ОЦЕНКА НА ПАРАМЕТРИ

§24. Точкови оценки (статистики). Видове оценки – неизместена и изместена, състоятелна и ефективна.

Една от важните задачи на математическата статистика се състои по данните от извадката да се намерят някои от неизвестните параметри на изследваната случайна величина X.

Параметър на разпределението на случайна величина е неслучайно число, което остава неизменно за тази случайна величина.

Такива са всички числени характеристики на една случайна величина като математическо очакване, дисперсия, мода, медиана и т.н. Обикновено законът на една величина зависи от един (например, разпределението на Поасон), два (биномно $B(p,n)$, нормално $N(a,\sigma)$) или повече параметри (хипергеометрично).

Когато някой от параметрите на разпределението е неизвестен, то за определянето му се използват наличните данни от извадката като тук стоят три основни въпроса:

- по каква формула да се пресметне приближена стойност на параметъра;
- дали намерената стойност действително е оценка на параметъра;
- каква е точността на намерената оценка.

Нека е неизвестен параметърът θ на случайна величина X и нека за намирането му са направени n наблюдения като са наблюдавани стойностите $(x^{(1)},\ldots,x^{(n)})$, т.е. получена е извадка $(X_1,\ldots,X_n)$ с обем n.

Точкова оценка θ^* на параметъра θ наричаме приближената стойност на този параметър, получена от извадката.

Оценката се нарича точкова, защото по определен начин от числата $(x^{(1)},\ldots,x^{(n)})$ сме получили само едно число (една точка от числовата права).

Очевидно, θ^* е функция на наблюдаваните стойности, т.е.

$$\theta^* = \theta^*(x^{(1)},\ldots,x^{(n)}).$$

Ще припомним, че:

- Резултатът $x^{(i)}$ от i-тото наблюдение, може да се тълкува като реализация на случайната величина X_i ={наблюдавана стойност на X в i-тото измерване} ($i=1,\ldots n$).
- Предполагаме, че величините X_i са независими, приемат възможните стойности на величината X и имат закон, съвпадащ със закона на X.

Следователно, може да разглеждаме θ^* като реализация на случайната величина

$$\Theta^* = \Theta^*(X_1,\ldots,X_n),$$

зависеща неслучайно (т.е. по точно определено правило) от величините $X_1,\ldots,X_n$.

.

Произволна функция $f(X_1,\ldots,X_n)$ се нарича статистика на извадката $(X_1,\ldots,X_n)$. Изчислената стойност $f(x^{(1)},\ldots,x^{(n)})$ също се нарича статистика. Тя зависи от обема n на извадката и наблюдаваните стойности $x^{(1)},\ldots,x^{(n)}$ и се разглежда като реализация на $f(X_1,\ldots,X_n)$.

Следователно, *точковите оценки на генералната съвкупност (и всички получени в* **§§22-23** *характеристики на извадката) са нейни статистики*.

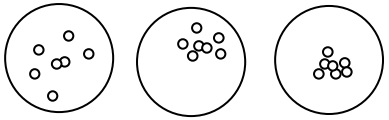
Най-простият начин за оценка на неизвестните параметри е неизвестната характеристика θ на генералната съвкупност да се оценява по съответствуващата ѝ характеристика θ^*, взета от извадката (извадъчна характеристика). Например,

$\bar{x}$ *е оценка за* EX,

s_x^2 *е оценка за* DX,

$\nu_i = \frac{m_i}{n}$ *е оценка за* $p_i = P(X = x_i)$ и т.н.

Видове оценки. За да считаме, че изчислената стойност θ^* е добра оценка на неизвестния параметър θ, тя трябва да има определени качества. Например, ако резултатът от стрелба по мишена са представени на фиг. 24.1, то в случая а) казваме, че стрелбата е неефективна, в случая б) че е отместена (допуска се систематична грешка), докато в случая в) стрелбата е и ефективна, и неотместена. Естествено, най-добър е резултатът в случая в).



а) б) в)

Фиг.24.1.

Подобни качества за оценката θ^* на неизвестния параметър θ може да осигурим, ако случайната величина $\Theta^* = \Theta^*(X_1,\ldots,X_n)$, отразяваща правилото, по което от извадката е изчислена оценката θ^*, притежава определени свойства.

Казваме, че оценката θ^* е *неотместена, ако* $E\Theta^* = \theta$.

Казваме, че оценката θ^* е *отместена, ако* $E\Theta^* \neq \theta$.

Пример 24.1. Тъй като $\bar{x} = \frac{1}{n}(x^{(1)} + \ldots + x^{(2)})$ е наблюдаваната стойност на величината $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ и $EX_1 = \ldots = EX_n = EX$, то по свойства (15.1) имаме $E\bar{X} = \frac{1}{n}(EX_1 + \ldots + EX_n) = EX$, т.е. *средната на извадката* $\bar{x}$ *е неотместена оценка на математическото очакване* EX (генералната средна). ♦

Разликата $E\Theta^* - \theta$ се нарича *отместване*.

От свойствата на математическото очакване следва също, че:

1) Ако $E\Theta^* = \theta + a$, *където* $a = const$, *то оценката* $\Theta^* - a$ *е неотместена.*

2) Ако $E\Theta^* = a\theta$, *то оценката* $\frac{1}{a}\Theta^*$ *е неотместена оценка.*

Действително, $E(\Theta^* - a) = E\Theta^* - a = (\theta + a) - a = \theta$.

Аналогично се доказва свойство 2).

Както знаем, $E\Theta^*$ е средната стойност, около която се колебаят възможните стойности на Θ^*. Затова условието $E\Theta^* = \theta$ *означава, че* θ^* *оценява неизвестната стойност на* θ *без систематическа грешка.*

2. *Казваме, че оценката* θ^* *е състоятелна, ако за произволно* $\varepsilon > 0$ $\lim\limits_{n\to\infty} P(|\Theta^* - \theta| < \varepsilon) = 1$.

Например, от теорема **16.3** за големите числа следва, че оценката $\bar{x}$ е състоятелна оценка на EX.

Ако една оценка е състоятелна, то колкото по-голям е обемът на извадката, толкова разликата $|\theta^* - \theta|$ е по-малка.

Пример 24.3. Ще покажем, че относителната честота $\nu_i = \frac{m_i}{n}$ е състоятелна оценка на вероятността $p_i = P(X = x_i)$.

Действително, ν_i е относителната честота на събитието $X = x_i$ от n направени наблюдения. По теоремата на Бернули **16.4** за произволно $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{m_i}{n} - p_i\right| < \varepsilon\right) = 1,$$

следователно $\nu_i = m_i / n$ е състоятелна оценка на вероятността p_i

3. За един и същ параметър θ на случайната величина X може да се получат няколко оценки от дадената извадка.

Казваме, че оценката θ^* *е ефективна, ако за всяка друга оценка* θ_1^* *имаме* $E(\Theta^* - \theta)^2 \le E(\Theta_1^* - \theta)^2$

Ефективната оценка се характеризира с най-малкото разсейване около неизвестния параметър.

§25. Точкови оценки на генералните математическо очакване и дисперсия.

Да разгледаме една извадка $(X_1, X_2, \ldots, X_n)$ на неизвестния признак X на генералната съвкупност, т.е. наблюдавани са случайните събития $X_i = x^{(i)}$, където случайните величини X_i имат разпределение, еднакво с разпределението на X.

Точкова оценка на математическото очакване EX **.**

Средната $\bar{x} = \frac{1}{n}(x^{(1)} + \ldots + x^{(n)})$ на извадката е една от възможните стойности на случайната величина $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ и е оценка за математическото очакване на величината X.

За да определим свойствата на тази оценка, ще намерим:

$$E\bar{X} = \frac{1}{n}(EX_1 + \ldots + EX_n) = \frac{1}{n} n.EX = EX \ . \qquad (25.1)$$

$$D\bar{X} = \frac{1}{n^2}(DX_1 + \ldots + DX_n) = \frac{1}{n^2} n.DX = \frac{1}{n} DX \ . \qquad (25.2)$$

От полученото равенство $E\bar{X} = EX$ следва, че $\bar{x}$ е *неотместена оценка на математическото очакване* EX. От теорема **16.3** на Чебишев

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \ldots + X_n}{n} - EX\right| < \varepsilon\right) = 1,$$

следва, че $\bar{x}$ *е състоятелна оценка на* EX.

Точкови оценки на дисперсията DX **.**

Оценка на дисперсията $DX = E(X^2) - (EX)^2$ на величината X е извадъчната дисперсия $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x^{(i)2} - (\bar{x})^2$, която е наблюдаваната стойност на величината $S_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2$ ($x^{(i)}$ е

резултатът от i-тото наблюдение). За да установим дали е отместена оценка, намираме математическото очакване на S_X^2 като използваме, че за еднакво разпределените величини X_i и X имаме $E(X_i^2)=E(X^2)$:

$$E(S_X^2)=E\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\overline{X}\right)^2\right]=\frac{1}{n}\sum_{i=1}^{n}E(X_i^2)-E(\overline{X}^2)=\frac{1}{n}\sum_{i=1}^{n}E(X^2)-E(\overline{X}^2).$$

Но $\frac{1}{n}\sum E(X^2)=\frac{1}{n}nE(X^2)=E(X^2)$, а като прибавим и извадим $(EX)^2=(E\overline{X})^2$ (виж формула (25.1)), получаваме

$$E(S_X^2)=E(X^2)-(EX)^2+(EX)^2-E(\overline{X}^2)=DX-[E(\overline{X}^2)-(E\overline{X})^2]=DX-D\overline{X}.$$

Накрая, съгласно (25.2), имаме

$$E(S_X^2)=DX-\frac{1}{n}DX=\frac{n-1}{n}DX,$$

което означава, че извадъчната дисперсия s_x^2 е отместена оценка на дисперсията DX.

Не е трудно да се провери, че статистиката

$$\tilde{s}_x^2=\frac{n}{n-1}s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

е *неотместена оценка на дисперсията DX. Тя се нарича <u>поправена извадъчна дисперсия</u>*.

Очевидно, $\tilde{s}_x^2>s_x^2$, но за обем на извадката $n>30$, разликата между двете числа е незначителна и може да приемем, че $\tilde{s}_x^2\approx s_x^2$.

3. Оценка на средно квадратичното отклонение σX.

Числото $\tilde{s}_x=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$

се нарича <u>поправено средно квадратично отклонение</u>, <u>стандарт</u> или <u>стандартна грешка</u> и е неотместена оценка на средно квадратичното отколонение $\sigma=\sqrt{DX}$.

И тук за големи стойности на обема на извадката $\tilde{s}_x\approx s_x$.

Пример 25.1. Да се намерят точковите оценки на генералното математическо очакване EX и дисперсия DX на неизвестния признак X по дадената извадка

x_i	0,01	0,04	0,08
m_i	5	3	2

Решение. Неотместена оценка за EX е средната на извадката

$$\overline{x}=\frac{1}{10}(0{,}01.5+0{,}04.3+0{,}08.2)=0{,}033.$$

За дисперсията DX имаме две оценки:

1) отместена оценка на DX е дисперсията на извадката $s_x^2=\overline{x^2}-(\overline{x})^2$, където

$$\overline{x^2}=\frac{1}{10}(0{,}01^2.5+0{,}04^2.3+0{,}08^2.2)=0{,}00181\Rightarrow s_x^2=0{,}00181-0{,}033^2=0{,}0007$$

2) неотместена оценка на DX е

$\tilde{s}_x^2=\frac{n}{n-1}s_x^2=\frac{10}{9}0.0007\approx 0{,}0008$ (поправена дисперсия).

Упражнения.

1. За изучаване на броя X на лицата в едно домакинство са получени данните:

x_i	1	2	3	4	5	6
m_i	8	17	11	10	2	2

Да се намерят неотместените точкови оценки на величината X.

2. Полученото месечно възнаграждение на 15 случайно избрани работници от дадено предприятие (в лв):

200, 200, 300, 300, 300, 400, 400, 400, 400, 400, 500, 500, 500, 500, 700.

Да се оцени по извадката средната работна заплата и отклонение от средната заплата в предприятието.

3. Времето за подготовка на домашната работа за случайно избрани 80 ученика е както следва

(x_{i-1},x_i)	(15,20)	(20,30)	(30,40)	(40,50)	(50,60)
m_i	2	6	26	30	16

Да се намери средното време, необходимо за подготовка на домашна работа.

4. Дадена е следната извадка 0, 3, 2, 4, 1, 0, 2, 3, 2, 0, 1, 4, 5, 6, 2, 1, 2, 1, 0, 0, 3, 3, 2,1, 0. Да се състави статистическото разпределение на извадката, да се намерят медианата и горният квартил и да се начертае полигонът на относителната честота. Да се намерят точковите оценки на генералната съвкупност, от която е извлечена извадката.

§26. Начини за намиране на оценки на параметри - метод на моментите. Метод на най-голямото правдоподобие.

Нека за изучаването на неизвестната случайна величина X е получена извадка с обем n. В **§25** получихме точкови оценки за математическото очакване и дисперсията. Поставяме сега по-общия въпрос: как с помощта на наличните данни да намерим оценки на други неизвестни параметри на разпределението на величината X.

Означаваме неизвестните параметри на разпределението с $\theta_1,\ldots,\theta_k$ (общ брой на неизвестните k). Ще разгледаме ***метода на моментите*** за намирането им, който използва факта, че неизвестните параметри могат да се изразят чрез моментите на разпределението. Например, за величината $X\sim B(p,n)$ (**§11**) е известно, че $EX=pn$, $DX=npq=np(1-p)$. Следователно, ако търсим оценка на параметъра p, е

достатъчно от извадката да намерим оценката $\bar{x}$ за математическото очакване EX. Тогава оценката за p е $p^*=\frac{\bar{x}}{n}$

***Методът на моментите** се състои в следното:*
1) *От извадката изчисляваме необходимите извадъчни моменти, които оценяват съответните моменти на разпределението на величината X.*
2) *Търсените оценки на параметрите $\theta_1,\ldots,\theta_k$ се определят като се реши системата уравнения, отразяващи зависимостта между тези параметри и моментите на разпределението.*

Пример 26.1. Известно е, че величината X е равномерно разпределена в интервала $[a,b]$ (**§12**). Да се намерят оценки за параметрите a и b по дадената извадка $\begin{array}{c|cccc}(x_{i-1},x_i) & (0,2) & (2,4) & (4,6) & (6,8)\\ \hline m_i & 6 & 5 & 8 & 5\end{array}$.

Решение: Ще използваме, че за равномерното разпределение имаме

$$E\xi=\frac{a+b}{2},\ D\xi=\frac{(b-a)^2}{12}.$$

Заменяйки интервалите със средите им, получаваме таблицата $\begin{array}{c|cccc}x_i & 1 & 3 & 5 & 7\\ \hline m_i & 6 & 5 & 8 & 5\end{array}$, от която за оценките $\bar{x}$ и $\tilde{s}_x^2$ на EX и DX получаваме:

$$\bar{x}=\frac{1}{24}(1.6+3.5+5.8+7.5)=4;$$

$$s_x^2=\overline{x^2}-(\bar{x})^2=\frac{1}{24}(1.6+9.5+15.8+49.5)-16=\frac{14}{3}\ \Rightarrow\tilde{s}_x^2=\frac{24}{23}s_x^2=\frac{112}{23}$$

Следователно, за оценките a^* и b^* получаваме системата

$$\begin{cases}\dfrac{a^*+b^*}{2}=4\\ \dfrac{(a^*-b^*)^2}{12}=\dfrac{112}{23}\end{cases}$$ с решение $a^*=0{,}18$, $b^*=7{,}82$. ♦

Един от най-разпространените методи за намиране на оценки на параметрите на разпределението на случайна величина е *<u>методът на максималното правдоподобие</u>*. Основната идея на този метод се състои в следното: ако $x^{(1)},\ \ldots,\ x^{(n)}$ са наблюдавани стойности на величината X, да се намери такава функция, наречена функция на правдоподобието $L(\theta_1,\theta_2,\ldots,\theta_k,x^{(1)},\ldots,x^{(n)})$ и зависеща от оценяваните парметри $\theta_1,\ldots,\ \theta_k$, за която търсените оценки $\theta_1^*,\ldots,\theta_k^*$ да са стойностите, при които функцията приема максимална стойност, т.е.

$$L(\theta_1^*,\theta_2^*,\ldots,\theta_k^*,x^{(1)},\ldots,x^{(n)})=\max.$$

Ще видим как се получава функцията на най-голямо правдоподобие за величина X, за определеност дискретна, с един неизвестен параметър θ на разпределението. Плътността на разпределение зависи от неизвестния параметър θ, т.е. $p_X=p_X(x,\theta)$, и тъй като наблюдаваните стойности $x^{(1)},\ \ldots,x^{(n)}$ са възможни стойности на величината, то за вероятностите на събитията $X=x^{(i)}$ имаме $P(X=x^{(i)})=p_X(x^{(i)},\theta)$, т.е. зависят от параметъра θ. Образуваме функцията на променливата е θ

$$L(x^{(1)},\ldots x^{(k)},\theta)=P[(X=x^{(1)})\cap\ldots\cap(X=x^{(N)})]=\prod_{i=1}^{N}p_X(x^{(i)},\theta),$$

която се нарича <u>функция на правдоподобието</u>.

Събитията $X=x^{(i)}$ са настъпили, затова най-достоверна ще е тази стойност θ^* на параметъра θ, за която вероятностите $P(X=x^{(i)})$ са възможно най-големи (най-близки до 1). Следователно като оценка на параметъра θ се приема тази стойност θ^*, за която

$$L(x^{(1)},\ldots,x^{(N)},\theta)=\max,$$

т.е. θ^* се определя от уравнението $\frac{dL}{d\theta}=0$.

Използва се също и

<u>логаритмичната функция на правдоподобието</u>

$$\Lambda(x^{(1)},\ldots x^{(k)},\theta)=\ln L(x^{(1)},\ldots x^{(k)},\theta)=\sum_{i=1}^{N}\ln p_X(x^{(i)},\theta),$$

за която θ^* е решение на по-удобното за получаване уравнение $\frac{d\Lambda}{d\theta}=0$.

Пример 26.1. Да се намери по метода на най-голямото правдоподобие оценка на математическото очакване на величина, която има разпределение на Поасон.

Решение. Ако $X\sim Po(\lambda)$, то $P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$, $k=0,1,2,\ldots$ и $EX=\lambda$. Търсим оценка на неизвестния параметър $EX=\lambda$ по дадени наблюдавани стойности $x^{(1)},\ \ldots,x^{(n)}$. Образуваме логаритмичната функция на правдоподобие:

$$\Lambda(x^{(1)},\ldots x^{(N)},\lambda)=\sum_{i=1}^{N}\ln\frac{\lambda^{x^{(i)}}e^{-\lambda}}{x^{(i)}!}=\sum_{i=1}^{N}\left(x^{(i)}\ln\lambda-\lambda-\ln(x^{(i)}!)\right)$$

Изчисляваме $\frac{d\Lambda}{d\lambda}=\sum_{i=1}^{N}\left(x^{(i)}\frac{1}{\lambda}-1\right)=\frac{1}{\lambda}\sum_{i=1}^{N}x^{(i)}-n$ и решаваме уравнението

$$\frac{1}{\lambda}\sum_{i=1}^{N}x^{(i)}-n=0 \;\Rightarrow\; \lambda=\frac{1}{n}\sum_{i=1}^{N}x^{(i)}\,.$$

Следователно, получената по метода на най-голямото правдоподобие оценка на математическото очакване на величината $X\sim Po(\lambda)$ е отново извадъчната средна $\bar{x}$ ♦.

§27. Интервални оценки. Доверителен интервал и доверителна вероятност.

Нека θ е неизвестен параметър на случайната величина X и нека от извадка с обем n е изчислена негова точкова оценка θ^*. Поставяме въпроса каква е точността на статистиката θ^*. Очевидно, колкото разликата $\theta^*-\theta$ е по-близка до нула или отношението $\frac{\theta^*}{\theta}$ е по-близко до единица, толкова по-точна е оценката θ^*. Тези две нови статистики дават представа за точността на θ^*, но не са приложими, защото нямат точно определени закони на разпределение, а и самият параметър θ е неизвестен. С тяхна помощ, обаче, за всяка точкова оценка се намира подходяща статистика, законът на която е предварително известен.

Например, ако $X\sim N(EX,\sigma X)$, то за определяне на точността на оценката $\bar{x}$ на математическото очакване EX вместо статистиката $\bar{x}-EX$ се използва статистиката

$$Z=\frac{\overline{X}-EX}{\frac{\sigma}{\sqrt{n}}}\text{, където } \overline{X}=\frac{1}{n}(X_1+\ldots+X_n)\text{, } \sigma=\sigma X\text{,} \qquad (27.1)$$

разпределението на която, съгласно свойствата на нормално разпределените величини, е нормално разпределение (виж забележка **15.2**), при това $EZ=0$, $\sigma Z=1$, т.е. $Z\sim N(0,1)$.

Забележка 27.1. От централната гранична теорема и формули (15.3) следва, че при големи стойности на обема на извадката величината Z има разпределение, близко до стандартното нормално разпределение $N(0,1)$ и за величини, които нямат нормално разпределение.

Следователно величината (27.1) има стандартно нормално разпределение ако:

- *величината X има нормално разпределение, независимо от обема n на извадката;*

- *величината X има друго разпределение, но обемът n на извадката е голям.*

По същия начин се въвеждат и други удобни статистики, които ще използваме по-нататък. По-важните от тях, както и теоретичните им разпределения са:

$Z=\frac{\bar{x}-EX}{\frac{\sigma}{\sqrt{n}}}\to N(0,1)$ (Z-***разпределение***)

$t=\frac{\bar{x}-EX}{\frac{\tilde{s}_x}{\sqrt{n}}}\to t(n-1)$ (***t-разпределение на Стюдънт*** с n степени на свобода)

$\chi^2=\frac{n\tilde{s}_x^2}{\sigma^2}\to\chi^2(n)$ (χ^2 -***разпределение*** с n степени на свобода)

$F=\frac{\tilde{s}_x^2/\sigma_x^2}{\tilde{s}_y^2/\sigma_y^2}=\frac{\tilde{s}_x^2/\tilde{s}_y^2}{\sigma_x^2/\sigma_y^2}\to F(n_1-1,n_2-1)$ (***F-разпределение на Фишер*** с n_1-1 и n_2-1 степени на свобода).

Определенията на цитираните теоретични разпределения са разгледани в **§14.**

С помощта на тези статистики се намира интервал, наречен доверителен, в който с дадена вероятност се намира истинската стойност на оценявания параметър.

Нека се оценява неизвестният параметър θ и нека е дадено число $0<\gamma<1$.

<u>Доверителен интервал</u> на параметъра θ <u>с доверителна вероятност (надеждност)</u> γ наричаме интервала (θ_1,θ_2), за който $P(\theta_1<\theta<\theta_2)=\gamma$.

Числото $\alpha=1-\gamma$ наричаме <u>коефициент на доверие</u>, <u>риск</u> или <u>ниво на значимост (на състоятелност)</u>. То е равно на вероятността изследваният параметър да се намира вън от доверителния интервал.

Радиусът $\delta=\frac{\theta_2-\theta_1}{2}$ на доверителният интервал се нарича <u>представителна грешка</u> на оценката с доверителна вероятност γ.

Доверителната вероятност (надеждност) γ се избира в зависимост от конкретните условия и обикновено има стойности от порядъка на 0,9, 0,95, 0,99.

При дадена извадка с обем n намирането на доверителния интервал на параметъра θ по зададена доверителна вероятност γ се извършва по следния начин:

1. От извадката се изчислява точкова оценка θ^* на параметъра θ.

2. Избира се подходяща статистика $Y=Y(\theta^*,\theta)$, която да оценява различието между θ^* и θ. Разпределението на тази величина трябва да е известно и да не зависи от параметъра θ.

3. В зависимост от зададената доверителна вероятност γ се определя интервал (y_1,y_2), за който $P(y_1<Y(\theta^*,\theta)<y_2)=\gamma$.

4. Границите на доверителния интервал на параметъра θ се определят като се реши относно θ неравенството

$$y_1<Y(\theta^*,\theta)<y_2 .$$

§28. Доверителни интервали за математическото очакване и дисперсията на нормално разпределена случайна величина

Ще намерим доверителния интервал за математическото очакване $EX=\mu$ на ***нормално разпределена величина*** X, *средно квадратичното отклонение на която $\sigma X=\sigma$ е известно*.

Съгласно **§27**:

1) По дадената извадка с обем n намираме извадъчната средна $\bar{x}$, която е оценка на $EX=\mu$.

2) От забележка **27.1** следва, че $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$, т.е. разпределението на случайната величина Z е известно и не зависи нито от $\bar{x}$, нито от μ,. Следователно, Z може да се използува като статистика, оценяваща отклонението на величината от математическото й очакване.

3) При доверителната вероятност γ можем да определим такова число ε, че вероятността на събитието $-\varepsilon<Z<\varepsilon$ да бъде равна на γ. Действително, $P(-\varepsilon<Z<\varepsilon)=P(|Z|<\varepsilon)=2F(\varepsilon)-1=\gamma$, откъдето $F(\varepsilon)=\dfrac{1+\gamma}{2}$. Следователно, $\varepsilon=Z_{\frac{1+\gamma}{2}}$, т.е. ε е квантилът от ред $\dfrac{1+\gamma}{2}$ на стандартното нормално разпределение.

4) Накрая, от неравенството $-\varepsilon<\dfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}<\varepsilon \;\Rightarrow\; -\varepsilon\dfrac{\sigma}{\sqrt{n}}<\bar{x}-\mu<\varepsilon\dfrac{\sigma}{\sqrt{n}}$

получаваме, че $\bar{x}-Z_{\frac{1+\gamma}{2}}\dfrac{\sigma}{\sqrt{n}}<\mu<\bar{x}+Z_{\frac{1+\gamma}{2}}\dfrac{\sigma}{\sqrt{n}}$ и това е доверителният

интервал за математическото очакване.

Радиусът на интервала е $\delta=Z_{\frac{\gamma+1}{2}}\dfrac{\sigma}{\sqrt{n}}$.

Аналогично се постъпва и в случая, *когато не е известно средно квадратичното отклонение на X*, но този път като статистика се използува величината $\dfrac{\bar{x}-EX}{\tilde{s}_x/\sqrt{n}}$, която има разпределение на Стюдънт (**§14**) с $n-1$ степени на свобода. За определянето на доверителния интервал $(\bar{x}-\delta,\bar{x}+\delta)$, се пресмята квантилът $t_{\frac{1+\gamma}{2}}(n-1)$ от ред $\dfrac{1+\gamma}{2}$ на разпределението на Стюдънт (*t*-разпределение) с $n-1$ степени на свобода. Таблица със стойностите на квантилите на *t*-разпределението е дадена на стр. 169.

Така се получават следните правила за намиране на доверителни интервали:

Доверителен интервал за математическото очакване EX ***нормално разпределена величина*** X ***с доверителна вероятност*** γ ***по извадка с обем*** n*:*

ако $\sigma X=\sigma$ е известно: $\bar{x}-Z_{\frac{1+\gamma}{2}}\dfrac{\sigma}{\sqrt{n}}<\mu<\bar{x}+Z_{\frac{1+\gamma}{2}}\dfrac{\sigma}{\sqrt{n}}$; (28.1)

ако σX не е известно: $\bar{x}-t_{\frac{1+\gamma}{2}}(n-1)\dfrac{\tilde{s}_x}{\sqrt{n}}<\mu<\bar{x}+t_{\frac{1+\gamma}{2}}(n-1)\dfrac{\tilde{s}_x}{\sqrt{n}}$. (28.2)

Понякога е необходимо да се определи какъв би трябвало да е обемът на извадката, че представителната грешка с дадена доверителна вероятност γ да бъде не по-голяма от дадено число δ_0. Като се използва, че *t*-разпределението с $n=\infty$ съвпада с разпределението $N(0,1)$, се определя, че

$n>\left(\dfrac{\sigma Z_{\frac{\gamma+1}{2}}}{\delta_0}\right)^2$, *ако $\sigma X=\sigma$ е известно, или* $n>\left(\dfrac{\tilde{s}_x.Z_{\frac{\gamma+1}{2}}}{\delta_0}\right)^2$, *ако σX не е известно и се оценява с $\tilde{s}_x$.*

Пример 28.1. Дадена е извадката

x_1	−2	1	2	3	4	5
m_1	2	1	2	2	2	1

а) Да се намерят доверителният интервал на математическото очакване EX и представителната грешка с доверителна вероятност $\gamma=0{,}95$. б) Колко трябва да е обемът на извадката, че представителната грешка за EX с доверителна вероятност 0,95 да е не по-голяма от 0,5?

Решение. От извадката изчисляваме точковите оценки на EX и DX:

$$\bar{x}=\frac{1}{n}\sum x_i m_i=2\ ,\ \tilde{s}_x^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2 m_i=\frac{52}{9}\ ,\ \tilde{s}_x\approx 2{,}4$$

а) Тъй като σX не е дадено, прилагаме формула (28.2). Изчисляваме $\frac{\gamma+1}{2}=\frac{1+0{,}95}{2}=0{,}975$. От таблицата за *t*-разпределението при степени на свобода $n-1=10-1=9$ определяме квантила $t_{0{,}975}(9)=2{,}26$.

Следователно, представителната грешка е

$$\delta=t_{\frac{\gamma+1}{2}}(n-1).\frac{\tilde{s}_x}{\sqrt{n}}=2{,}26\frac{2{,}4}{\sqrt{10}}=1{,}72\ ,$$

а доверителният интервал за EX е $(0{,}8;\ 3{,}72)$.

б) За да определим обема на извадката, ще използваме отново формулата за δ, но този път ще вземем $t_{0{,}975}(\infty)=Z_{0{,}975}=1{,}96$. Решаваме неравенството $0{,}5\geq 1{,}96\frac{2{,}4}{\sqrt{n}}$, откъдето получаваме $n\geq 88$. ♦

По подобен начин се намират и доверителни интервали за дисперсията на нормално разпределена случайна величина, като в този случай се използва статистиката

$$\chi^2=\frac{(n-1)\tilde{s}_x^2}{\sigma^2}\ ,$$

която има χ^2-разпределение с $n-1$ степени на свобода (стр.170). Така се получават следните *доверителни интервали за математическото очакване и дисперсията на нормално разпределена величина:*

Доверителен интервал за дисперсията DX ***на нормално разпределена величина*** X ***с доверителна вероятност*** γ ***по извадка с обем*** n*:*

• *ако* EX *е известно:* $\frac{ns_0^2}{\chi^2_{\frac{1+\gamma}{2}}(n)}<DX<\frac{ns_0^2}{\chi^2_{\frac{1-\gamma}{2}}(n)}$, $s_0=\frac{1}{n}\sum_{i=1}^{k}(x_i-\mu)^2 m_i$ (28.3)

• *ако* EX *не е известно:* $\frac{(n-1)\tilde{s}_x^2}{\chi^2_{\frac{1+\gamma}{2}}(n-1)}<DX<\frac{(n-1)\tilde{s}_x^2}{\chi^2_{\frac{1-\gamma}{2}}(n-1)}$ (28.4)

Тук $\chi^2_{\frac{1+\gamma}{2}}(n)$ *и* $\chi^2_{\frac{1-\gamma}{2}}(n)$ *са квантили на* χ^2*-разпределението с* n *степени на свобода, а* $\chi^2_{\frac{1+\gamma}{2}}(n-1)$ *и* $\chi^2_{\frac{1-\gamma}{2}}(n-1)$ *- квантили на* χ^2*-разпределението с* $n-1$ *степени на свобода*

Пример 28.2. Да се намери доверителният интервал за DX с надеждност 0,95 по извадката от пример **28.1.** ♦

Решение. Прилагаме формула (28.4). Изчисляваме квантилите на χ^2-разпределението със степени на свобода $n-1=9$

$$\chi^2_{\frac{1+\gamma}{2}}(n-1)=\chi^2_{0{,}975}(9)=19{,}02\ ,\ \chi^2_{\frac{1-\gamma}{2}}(n-1)=\chi^2_{0{,}025}(9)=2{,}70$$

и определяме границите на доверителния интервал

$$\frac{9.2{,}4^2}{19{,}02}<DX<\frac{9.2{,}4^2}{2{,}70}\ .$$

Следователно, $2{,}73<DX<19{,}26$. ♦

Пример 28.3. По данни на администрацията на дадено предприятие средната продължителност на работната седмица е 40 часа. Избрани са 10 работника и за тях е изчислено, че работната седмица е средно 42 часа със средно отклонение 6 часа (поправено средно квадратично отклонение).

а) С каква доверителна вероятност може да приемем, че средната продължителност на работната седмица е 40 часа?

б) За по-нататъчни изследвания е необходимо да се определи средната седмична заетост с точност $\pm$ 1 час и доверителна вероятност 0,99. Колко работници трябва да съдържа извадката?

Решение.

а) Разликата между генералната средна EX=40 и средната $\bar{x}=42$ на извадката е равна на 2, т.е. 40 попада в доверителния интервал, ако радиусът му δ е по-голям от 2. От формулата за представителната грешка намираме

$$t_{\frac{1+\gamma}{2}}(9)=\frac{\delta\sqrt{n}}{\tilde{s}_x}>\frac{2.\sqrt{10}}{6}=1{,}05\ .$$

От таблицата за квантилите определяме, че $t_{0{,}84}(9)\approx 1{,}05$. Следователно, приетата средна продължителност попада в доверителния интервал, определен по дадената извадка, с доверителна вероятност, не по-малка от 0,84.

б) Тук доверителната вероятност е $\gamma=0{,}99$. Ще изчислим какъв трябва да бъде обемът на извадката, че представителната грешка да бъде $\delta=1$.

Изчисляваме

$$Z_{\frac{1+\gamma}{2}}=Z_{0{,}995}=2{,}5758$$

и определяме n от неравенството $n>\left(\frac{Z_{\frac{1+\gamma}{2}}.\tilde{s}_x}{\delta}\right)^2$, т.е.

$$n > \left(\frac{2{,}575.6}{1}\right)^2 \Rightarrow n > 238{,}85 .$$

Следователно, обемът на извадката трябва да е около 240. ♦

Общи задачи.

1. За изследването на нормално разпределена случайна величина *X* е получена извадката 36 34 33 36 38 39 40 32.

а) Да се намерят точковите оценки на математическото очакване EX и дисперсията DX .

б) Да се намерят доверителните интервали с доверителна вероятност $\gamma = 0{,}95$ за EX и DX .

2. Наблюдава се работата на автомат, който произвежда детайли по даден стандарт като за определяне на точността му са взети 30 детайла и е изчислена дисперсията на извадката $s_x^2 = 2{,}9$.

а) Коя от числените характеристики на извадката оценява точността на автомата?

б) Да се намери доверителният интервал за точността на автомата с доверителна вероятност $\gamma = 0{,}95$.

3. Автомат пълни бутилки като съдържанието им по стандарт е 300 мл. За определяне на точността на дозировката са проверени 10 бутилки и са получени данните 299, 276, 283, 301, 297, 281, 300, 291, 295, 291. а) Да се намери 90%-доверителен интервал за генералното средно квадратично отклонение на количеството течност в бутилките. б) Ако отклонението е по-голямо от 10мл, то автоматът трябва да се настрои. Има ли основание за пренастройка на апарата?

4. При измерване на физична величина X са получени резултатите -2, 0, -1, 1, 3. Да се определи с надеждност $\gamma = 0{,}95$ доверителният интервал за стойността на величината X , ако е известно, че точността на измерването е $\pm 1{,}9$.

5. В резултат на 5 измервания са получени резултатите 10, 9, 12, 15, 10. Да се намерят представителната грешка с доверителна вероятност $\gamma = 0{,}99$ на оценката за генералната средна на измерваната величина. Какъв трябва да е обемът на извадката, че представтелната грешка да се намали три пъти?

6. Ако се приеме, че са дадени извадки от нормално разпределени величини, да се намерят доверителни интервали за:

а) средното време , необходимо за почистване на дома, по извадката от зад. 1, **§22**;

б) средната стойност и средното отклонение от тази стойност на броя X на дефектните изделия по данните от зад. 3, **§22**;

в) средната стойност и средното отклонение от тази стойност на броя X на пътниците в един автобус по данните от зад. 5, **§21**.

7. От автомат за пакетиране на брашно са взети за контролно измерване 15 пакета и е изчислено $\bar{x} = 956g$.

а) Ако точността на дозиране е 20 g; да се оцени с надеждност $\gamma = 0{,}9$ средното тегло на пакет брашно;

б) Ако извадъчната дисперсия е $\tilde{s}_x = 20g$ да се оцени точността на автомата.

8. По проект дължината X на детайл трябва да бъде 50,0 см. За контрол на качеството е получена извадката 4,99 4,91 4,95 4,97 5,00 5,01 4,83 4,91 4,83 4,76. Да се оцени дължината на произвежданите детайли по дадената извадка. Какъв трябва да е обемът на извадката, че тази дължина да е оценена с грешка, по-малка или равна на 0,5 см.