

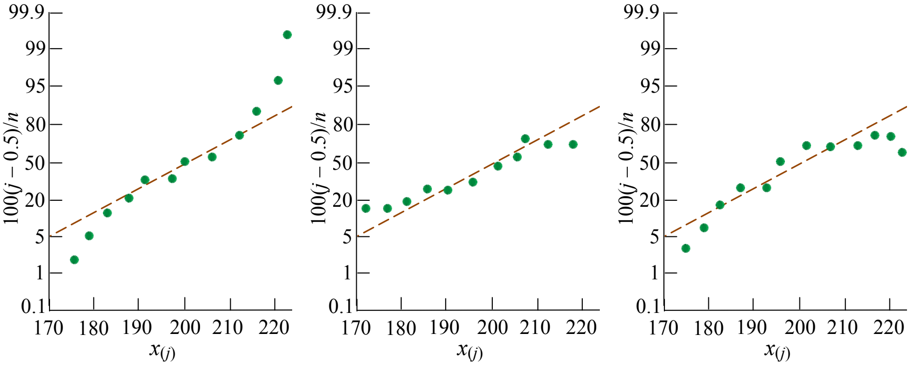
б) Таа е симетрична, но краевите се "помалку стрмни" отколку кај нормалната распределба;

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 (податоците < 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 Ƨ. 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 3.1 *Три примери на веројатносни дијаграми што индицираат не-нормална 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 ссучај, 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) Пресметај го просекот, стандардната девијација и медијаната;

   б) 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	133.2	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	

   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. 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.

6. 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?

7. 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?

8. 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.

9. 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.

4

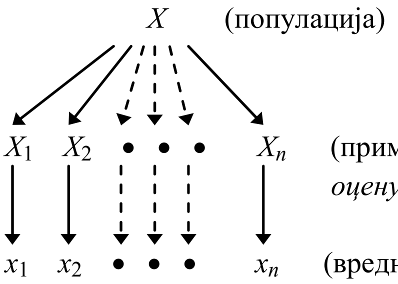
Оценки на непознати параметри

Секоја статистичка анализа се базира на две основни компоненти, статистички модел даден со двојката (Φ, (X_1, X_2, …, X_n)), каде што Φ е веројатносниот модел, а (X_1, X_2, …, X_n) е модел на примерокот и множество податоци (x_1, x_2, …, 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 $\Phi = \{f(x;\,\Theta) \mid x \in \mathrm{R}\}$. 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 $\hat{\Phi} = \{f(x; \hat{\theta}_1, \hat{\theta}_1, \ldots, \hat{\theta}_k) \mid x \in \mathrm{R}\}$. Вообичаено секој од параметрите се оценува индивидуално.

Оценката на параметарот θ е пресликување $h(\cdot)$ од просторот на примерокот што е подмножество $\mathrm{V} \subseteq \mathrm{R}^n$ во множеството параметри Θ,

$$h(\cdot)\colon \mathrm{V} \to \Theta.$$

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.



(примерок)

оценувач: $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$

(вредности на примерокот)

оценка: $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$

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 4.1 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,

$\Phi = \{f(x;\Theta) = Z(\mu, \sigma)\}, x \in \mathrm{R}, \mathrm{R}_\Theta = \mathrm{R} \times \mathrm{R}_+, \Theta = \{\mu, \sigma\}$,

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 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

$$G(y) = p(\hat{\theta} < y) = \underset{(\mathrm{P}^n : h(x_1, x_2, \ldots, x_n) < y)}{\int \cdots \int} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n .$$

ПРИМЕР 4.2 Да ги најдеме распределбите на различните оценувачи $\hat{\theta}$ од примерот 4.1.

Ако случајните променливи $X1, X2, \ldots, Xn$ се независни со бернулиева распределба, тогаш $\sum_{i=1}^{n} X_i$ има биномна распределба 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 $\frac{1}{n}\sum_{i=1}^{n} X_i$.

4.1. 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.

4.1.1. 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.

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 $n \to \infty$.

4.1.2. 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

$$\begin{aligned}S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big((X_i-\mu)-(\overline{X}-\mu)\big)^2\\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left((X_i-\mu)-\frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)\right)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}(X_i-\mu)(X_j-\mu)\,.\end{aligned}$$

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.

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.

4.1.3. Пропорција во примерокот

Статистиката

$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$, каде што секое X_i е 0 или 1 и означува дали некој настан не се случил 0 или се случил 1, се нарекува пропорција на примерокот земен од популацијата X. Во основа $\hat{P} = \frac{m}{n}$ каде што е број на случувања т.е. релативна честота на појавување на некој настан.

Нека просекот и дисперзијата на популацијата бидат $EX = p$ и $DX = \sigma^2$. Просекот и дисперзијата на статистиката $\hat{P}$ се

$$E\hat{P} = \frac{1}{n}\sum_{i=1}^{n} EX_i \ = \frac{1}{n}(np) = p$$

$$D\hat{P} = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Јасно е дека $\hat{P}$е најдобар оценувач за непознатата веројатност (прпорција) во популацијата.

4.2. 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 биземал 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

$$E\hat{\theta} = \theta \text{ и } D\hat{\theta} = 0.$$

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 оценувачот, т.е 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.

4.2.1. Центрираност

Дефиниција 4.1 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 нецентриран 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 $E\hat{\nu} \neq \nu$.

4.2.2. 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 4.2 $\hat{\theta}$ 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 4.3 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

$$D\overline{X} = \frac{\sigma^2}{n}, \text{ т.е. стандардната грешка е } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

што очигледно опаѓа со растењето на n. ■

ПРИМЕР 4.4 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 ваквите прашања поврзани со наоѓање оценувачи со најмала дисперзија, од огромна полза е теоремата на Крамер-Рао.

Теорема 4.1 **(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.

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 давеме некои поважни забелешки во врска со неравенството на Крамер-Рао:

а) Неравенството може да се напише во еквивалентен облик

$$D\hat{\theta} \geq -\left(E\left(\frac{\partial^2 \ln f(x_1, x_2, \ldots, x_n;\theta)}{\partial\theta^2} \right) \right)^{-1},$$

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

$$D\hat{\theta} \geq \left(\frac{\partial E\hat{\theta}}{\partial\theta}\right)^2 \left(E\left(\frac{\partial \ln f(x_1, x_2, \ldots, x_n; \theta)}{\partial\theta}\right)^2\right)^{-1};$$

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 4.5 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. ■

ПРИМЕР 4.6 Во Бернулиевиот модел:

1) Веројатностен модел, $\Phi = \{f(x;\Theta) = \theta^x(1-\theta)^{1-x} \mid 0 \leq \theta \leq 1, x = 0, 1\}$,

2) Модел на примерок, $(X_1, X_2, \ldots, X_n) : \Omega \rightarrow \{0, 1\}^n$,

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. ■

4.2.3. Конзистентност

Веќе кажавме дека идеален оценувач θ^* за кој важи $p(\theta^* = \theta) = 1$ не е возможен за примерок со конечечна големина n, па оттука е логично од оценувачот да бараме ваква карактеристика кога тоа е можно, т.е. кога n расте до бесконечност.

Дефиниција 4.3 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 пиближува 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 $h(X_1, X_2, \dots, X_n)$, 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 4.9 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

$$\lim_{n\to\infty} p(|S^2 - \sigma^2| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{1}{\varepsilon^2} \left(\frac{2\sigma^4}{n-1} \right) = 0 \cdot$$

Значи S^2 е конзистентен оценувач на σ^2. ■

Овој пример ја инспирира следната теорема.

Теорема 4.2 Нека $\hat{\theta}$ е оценувач на θ базирана на примерок со големина n. Тогаш, ако

$\lim\limits_{n\to\infty} E\hat{\theta} = \theta$ и $\lim\limits_{n\to\infty} D\hat{\theta} = 0$, 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 4.10 Спореди ги карактеристиките на оценувачите на μ кај нормалниот модел од примерот 11.1.

Решение

Карактеристиките на овие оценувачи се дадени 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.

4.3. 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 *максимална подобност* и методот на *најмали квадрати*.

4.3.1. 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

$$L(\theta; X_1, X_2, \ldots, X_n) : \theta \to [0, \infty), \text{ а}$$

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 4.18 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

$$\ln L(\mu, \sigma^2; X_1, X_2, \ldots, X_n) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 - \frac{1}{2} n \ln \sigma^2 - \frac{1}{2} n \ln 2\pi \cdot$$

Означуваме $\theta_1 = \mu$ и $\theta_2 = \sigma^2$ и го добиваме системот равенки од

$$\frac{\partial \ln L}{\partial \theta_1} = \frac{1}{\theta_2} \sum_{i=1}^{n} (X_i - \theta_1) = 0,$$

$$\frac{\partial \ln L}{\partial \theta_2} = \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (X_i - \theta_1)^2 - \frac{n}{2\theta_2} = 0 \cdot$$

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 4.19 Користејќи го методот на максимална подобност најди оценувач на θ кај рамномерната распределба.

Решение

Имајќи предвид дека густината на рамномерната распределба е

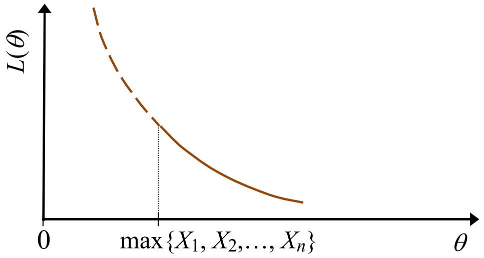
$$f(x;\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0, & \text{во спротивно} \end{cases}$$

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.

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.

4.3.2. 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 функцијта $h(x) = \sum_{i=0}^{k} a_i \phi_i(x)$,

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 4.20 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

$$z(a_0, a_1) = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 .$$

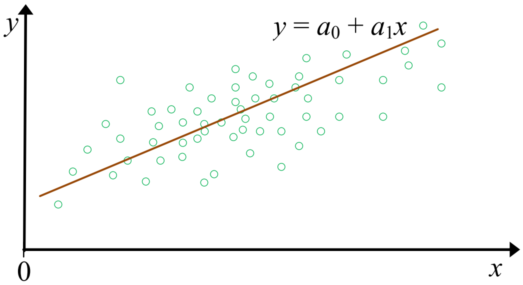
Функцијата $z(\cdot)$ е диференцијабилна, па минимумот го бараме со изедначување на изводот на 0,

$$\frac{\partial z}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0 , \quad \frac{\partial z}{\partial a_1} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) x_i = 0 .$$

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.

Генерално, методот на најмали квадрати априори не обезбедува оценки со така добри особини како методот на максимална подобност. 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 $f(x;\theta) = \dfrac{1}{\theta} e^{-x/\theta}$, $x \geq 0$, каде што θ е непознат параметар. Предложени се два оценувачи на θ, $\hat{\theta}_1 = (X_1 + X_2)/2$ и $\hat{\theta}_2 = \dfrac{4}{\pi}\sqrt{X_1 X_2}$. Кој од оценувачите е подобра во однос на центрираност и помала дисперзија?

6. 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 $n \to \infty$?

7. 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;

8. Нека $X_1, X_2, \ldots, X_n$ е примерок земен од распределбата на Реили (Rayleigh)

   $f(x;\theta) = \dfrac{x}{\theta} e^{-x^2/(2\theta)}$, $x > 0$.

   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.

9. Во следните 4 распределби, користејќи го неравенството на Крамер-Рао, определи ја долната граница на дисперзијата на оценувачите на непознатиот параметар θ:

   а) $f(x;\theta)=\frac{1}{\theta}e^{-x/\theta}, x\ge 1;$ б) $f(x;\theta)=\theta\cdot x^{\theta-1}, 0\le x\le 1, \theta>0;$

   в) $f(x;\theta)=\theta^{x}(1-\theta)^{1-x}, x=0, 1;$ г) $f(x;\theta)=\frac{\theta^{x}e^{-\theta}}{x!}, x=0, 1, 2, \dots .$

10. 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 $f(k;p)=(1-p)p^{k-1}$, $k=1, 2, \dots$. Ако $X_1, X_2, \dots, X_n$ 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 $p(X>9)=\sum_{k=1}^{9}(1-p)p^{k-1}$.

11. Нека X има поместена експоненцијална распределба $f(x;a)=e^{\alpha-x}$, $x\ge a$. Врз база на примерок со големина n определи ги оценувачот со максимална подобност за параметарот a.

12. Нека примерокот $X_1, X_2, \dots, X_n$ е земен од популација со поместена експоненцијална распределба со густина

$$f(x;\lambda,a)=\begin{cases}\lambda e^{-\lambda(x-a)}, & x\ge a\\ 0, & \text{во спротивно}\end{cases}.$$

    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.

13. Екологист избира n дисјунктни региони $R_1, R_2, \dots, R_n$ 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.

5

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 $\overline{x} = 6.4$ литри (на 100 километри). Поради варијабилноста на примерокот, практично никогаш нема да се добие $\mu = \overline{x}$, а самата точкаста оценка $\overline{x}$ 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 5.1 Нека $L_1 = L_1(X_1, X_2, \dots, X_n)$ и $L_2 = L_2(X_1, X_2, \dots, X_n)$ се две статистики од примерокот $X_1, X_2, \dots, X_n$ 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;

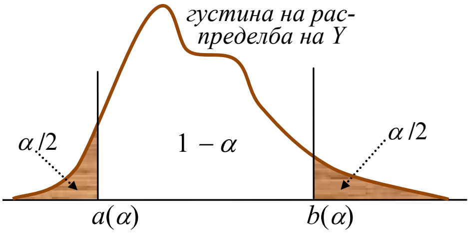
3) За дадена доверба, јасно е дека најдобар интервал е најтесниот, а бидејќи ширината на интервалот $L = L_2 - L_1$ е случајна променлива, логично би можело да се бара "минималната очекувана ширина" како оптимум. Проблемот е што таков минимум не мора да постои за сите вредности на θ.

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

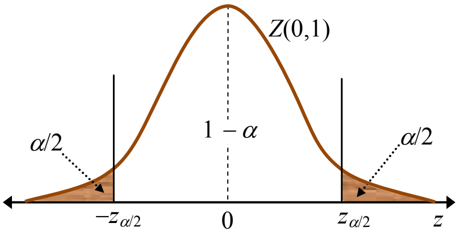
$p\left(-z_{\alpha/2} < \dfrac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) = 1-\alpha$, што понатаму директно дава

$p(\hat{\theta}-z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta}+z_{\alpha/2}\sigma_{\hat{\theta}}) = 1-\alpha$ – 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 $S_{\hat{\theta}}$.

5.1. 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) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ има нормална распределба $Z(\mu, \frac{\sigma}{\sqrt{n}})$. 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 $\overline{x}$.

ПРИМЕР 5.1 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 $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$, добиваме

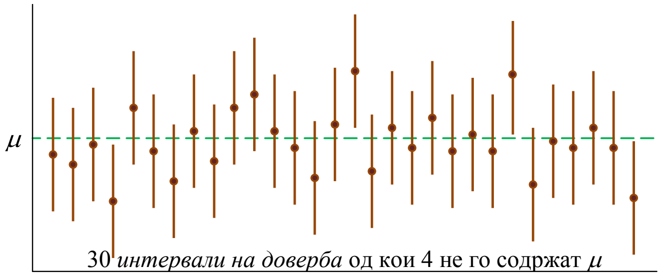
$$p\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = p\left(118.6 - 1.96\frac{25}{\sqrt{28}} < \mu < 118.6 + 1.96\frac{25}{\sqrt{28}}\right)$$

$$= p(109.34 < \mu < 127.86) = p(\mu \in [109.34, 127.86]) = 0.95.$$

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 $2 \cdot 1.96 \cdot 25 / \sqrt{n} \leq 10$, што дава неравенка по n, т.е. $\sqrt{n} \geq 2 \cdot 1.96 \cdot 25/10 = 9.80 \Rightarrow n \geq 96.04$, 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.

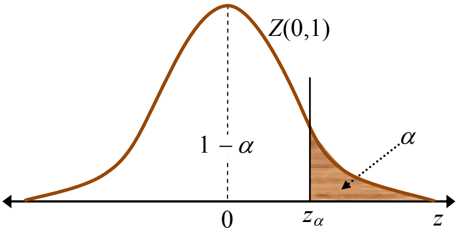


30 *интервали на доверба* од кои 4 не го содржат μ

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 – α (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 –∞. 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 еднострани 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 познато. 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 5.2 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

$$p\left(0.5250-2.5758\frac{0.3486}{\sqrt{53}}<\mu<0.5250+2.5758\frac{0.3486}{\sqrt{53}}\right)=$$

$$=p(0.4017<\mu<0.6483)=$$

$$=p(\mu\in(0.4017,0.6483))=0.99.$$

За ширина од 10%, $n \geq (2\cdot 2.5758\cdot 0.3486/(0.0525/2))^2 \approx 706$. ■

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 $\sum(\bar{X}-X_i)=0$. Сега случајната променливата $T = Z\sqrt{n}/\sqrt{\chi^2}$ $=$ $\dfrac{X-\mu}{\sigma/\sqrt{n}}\sqrt{n-1}/\sqrt{(n-1)S^2/\sigma^2}$ $=\dfrac{X-\mu}{S}\sqrt{n}$ ќе има студентова распределба со $n-1$ степени на слобода. Оттука веднаш следува

$$p\left(-t_{\alpha/2}<\frac{\bar{X}-\mu}{S/\sqrt{n}}<t_{\alpha/2}\right)=1-\alpha\text{, што дава интервален оценувач}$$

$$p\left(\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}}<\mu<\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right)=1-\alpha\ .$$

ПРИМЕР 5.3 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 $t_{\alpha/2}\frac{8.94}{\sqrt{14}} \le 3$, т.е. $t_{\alpha/2} \le 3\frac{\sqrt{14}}{8.94} = 1.2555.$

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). ■

5.2. 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 $\bar{X}$, од што следува дека грешката на предвидувањето е $\bar{X} - X_{n+1}$. Очекувана вредност на грешката е $E(\bar{X} - X_{n+1}) = E\bar{X} - EX_{n+1} = \mu - \mu = 0$. Бидејќи X_{n+1} е независна од $X_1, X_2, \ldots, X_n$, таа е независна и од $\bar{X}$, па дисперзијата на грешката на предвидувањето е

$$D(\bar{X} - X_{n+1}) = D\bar{X} + DX_{n+1} = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right).$$

Грешката на предвидувањето $\bar{X} - X_{n+1}$, 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 5.4 Количеството маснотија во примерок од 10 сендвичи со виршла 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 $\overline{x} = 21.90$ 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 (μ – $z_{\alpha/2}\sigma$, $\mu + z_{\alpha/2}\sigma$). ■

Интервалите на доверба за μ во случај на мал примерок се базирани на студентовата распределба и не се многу поуздани при отстапувања од нормалната распределба. Ако n е мало и распределбата на популацијата е "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].

5.3. 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 со закон $p(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ 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

$p\in\dfrac{\hat{P}+z_{\alpha/2}^2/2n}{1+z_{\alpha/2}^2/n}\pm z_{\alpha/2}\dfrac{\sqrt{\hat{P}(1-\hat{P})/n+z_{\alpha/2}^2/4n^2}}{1+z_{\alpha/2}^2/n}$. Во пракса, сметаме дека за доволно големо n, $z_{\alpha/2}^2/n$ 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 5.5 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

$$\hat{p}=\frac{m}{n}=\frac{58}{90}=0.64\,,\ \ z_{\alpha/2}=z_{0.025}=1.96\ ,\ \text{па}$$

$$p\left(0.64-1.96\sqrt{\frac{0.64\cdot 0.36}{90}}<p<0.64+1.96\sqrt{\frac{0.64\cdot 0.36}{90}}\right)=p(0.541<p<0.739)=$$

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,

$$p\left(p>\hat{P}-z_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)=1-\alpha\quad \text{и}\quad p\left(p<\hat{P}+z_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)=1-\alpha\,.$$

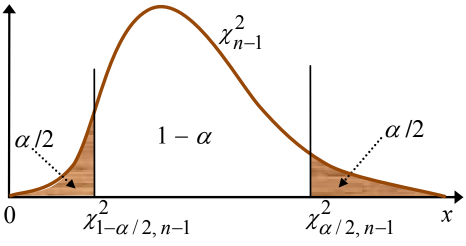
5.4. Интервални оценки за дисперзијата

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

$$\frac{(n-1)S^2}{\sigma^2}$$ 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

$$p\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha .$$ За оценка на стандардната девијација σ, треба само да се коренуваат двете страни на неравенството.

ПРИМЕР 5.6 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 5.7 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 $t_{\alpha,n\text{-}1}$ 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 милиметри 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 миљи 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 пакoвка 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?

6

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 $h(X_1, X_2, \ldots, X_n)$. Просторот на примерокот V го делиме на две дисјунктни множества A и B = V – A. Ако вредноста на статистиката $q = h(x_1, x_2, \ldots, x_n) \in$ 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 $p(q \in \mathrm{B} \mid H_0) = 0$ (никогаш не се отфрла H_0 кога таа е точна), и $p(q \in \mathrm{A} \mid H_A) = 0$ (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:

$p(q \in \mathrm{B} \mid H_0) \leq \alpha$, 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

$p(q \in \mathrm{A} \mid H_A) = \beta$, каде што β се нарекува грешка од тип 2, и ја дава веројатноста на прифаќање на H_0 кога таа не е точна. Вредноста $1 - \beta = p(q \in \mathrm{B} \mid H_A)$ 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.

6.1. 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 6.1 (Нојман-Пирсон)**.** Ако постои област B во R^n и број c таков што

$$\left(\frac{\prod_{i=1}^{n} f(x_i,\theta_0)}{\prod_{i=1}^{n} f(x_i,\theta_1)} \right) \begin{matrix} < c \\ \geq c \end{matrix} \quad \text{кога } (x_1, x_2, \ldots, x_n) \begin{matrix} \in \mathrm{B} \\ \notin \mathrm{B} \end{matrix},$$

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 6.1 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 $f(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i - \mu)^2}$, што дава

$$\left(\frac{e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i-\mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i-\mu_1)^2}} \right) < c \quad \underset{\textit{и средување}}{\overset{\textit{со логаритмирање}}{\Rightarrow}} \quad \frac{1}{n}\sum_1^n x_i < \overset{\text{ново } c}{\boxed{\frac{2\sigma^2 \ln c + n(\mu_0^2 - \mu_1^2)}{2n(\mu_0 - \mu_1)}}} \text{ и сега}$$

$$p(\bar{x} < c / H_0) = \alpha \;\Rightarrow\; p\Big(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < \overbrace{\frac{c - \mu_0}{\sigma/\sqrt{n}}}^{z_\alpha} / H_0\Big) = \alpha \text{. Значи ако } \overbrace{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}^{\textit{статистика}} < z_\alpha,$$

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) $p(q \in \mathrm{B} \mid H_0 \text{ е точна}) \leq \alpha$; и

б) $p(q \in \mathrm{A} \mid H_A \text{ е точна}) = \beta(\theta)$ е минимална.

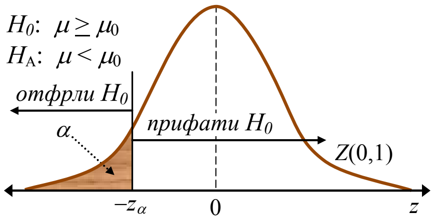
Секој параметарски тест може да се опише низ следните чекори:

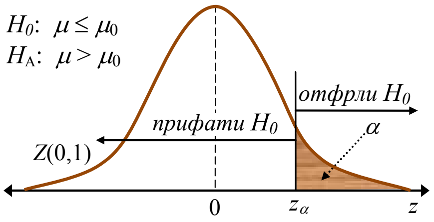
1. Идентификувај ги параметрите што се од интерес;
2. Определи ја нултата H_0, и алтернативната хипотеза H_A;
3. Избери ниво на значајност α и според него областите на отфрлање и порифаќање на H_0;
4. Определи ја статистиката за тестот, евентуалните непознати параметри и пресметај ја вредноста на статистиката;
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.

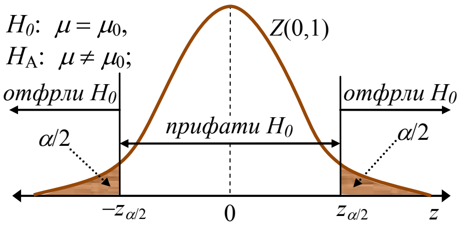
Забележи дека чекорите 1-3 може да се комплетираат пред обезбедувањето на примерокот. *P*-вредноста на тестот ќе ја дискутираме понатаму.

6.2. 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 $z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$ за познато σ, и $z = \dfrac{\overline{x} - \mu_0}{s / \sqrt{n}}$ за непознато σ (за $n \geq 30$, s е добра оценка за σ).

ПРИМЕР 6.2 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 $z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{131.08 - 130}{1.4 / \sqrt{9}} = 2.16$, а од таблицата за нормална распределба имаме $z_{0.005}$ = 2.96. Поради тоа што z =

2.16 $\in$ [–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

$1-\beta = p(q \in \mathrm{B} \mid H_\mathrm{A}) = p\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha \mid H_\mathrm{A}\right)$. Но $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ 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

$$p\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} \mid H_\mathrm{A}\right) = p\left(z < -z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right).$$

Така може да се дојде до изразите за β за различните случаи:

$$H_\mathrm{A}:\ p > p_0 \ \Rightarrow\ \beta = \Phi\left(z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

$$H_\mathrm{A}:\ p < p_0 \Rightarrow\ \beta = 1-\Phi\left(-z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

$$H_\mathrm{A}:\ p \neq p_0 \ \Rightarrow \beta = \Phi\left(z_{\alpha/2} + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

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,

$$\text{за } H_\mathrm{A}:\ \mu > \mu_0 \text{ и } H_\mathrm{A}:\ \mu < \mu_0 \ \Rightarrow\ n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0-\mu_1}\right)^2$$

$$\text{за } H_\mathrm{A}:\ \mu \neq \mu_0 \ \Rightarrow\ n = \left(\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0-\mu_1}\right)^2.$$

ПРИМЕР 6.3 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 $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{74.1 - 72}{13.3/\sqrt{50}} = 1.1648$, 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 $1 - \beta = p(h \in \mathrm{B} / H_A) = p\left(\frac{\bar{x} - 72}{13.3/\sqrt{50}} > 1.28 / H_A\right) =$ $p\left(\frac{\bar{x} - 78}{13.3/\sqrt{50}} > 1.28 + \frac{72 - 78}{13.3/\sqrt{50}} / H_A\right) = p(z > -1.91) = 0.9719$. 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, $1 - \beta =$ $p\left(z > 1.28 + \frac{72 - 74}{13.3/\sqrt{50}} / H_A\right) = p(z > 0.22) = 0.4129$, 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 $n = \left(\frac{13.3(1.28 + 1.28)}{72 - 74}\right)^2 \approx 290$. ■

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

$T = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$ 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 6.4 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 $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{9.15 - 8.5}{1.16 / \sqrt{20}} = 2.506$, 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 B (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".

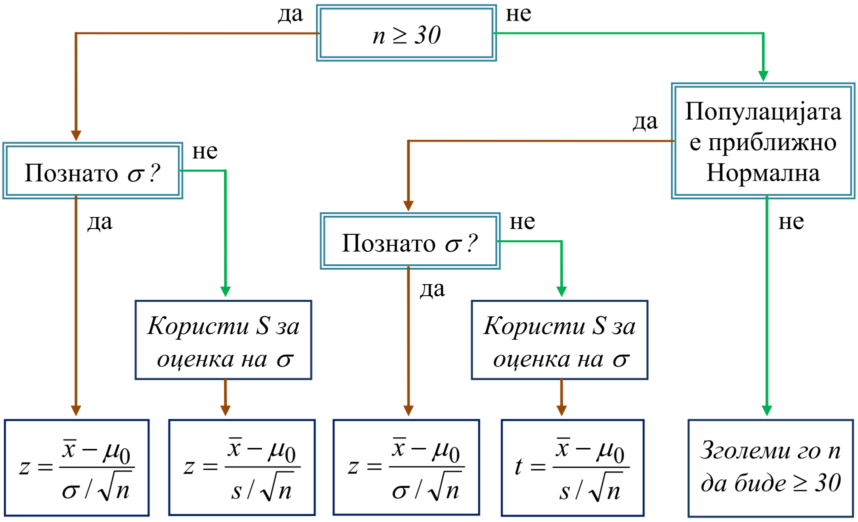
Јачината на тестот $1-\beta$ (веројатност за коректно отфрлање на H_0) е многу информативна и концизна мера на чуствителноста на тестот (способност да ги детектира разликите). Во случај на слаба јачина, единствено што останува е да се зголеми α или најдобро, големината на примерокот n.

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.



6.3. *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 6.5 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

$$P\text{-вредност} = p(z > 3.04 \mid H_0) = \text{плоштината под густината надесно од } 3.04$$
$$= 1 - \Phi(3.04) = 0.0012.$$

Кои P-вредности обезбедуваат доволно докази против H_0?

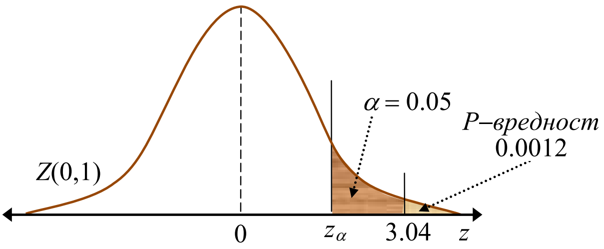
Кога P-вредноста = 0.0012 значи дека само 0.12% од сите можни вредности на статистиката се контрадикторни на H_0 најмалку онолку, колку и нашата статистика. Така, примерокот силно сугерира отфрлање на H_0.

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 $\leq \alpha$ *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 6.6 Ефикасноста на горивото (миљи по галон - mpg) варира од возило до возило за ист производител. Нека μ биде вистинскиот просек на ефикасноста 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. ■

Секоја статистика е случајна променлива па и P-вредноста е непрекината случајна променлива што зема вредности во интервалот [0, 1] (е веројатност). Овој пример 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).

6.4. 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

$$p(\text{грешка од тип 1}) = p(H_0 \text{ е отфрлена кога е точна}) =$$
$$= p(Z \geq z_\alpha \text{ кога } Z \text{ има приближно } Z(0,1) \text{ распределба}) \approx \alpha.$$

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: $p < p_0$	$z \leq z_\alpha$
H_A: $p > p_0$	$z \geq z_\alpha$
H_A: $p \neq p_0$	$z \leq -z_{\alpha/2}$ или $z \geq z_{\alpha/2}$

при што n треба да е доволно големо, т.е. $n \cdot p \geq 10$ и $n \cdot (1 - p) \geq 10$.

ПРИМЕР 6.7 Тапите од плута кај шишињата вино се подложни на деградација што води до намалување на квалитетот на виното. Во еден чланак 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 $n \cdot p_0 = 91 \cdot 0.15 \geq 10$ и $n \cdot (1 - p_0) = 91 \cdot 0.85 \geq 10$ користиме стандарден z тест. Од $\hat{p} = 16/91 = 0.1758$, за вредноста на статистиката добиваме

$$z = \frac{0.1758 - 0.15}{\sqrt{0.15 \cdot 0.85/91}} = 0.6898 < 1.28 = z_{0.1}, \text{ и така } H_0 \text{ не се отфрла. } \blacksquare$$

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,

$EZ = \dfrac{p_1 - p_0}{\sqrt{p_0(1-p_0)/n}}$, $\quad DZ = \dfrac{p_1(1-p_1)}{p_0(1-p_0)}$. 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 6.8 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.

Од $z_\alpha = z_\beta = z_{0.01} = 2.33 \Rightarrow n = \left(\dfrac{2.33\sqrt{0.9\cdot 0.1}+2.33\sqrt{0.8\cdot 0.2}}{0.8-0.9}\right)^2 \approx 266.$ ■

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 $\alpha = p(X \le c \mid H_0) = Bin(c;\, p_0,\, n)$ = 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 6.9 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. ■

6.5. 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:

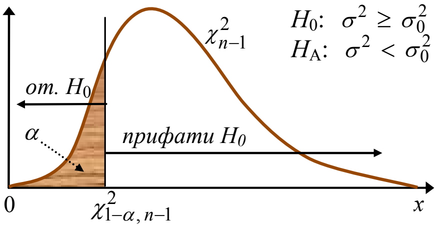
H_0: $\sigma^2 \geq \sigma_0^2$, наспроти H_A: $\sigma^2 < \sigma_0^2$,

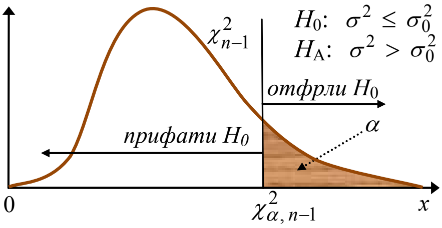
H_0: $\sigma^2 \leq \sigma_0^2$, наспроти H_A: $\sigma^2 > \sigma_0^2$,

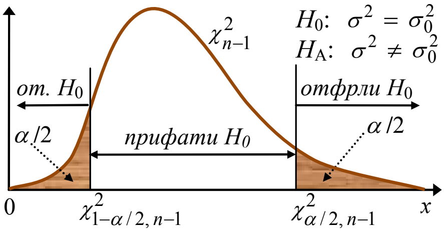
H_0: $\sigma^2 = \sigma_0^2$, наспроти H_A: $\sigma^2 \neq \sigma_0^2$.

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 6.10 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 χ_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

$$1-\beta = p(\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha,n-1} \,|\, H_A) = p(\frac{(n-1)S^2}{\sigma_1^2} > \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{\alpha,n-1}) = (\chi^2 > \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{\alpha,n-1})$$

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 $n = 61$.

6.6. Статистичка или 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 отстапуцање 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.

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: $\lambda \le 0.01$, 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 $\leq$ 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 α = 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.

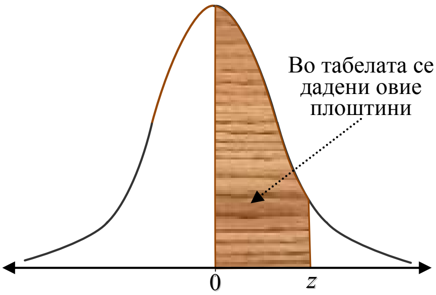
13. Производител на автомобилски гуми го испитува животниот век на гуми со нов тип каучук. За таа цел се направени 16 гуми и тестирани на пат, при што е добиен просечен животен век од 60139.7 километри со стандардна девијација од 3645.94 километри.

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.

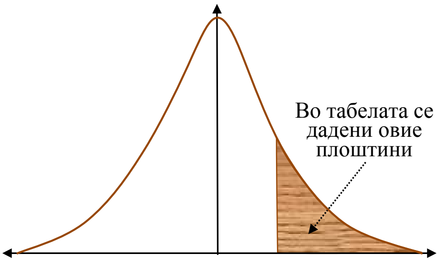
Табели на распределби

Табела на стандардна нормална распределба



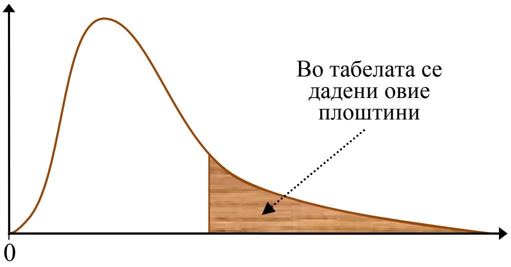
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	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. Податоците ги имаме групно, па работиме со средините на интервалите:

x	$x \le 100$	$100 < x \le 150$	$150 < x \le 200$	$200 < x \le 250$	$250 < x \le 300$	$300 < x \le 350$	$350 < x \le 400$	$400 < x \le 450$	$450 < x \le 700$	$x > 700$
$F_n(x)$	0	1/40	6/40	15/40	21/40	29/40	35/40	37/40	39/40	1

2. Точно ... .
3. Примерок со големина n се n случајни променливи $X_1, X_2, \dots, X_n$ или идентично, случаен вектор $(X_1, X_2, \dots, X_n)$. Реализација на примерок се n конкретни вредности $x_1, x_2, \dots, x_n$ на случајните променливи $X_1, X_2, \dots, X_n$ (x_j е една вредност на случајната променлива X_j), т.е обичен вектор на броеви $(x_1, x_2, \dots, x_n)$.
4. Распределба на примерок е заедничката распределба на случајни променливи $X_1, X_2, \dots, X_n$ или идентично, распределба на случајниот вектор $(X_1, X_2, \dots, X_n)$.
5. Кај набљудуваните податоци немаме никакво влијание.
6. Сѐ што имаме е примерок и знаење од веројатност и статистика. Ако дополнително ја имаме (ја добиеме барем приближно) и распределбата, имаме сѐ што ни треба.
7. 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 3

1. а) $\overline{x} = 7.184$, $s^2 = 0.0004268$, $s = 0.02066$;   б) 7.18, 7.20.
2. а) $\overline{x} = 65.85$, $s = 12.16$, 58;   б) Цртеж;   в) $\overline{x} = 66.86$, $s = 10.74$, 60.

3. а) $\bar{x} = 89$, $s = 2.8$, 90;

   б) 22/40.

4. Блиска до ѕвонест облик (нормален) со очекување ≈ 135 и релативно значајна дисперзија.

5. 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].

6. 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

Кај ваквите $\ln(x)$ вредности имаме многу поголема симетрија и помала варијабилност. Исто така има помали празнини низ податоците.

5.25 - 5.74	4	0.08
5.75 - 6.25	3	0.06

г) 0.38, 0.14.

7. Да.

8. Да.

9. а) 5.42, $\sqrt{0.33} = 0.57$, 5.46;

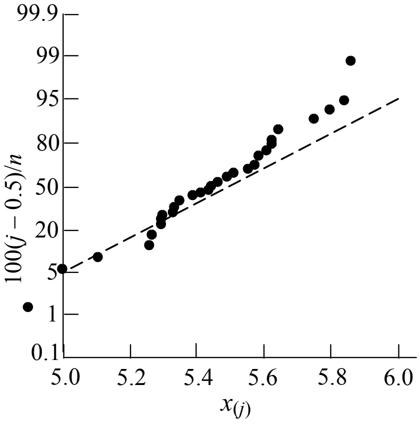
   б) Веројатно да, бидејќи податокот 4.07 се чини нереално мал, а тој влегува во просекот но не и во медијаната (денеска знаеме дека вистинската вредноста е приближно 5.52);

   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 4

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. Не е центриран. Кога $n \to \infty$ е центриран.

7. а) Следува од $EX_i = n_i p_i$, за $i = 1, 2$; б) $\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}$;
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.

8. а) $\hat{\theta} = \sum_{j=1}^{n} X_j^2 / (2n)$; б) 74.505.

9. а) θ^2/n; б) θ^2/n; в) $\theta(1-\theta)/n$; г) θ/n.

10. а) $1 - \frac{1}{\overline{X}}$; б) $1 - \sum_{k=1}^{n} \frac{1}{\overline{X}} \left(1 - \frac{1}{\overline{X}}\right)^{k-1}$

11. $X_{(1)} = \min(X_i)$, $\overline{X} - 1$.

12. $\hat{\lambda} = \dfrac{n}{\sum (X_i - \min(X_i))}$, 0.202.

13. 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 $\lambda = \dfrac{\sum x_i}{\sum a_{R_i}}$, т.е. оценувач $\hat{\lambda} = \dfrac{\sum \xi_i}{\sum a_{R_i}}$.

Глава 5

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 6

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.

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.