Глава 12

Оценки на параметри на разпределение

12.1 Увод

По вида на полигона на честотите или на хистограмата се прави предположение за вида на разпределението на с.в. X. Понякога видът на разпределението е известен предварително по някакви съображения (нормално, равномерно, биномно, на Поасон и др.). За да се доуточни този закон, трябва да се намерят някои негови параметри, които го определят еднозначно. Определящите параметри за някои разпределения са дадени по-долу заедно с математическите им очаквания и дисперсиите:

разпределение	параметри	EX	DX
нормално	m, σ	$EX = m$	$DX = \sigma^2$
експоненциално	μ	$EX = 1/\mu$	$DX = 1/\mu^2$
на Поасон	λ	$EX = \lambda$	$DX = \lambda$
равномерно	a, b	$EX = (a+b)/2$	$DX = (b-a)^2/12$
биномно	p, n	$EX = np$	$DX = np(1-p)$

Понеже между тези параметри и EX и DX съществува тясна зависимост, то често уточняването на закона за разпределение се свежда до оценка на EX и DX по получената извадка $x_1, x_2, \ldots, x_n$.

Използват се два вида оценки на параметрите на разпределението:

- **точкови оценки**;
- **интервални оценки**.

12.2 Точкови оценки на параметри

Нека в закона за разпределение на с.в. X се съдържа параметър θ и трябва да оценим този параметър по стойностите на X, получени от извадката:

$$x_1, x_2, \ldots, x_n\,. \tag{12.1}$$

Всяка точкова оценка $\overline{\theta}$ на параметъра θ е някаква статистика, т.е. функция на тези стойности:

$$\overline{\theta} = \overline{\theta}(x_1, x_2, \ldots, x_n)\,.$$

Следователно $\overline{\theta}$ е случайна величина, получаваща своите стойности в резултат на n наблюдения над с.в. X (получаване на извадка с обем n). Естествени изисквания към оценката са следните:

- Функцията $\overline{\theta}$ е симетрична относно променливите си, т.е. резултатът от пресмятането на стойността ѝ не се променя, ако променим реда на променливите $x_1, x_2, \ldots, x_n$;
- Желателно е, когато използваме $\overline{\theta}$ вместо θ, да не се получават систематични грешки нито към занижаване, нито към завишаване на θ, т.е.
$$E(\overline{\theta}) = \theta.$$
Оценка, която удовлетворява това условие се нарича **неизместена**;
- Желателно е с увеличаването на обема на извадката n оценката $\overline{\theta}$ да се концентрира около θ, т.е.
$$D(\overline{\theta}) = D(\overline{\theta}(x_1, x_2, \ldots, x_n)) \to 0, \quad \text{при} \;\; n \to +\infty.$$
Оценка, която удовлетворява това условие се нарича **състоятелна**.

12.3 Оценка за математическото очакване EX

Нека $EX = m$. Оценка на параметъра m е **статистическото средно**

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k\,, \tag{12.2}$$

където x_k са независими и еднакво разпределени с X. Имаме, че

$$E(\overline{x}) = m, \quad D(\overline{x}) = DX/n\,,$$

т.е. оценката е неизместена и състоятелна.

12.4 Оценка за дисперсията DX

Нека $DX = D$. Оценка на параметъра D е величината

$$D_n = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2\,, \tag{12.3}$$

която се нарича **изместена статистическа дисперсия**. Оказва се, че D_n е състоятелна оценка за D, но е изместена относно D, т.е. $E(D_n) \neq D$. По-точно,

$$E(D_n) = \frac{n-1}{n}D\,,$$

откъдето следва, че неизместената оценка за дисперсията $DX = D$ е величината

$$\overline{D} = \frac{n}{n-1}D_n \tag{12.4}$$

или

$$\overline{D} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2\,, \tag{12.5}$$

която се нарича **неизместена статистическа дисперсия**.

Забележка 12.1. *При пресмятането на D_n е по-удобно вместо* (12.3) *да се използва формулата*

$$D_n = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - \overline{x}^2\,. \tag{12.6}$$

Различните начини за пресмятане на $\overline{x}$ и D_n са разгледани подробно в раздел 10.2 .

12.5 Интервални оценки на параметри. Доверителен интервал

Разгледахме задачата за оценка на параметъра θ с едно число (точкова оценка). При големи n такава оценка е близка до неизвестния параметър $\overline{\theta}$. Обаче, когато n не е голямо, случайният характер на $\overline{\theta}$ може да доведе до съществена разлика между θ и $\overline{\theta}$. Възниква задачата за оценка на θ не с едно число, а с цял интервал (θ_1, θ_2) така, че вероятността за поглъщане (покриване) на θ от този интервал да не е по-малка от дадена стойност $p = 1 - \alpha$, т.е.

$$P(\theta_1(x_1, x_2, \ldots, x_n) < \theta < \theta_2(x_1, x_2, \ldots, x_n)) \geq p\,, \tag{12.7}$$

където $p \in (0, 1)$. Тук величината p се нарича **доверителна вероятност**; θ_1 и θ_2 се наричат **доверителни граници**, интервалът (θ_1, θ_2)– **доверителен интервал**, а положителното число $\alpha = 1 - p$ се нарича **ниво на значимост**. Ако нивото на значимост α е избрано малко (α=0.10, α=0.05, α=0.01), то доверителната вероятност $p = 1 - \alpha$ е близка до 1 ($p = 0.90, p = 0.95, p = 0.99$) и тогава събитието $(\theta_1 < \theta < \theta_2)$ практически е достоверно.

Получаването на интервална оценка за параметъра θ се състои в намирането на доверителен интервал $I_p = (\theta_1, \theta_2)$ при зададена доверителна вероятност p (или ниво на значимост α). Границите θ_1 и θ_2 на доверителния интервал I_p се определят от условието

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha. \tag{12.8}$$

Често се използват **едностранни доверителни интервали**, границите на които се определят от условията

$$P(\theta < \theta_2) = 1 - \alpha \quad \text{или} \quad P(\theta_1 < \theta) = 1 - \alpha.$$

Тези интервали се наричат съответно **едностранен ляв** и **едностранен десен** доверителен интервал.

Ще разгледаме няколко най-често срещани случаи на получаване на интервални оценки. Преди запознаването с тях е желателно читателят да прегледа внимателно темата за квантилите от глава 11.

12.6 Доверителен интервал за центъра m = EX на нормално разпределение

Нека $X \in N(m, \sigma)$. Ще покажем как се извършва интервална оценка за $m = EX$ в следните два случая:

А) **При известно стандартно отклонение $\sigma_X = \sigma$.**

В този случай, намирането на доверителния интервал се основава на факта, че случайната величина

$$Y = \frac{\overline{x} - m}{\sigma/\sqrt{n}}$$

има стандартно нормално разпределение $U = N(0, 1)$. Тогава от формула (11.1), в която $R = U$ и $x_p = u_p$, получаваме, че

$$P(u_{\alpha/2} < Y < u_{1-\alpha/2}) = 1 - \alpha.$$

Като решим неравенството

$$u_{\alpha/2} < \frac{\overline{x} - m}{\sigma/\sqrt{n}} < u_{1-\alpha/2}$$

относно m, получаваме, че с вероятност $p = 1 - \alpha$ се сбъдва събитието

$$\overline{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < m < \overline{x} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}\,.$$

Понеже $u_{\alpha/2} = -u_{1-\alpha/2}$, то полученият доверителен интервал за m може да се запише по следния начин:

$$\overline{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < m < \overline{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\,.$$

Затова намирането на доверителния интервал за $m = EX$ по зададено ниво на значимост α се извършва в следния ред:

1. Намираме точковата оценка за m:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k;$$

2. От Таблица 4 определяме квантила $x_\alpha = u_{1-\frac{\alpha}{2}}$ на стандартното нормално разпределение U и пресмятаме точността

$$\Delta = \frac{\sigma}{\sqrt{n}} x_\alpha;$$

3. Тогава доверителният интервал за m е $I_p = (\overline{x} - \Delta, \overline{x} + \Delta)$ и m попада в този интервал с доверителна вероятност $p = 1 - \alpha$.

Пример 12.1. *Нека с.в. X – ръстът на $n = 1000$ ученици от пример* 10.2 *е разпределена по нормалния закон $N(m, \sigma)$ и стандартното отклонение σ_X е известно: $\sigma_X = \sigma = 6$ см. Да се намерят доверителните интервали за $EX = m$ с доверителна вероятност $p = 0.95$ и $p = 0.99$.*

В пример 10.2 *е пресметнато, че $\overline{x} = 165.5$ см. При $p = 0.95$ нивото на значимост е равно на $\alpha = 1 - 0.95 = 0.05$. От Таблица* 4 *определяме квантила $x_\alpha = u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.96$. Тогава*

$$\Delta = \frac{6 \cdot 1.96}{\sqrt{1000}} \approx 0.37$$

и доверителният интервал е

$$I_{0.95} = (165.5 - 0.37, 165.5 + 0.37) = (165.13, 165.87).$$

При $p = 0.99$ нивото на значимост е равно на $\alpha = 1 - 0.99 = 0.01$. От Таблица 4 *определяме квантила $x_\alpha = u_{1-\frac{\alpha}{2}} = u_{0.995} = 2.58$. Тогава*

$$\Delta = \frac{6 \cdot 2.58}{\sqrt{1000}} \approx 0.49$$

и доверителният интервал е

$$I_{0.99} = (165.5 - 0.49, 165.5 + 0.49) = (165.01, 165.99).$$

Забележка 12.2. *Ако увеличим доверителната вероятност, то доверителният интервал се разширява.*

Забележка 12.3. *В разглеждания случай, едностранните доверителни интервали за $m = EX$ при ниво на значимост α се задават с неравенствата:*

$$m < \overline{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \tag{12.9}$$

(*за ляв доверителен интервал*);

$$m > \overline{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \tag{12.10}$$

(*за десен доверителен интервал*),
където квантилът $u_{1-\alpha}$ *се определя от таблицата за квантилите на стандартното нормално разпределение* U (*Таблица* 4).

Б) **При неизвестно стандартно отклонение** σ_X**.**

В този случай, намирането на доверителния интервал се основава на факта, че случайната величина

$$Y = \frac{\overline{x} - m}{\overline{s}/\sqrt{n}}$$

има разпределение на Стюдънт $T(n-1)$ с $(n-1)$ степени на свобода. Тогава от формула (11.1), в която $R = T(n-1)$ и $x_p = t_p(n-1)$, следва, че

$$P(t_{\alpha/2}(n-1) < Y < t_{1-\alpha/2}(n-1)) = 1 - \alpha.$$

Процедирайки както по-горе, получаваме, че доверителният интервал за m се задава с неравенствата:

$$\overline{x} - \frac{\overline{s}}{\sqrt{n}} t_{1-\alpha/2}(n-1) < m < \overline{x} + \frac{\overline{s}}{\sqrt{n}} t_{1-\alpha/2}(n-1)\,.$$

Затова намирането на доверителния интервал за $m = EX$ по зададено ниво на значимост α се извършва в следния ред:

1. Намираме точковата оценка за m:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k;$$

2. Намираме

$$D_n = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})^2, \quad \overline{D} = \frac{n}{n-1} D_n;$$

3. Определяме $\overline{s} = \sqrt{\overline{D}}$;

4. От Таблица 6 определяме квантила $x_\alpha = t_{1-\frac{\alpha}{2}}(n-1)$ на разпределението на Стюдънт $T(n-1)$ с $(n-1)$ степени на свобода и пресмятаме точността
$$\Delta = \frac{x_\alpha}{\sqrt{n}}\overline{s};$$

5. Тогава доверителният интервал за m е $I_p = (\overline{x} - \Delta, \overline{x} + \Delta)$ и m попада в този интервал с доверителна вероятност $p = 1 - \alpha$.

Пример 12.2. *Нека с.в. X – ръстът на $n = 1000$ ученици от пример* 10.2 *е разпределена по нормалния закон $N(m, \sigma)$ и стандартното отклонение σ_X не е известно. Да се намери доверителният интервали за $EX = m$ с доверителна вероятност $p = 0.90$.*

В пример 10.2 *е пресметнато, че $\overline{x} = 165.5$ см и $\sigma_X \approx \overline{s} = 6$ см. При $p = 0.90$ нивото на значимост е равно на $\alpha = 1 - 0.90 = 0.10$. От Таблица* 6 *определяме квантила $x_\alpha = t_{1-\frac{\alpha}{2}}(n-1) = t_{0.95}(999) = 1.645$. Тогава*
$$\Delta = \frac{6 \cdot 1.645}{\sqrt{1000}} \approx 0.31$$
и доверителният интервал е
$$I_{0.90} = (165.5 - 0.31, 165.5 + 0.31) = (165.19, 165.81)\,.$$

Забележка 12.4. *В разглеждания случай, едностранните доверителни интервали за $m = EX$ при ниво на значимост α се задават с неравенствата:*
$$m < \overline{x} + \frac{\overline{s}}{\sqrt{n}} t_{1-\alpha}(n-1), \tag{12.11}$$
(за ляв доверителен интервал);
$$m > \overline{x} - \frac{\overline{s}}{\sqrt{n}} t_{1-\alpha}(n-1), \tag{12.12}$$
(за десен доверителен интервал),
където квантилът $t_{1-\alpha}(n-1)$ се определя от таблицата за квантилите на разпределението на Стюдънт $T(n-1)$ с $(n-1)$ степени на свобода (Таблица 6*).*

12.7 Доверителен интервал за неизвестна вероятност

Точкова оценка за неизвестна вероятност $p = P(A)$ за сбъдване на дадено събитие A е относителната честотата w на настъпване на това събитие

$$\overline{p} = \frac{\overline{x}}{n} = w\,,$$

където n е броят на опитите, а $\overline{x}$ – броят на успешните опити, при които се е сбъднало събитието A.

Ако са изпълнени условията

$$n > 50,\ nw > 5,\ n(1-w) > 5, \tag{12.13}$$

то разпределението на случайната величина

$$Z = \frac{w-p}{\sqrt{\frac{p(1-p)}{n}}}$$

достатъчно точно се апроксимира със стандартното нормално разпределение U. Като отчетем този факт, може да получим следните формули за границите на доверителния интервал (p_1, p_2):

$$p_{1,2} \approx w \mp \frac{x_\alpha^2}{2}\sqrt{w(1-w)}, \tag{12.14}$$

$$p_{1,2} \approx \frac{1}{n+x_\alpha^2}\left(\overline{x} + \frac{x_\alpha^2}{2} \mp x_\alpha\sqrt{\frac{\overline{x}(n-\overline{x})}{n} + \frac{x_\alpha^2}{4}}\right), \tag{12.15}$$

където α е избраното ниво на значимост а $x_\alpha = u_{1-\frac{\alpha}{2}}$ е квантилът на стандартното нормално разпределение U, определен от Таблица 4.

Формула (12.15) дава по-точна оценка на доверителните граници, отколкото формула (12.14).

Забележка 12.5. *Едностранните доверителни интервали за неизвестната вероятност $p = P(A)$ при ниво на значимост α се задават с неравенствата:*

$$0 \le p < p_2 \quad (\textit{за ляв доверителен интервал}), \tag{12.16}$$

$$p_1 < p \le 1 \quad (\textit{за десен доверителен интервал}), \tag{12.17}$$

където доверителните граници p_1 *и* p_2 *се пресмятат по формулите* (12.14) *или* (12.15), *в които квантилът е равен на* $x_\alpha = u_{1-\alpha}$ *и се определя от Таблица* 4.

Пример 12.3. *При проверка на* 100 *детайла от голяма партида са намерени* 10 *дефектни. Да се намери* $95\%-$*ен доверителен интервал за относителния дял* p *на дефектните детайли в цялата партида.*

Имаме, че $n = 100$, $\overline{x} = 10$, $w = \dfrac{10}{100} = \dfrac{1}{10}$ *и* $1 - \alpha = 0.95$, *т.е.* $\alpha = 0.05$. *Понеже* $n > 50$, $nw = 10 > 5$, $n(1-w) = 90 > 5$, *то за намирането на границите на доверителните интервали може да се използват формулите* (12.14) *или* (12.15). *От Таблица* 4 *определяме квантила* $x_\alpha = u_{1-\alpha/2} = u_{0.975} = 1.96$. *Доверителният интервал по формула* (12.14) *ще бъде*

$$0.041 < p < 0.159.$$

По-точната формула (12.15) *дава малко по-друг резултат:*

$$0.055 < p < 0.174.$$

Забележка 12.6. *Ако някое от условията* (12.13) *е нарушено, то границите на доверителния интервал* (p_1, p_2) *се определят с използване на разпределението на Фишер* F *по формулите*

$$p_1 = \frac{\overline{x} F_{\alpha/2}(m_1, m_2)}{n - \overline{x} + 1 + \overline{x} F_{\alpha/2}(m_1, m_2)}\,, \tag{12.18}$$

където $m_1 = 2\overline{x}, m_2 = 2(n - \overline{x} + 1)$;

$$p_2 = \frac{(\overline{x}+1) F_{1-\alpha/2}(k_1, k_2)}{n - \overline{x} + (\overline{x}+1) F_{1-\alpha/2}(k_1, k_2)}\,, \tag{12.19}$$

където $k_1 = 2(\overline{x}+1), k_2 = 2(n - \overline{x})$.

В следващите три глави са изложени основните елементи на теорията за проверка на статистически хипотези, изградена предимно в трудовете на К. Пирсън, Р. Фишер, Е. Пирсън[1] и Е. Нейман[2].

На първо четене тези глави може да се пропуснат и да се премине направо към запознаването с глава 15, Анализ на зависимости.

[1] Пйрсън (Pearson) Егон (1895 – 1980) – английски математик, син на К. Пирсън
[2] Нèймам (Neyman) Ежи (1894 – 1981) – американски математик и статистик