Лекция 17

§17. Случайни извадки. Оценка на параметрите

1. Основни определения. Изходен материал за статистическите изследвания се явява съвкупността от резултатите от някакви наблюдения, които се оформят като таблици от данни. Статистиката представлява синтетична математическа дисциплина, основана върху теорията на вероятностите. Статистиката борави с привидно елементарни понятия, които могат да бъдат усвоени на сравнително добро ниво на абстракция. Изложението на материала ще бъде направено въз основа на подробен анализ на различни примери.

Пример 17.1. Извадка от $n=175$ на брой курсанти е подложена на психологическо изследване посредством тест за установяване нивото на положително отношение към армията. Отчитането на резултатите става въз основа на суровия бал от теста, при което са получени следните резултати.

105, 85, 113, 91, 137, 119, 111, 127, 70, 115, 90, 117, 129, 119, 128, 133, 136, 121, 116, 134, 109, 117, 138, 76, 102, 103, 107, 103, 112, 120, 97, 86, 96, 95, 119, 141, 131, 119, 131, 106, 115, 110, 84, 125, 111, 124, 136, 142, 106, 116, 107, 106, 113, 125, 129, 121, 145, 91, 100, 120, 150, 113, 150, 118, 122, 139, 132, 106, 120, 116, 123, 135, 131, 115, 107, 86, 129, 133, 111, 126, 129, 132, 144, 93, 131, 110, 127, 137, 135, 124, 129, 144, 128, 102, 129, 122, 115, 84, 120, 117, 86, 88, 142, 138, 62, 81, 128, 127, 72, 111, 123, 99, 126, 68, 88, 105, 133, 87, 130, 128, 134, 108, 127, 110, 120, 98, 100, 126, 107, 91, 94, 124, 96, 95, 90, 101, 137, 103, 102, 108, 125, 118, 124, 142, 79, 103, 143, 119, 130, 126, 100, 141, 102, 122, 106, 118, 138, 124, 110, 105, 65, 140, 100, 104, 116, 85, 117, 102, 122, 99, 125, 126, 129, 105, 130

Тези резултати, които се наричат още ***наблюдения*** и се бележат например с x_1, x_2, ..., x_n, ще бъдат подложени на типичен статистически анализ, в основата на който стои изграждането на подходящ вероятностен модел. За тази цел ще въведем случайна величина X, обвързана с тестовите резултати, а всеки отделен резултат x_k, $k=1,2,\ldots,n$, ще разглеждаме като ***случайна реализация*** на X. Всяко наблюдение x_k всъщност се разглежда като реализация на случайната величина X_k, която се разглежда като идентично копие на X. Основната задача тук е определяне вида на разпределението на X въз основа на данните от експеримента. Ще предполагаме, че X следва някакво ***нормално разпределение*** със средно μ и дисперсия σ^2, $X\in N\left(\mu,\sigma^2\right)$. (Такова предположение се оказва добре обосновано независимо някои уточнения, които трябва да бъдат направени във връзка със стойностите на X, които имат дискретен характер.) Следователно всяка величина X_k има плътност на разпределение

$$f\left(x_k\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(x_k-\mu\right)^2}{2\sigma^2}},\ k=1,2,\ldots,n.$$

По този начин резултатът от експеримента може да бъде разглеждан като един ***единствен*** изход от наблюдение над случайния вектор $\mathbf{X}=\left(X_1,X_2,\ldots,X_n\right)$ с възможни изходи $\mathbf{x}=\left(x_1,x_2,\ldots,x_n\right)$. Резултатите от отделните наблюдения се разглеждат като ***независими***, следователно случайните величини X_k са независими по съвкупност и тяхната ***съвместна плътност*** $f\left(\mathbf{x}\right)$ се явява ***произведение*** от отделните плътности

$$(17.1)\quad f\left(\mathbf{x}\right)=\prod_{k=1}^{n}f\left(x_k\right)=\frac{1}{\sigma^n}\frac{1}{\left(\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left(x_k-\mu\right)^2}.$$

За да избегнем употребата на много символи, възможните стойности на X_k и техните конкретни стойности от наблюденията се означават по един и същи начин x_k. (В някои

източници даже не се прави разлика в означенията за случайна величина, нейните възможни стойности и конкретни реализации.) Съвместната плътност от (17.1), когато се разглежда като функция на параметрите μ и σ се нарича ***функция на правдоподобие*** (***likelihood function***) и се бележи с $L = L(\mu, \sigma)$. Нейният логаритъм $l = \ln L$ се нарича ***логаритмична функция на правдоподобие (logarithmic likelihood function)***. По този начин имаме

(17.2) $$l = \ln L = \ln f(\mathbf{x}) = -\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - \mu)^2 - n\ln\sigma - \frac{n}{2}\ln 2\pi .$$

Оценяването на параметрите μ и σ^2 се извършва обикновено следвайки ***принципа на максимално правдоподобие***, според който параметрите се избират с оглед достигане максимума на функцията на правдоподобие L, което е все едно да бъдат избрани да максимизират логаритмичната функция на правдоподобие l.

Принципът на максимално правдоподобие отразява философията, че в природата се реализират най-вероятните възможности и в този смисъл се явява самоочевиден. Съществуват и други подходи за оценяване на параметрите, като например метода на моментите. Оценки от този вид се наричат ***точкови оценки (point estimate)*** за параметрите.

За да намерим максимума на $l = l(\mu, \sigma)$ анулираме двете частни производни

(17.3) $$\frac{\partial}{\partial\mu} l = 0 \text{ и } \frac{\partial}{\partial\sigma} l = 0 ,$$

които уравнения се наричат ***уравнения на максимално правдоподобие***. Уравненията (17.3) имат следния конкретен вид

$$-\frac{1}{2\sigma}\sum_{k=1}^{n}(x_k - \mu) = 0 \text{ и } \frac{1}{2\sigma^3}\sum_{k=1}^{n}(x_k - \mu)^2 - \frac{n}{\sigma} = 0 .$$

Първото има решение

(17.4) $$\mu = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} ,$$

след което за второто намираме

(17.5) $$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} .$$

За оценките на максимално правдоподобие има запазено означение с шапка над символа на параметъра, а $\bar{x}$ е запазено означение за аритметично средно. Сега (17.4) и (17.5) дават основание да запишем следните максимално правдоподобни оценки за средното μ и дисперсията σ^2

(17.6) $$\hat{\mu} = \frac{X_1 + X_2 + \cdots + X_n}{n} ,$$

(17.7) $$\hat{\sigma}^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n} ,$$

където

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

се нарича ***средно на извадката (sample mean)***. Поради причини, които ще изложим по-нататък, вместо оценката (17.7) за дисперсията се предпочита следната оценка

(17.8) $$s^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} ,$$

която се нарича ***дисперсия на извадката (sample variance)***. От формулите (17.7) и (17.8) се вижда, че при достатъчно големи n, различието между $\hat{\sigma}^2$ и s^2 е несъществено.

Като заместим във формулите с конкретните стойности от наблюденията се получават техните ***стойности от експеримента***. За примера (17.1) се получава

$\hat{\mu} = 114.726$ и $s^2 = 335.269$.

Формулите (17.6-8) представляват частни случаи на ***статистики***, общото определение на които ще дадем след малко.

По този начин за случайната величина X от пример 17.1 определихме основен вероятностен модел, според който X се подчинява на нормално разпределение със средно 114.726 и дисперсия 335.262.

Анализираният пример показва някои от основните понятия, свързани със статистическия анализ и дава основание за следните определения.

Определение 17.1. Нека X_1, X_2, ..., X_n са n на брой независими по съвкупност и еднакво разпределени случайни величини с плътност $f(x)$. Тогава се казва, че X_1, X_2, ..., X_n представляват ***проста случайна извадка*** (***i.i.d. sample***) с обем n от разпределение с вероятностна плътност $f(x)$.

На практика понятието извадка се употребява и за означаване съвкупността от стойностите на наблюденията.

Неизвестните параметри на разпределението $f(x)$ се оценяват посредством функции от величините на извадката.

Определение 17.2. ***Статистика*** се нарича функция на случайните величини от извадката, която функция не зависи от параметри.

Нека X е дискретна случайна величина с разпределение $P(X = x_j) = p_j$. В този случай е удобно да говорим за ***плътност*** на разпределение $f(x)$, определена само върху възможните стойности на X по формулата $f(x_j) = p_j$. Тази уговорка позволява единни означения за непрекъснатите и дискретните величини.

Следващият пример е свързан с повтарящи се опити, при всеки от които се сбъдва някакво събитие с вероятност p или неговото противоположно с вероятност $q = 1 - p$. Такива опити разгледахме при схемата на Бернули. Тук се определя естествено дискретна случайна величина X, приемаща стойности 1 или 0 съответно при сбъдване или не на въпросното събитие. Тази величина има разпределение $P(X = 1) = p$ и $P(X = 0) = 1 - p$. Нейната плътност се задава чрез формулата

(17.9) $f(x) = p^x (1-p)^{1-x}$, $x = 0,1$.

Пример 17.2. Нека X_1, X_2, ..., X_n е случайна извадка от разпределение (17.9). Тогава величината X_k, $k = 1,2,\ldots,n$, има плътност

$$f(x_k) = p^{x_k} (1-p)^{1-x_k},$$

където x_k е означение за възможните стойности на X_k, които са 1 или 0. Тук функцията на правдоподобие има вида

$$L = \prod_{k=1}^{n} f(x_k) = p^{\sum_{k=1}^{n} x_k} (1-p)^{n - \sum_{k=1}^{n} x_k},$$

а логаритмичната функция на правдоподобие е

$$l = \ln p \sum_{k=1}^{n} x_k + \ln(1-p)\left(n - \sum_{k=1}^{n} x_k \right).$$

Оценка на максимално правдоподобие за параметъра p се получава след решаване на уравнението на правдоподобие

$$\frac{\partial l}{\partial p}=0\,,\ \frac{1}{p}\sum_{k=1}^{n}x_k-\frac{1}{1-p}\left(n-\sum_{k=1}^{n}x_k\right)=0\,,$$

чието решение е

$$p=\frac{x_1+x_2+\cdots+x_n}{n}\,.$$

Следователно оценката на максимално правдоподобие $\hat{p}$ се задава от статистиката

(17.10) $$\hat{p}=\frac{X_1+X_2+\cdots+X_n}{n}\,.$$

Последната формула всъщност определя $\hat{p}$ като относителния брой на сбъдвания на събитието, което лежи в основата на единичния опит.

Определение 17.3. Казва се, че статистиката $\varphi(X_1,X_2,\ldots,X_n)$ представлява ***неизместена (unbiased)*** точкова оценка за параметъра θ, когато нейното математическо очакване е равно на θ, т.е. $\mathbf{E}[\varphi(X_1,X_2,\ldots,X_n)]=\theta$. В противен случай статистиката се нарича изместена оценка.

Пример 17.3. Нека X_1, X_2, ..., X_n е случайна извадка от нормално разпределение $N(\mu,\sigma^2)$. Тогава статистиката (17.6) представлява неизместена оценка за средното μ, понеже

$$\mathbf{E}[\overline{X}]=\mathbf{E}\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]=\frac{1}{n}\left(\mathbf{E}[X_1]+\mathbf{E}[X_2]+\cdots+\mathbf{E}[X_n]\right),$$

$$\mathbf{E}[\overline{X}]=\frac{1}{n}(\mu+\mu+\cdots+\mu)=\mu\,.$$

Статистиката (17.7) обаче се оказва изместена оценка за дисперсията σ^2, понеже пресмятанията показват

$$\mathbf{E}\left[\frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n}\right]=\frac{n}{n-1}\sigma^2\,.$$

Последното именно представлява причината, поради която за оценка на дисперсията се предпочита неизместената оценка s^2 от формулата (17.8).

Пример 17.4. Нека X_1, X_2, ..., X_n е случайна извадка разпределение на Поасон с параметър $\lambda>0$, което има плътност

$$f(x)=P(X=x)=e^{-\lambda}\frac{\lambda^x}{x!}\,,\ x=0,1,2,\ldots.$$

Функцията на правдоподобие има вида

$$L=\prod_{k=1}^{n}f(x_k)=e^{-n\lambda}\frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1!x_2!\cdots x_n!}\,,$$

а логаритмичната функция на правдоподобие $l=\ln L$ има вида

$$l=-n\lambda+(x_1+x_2+\cdots+x_n)\ln\lambda-\ln x_1!x_2!\cdots x_n!\,,$$

откъдето веднага намираме, че средното аритметично представлява оценка на максимално правдоподобие за параметъра λ,

$$\hat{\lambda}=\frac{X_1+X_2+\cdots+X_n}{n}\,,$$

която оценка освен това е неизместена.

2. Извадка и популация. В статистическите изследвания важна роля играе понятието ***популация*** (***population***) и връзката между извадка и популация. Данните обикновено се получават в резултат на наблюдения на определен брой ***статистически единици***, които например в социологията представляват отделни индивиди, в икономиката различни икономически субекти и т.н.. Групата от статистически единици, над които е извършено наблюдението образува самата извадка. Извадката се разглежда като част от по-голяма група, наречена ***генерална съвкупност*** или още ***популация***. Физическата популация е съвкупността от реално съществуващи статистически единици, които са хомогенни относно наблюдаваните величини. Ролята на популацията в статистическите изследвания е твърде специфична както от гледна точка на предметната област на изследването, така и от гледна точка на статистическите процедури, които се прилагат за различните цели на анализа, и по тази причина детайлите в познавателното значение на понятието популация винаги се разкрива в определен контекст. Ще отбележим преди всичко, че популацията има винаги в някаква степен хипотетичен характер и методите, които ние ще използваме я разглеждат като концептуално безкрайна (даже когато физическата популация практически съвпада със самата извадка).

При всички ситуации в една или друга степен целта на статистическото изследване е да се направят изводи за състоянието на популацията според данните от извадката, което изисква извадката да отразява пропорционално основните черти на популацията. Такива извадки се наричат ***представителни***. В повечето случаи популацията е ясно дефинирана еднородна група от статистически единици по основен формален признак, което не изключва възможността вътре в популацията да се обособяват съществени еднородни групи по някакви други важни за изследването признаци. Например нека целта на изследването е характеризиране по някакви показатели на децата от средно училищна възраст в република България. В този случай популацията представлява напълно определена група деца на възраст между 11 и 15 години. Тази популация обаче съдържа обособени групи от момичета и момчета, както и обособени групи по териториален признак. Тук могат да се разгледат и други такива признаци. В този смисъл ***идентифицирането на популацията зависи много съществено от контекста на изследователската задача от етапа на нейното формулиране до етапа на интерпретация на резултатите***. Ако задачата е изследване отношението на горната популация към въвеждането на вечерен час, то е твърде правдоподобно да си мислим, че не съществуват съществени различия по пол и териториален признак. По този начин една представителна извадка от 300 души може да се образува на базата на случаен (непреднамерен) подбор от един достатъчно голям град. В статистиката понятията представителна и непреднамерена извадка са синоними. Непреднамереността означава, че изследователят не влага специална умисъл при съставянето на извадката.

Ако задачата се състои в изследване на евентуалните различия между момчетата и момичетата, то фактически става дума за сравняване характеристиките на две отделни популации посредством данните от две извадки – по една за популацията на момчетата и момичетата. В този случай и двете извадки трябва да бъдат представителни за своите популации.

Нека например целта на изследването е установяване отношението на населението на република България към Европейския Съюз. Тук популацията се състои от всички граждани в активна съзнателна възраст. Да предположим условно, че става дума за 6000000 лица. Понеже проблемът тук има преди всичко икономическа природа, редно е да разделим популацията на относително еднородни групи по икономически признак, които групи споделят приблизително едно и също отношение. Нека за

простота и определеност да предположим, че въпросната група съдържа 1000000 заети в частния бизнес, 2000000 заети в държавната администрация и 3000000 наемни работници. В такъв случай една представителна извадка от 600 души трябва да съдържа 100 случайно подбрани лица от сферата на частния бизнес, 200 случайно подбрани лица от държавната администрация и 300 случайно подбрани лица – наемни работници. Истинското решаване всъщност на последната задача изисква точни данни за структурата на популацията, както и по-съдържателно структуриране.

На практика липсата на явна преднамереност в дадена извадка се разглежда като достатъчен белег за нейната представителност и в много случаи това наистина е така. Както се вижда обаче от приведените примери, проблемът за представителността е по-сложен отколкото изглежда първоначално. Една извадка може да бъде представителна по отношение на дадена цел и да не бъде представителна по отношение на друга.

Основният поток на статистическите методи за анализ се отнася към физически популации с голямо количество единици, които популации формално се интерпретират като безкрайни. На практика безкрайността означава наличието на достатъчно много статистически единици в реално съществуващата физическа популация, например ако съдържа повече от 10000 такива.

В някои случаи популацията се характеризира със сравнително малко на брой физически представители, които могат да бъдат обхванати изцяло в дадено конкретно изследване, като по този начин извадката на практика съвпада с физическата популация. В тези случаи прилагането на статистически анализи, характерни за безкрайни популации не води до съществени промени в логиката на предметното разсъждение.

Всяка извадка е представителна за някаква хипотетична популация. От такава гледна точка въпросът за представителността се състои в това, доколко тази популация е добре идентифицирана, което обикновено предхожда статистическия анализ и преди всичко доколко след като бъде идентифицирана тя изобщо представлява интерес за изследване. Очевидно всяко сериозно приложно изследване трябва да започва с подробно характеризиране на своя обект и своите цели, а не да се търси после обект на приложения на моделите.

Представителността на извадката в повечето случаи представлява единственият спорен пункт на дадено статистическо изследване.

В отделни статистически изследвания смисълът на понятието популация е твърде отдалечен от неговото първоначално предназначение и се запазва само за съгласуване на терминологията.

Характеризирането на популацията въз основа на данните от извадката представлява главната цел на статистическия анализ. По тази причина параметрите, които описват теоретичното разпределение на извадката се наричат още популационни параметри в контраст с техните извадкови аналози, които се наричат ***статистики***. За примера 17.1 параметърът μ представлява популационното средно с извадков аналог $\bar{x}$, а σ^2 представлява популационната дисперсия с извадков аналог s^2. Аналогична е ситуацията и с другите възможни популационни параметри. С нарастване обема на извадката, извадковите стойности стават все по близки до "истинските" стойности на популационните параметри.

Терминът извадков параметър не е много удачен от методическа гледна точка понеже понятието параметър трябва да се отнася само към популацията.

3. Видове величини. Величините от интерес за изследователя могат да имат разнообразна природа. Според количествената информация, заложена в различните величини, е приета класификацията на Стивънс, при която условно различаваме четири типа величини и съответно четири типа скали за измерване.

Номинални величини. Такива са величините, при които имаме определен брой категории, които се различават чрез техните имена, като всеки индивид се отнася точно към една от тези категории. Например в социологическите изследвания, номинални са величините "пол" и "раса", величините, които показват религиозна или политическа идентификация към някоя обособена група. В този случай за величините се казва, че са измерени по ***номинална скала*** (скала на наименованията). Номиналните величини участват по естествен начин като фактори в дисперсионния анализ.

Ординални величини. При тях резултатите от наблюденията могат да се сравняват в термините повече или по-малко. Особеното при ординалните величини е, че за тях се предполага единствено възможността за сравнение между резултатите на различните статистически единици. В този случай за величините се казва, че са измерени по ***ординална скала***. Ординални са формално погледнато и всички величини, свързани с постижения по отделни учебни дисциплини, когато тези постижения са отчетени в петобалната система (слаб 2, среден 3, добър 4, много добър 5, отличен 6). Статистическите процедури, които се отнасят за чисто ординални величини работят фактически с ***ранговете*** на отделните статистически единици вместо с техните натурални стойности.

Интервални величини. Интервали са онези ординални величини, за които интервалите между отделните категории могат да се интерпретират. Въпросът дали дадена величина е интервална или ординална в много случаи е дискусионен, като в този случай критерият е наличието на достатъчна ***изменчивост*** (***вариабелност***) на величината (поне седем различни стойности), но всъщност основният критерий е достатъчният по големина обем на извадката и формата на разпределението, за което ще стане дума по-нататък. В този случай за величините се казва, че са измерени по ***интервална скала***.

Абсолютни величини. Абсолютни са тези интервални величини, при които има нулева точка на измерването. Всички величини, за които съществуват физически мерни единици имат абсолютен характер. Например различни физиологически показатели като тегло, ръст, кръвно налягане, време за психомоторни реакции и т.н. В този случай за величините се казва, че са измерени по абсолютна скала или още по ***скала на отношенията***.

Горната класификация отчита преди всичко в каква степен върху резултатите от даден тип величина ***могат да се извършват различни математически операции*** като сравняване и (претеглено) събиране. Категориите на чисто номиналните величини подлежат само на броене. Резултатите от наблюдението по ординална величина позволява извадката да се подреди (ранжира) според степента на притежаване на измерваната характеристика. Когато величината е интервална или абсолютна, могат да се пресмятат средни величини и дисперсии.

От процедурно-изчислителна гледна точка, величините могат да се разделят на три вида ***номинални***, ***ординални*** и ***метрични (скали)***. В този смисъл метрични са величините, определени по-горе като интервални и абсолютни. От гледна точка на количественото представяне, величините се делят условно на ***непрекъснати*** и ***дискретни***. Дискретни са тези величини, които могат да приемат предварително известни фиксирани стойности. Непрекъснатите величини могат да приемат стойности от цял интервал. Някои по природа метрични величини след наблюдение се отразяват като дискретни, което не променя съществено тяхната природа. Номиналните и ординалните величини се наричат понякога ***категорийни***.

Статистическият анализ, който ще разглеждаме по-нататък се отнася преди всичко за метрични величини. За тях е разработен съдържателен и сложно структуриран математически апарат. За категорийните величини също е разработен

съдържателен математически апарат за търсене на статистически връзки с други величини, категорийни или метрични.

4. Емпирични характеристики. В този раздел ще съсредоточим вниманието върху стойностите от наблюденията x_1, x_2, ..., x_n, получени в резултат на проста случайна извадка на непрекъсната случайна величина X с плътност $f(x)$. Ако подредим наблюденията във възходящ ред $\xi_1 \le \xi_2 \le \cdots \le \xi_n$ ще получим ***вариационния ред*** на извадката. Вариационният ред може да бъде обработен по различни начини. Например функцията $F_n(x)$, определена както следва

$$F_n(x) = \begin{cases} 0 \text{ за } x \le \xi_1 \\ \dfrac{k}{n} \text{ за } \xi_k < x \le \xi_{k+1} \\ 1 \text{ за } x > \xi_n \end{cases}$$

се нарича ***емпирична (извадкова) функция на разпределение*** за величината X. Функцията $F_n(x)$ представлява всъщност истинска функция на разпределение за случайна величина, разпределена равномерно върху стойностите x_k, $k = 1,2,\ldots,n$.

Нека (a,b) е интервал, който съдържа всичките наблюдения. Да разделим този интервал на m части с помощта на междинни точки $a = c_0 < c_1 < c_2 < \cdots < c_{m-1} < c_m = b$ и нека n_j (***observed frequencies***) означава броят на наблюденията попадащи в интервала $(c_{j-1}, c_j]$, $j = 1,2,\ldots,m$. Очевидно $n_1 + n_2 + \cdots + n_m = n$. Обикновено интервалите се избират с ***равни*** дължини. Сега ако над всеки интервал Δ_j издигнем стълб с височина, равна на n_j, ще получим фигура, която се нарича ***хистограма*** (***histogram***) на наблюденията. Ако изберем височината на стълбовете да бъде

$$\frac{n_j}{n}\frac{1}{c_j - c_{j-1}},\ j = 1,2,\ldots,m,$$

то горните страни на стълбовете образуват графика на функция, която удовлетворява условията за плътност на разпределение, което представлява някакво приближение на $f(x)$. Изследователят притежава свобода при избора на вида и броя на интервалите.

За примера 17.1 имаме следната таблица на честотите.

Таблица 17.1.

интервал	наблюдавани честоти	натрупани наблюдавани честоти	очаквани честоти	натрупани очаквани честоти	разлики
x<= 66.7	2	2	1	1	1
66.7<x<=73.3	3	5	1	2	2
73.3<x<=80.0	2	7	3	5	-1
80.0<x<=86. 7	8	15	6	11	2
86. 7<x<=93.3	9	24	10	21	-1
93.3<x<=100.0	13	37	16	37	-3
100.0<x<=106. 7	20	57	21	58	-1
106. 7<x<=113. 3	19	76	24	82	-5
113. 3<x<=120.0	25	101	25	107	-0
120.0<x<=126. 7	22	123	23	130	-1
126. 7<x<=133.3	27	150	18	148	9

133.3<x<=140.0	14	164	12	160	2
140.0<x<=146. 7	9	173	8	168	1
146. 7<x<=153.3	2	175	4	172	-2

Въз основа на таблицата получаваме следната хистограма.

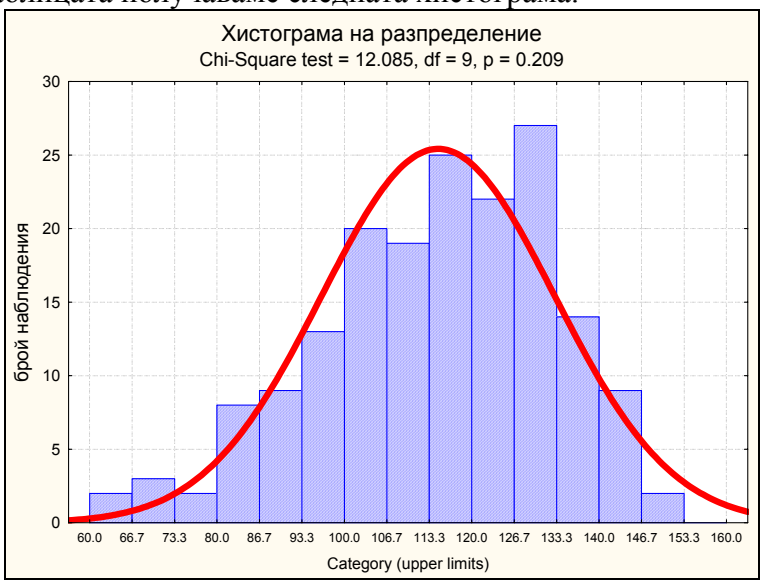

Рис. 17.1.

Тук върху хистограмата е изчертана и кривата на нормално разпределение, уточнена в предишния раздел. ***Очакваните честоти*** (***expected frequencies***) се получават като се умножи площта над съответния интервал, разположена под нормалната крива, по броя на всичките наблюдения. Доброто съвпадение между наблюдавани и очаквани честоти представлява белег в подкрепа на предложението за нормално разпределение на наблюдаваната величина.

От формална гледна точка предположението, че плътността на разпределение на извадката има даден конкретен вид $f(x)$ (за случая на непрекъснати величини) изцяло се покрива с това, че с нарастване обема на извадката n, относителната честота на наблюденията, попаднали в някакъв (кой да е) зададен интервал (a,b), се стреми към съответната площ под кривата на плътност, т.е.

$$\lim_{n\to\infty}\frac{\nu}{n}=\int_a^b f(x)dx\,,$$

където ν е наблюдаваната честота за интервала (a,b), а n е обемът на извадката.

Съгласно една ***теорема на Гливенко-Кантели***, ако $F(x)$ е истинската функция на разпределение на случайната величина X, а $F_n(x)$ е емпиричната функция на разпределение на случайна извадка от n наблюдения, то с нарастването на обема на извадката n вероятността за отклонение на $F_n(x)$ от $F(x)$ клони към нула.

Следващата диаграма представлява хистограма с натрупване – ***кумулативна хистограма***.

Рис. 17.2.

Обикновената хистограма показва приблизителната форма на плътността на разпределение, докато кумулативната хистограма показва приблизителната форма на функцията на разпределение.

Тук допълнително е проведен χ^2***-тест за нормалност***, който дава добри статистически аргументи в полза на хипотезата за нормално разпределение на наблюдаваната величина.

Разликата между най-малкото наблюдение $x_{\min}$ и най-голямото $x_{\max}$ се нарича ***размах*** (***range***) на извадката. За пример 14.1 имаме $x_{\min}=62$, $x_{\max}=150$, следователно за размаха имаме $range = x_{\min} - x_{\max} = 88$.

Основните отклонения от нормалното разпределение се изразяват чрез термините на ***изкривяване*** (***skewness***) и ***ексцес (kurtosis)***, за които въз основа на данните от извадката се пресмятат следните статистики

$$skewness = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} \left(\frac{x_k - \bar{x}}{s} \right)^3 ,$$

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{k=1}^{n} \left(\frac{x_k - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} .$$

Стойност на коефициента *skewness* близка до нула показва, че разпределението е симетрично, а стойност на коефициента *kurtosis* близка до нулата показва, че центърът на разпределението е "полегат" по характерния за нормално разпределение начин. В този смисъл стойности близки до нула на тези два коефициента представляват белег за нормално разпределение. За примера 17.1 имаме $skewness = -0.522$ и $kurtosis = -0.125$.

Квантилът (***quantile***) е число върху оста на наблюденията, което разделя техния брой в определена пропорция. Когато се казва, че k_α ($0<\alpha<1$) е α-квантил (или още $\alpha.100\%$ - процентов квантил) за наблюдаваната величина, се има предвид, че (приблизително) $\alpha.100\%$ от броя на наблюденията лежат вляво от k_α, а останалите $(1-\alpha).100\%$ са разположени вдясно от него. Квантилът за $\alpha = 0.5$ се нарича ***медиана*** (***median***) на разпределението, $Me = k_{0.5}$. По определение медианата разделя наблюденията на две (приблизително) равни по брой части. За пример 17.1 имаме $Me = 117$. Квантилът за $\alpha = 0.25$ се нарича ***долен квартил*** (***lower quartile***), а квантилът за $\alpha = 0.75$ се нарича ***горен квартил*** (***upper quartile***). Долният квартил, медианата и

горният квартил разделят наблюденията на четири части, всяка от които съдържа приблизително 25% от броя на наблюденията. За примера 17.1 имаме $k_{0.25}=103$ и $k_{0.75}=129$. Интервалът $[k_{0.25},k_{0.75}]$ съдържа около 50% от случаите. За разглеждания пример 17.1 можем да направим заключението, че вероятността случайно избран индивид да има резултат по скалата между 103 и 129 е около 50%. Последното твърдение представлява характерен извод на статистическия анализ. Други специални квантили са ***процентилите*** (***percentiles***) $P_m=k_{\frac{m}{100}}$, за всевъзможните стойности $m=1,2,\ldots,99$. Програмните среди за статистическа обработка притежават заложени формули за пресмятане на различните квантили.

Тук също различаваме ***популационен квантил и извадков квантил***.

Централна тенденция и изменчивост. Когато за дадена метрична величина трябва да се покаже общото за цялата извадка чрез единствено число, се използват мерки за централна тенденция. Терминът централна тенденция е достатъчно показателен за съдържанието си. Има се предвид някоя типична обобщаваща числова характеристика на разпределението на наблюдаваната величина, която отразява адекватно количественото поведение на резултатите от цялата извадка.

Най-важната мярка за централна тенденция е извадковото средно

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}.$$

Едва ли е необходимо да се обяснява дълго защо средното дава достатъчно обща характеристика на извадката, в съответствие с нашия опит от ежедневието, където присъства изобилие от средни величини като средни доходи, средни постижения и т.н. Като мярка за изменчивост относно средното служи извадковата дисперсия

$$s_x^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}=\frac{\sum(x_k-\bar{x})^2}{n-1}$$

и свързаните с нея ***стандартно отклонение*** (***standard deviation***)

$$SD=s_x=\sqrt{s_x^2}=\sqrt{\frac{\sum(x_k-\bar{x})^2}{n-1}}$$

и ***стандартна грешка*** (***standard error***)

$$SE=\frac{SD}{\sqrt{n}}=\frac{s_x}{\sqrt{n}}.$$

По-малката дисперсия означава по-голяма групиране на данните около средното и обратно. Стандартната грешка има важно значение при образуване на доверителни интервали и проверка на хипотези, което ще бъде разисквано по-нататък.

Другата основна мярка за централна тенденция е ***медианата*** $Me=k_{0.5}$, чиято роля в този смисъл също е практически очевидна. Медианата замества средното като централна тенденция в различни сравнителни анализи. В повечето случаи средното $\bar{x}$ и медианата Me имат равноценно познавателно значение. Не се определят мерки за изменчивост относно медианата. За илюстрация да пресметнем медианата на величина x със стойности 1, 5, 2, 7, 3. Отначало подреждаме наблюденията в нарастващ ред (***вариационен ред***) и получаваме 1, 2, 3, 5, 7. Тук медианата се средното по разположение наблюдение $Me=x_3=3$. Ако броят на наблюденията е четно число, например 1, 9, 5, 2, 7, 3, то вариационният ред има вида 1, 2, 3, 5, 7, 9, а медианата е полусборът на средните две, $Me=\frac{x_3+x_4}{2}=\frac{3+5}{2}=4$.

Модата (***mode***) Mo представлява мярка за централна тенденция, показваща струпването на наблюденията около някоя модална стойност. За примера 17.1 имаме $Mo = 129$. Струпването около тази стойност добре се забелязва от хистограмата на диаграма 17.1. Има величини, чиито разпределения имат две (***бимодални разпределения***) или повече моди. В този случай разпределението сигурно не е нормално. Разглеждана като популационен параметър, модата представлява максимума на плътността на разпределението.

Трите описани мерки за централна тенденция не се различават много. Такова състояние на нещата е характерно в типичния случай. При малък брой наблюдения, наличието на големи отклонения в някои от стойностите обаче води до забележими промени в средното, докато медианата не се променя, което прави медианата предпочитана мярка за централна тенденция в случая на извадки с малък обем. При голям брой наблюдения, когато величината се подчинява по същество на нормално разпределение, различието между средното и медианата се получава пренебрежимо.

Мярка за изменчивост като цяло в извадката е размахът $R = x_{\max} - x_{\min}$. Полезен е също размахът между 90% и 10% квантил $D = k_{0.9} - k_{0.1}$, който представлява по-устойчива мярка в сравнение с обикновения размах, понеже при него не се разглеждат изключителните стойности (***outliers***) (много малките или много големите стойности), чиито носители обикновено са маргинални индивиди със силно отклоняващи се характеристики от общите за извадката и респективно за популацията.