ИЗБРАНИ ЛЕКЦИИ

по

СТАТИСТИКА

Автор:

Драго Йорданов Михалев

П Ъ Р В А Г Л А В А

ЗАКОН ЗА ГОЛЕМИТЕ ЧИСЛА. ГРАНИЧНИ ТЕОРЕМИ

§1. ВИДОВЕ СХОДИМОСТ НА РЕДИЦИ ОТ СЛУЧАЙНИ ВЕЛИЧИНИ

Определение. Редицата $\{X_n\}$ се нарича *сходяща по вероятност* към X и се записва $X_n \xrightarrow{P} X$, ако за всяко $\epsilon > 0$ имаме:

$$\lim_{n\to\infty} P\{|X_n - X| \geq \epsilon\} = 0 \ .$$

Забележка. Горното равенство означава, че редицата от вероятности на събитията $A_n = \{\omega : \ |X_n(\omega) - X(\omega)| \geq \epsilon\}$ $n = 1, \ 2, \ldots,$ клони към нула, т.е. $P(A_n) \to 0$ при $n \to \infty$.

Определение. Редицата $\{X_n\}$ се нарича *сходяща почти сигурно (с вероятност единица)* към X и се записва $X_n \xrightarrow{\text{П.С.}} X$, ако имаме:

$$P\{\lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1 \ ,$$

т.е. ако $X_n(\omega) \to X(\omega)$ при $n \to \infty$ за всяко $\omega \in \Omega$ с изключение може би на елементарни събития $\omega \in N$ от множество N с нулева вероятност $(P(N) = 0)$.

Определение. Редицата $\{X_n\}$ се нарича *сходяща по разпределение* към X и се записва $X_n \xrightarrow{d} X$, ако редицата от функции на разпределение $F_n(x) = P(X_n < x), \ n = 1, \ 2, \ldots,$ клони към $F(x) = P(X < x)$, т.е.

$$\lim_{n\to\infty} F_n(x) = F(x),$$

за всяка точка x на непрекъснатост за $F(x)$. Разпределението $F(x)$ се нарича *гранично или асимптотично* за X_n при $n \to \infty$.

Определение. Редицата $\{X_n\}$ се нарича *сходяща средно квадратично* към X и се записва $X_n \xrightarrow{\text{С.К.}} X$, ако имаме:

$$\lim_{n\to\infty} M((X_n - X)^2) = 0 \ .$$

Забележка. Доказва се, че от сходимостта по вероятност следва сходимост по разпределение, т.е. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$. Ако граничната случайна величина е константа C, то тези два вида сходимост са еквивалентни, т.е. $X_n \xrightarrow{P} C \iff X_n \xrightarrow{d} C$.

§2. НЕРАВЕНСТВО НА ЧЕБИШЕВ

Теорема 2.1. (неравенство на Чебишев) *Ако случайната величина X има математическо очакване $M(X) = E(X) = m$ и дисперсия $D(X) = \sigma^2$, то за всяко $\epsilon > 0$ е в сила:*

$$P\{|X - M(X)| \geq \epsilon\} \leq \frac{D(X)}{\epsilon^2}.$$

Доказателство. Нека непрекъснатата случайната величина X има плътност $f(x)$. Тогава имаме:

$$P\{|X - M(X)| \geq \epsilon\} = \int_{|x-M(X)|\geq\epsilon} f(x)dx \ .$$

За стойностите на x, за които е изпълнено $|x - M(X)| \geq \epsilon$ имаме $(x - M(X))^2 \frac{1}{\epsilon^2} \geq 1$ и тогава $f(x) \leq \frac{1}{\epsilon^2}(x - M(X))^2 f(x)$. Оттук получаваме :

$$P\{|X - M(X)| \geq \epsilon\} = \int_{|x-M(X)|\geq\epsilon} f(x)dx \ \leq$$

$$\leq \int_{|x-M(X)|\geq\epsilon} \frac{1}{\epsilon^2}(x - M(X))^2 f(x)dx \leq$$

$$\leq \frac{1}{\epsilon^2} \int_{-\infty}^{\infty} (x - M(X))^2 f(x)dx = \frac{D(X)}{\epsilon^2}.$$

Когато случайната величина е дискретна доказателството е аналогично на горното, но интегралът се заменя със сума. Тъй като е изпълнено

$$P\{|X - M(X)| \geq \epsilon\} + P\{|X - M(X)| < \epsilon\} = 1 \ ,$$

тогава получаваме:

$$P\{|X - M(X)| < \epsilon\} \geq 1 - \frac{D(X)}{\epsilon^2}.$$

Следствие 2.2. *Ако случайната величина X има математическо очакване $M(X) = E(X) = m$ и дисперсия $D(X) = \sigma^2$, то са в сила:*

$$P\{|X - M(X)| \geq t\sigma\} \leq \frac{1}{t^2}, \qquad P\{|X - M(X)| < t\sigma\} \geq 1 - \frac{1}{t^2},$$

$$P\{|X - M(X)| \geq 3\sigma\} \leq \frac{1}{9}, \qquad P\{|X - M(X)| < 3\sigma\} \geq \frac{8}{9}.$$

Предпоследното неравенство е често използвано и дава груба оценка за известното като „правило на трите сигми“ неравенство за вероятността $P\{|X - M(X)| \geq 3\sigma\}$. При различните разпределения, то може да се засили и дооточни.

Пример 2.1. Случайната величина X е разпределена по показателния закон с плътност

$$f(x) = \begin{cases} 0\ , & x \leq 0\ , \\ ae^{-ax}\ , & x > 0\ . \end{cases}$$

, т.е. има математическо очакване m и средно квадратично отклонение σ равни на $\frac{1}{a}$. Вероятността за изпълване на „правилото на трите сигми“ тогава е:

$$P\{X - m \geq 3\sigma\} = P\{X \geq 4\sigma\} = 1 - P\{X < \frac{4}{a}\} = 1 - F(\frac{4}{a}),$$

където $F(x) = 1 - e^{-ax}$ е функцията на разпределение на с.в.X. Тогава имаме:

$$P\{X - m \geq 3\sigma\} = 1 - F(\frac{4}{a}) = 1 - (1 - e^{-\frac{4a}{a}}) = e^{-4} \approx 0,0183.$$

Пример 2.2. Случайната величина X е нормално разпределена $N(m, \sigma)$, т.е. има математическо очакване m и дисперсия σ^2. Вероятността за изпълване на „правилото на трите сигми“ тогава е:

$$P\{|X - m| \geq 3\sigma\} = 1 - P\{|X - M(X)|\} < 3\sigma\} = 1 - 2\Phi(\frac{3\sigma}{\sigma}) =$$

$$= 1 - 2\Phi(3) \approx 1 - 2.0,49865 = 0,0027,$$

където $\Phi(x)$ е функцията на Лаплас. За нормалното разпределение само незначителна част от случайната величина (около 3 %) е извън интервала $[m - 3\sigma, m + 3\sigma]$.

§3. ЗАКОН ЗА ГОЛЕМИТЕ ЧИСЛА

Нека ни е зададена редицата от случайни величини:

$$X_1,\ X_2, \ldots,\ X_n, \ldots, \tag{3.1}$$

определени в едно и също вероятностно пространство. В частност може горната редица да е от независими случайни величини. Да положим

$$Y_n = \frac{X_1 + X_2 + \cdots + X_n}{n},\quad n = 1,\ 2, \ldots,$$

т.е. с.в. Y_n е средно аритметично на $X_1,\ X_2, \ldots,\ X_n$.

Определение. Ще казваме, че за редицата $\{X_n\}_{n=1}^{\infty}$ е в сила *законът за големите числа*, ако при $n \to \infty$ е изпълнено $Y_n - M(Y_n) \xrightarrow{P} 0$, т.е. ако за всяко $\epsilon > 0$ имаме:

$$\lim_{n\to\infty} P\{|Y_n - M(Y_n)| \geq \epsilon\} = \lim_{n\to\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} M(X_k)| \geq \epsilon\} = 0\ .$$

Определение. Ще казваме, че за редицата $\{X_n\}_{n=1}^{\infty}$ е в сила *усиленият закон за големите числа*, ако при $n \to \infty$ е изпълнено, че $Y_n - M(Y_n) \xrightarrow{\text{П.С.}} 0$, т.е. ако имаме:

$$P\{\lim_{n\to\infty} |Y_n - M(Y_n)| = 0\} = 1\ .$$

Ще отбележим, че за редицата $\{X_n\}_{n=1}^{\infty}$ от еднакво разпределени случайни величини с математическо очакване $M(X_n) = E(X_n) = m,\ n = 1,\ 2, \ldots,$ и от дадената дефиниция на Y_n следва, че е изпълнено $M(Y_n) = E(Y_n) = m$. В този случай законът за големите числа означава, че $Y_n \xrightarrow{P} M(Y_n) = m$.

Ще разгледаме някои достатъчни условия, при които е в сила законът за големите числа.

Теорема 3.1. (теорема на Чебишев) *Ако за редицата от независими случайни величини $\{X_n\}_{n=1}^{\infty}$ с дисперсии $D(X_n)$, които са ограничени от една и съща константа C , т.е. $D(X_n) \leq C,\quad n = 1,\ 2, \ldots,$ то за тази редица е в сила законът за големите числа, т.е. имаме:*

$$\lim_{n\to\infty} P\{|Y_n - M(Y_n)| \geq \epsilon\} = 0. \tag{3.2}$$

Доказателство. За дисперсията на $Y_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ имаме $D(Y_n) = \dfrac{1}{n^2}(X_1 + X_2 + \cdots + X_n) \leq \dfrac{1}{n^2}.nC = \dfrac{C}{n}, \quad n = 1,\ 2, \ldots .$ Прилагаме неравенството на Чебишев и получаваме:

$$P\{|Y_n - M(Y_n)| \geq \epsilon\} \leq \frac{D(Y_n)}{\epsilon^2} \leq \frac{C}{n\epsilon^2}.$$

Тъй като за всяко $\epsilon > 0$ е в сила, че $\dfrac{C}{n\epsilon^2} \to 0$ при $n \to \infty$, то следва :

$$\lim_{n\to\infty} P\{|Y_n - M(Y_n)| \geq \epsilon\} = 0.$$

Ще отбележим едно твърдение, доказано от Бернули(1654-1705) и наречено от него закон за големите числа, което е важно следствие от теоремата на Чебишев.

Теорема 3.2. (теорема на Бернули) *Ако ν_n е броят на случването на събитието A при n независими опита и $p = P(A)$ е вероятността за сбъдване на събитието A при всеки отделен опит. Тогава за всяко $\epsilon > 0$ е в сила:*

$$\lim_{n\to\infty} P\{|\frac{\nu_n}{n} - p| < \epsilon\} = 1\ . \tag{3.3}$$

Доказателство. Граничното равенство 3.3 е еквивалентно на

$$\lim_{n\to\infty} P\{|\frac{\nu_n}{n} - p| \geq \epsilon\} = 0$$

и означава, че относителната честота на събитието A клони към вероятността $p = P(A)$ при $n \to \infty$, т.е.

$$W_n(A) = \frac{\nu_n}{n} \xrightarrow{P} p = P(A)\ .$$

За всяко фиксирано n случайната величина ν_n има биномно разпределение и може да се представи като сума на n независими случайни величини $\nu_n = \sum_{k=1}^{n} X_k$, където X_k е броят на сбъдването на събитието A при k-тия опит. Тъй като $P(X_k = 1) = p, \quad P(X_k = 0) = 1 - p = q,$ то $M(X_k) = p, \quad D(X_k) = pq\ .$

Тогава за $Y_n = \frac{\nu_n}{n}$ имаме $M(Y_n) = \frac{np}{n} = p$ и понеже $D(X_k) = pq < 1$, от теоремата на Чебишев имаме

$$\lim_{n\to\infty} P\{|Y_n - M(Y_n)| \geq \epsilon\} = \lim_{n\to\infty} P\{|\frac{\nu_n}{n} - p| \geq \epsilon\} = 0,$$

т.е. в сила е:

$$\lim_{n\to\infty} P\{|\frac{\nu_n}{n} - p| < \epsilon\} = 1 \ .$$

Теоремата на Бернули предполага , че вероятността за появяване на събитието A при всеки опит е една и съща и е равна на $p = P(A)$, т.е. предполага се, че условията са постоянни. Ако условията се променят и вероятността за сбъдване на събитието A при k-тия опит означим с $p_k, \quad k = 1,\ 2, \ldots,$ то е в сила следното твърдение.

Теорема 3.3. (теорема на Поасон) *Ако ν_n е броят на случването на събитието A при n независими опита и p_k е вероятността за сбъдване на събитието A при при k-тия опит. Тогава за всяко $\epsilon > 0$ е в сила:*

$$\lim_{n\to\infty} P\{|\frac{\nu_n}{n} - \frac{1}{n}\sum_{k=1}^{n} p_k| < \epsilon\} = 1 \ . \tag{3.4}$$

Теорема 3.4. (теорема на Хинчин) *Ако за редицата от независими случайни величини $\{X_n\}_{n=1}^{\infty}$, които са еднакво разпределени с математическо очакване $M(X_n) = E(X_n) = m, \quad n = 1,\ 2, \ldots,$ то за тази редица е в сила законът за големите числа, т.е. имаме :*

$$\lim_{n\to\infty} P\{|Y_n - m)| \geq \epsilon\} = 0 \Longleftrightarrow \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{P} m.$$

Забележка. Теоремата на Бернули ни дава възможност приближено да определяме вероятността на събитието в опитите без схема от случаите, а по честотата на тези събития при достатъчно голям брой опити.

Теоремата на Поасон ни дава възможност приближено да намираме средната вероятност на събитието A в серия от опити, при които еднаквостта на условията е нарушена.

Законът за големите числа във всички негови форми има голямо значение в практическото приложение на вероятностните методи и в частност в инженерната практика.

§4. ЦЕНТРАЛНА ГРАНИЧНА ТЕОРЕМА

Нека ни е зададена редицата от независими случайни величини: $X_1,\ X_2, \ldots,\ X_n, \ldots$. Да положим:

$$S_n = X_1 + X_2 + \cdots + X_n, \quad Z_n = \frac{S_n - M(S_n)}{\sqrt{D(S_n)}}, \quad n = 1,\ 2, \ldots,$$

т.е. с.в. Y_n е средно аритметично на $X_1,\ X_2, \ldots,\ X_n$. Нека случайната величина Z има нормално разпределение $N(0,1)$ и за функция й на разпределение $\bar{F}(x)$ имаме :

$$\bar{F}(x) = \frac{1}{2} + \Phi(x), \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}}\, dt.$$

Определение. Всяко твърдение, съгласно което при определени условия редицата $\{Z_n\}$ е сходяща по разпределение към Z ($Z_n \xrightarrow{d} Z$) се нарича *централна гранична теорема* за дадената редица от случайни величини $\{X_n\}$.

От въведените горе полагания следва, че такава теорема е в сила, когато за всяко x имаме:

$$\lim_{n\to\infty} P\{\frac{S_n - M(S_n)}{\sqrt{D(S_n)}} < x\} = \bar{F}(x). \tag{4.1}$$

В такъв случай при достатъчно голямоn сумата $S_n = X_1 + X_2 + \cdots + X_n$ има приблизително (асимптотично) нормално разпределение. Ще представим следната централна гранична теорема, която се отнася за независими и еднакво разпределени случайни величини.

Теорема 4.1. *Ако за редицата от независими случайни величини* $\{X_n\}_{n=1}^{\infty}$, *които са еднакво разпределени с математическо очакване* $M(X_n) = E(X_n) = m$ *и имат крайна дисперсия* $D(X_n) = \sigma^2 > 0$ *и нека* $S_n = X_1 + X_2 + \cdots + X_n$, *тогава за тази редица е в сила централната гранична теорема, т.е. имаме :*

$$\lim_{n\to\infty} P\{\frac{S_n - m}{\sigma\sqrt{n}} < x\} = \bar{F}(x), \qquad x \in (-\infty, \infty).$$

Ще отбележим още един много важен частен случай на централната гранична теорема, който се отнася за биномното разпределение.

Теорема 4.2. (теорема на Моавър-Лаплас) *Ако ν_n е броят на случването на събитието A при n независими опита и $p = P(A)$ е вероятността за сбъдване на събитието A при всеки отделен опит. Тогава за произволни числа α и β $(\alpha < \beta)$ е в сила:*

$$\lim_{n\to\infty} P\{\alpha \le \frac{\nu_n - np}{\sqrt{npq}} < \beta\} = \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta e^{-\frac{t^2}{2}} dt =$$

$$= \bar{F}(\beta) - \bar{F}(\alpha) = \Phi(\beta) - \Phi(\alpha), \tag{4.2}$$

където $q = 1 - p$ и $\Phi(x)$ е функцията на Лаплас.

Ще представим и централната гранична теорема за редица от независими , но не непременно еднакво разпределени случайни величини.

Теорема 4.3. (теорема на Ляпунов) *Ако за редицата от независими случайни величини $\{X_n\}_{n=1}^\infty$, които имат математическо очакване $M(X_n) = E(X_n) = a_n$ и дисперсия $D(X_n) = b_n^2$ и $M(|X_n - a_n|^3) = c_n^3, \quad n = 1, 2, \ldots$. Тогава ако $A_n = \sum_{k=1}^n a_k, \quad B_n^2 = \sum_{k=1}^n b_k^2, \quad C_n^3 = \sum_{k=1}^n c_k^3$ и още $\lim_{n\to\infty} \frac{C_n}{B_n} = 0$, тогава е в сила централната гранична теорема, т.е. имаме :*

$$\lim_{n\to\infty} P\{\frac{X_1 + X_2 + \cdots + X_n - A_n}{B_n} < x\} = \bar{F}(x), \qquad . \tag{4.3}$$

Забележка. Горното равенство (4.3) означава, че при достатъчно големи n, случайната величина $S_n = X_1 + X_2 + \cdots + X_n$ има приблизително (асимптотично) нормално разпределение с параметри $M(S_n) = E(S_n) = A_n$ и $\sigma = \sqrt{D(X_n)} = B_n$.

Ще приведем без доказателство често използвана в статистиката формулировка на централната гранична теорема.

Теорема 4.4. *За всяка генерална съвкупност със средна μ и стандартно отклонение σ, при нарастване на обема на извадката n, извадковото разпрледеление на средната се доближава до нормалното разпределение със средна съвпадаща със средната на генералната съвкупност $\mu_{\bar{X}} = \mu$ и стандартно отклонение $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$, където N е обема на генералната съвкупност.*

Разгледаните гранични теореми имат голямо теоретично и приложно значение. Те дават обяснение на факта, че при моделиране на реални явления често се използва нормално разпределение.Това се отнася особено за случаите, когато върху даден параметър на явлението оказват влияние много случайни фактори със съществен сумарен ефект, но всеки отделен фактор има малко влияние върху параметъра.

Пример 4.1. Стан с числово-програмно управление произвежда за смяна $n = 1000$ изделия, от които средно 2 % дефектни. Да се намери приблизителната вероятност, че за смяна да бъдат произведени поне 970 недефектни изделия, ако изделията се оказват недефектни независимо едно от друго.

Решение. Вероятността за производство на доброкачествено (нормално) изделие е $p = 0,98$, Y е броя на нормалните изделия при n независими опита. Намираме:

$$M(Y) = np = 980; \qquad \sigma(Y) = \sqrt{npq} \approx 4,43;$$

$$np - 3\sigma \approx 980 - 13,3 > 0; \qquad np + 3\sigma < 1000.$$

Ползвайки нормалния закон за разпределение според теоремата на Моавър- Лаплас намираме:

$$P(Y \geq 970) = P(970 \leq Y < \infty) \approx 0,5 - \Phi(\frac{970 - 980}{4,43}) \approx 0,988.$$

Търсената вероятност е доста голяма (тя е 0,988), но все пак с вероятност 0,012 може да се очаква, че броят на нормалните изделия за смянаможе да бъде по-малък от 970.

Пример 4.2. Влакова композиция се състои от n вагона, теглото на всеки вагон в тонове е случайна величина X с математическо очакване m_X и стандартно отклонение σ_X. Локомотивът може да тегли композиция с тегло не повече от q тона; ако теглото на композицията надвишава q тона, то се налага да се сложи втори локомотив. Да се намери вероятността, че един локомотив да не е достатъчен, за да тегли композицията.

Решение. Означаваме с $Q = \sum_{k=1}^{n} X_k$ теглото на композицията. От централната гранична теорема имаме при достатъчно голямо n, че с.в. Q е разпределена приблизително по нормалния закон с параметри $m = nm_X$; $\sigma = \sqrt{n}\sigma_X$.

Търсената вероятност е равна на едно минус функцията на разпределение на случайната величина Q:

$$P(Q > q) = 1 - (0,5 + \Phi(\frac{q - nm_X}{\sqrt{n}\sigma_X}) \ .$$

Пример 4.3. Влакова композиция се състои от n_1 вагона, n_2 платформи и n_3 цистерни с тегло в тонове случайни величини съответно: за вагон- X_1 с математическо очакване m_1 и стандартно отклонение $\sigma_1 = \sqrt{D}_1$; за платформа- X_2 с математическо очакване m_2 и стандартно отклонение $\sigma_2 = \sqrt{D}_2$; за цистерна- X_3 с математическо очакване m_3 и стандартно отклонение $\sigma_3 = \sqrt{D}_3$; Предполагаме, че величините имат един и същи порядък и n_1, n_2, n_3 са достатъчно големи. Локомотивът може да тегли композиция с тегло не повече от q тона; ако теглото на композицията надвишава q тона, то се налага да се сложи втори локомотив. Да се намери вероятността, че един локомотив да не е достатъчен, за да тегли композицията.

`Решение.` Означаваме с Q теглото на композицията. По теоремата на Ляпунов можем да твърдим, че при достатъчно голямо n $(n = n_1 + n_2 + n_3)$ случайната величина Q има приблизително нормално разпределение с параметри :

$$m_Q = \sum_{k=1}^{3} n_k m_k; \qquad D_Q = \sum_{k=1}^{3} n_k D_k; \qquad \sigma_Q = \sqrt{D}_Q.$$

Вероятността един локомотив да не може да тегли сам композицията, приблизително пресмятаме по формулата:

$$P(Q > q) \approx 0,5 - \Phi(\frac{q - m_Q}{\sigma_Q}) \ .$$

Пример 4.4. Нека случайната величина X е разпределена по закона на Поасон с параметър a. Тогава, при достатъчно голямо a, вероятността $P_k = P(X = k)$ може да се смята приблизително по формулата:

$$P_k = P(X = k) = \frac{a^k}{k!} e^{-k} = P(k, a) \approx$$
$$\approx \Phi(\frac{k + 0,5 - a}{\sqrt{a}}) - \Phi(\frac{k - 0,5 - a}{\sqrt{a}}).$$

ВТОРА ГЛАВА
СЪЩНОСТ. ОСНОВНИ ПОНЯТИЯ. ЕТАПИ

§5. ВЪЗНИКВАНЕ, ПРЕДМЕТ, ОСНОВНИ ПОНЯТИЯ НА СТАТИСТИКАТА

1. Възникване и предмет на Статистиката

Статистиката като практика се е появила, за да отговори на определени потребности на човешкото общество. Статистиката има история и предистория още в дълбока древност от развитието на човечеството. Още в древността хората са се нуждаели от определени знания, които днес наричаме статистически. Има сведения, че около 3500 години преди н.е. в Египет е било извършвано преброяване на населението. Такова преброяване има в Китай 2000 години преди н.е., в Римската империя и др. Римският писател Тацит, живял по времето на император Август говори за преброяване на войската, на държавните разходи, на различните запаси, на богатството, включително корабите и др. Управлението на големите държави се е нуждаело от такава статистическа информация.

Статистиката като наука възниква по-късно, като много автори свързват статистиката с името на немския учен Херман Конринг, живял през XVII век и въвел през 1660 г. нова университетска дисциплина, наричана държавоведение. Слага се началото на една описателна наука, наречена статистика, която на основата на набраните сведения е трябвало да разказва за състоянието на отделните държави.

Днес статистическата наука и практика с развитието на обществото и информатиката достига висока степен на развитие и има широко приложение в различни сфери за описване, анализ и прогнозиране на явленията и процесите в заобикалящата ни действителност.

Ще дадем едно от определенията за статистиката , като наука.

Определение. *Статистика* се нарича наука за събиране, организация, представяне, анализ и интерпретация на данни с цел да се подпомогне вземането на решение.

Обект на всяко статистическо изследване са една или повече статистически съвкупности и съществените белези, свойства, закономерности

на тяхното проявление.

2. Статистически съвкупности. Видове

Определение. *Статистическа съвкупност* е множество от еднородни единици(случаи) по дефиниционни признаци, чрез които се проявява дадено масово явление, изучавано статистически в определени пространствени и времеви граници.

Статистическите съвкупности биват различни видове по различни признаци.

Моментни статистически съвкупности- изучаваните единици съществуват към определен момент, напр. населението на България към 31.12 2007г.

Периодични статистически съвкупности- изучаваните единици възникват към определен момент, но формират съвкупността за определен период от време, напр. произведения през 2007г брутен вътрешен продукт.

Общи (генерални) статистически съвкупности- обхващат всички единици на изучаваната съвкупност.

Представителни статистически съвкупности- обхващат определена част от единиците на общата (генерална) съвкупност.

Реални статистически съвкупности-обхваща действително съществуващи в определен период, място и време единици. Напр. обработваемата земя в България към 01.07.2001г е 49659хил.дка.

Хипотетични статистически съвкупности-безкрайна съвкупност, тя е логическа категория и има място при извадкови изучавания. Напр. потребителите на вносни стоки в България, очаквани раждания.

Интегрални статистически съвкупности- включват едни и същи единици за определен период от време и определено място. Напр. домакинствата в България през 2007г., населените места в България през 2007г.

Диференциални статистически съвкупности-включват различни единици през различни периоди. Например произведената продукция в България по години, сключените бракове по години.

3. Статистически единици. Признаци. Видове

Определение. *Статистическите единици* са елементи (отделни случаи или събития) от които се състои съвкупността. Те са еднородни, неделими и притежават определени качествени особености.

Статистическата единица може да има материален характер или да бъде събитие, факт, форма на организация. Биват прости и сложни.

Определение. *Статистическите признаци* са качествените особености и свойства на статистическите единици.

Съществуват различни критерии за квалификация на статистическите признаци. Според критерия“измерване“ със съответна мерна система те биват: вариационни (метрирани) и категорийни (неметрирани).

Определение. *Вариационни признаци* са тези признаци, които могат да се измерят с мерни единици и да се изразят с числа: възраст, метри плат, брой изделия, цена на стоки.

Определение. *Категорийни признаци* са тези признаци, които не могат да се измерят с мерни единици и да се изразят с числа, а се изразяват словесно (описателно): семейно положение, пол, професия, образование.

Вариационните признаци също се делят от своя страна по особеностите си:
прекъснати и непрекъснати; първични и вторични;
постоянни и променливи; факторни и резултативни.

Прекъснатите признаци- приемат се само отделни стойности, например цели: брой машини, деца, градове.

Непрекъснатите признаци- приемат стойности в определен интервал: работна заплата, цена на стока, продължителност на живот и др.

Първичните признаци - непосредствено присъщи признаци: годишна печалба,цена на кредит, количество вредни вещества от източник.

Вторичните признаци- производни свойства на статистическите единици, съотношение между два първични признака: например средна цена на кв.м., брутен продукт на човек, дял на безработни и др.

Постоянни признаци - не се променят във времето: година на раждане.

Променливи признаци - променят се във времето: потребление на лице от домакинсво, физически обем на брутен вътрешен продукт.

Факторни признаци- играят роля на фактори по отношение на други признаци: материални активи,трудов стаж, инвестиции.

Резултативни признаци- взаимодействат си с други признаци- фактори: брутна добавена стойност, производителност на труда.

4. Статистическо емпирично разпределение. Статистически характеристики.

Определение. *Статистическото разпределение* е обобщено подреждане по определени правила на единиците на изучаваната съвкупност според значенията на един или повече признаци.

Изразява се в относителни или абсолютни величини. Изразява връзката (закономерността) между значенията на признака и неговите честоти. Представя се таблично или графично.

Определение. *Обобщаващите статистически характеристики* се получават в резултат на статистическия анализ и в числен израз представят общото (типичното) и закономерното за изучавания статистически признак в съвкупността.

Такива са средна заплата, коефициентът на безработица, годишната инфлация в % през годината в България.

§6. СТАТИСТИЧЕСКО ИЗУЧАВАНЕ И ФАЗИТЕ МУ.

Определение. *Статистическото изучаване* има за обект изучаването на еднородни единици по съществени признаци, проявяващи се в определени пространствени и времеви граници. То преминава по правило през три взаимносвързани фази:
1. статистическо наблюдение;
2. групировка на статистическите сведения;
3. статистически анализ, изводи и заключения.

За събираните в статистическо наблюдение данни и измерителните скали, върху които ги отразяваме имаме различни квалификации и изисквания.

1. Измерителни скали

В зависимост от това дали данните са вариационни или категорийни, то скалите биват съответно *силни* или *слаби*.

Силните скали са също няколко вида:

Интервални- върху тях се нанасят стойностите на данните, такива каквито са и нулата също е стойност на данни , а не липса на данни. Такива са например- измерена температура, скорост, продължителност

на животи др. Тя позволява да се определи кое значение е по-голямо и с колко, но не и отношение на данните в пъти.

Относителната (пропорционалната) скала е най-силната скала. Тя има нула , която не е избрана произволно, нулата показва липсата на свойство. Например брой кредити нула значи няма кредити, брой дефекти нула - няма дефекти и др. При относителните скали се позволява действие делене и умножение за разлика от интервалните. Например ако производителността на труда във фирма А е 120лв, а във фирма В е 60лв, то се прави заключение, че производителността в А е два пъти повече от В.

Слабите скали биват:

Номинална скала е за категорийни признаци. Това са единиците попаднали в дадената категория, категориите трябва да са взаимно изключващи се. Такива са: разпределение на фирми по банки, които ги обслужват; броя на редовните слушатели на различните радиостанции и др. Те позволяват само броене на единиците за всяко значение на признака и пресмятане на относителния дял, но не и на други характеристики.

Ординална(рангова) скала е такава скала, при която признаците са подредени по възходящ или низходящ ред и между тях може да се установи разлика степен в дадено качество свойство на единиците. Например оценка за качество на даден продукт, степен на удовлетвореност от обслужване в даден хотел, класация на спортистите за спортист на годината и др. Те съдържат повече информация от номиналните скали, но разликата между отделните категории не може да бъде измерена числово.

2. Изисквания за събираните данни

Колкото по-вярно и точно отразяваме при измерването реалните свойства на изучаваните явление толкова по-адекватен е измерителят , който използваме.

Определение. *Адекватно отразяване* на статистическите данни означава, съответствието между обективните свойства на изучаваните явления и тяхното отразяване в статистически измерители, да може да се разглежда като „еднозначно“.

Точното и адекватно отразяване на данните осигурява тяхната достоверност и пригодност на метода.

Достоверни са резултатите от онези емпирични измерители, които

отговарят на критериите за *обективност, надеждност и валидност.*

Определение. *Достоверни* са резултатите от измерване, ако резултатите не зависят от субекта, който ги е извършил, т.е при повторно измерване биха се получили същите резултати.

Определение. *Обективен* е този измерител, който осигурява еднозначност на резултатите.

Определение. *Надежден* е този измерител, който осигурява едни и същи резултатите в два различни момента в достатъчно кратък интервал от време.

Определение. *Валидност* на измерителя означава съответствие между резултата от измерването и предвиденото за измерване, което се изисква от практическата постановка на задачата.

Един измерител може да е надежден, но не достатъчно валиден, когато отразява това, което не трябва.

Според обхвата си статистическите изучавания биват:
изчерпателни- обхващат всички единици от статистическата съвкупност;
частични- обхващат само част от единиците на статистическата съвкупност, поради нецелесъобразност или невъзможност за изчерпателно изучаване, ограничени средства, време и др.

§7. СТАТИСТИЧЕСКО НАБЛЮДЕНИЕ.

1. Същност, форми и методи

Определение. *Статистическото наблюдение* е първата фаза на статистическото изучаване при която се извършва планомерен и научнообоснован процес на събиране на сведения за отделните единици по съществени признаци на изучаваната статистическа съвкупност.

Извършването на статистическо наблюдение, включва също изработване на план, финансов разчет и осигуреност, определяне на местото времето и групите за изследване, изработване на формуляри и статистически листове и широтата на извадката, методологическа и организационна подготовка и др. Важни са също формулировките на въпросите в анкетните карти. Те трябва да са ясни, неподвеждащи и да не водят до двусмислие, а до ясен и точен отговор. В противен случай рискуваме

да съберем непълна информация или още по-лошо двусмислена информация, която после ще носи белезите на субективност при тълкуването ѝ.

Статистическите наблюдения се провеждат под три форми-
статистическа отчетност;
частични статистически наблюдения;
специално организирани статистически наблюдения.

Времето- определя се според изучаваната статистическа съвкупност. То трябва да е такова, че изучаванатаСС да е в нормално състояние. Например преброяването на населението се извършва през зимата, хората не са по отпуски.

Определя се критичния момент, който разделя два периода, т.е. две съвкупности по време. Преброяването у нас се е извършило към 0 часа на 1 март 2001 година.

Период на регистрация е времето през което се извършва анкетирането. Стараем се този период да е най-кратък.

Методите на извършване на статистически наблюдения са три:
самонаблюдение- извършва се от самите анкетирани;
кореспондентски метод- определени лица кореспонденти дават сведения за статистическите единици;
експедиционен метод - изпращат се специално подготвени за наблюдението лица, които регистрират на място сведенията.

Източниците на сведения могат да бъдат:
документални източници-първичните документи, намиращи се във всяка отчетна единица ;
наблюдението- контролира се ефектът на един или друг фактор върху изучавания признак, например чрез експеримент;
интервюто с неговите разновидности-
персонално интервю- предпочита се защото дава представа от реакцията на хората;
телефонно интервю, макар и по-евтино то не винаги дава точна информация и често анкетираните не дават информация изобщо, защото считат, че това е намеса в живота им;
пощенска анкета- има същите недостатъци като телефонното интервю;
фокусни групи- събират се по 10-12 човека и анкетата се води от опитен модератор, който има грижата да зададе всички въпроси и да формулира хипотези и да насочи следващото изследване;
тест на място на продажба или промоция и др.

Мястото на изследването е територията, където се помещават статистическите единици. При специално организираните наблюдения се

извършва райониране сточно определени граници. Например при преброяване на населението, при избори и др.

2. Грешки, видове грешки

Грешките при събиране на данните се наричат грешки при регистрация. Те биват два вида: *систематични и случайни(стохастични).*

Системните грешки изкривяват истинските значения само в едната посока и оказват сериозно влияние за изкривяване на резултата само в тази посока и води до неверни изводи. Например хората си завишават образованието, занижават си годините и доходите и т.н. Те могат да се дължат и на неправилно определена генерална съвкупност или извадката е правена само в част, която се различава много от цялата общност. Например само в големи градове се вземат доходите.

Случайните(стохастичните) грешки са такива, които са ту в едната ту в другата посока и се компенсират при голям брой единици. Например при опити с извадки всеки път се вземат различни елементи, които дават различни отклонения, но при голям брой опити и голям размер на извадката отклоненията се компенсират. Това е същността на Закона за големите числа. Той гласи изказано в разговорен език, че натрупването на голям брой случайности води до закономерност, т.е. много малки грешки в различна посока водят до компенсирането им и до истинските (верните) стойности (значения).

Грешките биват също: съзнателни и несъзнателни грешки.

Съзнателните грешки са резултат от преднамерено скриване на истината и води до изкривена невярна информация.

Несъзнателните грешки са допуснати по невнимание, грешки от пресмятане, забравяне на факти и събития, без преднамерено премълчаване на истината.

За премахване на грешките и подобряване на организацията на наблюдението и статистическото изследване като цяло се извършват пробни (пилотни) изследвания.

Друг метод е постоянният и текущ контрол, който се осъществява и по време на самото наблюдение, например чрез дублирано събиране на информация. Прави се контрол дали няма дублиране на информация, т.е. дали от един обект не е взета два или повече пъти информация, която да се представя за различна. Прави се и аритметичен контрол, дали правилно е извършено пресмятането.

Т Р Е Т А Г Л А В А

СТАТИСТИЧЕСКО ГРУПИРАНЕ. ПРЕДСТАВЯНЕ НА ДАННИ.СРЕДНИ.

§8. СТАТИСТИЧЕСКО ГРУПИРАНЕ

1. Същност и етапи

При статистическото групиране се образуват групи, в които попадат единици с еднакви или близки значения по един или повече признаци. За извършване на групирането е необходимо да се познава същността на изучаваното явление, неговия строеж и целта на статистическото изучаване. Статистическата групировка има три етапа:
определяне на групите;
разпределянена единиците по групи;
преброяване на единиците в групите и записване на техния брой за всяка група.

Определяне на групите. Това е най-важният етап. Определят се признаците по които ще се делят групите, броят на групите, ширината и границите им, когато е необходимо.

Разпределяне на единиците по групи. В зависимост от притежаваното значение всяка единица трябва да попадне точно в една група.

Преброяване на единиците в групите. Тук може да се използва както и в предния етап компютър , но за осъществяване на първия етап техниката не е достатъчна- важна е ролята на специалиста.

2. Видове групировки

Статистическите групировки биват *прости* и *сложни*, според това дали групировката се извършва според един или повече от един статистически признаци. Сложните групировки позволяват по-пълно да се разкрие вътрешната структура и зависимост в съвкупността.

Статистическите групировки се различават според вида на групировъчните признаци , а имено по:
-вариационни признаци;
-категорийни признаци;
-териториални признаци;
- признаци по време.

Вариационна групировка-извършва според стойностите на вариационния признак. При голям брой значения на вариационния признак има ненужно раздробяване на съвкупността. В тези случаи се прилагат интервалните групировки. В основата им стои съотношението между броя на групите и ширината на интервалите. Интервалното групиране се извършва на три принципа: аритметичен; геометричен; целеви.

Отличителен белег на аритметичния принцип е еднаквата ширина на груповите интервали. Ако k е броя на групите, то ширината h се получава от отношението на разликата от максималното x_{max} и минималното значение x_{min} на признака и броя на групите:

$$h = \frac{x_{max} - x_{min}}{k}.$$

При съвкупности с голям обем (N) ширината на груповия интервал се определят по формулата на Стърджис:

$$h = \frac{x_{max} - x_{min}}{1 + 3,322.lgN}.$$

Аритметичният принцип има редица предимства и затова използването му е най-голямо.

Геометричният принцип - при него ширината на интервалите се изменя в геометрична прогресия. Използва се когато данните са с неравномерни значения в широки граници. Групите са качествено нееднородни, което ограничава аналитичните му възможности.

Целевите групировки също образуват интервали с различна ширина, но единиците са по еднородни, което съответства на вътрешния строеж на съвкупността. Намира широко приложение в икономиката.

Важно значение при вариационните групировки е точното разграничаване между долната и горна граница на два съседни интервала. При непрекъснатите величини обикновено важи правилото „над-до“. (Отворен интервал отдолу и затворен отгоре При прекъснатите величини обикновено важи правилото „от-до“.

Когато крайните групи съдържат малко на брой значения се образуват групи с отворена горна или долна граница „над“ и „до“.

Категорийна групировка- когато се извършва по признак отразен на номиналната или ординалната скала, т.е. значенията на признака се проявяват като определения(наименования) на качествен признак и групите са естествено обусловени. При малък брой качествени определения на признака(пол, семейно положение, вид населено място) групировката се образува непосредствено от определенията на признака.

Териториална групировка групировката се извършва по териториални поделения или райони, формирани по различни принципи(административно деление, избирателен район, географска област и др.).

Групировка по време- според времето на проявяване на значенията в по-кратки или по-широки интервали от време, те зависят от спецификата на явленията (годишна инфлация, месечна заплата, валежите по годишни сезони, злополуките по часове от работен ден и др.).

3. Абсолютни , относителни и кумулативни честоти.

Броят на единиците в отделните групи се нарича *абсолютна честота*, а делът на всяка група от общата численост на съвкупността се нарича *относителен дял (относителна честота).*

Кумулативната (натрупана) честота бива два типа отново: абсолютна и относителна, в зависимост от това дали ползват абсолютните или относителни честоти.

От друга страна те биват „позитивни“, когато принципа е „отгоре - надолу“ или негативни, когато принципа е „отдолу-нагоре“. *Кумулативната честота* съдържа (е сума на) честотите на всички предходни групи, включително и на честотата на настоящата група, според споменатите вече принципи.

При позитивните кумулативни честоти първата кумулативна честота е равна на честотата, а последната е равна на броя на единиците в съвкупността за абсолютните и на единица при относителните такива.

При негативните кумулативни честоти последната (най-долната) кумулативна честота е равна на честотата, а първата (най горната) е равна на броя на единиците в съвкупността за абсолютните и на единица при относителните такива.

Пример 8.1. Разполагаме с таблица от данни, в които е представено разпределението на 25 специалисти от фирма, според средното време за извършване на определена операция, а именно:

Време в минути	Абс. чест.	Отн. чест.	Кум.ч. абс.поз.	Кум.ч. абс.нег.	Кум.ч. отн.поз.	Кум. ч. отн.нег.
5,5-7,5	2	0,08	2	25	0,08	1,00
7,5-9,5	8	0,32	10	23	0,40	0,92
9,5-11,5	10	0,40	20	15	0,80	0,60
11,5-13,5	5	0,20	25	5	1,00	0,20
Общо	25	1,0	-	-	-	-

.

4. Статистически редове и таблици.

Статистическите редове са резултат от статистическите групировки.

Статистическите редове са подредени по определени правила статистически данни.

Статистическия ред изразява разпределението на единиците от изучаваната съвкупност според определенията на групировъчния признак. Всеки статистически ред съдържа задължително елементите: заглавие; основание; членове.

Заглавието изразява съдържанието по същество, място и време на статистическата съвкупност.

Основанието на статистическия ред са групите на съответния групировъчен признак.

Членовете на статистическите редове са числовите характеристики (честотите) съответстващи на отделните групи. Могат да бъдат относителни, абсолютни или кумулативни честоти или няколко от тях.

Статистическите таблици са удобна форма за представяне на статистическите данни, подредени по определени правила, т.е.на статистическите редове.

Статистическата таблица съдържа формални елементи, които описват макета й:

заглавие; челна колона;

заглавен ред; клетки.

Заглавието трябва да е кратко и ясно и да не дублира заглавния ред.

Заглавния ред е първия ред на таблицата и определя смисъла на съдържащите се колони от статистическите данни.

Челната колонае първата колона и определя смисъла на статистическите данни по редове.

Клетките на таблицата се запълват с данни, които изразяват комбинация между значенията на признаците по редове и колони на таблицата.

Например таблицата:

Заети по територия и пол през декември 2000 година (хиляди)

Територия	Общо	Мъже	Жени
Общо Б-я	2735,5	1453,1	1282,4
Северна Б-я	941,7	508,7	433,0
Южна Б-я	1270,3	678,7	591,6
София	523,5	265,7	257,8

.

§9. СРЕДНИ ВЕЛИЧИНИ

1. Същност и видове средни величини

Средните величини са свързани с онези признаци и свойства на единиците от статистическата съвкупност, които са количествено измерими и варират. Те уравновесяват различията и чрез тях се определя най-често срещащото се значение на даден признак.

Средните величини разкриват *общите, типичните, закономерно* проявяващи се свойства на съвкупността.

Съществена предпоставка за тази функция на средните са следните условия: *еднородност* на единиците по изучавания принцип; *достатъчно голям брой* единици от съвкупността.

Основно средните величини се делят на два типа: алгебрични и неалгебрични.

Алгебричните средни - при определянето им участват всички значения на осреднявания признак.

При *неалгебричните средни* - за определянето им участват само определени значения в зависимост от мястото или честота им.

Средните се деля още например на:
притеглени и непритеглени - в зависимост от това дали са групирани или не данните, съответно.
генерални и извадкови - в зависимост дали са взети генерални съвкупности или извадки от тях.
общи и групови - зависи от обхвата на съвкупността по която се изчисляват, за общите средни- цялата съвкупност, за груповите средни-значението на подсъвкупност или група от съвкупността.

2. Изчисляване на средните величини- формули и означения

I. Средни алгебрични.

1. Средна аритметична ($\bar{x}$).
а) непритеглена средна аритметична(при негрупирани данни):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \ , \tag{9.1}$$

б) притеглена средна аритметична(при групирани данни):

$$\bar{x} = \frac{\sum\limits_{i=1}^{k} x_i f_i}{\sum\limits_{i=1}^{k} f_i} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} \ , \tag{9.2}$$

където сме означили с :
x_i- индивидуалните значения на признака при негрупирани данни или средата на интервала при групирани данни;
f_i- брой на единиците (честотата), с която се срещат значенията на изучавания признак в i-тия интервал (група);
k- брой на интервалните поделения (на групите).

2. Средна хармонична ($\bar{x}_h$).

а) непритеглена средна хармонична(при негрупирани данни):

$$\bar{x}_h = \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \ , \tag{9.3}$$

б) притеглена средна хармонична(при групирани данни):

$$\bar{x}_h = \frac{\sum\limits_{i=1}^{k} f_i}{\sum\limits_{i=1}^{k} \frac{f_i}{x_i}} = \frac{f_1 + f_2 + \cdots + f_k}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_k}{x_k}} \ , \tag{9.4}$$

където сме означили с:
$\frac{1}{x_i}$ - реципрочните значения на изучавания вариационен признак.

3. Средна геометрична ($\bar{x}_G$), средногеометричен темп $\bar{T}$.
а) непритеглена (при негрупирани данни):

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1.x_2.\ldots.x_n} \ , \tag{9.5}$$

$$\bar{T} = \sqrt[n]{\prod_{i=1}^{n} T_i} = \sqrt[n]{T_1.T_2.\ldots.T_n} \ , \tag{9.6}$$

б) притеглена (при негрупирани данни):

$$\bar{x}_G = \sum f_i \sqrt{\prod_{i=1}^{k} {x_i}^{f_i}} = \sum f_i \sqrt{{x_1}^{f_1}.x_2^{f_2}.\ldots.x_k^{f_k}} \ , \tag{9.7}$$

$$\bar{T} = \sum t_i \sqrt{\prod_{i=1}^{k} {T_i}^{t_i}} = \sum t_i \sqrt{{T_1}^{t_1}.T_2^{t_2}.\ldots.T_k^{t_k}} \ , \tag{9.8}$$

I. Средни неалгебрични.

4. Медиана (средна по положение) (M_e)-разделя реда от подредени по големина единици на значенията на признака на две части.

а) при негрупирани данни:
при нечетен брой елементи $(n = 2k-1): \quad x_1,\ x_2, \ldots,\ x_k, \ldots,\ x_{2k-1}$;
$M_e = x_k$, т.е. ’
за номера на елемента имаме $i_{M_e} = \dfrac{n+1}{2} = k$;

при четен брой елементи $(n = 2k): \quad x_1,\ x_2, \ldots,\ x_k,\ x_{k+1}, \ldots,\ x_{2k}$;
$M_e = \dfrac{x_k + x_{k+1}}{2}$.
За номера на елемента отново имаме .

б) при групирани данни:

$$M_e = L_{M_e} + (\frac{\sum f + 1}{2} - C_{M_{e-1}})\frac{h}{f_{M_e}} \ , \tag{9.9}$$

където:
L_{M_e} - долна граница на медианния интервал;
i_{M_e}- номер на медианния интервал, носител на численото значение на

медианата, където $i_{M_e} = \dfrac{\sum f + 1}{2}$;
$C_{M_{e-1}}$- кумулативна честота в предмедианния интервал;
f_{M_e}- фактическа честота на медианния интервал;
h - ширина на интервала.

5. Квартили (средни по положение) ($Q1,\ Q2,\ Q3$)-разделят реда от подредени по големина единици на значенията на признака на четири равни части.

$$Qi = L_{Qi} + (\frac{\sum f + 1}{4} i - C_{Qi-1}) \frac{h}{f_{Qi}} \ , \tag{9.10}$$

където:
L_{Qi} - долна граница на квартилния интервал;
i- пореден номер на квартила $i = 1,\ 2,\ 3$;
C_{Qi-1}- кумулативна честота в предквартилния интервал;
f_{Qi}- фактическа честота на квартилния интервал;
h - ширина на интервала.

6. Персентили (средни по положение) ($P1,\ P2, \ldots,\ P99$)-разделят реда от подредени по големина единици на значенията на признака на сто равни части.

$$Pi = L_{Pi} + (\frac{\sum f + 1}{100} i - C_{Pi-1}) \frac{h}{f_{Pi}} \ , \tag{9.11}$$

където:
L_{Pi} - долна граница на персентилния интервал;
i- пореден номер на персентила $i = 1,\ 2, \ldots,\ 99$;
C_{Pi-1}- кумулативна честота в предперсентилния интервал;
f_{Pi}- фактическа честота на персентилния интервал;
h - ширина на интервала.

7. Мода (средна по честота) (M_o)- точката (точките) с най-голяма честота.

а) при негрупирани данни:

$$x_1,\ x_2,\ x_3,\ x_3, x_4,\ x_5,\ x_6,\ x_7,\ x_8 \quad | \qquad M_o = x_3,$$

б) при групирани данни:

$$M_o = L_{M_o} + \frac{f_{M_o} - f_{M_o-1}}{2f_{M_o} - f_{M_o-1} - f_{M_o+1}} h \ , \tag{9.12}$$

където:
L_{M_o} - долна граница на модалния интервал;
f_{M_o}- брой единици в модалния интервал (с най-висока честота в разпределението);
f_{M_o-1}- брой единици в предмодалния интервал;
f_{M_o+1}- брой единици в следмодалния интервал;
h - ширина на интервала.

Пример 9.1. Индивидуалните заплати в цех през януари 2002г за произволно взети 12 наети лица са:
210, 220, 220, 215, 150, 225, 225, 225, 110, 300, 320,450. Да се изчисли средната заплата, като се използват различни средни (средно аритметично, медиана и мода).

Решение. Ще пресметнем първо средното аритметично ($\bar{x}$) в лева , защото при пресмятането му участват всички стойности ($n = 12$) на признака (заплатата). Тогава по формула (9.1) имаме:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2870}{12} = 239,2.$$

За медианата M_e при негрупирани данни имаме, че тя се намира в средата на реда от възходящо подредени стойности на работната заплата. Номерът й се намира по формулата: $i_{M_e} = \frac{n+1}{2} = \frac{12+1}{2} = 6,5$.
Възходящият ред на заплатата има вида:
110, 150, 210, 215, 220, 220, 225, 225, 225, 300, 320, 450
Тогава за стойността на медианата M_e в лева имаме, че е сумата на шестия и седмия член, разделена на две, т.е. :

$$M_e = \frac{220 + 225}{2} = 222,5.$$

Модата при негрупирани данни е най-често срещаната стойност на признака (заплатата). В случая имаме, че 225 се среща най-много (три) пъти. Тогава имаме: M_o=225 лв.

Интерпретация. Неалгебричните средни дават приблизително еднакви размери на средната заплата така че няма особено значение коя от тях ще изберем за характеризациях на средната заплата, но те използват само част от данните. Средната аритметична използва всички данни и би трябвало да е по-точна. Ако се елиминират крайните стойности: 110, 150 оттдолу и 300, 320, 450 отгоре и от останалите 7 сметнем

средната аритметична ще получим средна от 220 лв., това е така, защото интервала от 210 до 225, в който се намират 7-те заплати съдържа по-малко различия, т.е. единиците са по-еднородни. (Заплатите в долния край не са характерни за повечето, работещи цеха (напр. на помощен персонал, чистачи, пазачи и др.), същото важи и за заплатите в горния край (напр. началник цех, началник смяна и др.).)

Тъй като броят на елементите е малък, а средната аритметична е подходяща при достатъчно голям обем на съвкупността когато диспропорциите в края се компенсират и изпъква най-често срещаното значение.

Извод. Средната работна заплата на работещите в цеха през януари 2002 година е 222,5 лева.

Пример 9.2. Разпределението на наетите във фирма през януари 2002г според размера на заплатит е както следва:

№	Запл. от-до y_i	брой f_i	кум. абс чест	отн. дял $P_i = \frac{f_i}{\sum f_i}$	ср. на инт. x_i	$x_i.f_i$	$x_i.P_i$
1	до 110	1	1	0,005	90	90	0,45
2	110-150	20	21	0,111	130	2600	14,43
3	150-190	30	51	0,167	170	5100	28,39
4	190-230	40	91	0,222	210	8400	46,62
5	230-270	50	141	0,278	250	12500	69,50
6	270-310	20	161	0,111	290	5800	32,19
7	310-350	10	171	0,056	330	3300	18,48
8	над 350	9	180	0,050	370	3300	18,50
-	сума	180	-	1,0	-	41120	228,56

.

Да се изчисли средната заплата, като се използват различни средни (средно аритметично, медиана и мода). Да се пресметнат и квартилите и персентилите P_3, P_{97} за съвкупността.

Решение. Ще пресметнем първо *претегленото средно аритметично* ($\bar{x}$) в лева , защото при пресмятането му участват всички среди на интервалите ($n = 8$) и честотите на признака (заплатата) в тях. То-

гава имаме за $\bar{x}$ в лева:

$$\bar{x} = \frac{\sum\limits_{i=1}^{k} x_i f_i}{\sum\limits_{i=1}^{k} f_i} = \frac{90 + 2600 + \cdots + 3300}{1 + 20 + \cdots + 9} = \frac{41120}{180} = 228,44 \ .$$

Същият резултат може да се получи и чрез използването на последната колона на таблицата, тогава ще получим: $\bar{x} = 228,56$, разликата от 0,12 лева е в резултат от предварителното закръгляне.

Средната заплата във фирмата през януари 2002 година е 228,44лева.

Модата на групирани данни в лева ще приесметнем като намерим интервала (над 230 до 270 в случая) с най-голяма честота (50 в случая) и приложим формула (9.12) :

$$M_o = L_{M_o} + \frac{f_{M_o} - f_{M_o-1}}{2f_{M_o} - f_{M_o-1} - f_{M_o+1}} h =$$

$$= 230 + \frac{50 - 40}{2.50 - 40 - 20}.40 = 230 + 10 = 240,00 \ .$$

Медианата на групирани данни в лева ще пресметнем като намерим интервала, чиято кумулативна честота е повече от половината , а на предходния е по-малка от половината, т.е. номера му определяме от $i_{M_e} = \frac{n+1}{2} = \frac{181}{2} = 90,5$ и това е интервала „над 190 до 230“. Тогава имаме:

$$M_e = L_{M_e} + (\frac{\sum f + 1}{2} - C_{M_{e-1}})\frac{h}{f_{M_e}} =$$

$$= 190 + (90,5 - 51)\frac{40}{40} = 190 + 39,5 = 229,5.$$

Резултатите показват, че медианната средна е по-голяма от средната аритметична и е по-малка от модалната средна, т.е.

$$\bar{x} = 228,4 < M_e = 229,5 < M_o = 240.$$

Тъй като медианната средна и средната аритметична се различават с около лев, то е почти без значение кое от тях ще вземем за средна заплата. Данните са достатъчно голям брой и понеже средната аритметична използва всички данни и дава състоятелна, неизместена и ефективна оценка и това й дава предимство. Средната заплата в фирмата, която разглеждаме е 228,4 лева.

Квартилите на групирани данни $Q1,\ Q2,\ Q3$ в лева ще пресметнем като намерим интервалите, чиято кумулативна честота е повече от четвъртината, от половината , от три четвърти, съответно, а в предходния интервал е по-малка от тези стойности. Ясно е, че $M_e = Q2 = 229,5$. Ще ни е необходимо още стойността на $\frac{\sum f+1}{4} = \frac{181}{4} = 45,25$. Тогава имаме :

$$Qi = L_{Qi} + (\frac{\sum f+1}{4}i - C_{Qi-1})\frac{h}{f_{Qi}},$$

$$Q1 = 150 + (45,25 - 21)\frac{40}{30} = 182,3\ ,$$

$$Q3 = 230 + (45,25.3 - 91)\frac{40}{50} = 265,8\ .$$

Персентилите на групирани данни $P_1,\ P_2, \ldots, P_{99}$ в лева ще пресметнем като намерим интервалите, чиято кумулативна честота е повече от, $1\%,\ 2\%, \ldots, 99\%$, а в предходния интервал е по-малка от тези стойности. Ще ни е необходимо още стойността на $\frac{\sum f+1}{100} = \frac{181}{100} = 1,81$. Тогава имаме:

$$Pi = L_{Pi} + (\frac{\sum f+1}{100}i - C_{Pi-1})\frac{h}{f_{Pi}}\ ,$$

$$P_3 = 110 + (5,43 - 1)\frac{40}{20} = 119\ ,$$

$$P_{10} = 110 + (18,1 - 1)\frac{40}{20} = 144,2\ ,$$

$$P_{97} = 350 + (181 - 5,43 - 171)\frac{40}{9} = 370,3\ ,$$

$$P_{90} = 310 + (1,81.90 - 161)\frac{40}{10} = 317,6\ .$$

Очевидно е, че са в сила следните равенства:

$$P_{25} = Q1 = 182,3;\quad P_{50} = Q2 = M_e = 229,5;\quad P_{75} = Q3 = 265,8.4$$

Персентилите са в основата за определяне числените граници, в които се намира всяка нормативна група: „силно изоставащи“, „ изоставащи“, „под нормата“, „в нормата“, „над нормата“, „изпреварващи“ и „силно изпреварващи“.

§10. ИЗМЕРИТЕЛИ НА РАЗСЕЙВАНЕ

Различията между единиците по изучавания статистически признак са обективно съществуващи и присъщи на статистическата съвкупност. Средните величини балансират (уравновесяват) индивидуалните различия между единиците по изучаваните статистически признаци и изразяват типичното най-често срещащо се значение под влияние на трайно действащите фактори.

Статистическото разсейване или вариацията изразяват колебанията, различията, отклоненията между единиците по изучавания статистически признак в статистическата съвкупност.

Измерителите на различията между единиците по изучавания статистически признак характеризират в обобщен вид степента на неоднородност на статистическата съвкупност.

Средните величини на равнище и средните измерители на статистическото разсейване са основните характеристики на емпиричните статистически разпределения.

1. Абсолютни и относителни мерки на разсейването

Абсолютните мерки на разсейването са: *размах, средно аритметично, средно квадратично огклонение, квартилно отклонение и средна разлика*

Относителни мерки на разсейването са ненаименовани величини, които изразяват в проценти делът на различията от средната аритметична. Известни са като *коефициенти на вариацията.* Те позволяват да бъдат сравнявани разноименни статистически признаци.

Всеки измерител на разсейване може да бъде представен като абсолютна или относителна величина.

1. Размах -ширина на вариацията- R.

Измерва разсейването с числената разлика между крайните значения на измервания признак- максималното x_{max} и минималното значение x_{min}, т.е. имаме:

$R = x_{max} - x_{min}$ - абсолютна мярка;

$V_R\% = \frac{x_{max} - x_{min}}{\bar{x}} 100$ - относителна мярка,

където $V_R\%$ е коефициент на вариация за размаха и отразява в проценти относителния дял на размаха от средната аритметична.

2. Средно аритметично отклонение- δ (делта).

По-точна мярка за измерване на разсейването от размаха. Изчисляването на абсолютната мярка на средното аритметично отклонение зависи от това групирани ли са или не данните и се изчислява чрез:

$$\delta = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad \text{- за негрупирани данни;}$$

$$\delta = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| f_i}{\sum f_i} \quad \text{- за групирани данни.}$$

За относителната мярка имаме:

$$V_\delta\% = \frac{\delta}{\bar{x}} 100 \quad \text{- относителна мярка,}$$

където $V_\delta\%$ е коефициент на вариация за средното аритметично отклонение и отразява в проценти относителния дял на средното аритметично отклонение от средната аритметична.

3. Дисперсията- σ^2 (сигма на квадрат).

Дисперсията е точен измерител на разсейването и се основава на това, че сумата от квадратите на отклоненията на индивидуалните значения от средната аритметична да е минимално, тогава имаме:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{1}{n}(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})$$

- за негрупирани данни;

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{\sum f_i} = \frac{1}{\sum f_i}(\sum_{i=1}^{n} x_i^2 f_i - \frac{(\sum_{i=1}^{n} x_i)^2}{\sum f_i})$$

- за групирани данни.

При обем на съвкупността, по-малък от 30 ($n = \sum f_i \leq 30$), делителят при изчисляване на дисперсията е $(n-1)$, респективно $(\sum f_i - 1)$.

4. Средно квадратично (стандартно) отклонение- σ.

Средно квадратично (стандартно) отклонение е точен и най-често използван измерител на разсейването. Пресмята се като корен квадратен от дисперсията, т.е. $\sigma = \sqrt{\sigma^2}$. Тогава имаме:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{1}{n}(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})}$$

- за негрупирани данни;

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{\sum f_i}} = \sqrt{\frac{1}{\sum f_i}(\sum_{i=1}^{n} x_i^2 f_i - \frac{(\sum_{i=1}^{n} x_i)^2}{\sum f_i})}$$

- за групирани данни.

При обем на съвкупността, по-малък от 30 ($n = \sum f_i \leq 30$), делителят при изчисляване на дисперсията е $(n-1)$, респективно $(\sum f_i - 1)$.

Средно квадратично (стандартно) отклонение и средната аритметична са основните характеристики на статистическите разпределения. Те се наричат *параметри* на статистическото разпределение.

В теорията на управление на пазарния риск и доходността на портфейла стандартното отклонение се разглежда като специфичен измерител на риска, известен като *волатилитет.*

За относителната мярка- коефициент на вариация на σ имаме :

$$V_\sigma\% = \frac{\sigma}{\bar{x}}100.$$

5. Квартилно отклонение- Q.

Смята се като полуразлика между третия и първия квартил, т.е. имаме:

$$Q = \frac{Q_3 - Q_1}{2}.$$

Относителната мярка се смята като отношение на квартилната полуразлика върху квартилната полусума, умножено по 100, т.е. за квартилния коефициент на вариация имаме:

$$V_Q\% = (\frac{Q_3 - Q_1}{2} : \frac{Q_3 + Q_1}{2})100 = \frac{Q_3 - Q_1}{Q_3 + Q_1}100.$$

6. Средна разлика - G.

Пресмята се като отношение на сумата от разликите на индивидуалните стойности на признака върху броя им. Броят на тези разлики винаги е равен на $(\frac{n(n-1)}{2})$. Ако например индивидуалните значения

на признака са три $x_1,\ x_2,\ x_3$, то имаме за сумата на разликите:

$$\sum d_i = (x_3 - x_2) + (x_3 - x_1) + (x_2 - x_1).$$

Средната разлика, тогава общо се определя от израза:

$$G = \frac{\sum_{i=1}^{n} d_i}{\frac{n(n-1)}{2}}$$

за негрупирани данни.

При групирани данни удобна за изчисляване на средната разлика е формулата:

$$G = \frac{\sum_{i=1}^{k} f_i(C_i^+ - C_i^-)}{\frac{\sum f_i(\sum f_i - 1)}{2}},$$

където C_i^+ и C_i^- са кумулативните честоти, съответно в нарастващ и намаляващ порядък, а k е броят на групите.

Коефициентът на вариация за G се изчислява по формулата:

$$V_G\% = \frac{G}{\frac{x_{max} + x_{min}}{n}}.$$

7. Съотношения между измерителите на разсейване.

Между измерителите на разсейване са в сила следните съотношения:

$$R \geq G \geq \sigma \geq \delta \geq Q,$$

$$V_R \geq V_G \geq V_\sigma \geq V_\delta \geq V_Q.$$

Вижда се, че размахът и средната разлика имат свойството да надценяват различията между отделните единици в съвкупността по изучавания признак, докато средното аритметичното и квартилно отклонение имат свойството да подценяват различията в сравнение със стандартното отклонение.

Стандартното отклонение играе ролята на най-точен и адекватен измерител на разсейването в съвкупността и за това е най-използван.