С Т А Т И С Т И К А

§20. Предмет и задачи на математическата статистиката. Генерална съвкупност и извадка.

Математическата статистика е приложен дял от математиката, основа на която е теорията на вероятностите. Тя е *наука за събиране, обработване и интерпретиране на числени данни*. Основната ѝ цел е с помощта на събраните данни да се даде отговор на реални въпроси като:

- има ли изменение в климата;
- какво е въздействието на новосъздадено лекарство;
- какви са тенденциите в икономиката на дадена страна;
- кой от два метода на обучение е по-ефективен,

т.е. въпроси, свързани с явления и обекти, които или не са изучени, или е невъзможно пълното им изследване. За да бъде намерен правилният им отговор са създадени

- методи за събиране на данни. Тук основен въпрос е с минимални разходи да се възможно най-достоверна съвкупност от данни;
- методи за обработка и интерпретация на данните. С помощта на тези методи се правят изводи за развитието и закономерностите на явлението.

Основен предмет на математическата статистика е изследването на съвкупности от обекти, които не можем да изучим напълно.

Някои основни понятия:

Генерална съвкупност (*популация)* - съвкупността от еднотипни обекти, която се изследва. Обикновено се изследва една или няколко характеристики (признаци) на генералната съвкупност, които получават различни стойности преминавайки от един обект към друг обект на съвкупността. Тези характеристики ще наричаме *променливи* и ще ги означаваме с главни букви X, Y, Z и т.н.

Променливите (признаците*)* могат да бъдат:

качествени, ако стойността им не може да се представи като число.
количествени, ако стойността им има смисъл на число (измерване или броене).

Например, изследват се жителите в даден град (генерална съвкупност) по ръст (количествен признак) и по цвят на очите (качествен признак).

Когато анализираме качествен признак (например изучаваме структурата на икономиката) обикновено се работи с брой или проценти – какъв процент от общата продукция се пада на селското стопанство, на промишлеността, на транспорта и т.н. В най-простия случай, качественият признак се свежда към количествен като се въведе величина с две възможни стойности $X=1$, ако даден обект от генералната съвкупност притежава качеството и $X=0$ в противен случай. В по-сложните случаи на определена стойност на качествената величина може да се присвои числена стойност, но действията между

такива числени стойности нямат никакъв реален смисъл. Например, в преброяването на населението са въведени следните категории: несемеен, семеен, разведен и вдовец, на които може да се припишат числени стойности, примерно 0, 1, 2 и 3 (тук сумата 2+1 не изразява нищо). Ще отбележим, че със същия успех може да се вземат и произволни други четири символа, например, ♣,♦, ♥, и ♠.

Количествените променливи (признаци) се делят на две групи:

Дискретни, ако множеството от всички възможни стойности е дискретно.

Непрекъснати, ако множеството от всички възможни стойности се състои от интервал.

Вниманието ни ще бъде насочено към количествените променливи като в зависимост от интересите ни, се разглеждат една, две или повече количествени характеристики на генералната съвкупност – *едномерен, двумерен и многомерен признак на генерална съвкупност.*

Изследването на генералната съвкупност може да протече по следните начини:

- Да се изследват всички обекти от генералната съвкупност. По този начин се описват съвсем точно характеристиките на генералната съвкупност. Този метод е много неефективен, а в някои случаи и невъзможен (например, ако изследването води до разрушаване на обекта).
- Да се изследва само част от генералната съвкупност (извадка). Предимство тук е спестяването на средства и време. Недостатък е, че не се получава абсолютно точно описание на генералната съвкупност. Затова е от голямо значение подборът на обектите за изследване.

Пример 20.1. Изследва се средният ръст на населението на дадена страна. Това може да стане, ако вместо цялото население се изберат отделни индивиди и да се изчисли средният ръст за избраната група. Очевидно, много важно е как ще се избират отделните индивиди - необходимо е групата да бъде достатъчно голяма, индивидите да са избрани от различни области и случайно (а не избирателно, например, които са по-високи).♦

Извадка наричаме частта от генералната съвкупност, която се изследва. Броят на елементите в извадката се нарича обем на извадката.

Репрезентативна (представителна) извадка е тази, която отразява достоверно генералната съвкупност. Ако изборът на елементите става по случаен начин, то извадката се нарича *случайна*.

От опит е установено, че ***репрезентативни са случайните извадки с голям обем.*** Затова ще разглеждаме само случайни извадки. Тъй като изборът на елементи от генералната съвкупност е случаен, то променливата която се изучава, ще разглеждаме като случайна величина X с неизвестен закон, а резултатите от всяко измерване (или броене) –

възможни стойности $x^{(1)}$, $x^{(2)}$,..., $x^{(n)}$ на случайни величини X_1, X_2,..., X_n със същия закон на разпределение. По този начин:

При изследване на количествената променлива X на генералната съвкупност:

- *величината X се счита за случайна величина с неизвестен закон;*
- *извадката се представя като многомерна случайна величина $(X_1,\ldots,X_n)$, чиито компоненти X_i ={резултат от i-тото наблюдение} имат разпределение, съвпадащо с разпределението на X;*
- *наблюдаваните стойности $x^{(1)}$, $x^{(2)}$,..., $x^{(n)}$ са реализация на величините X_1, X_2,..., X_n, т.е. наблюдавани са събитията $X_1=x^{(1)}$, $X_2=x^{(2)}$ и т.н.*

Следователно, теоретична основа на статистиката е теорията на вероятностите, с помощта на която се решават следните задачи:

Основни задачи на математическата статистика:

1. *Да се систематизират събраните данни **(статистическа обработка на данните)**.*
2. *Въз основа на данните да се оцени неизвестна количествена характеристика на съвкупността **(оценка на параметри)**.*
3. *Да се направят изводи, отнасящи се за цялата съвкупност и провери верността им (**проверка на хипотези**).*
4. *Ако се изследват зависимости между две или повече променливи, да се установи характера и степента им на зависимост една от друга (**корелационен или регресионен анализ**).*

СТАТИСТИЧЕСКА ОБРАБОТКА НА ДАННИ

§21. Статистическо разпределение, полигон и хистограма. Емпирична функция на разпределение.

Статистиката започва със систематизиране на събраните данни. Първата стъпка е данните да се представят в такава форма, че да може да се състави обща картина на изследванията. За да се интерпретира по-нататък тази съвкупност от данни, като се има предвид източника на данните и методите на събирането им, трябва да се отстранят тези от данните (ако има такива), които са попаднали случайно или по погрешка в изследваната съвкупност. До голяма степен това се прави по осмотрение на изследователя. Целта ни тук е да представим данните в такъв вид, че да може да се вземе решение по този въпрос и да се даде насока за по-нататъчните изследвания.

Нека изследваме количествения признак X на генералната съвкупност, за която разполагаме с резултати от n наблюдения, т.е. дадена ни е извадка $(X_1,X_2,\ldots,X_n)$ с обем n. Тъй като в резултат на тези

наблюдения получаваме в n стойности $(x^{(1)},x^{(2)},\ldots,x^{(n)})$ на променливата X, то тези стойности, се наричат ***варианти***. Подреждаме вариантите по големина (между тях може да има и повторения) . Така получаваме

Вариационен ред - наблюдаваните стойности, наредени по големина

$$\hat{x}_1\le\hat{x}_2\le\ldots\le\hat{x}_n. \qquad (21.1)$$

Първите оценки на съвкупността от данни са

Размах и среда на извадката: $x_{\max}-x_{\min}$ и $\frac{x_{\max}+x_{\min}}{2}$.

1. Статистическо разпределение. Ако нанесем в таблица само различните варианти и броят на повторенията им (честота), получаваме

Статистическо разпределение на честотите (честотно разпределение) - таблица от вида

$$\frac{X\mid x_1\ \ x_2\ \ \ldots\ \ x_k}{m_i\mid m_1\ \ m_2\ \ \ldots\ \ m_k},\qquad \sum_{i=1}^{k}m_i=n, \qquad (21.2)$$

където m_i е честота на вариантата x_i.

Отношението $\nu_i=\frac{m_i}{n}$ наричаме относителна честота на вариантата x_i. Таблицата

$$\frac{X\mid x_1\ \ x_2\ \ \ldots\ \ x_k}{\nu_i\mid \nu_1\ \ \nu_2\ \ \ldots\ \ \nu_k},\qquad \sum_{i=1}^{k}\nu_i=1 \qquad (21.3)$$

наричаме статистическо разпределение на относителните честоти.

Графически статистическите разпределения се представят с:

Полигон на честотите: начупена линия, съединяваща точките с координати (x_i,m_i) или полигон на относителните честоти: - начупена линия, съединяваща точките с координати (x_i,ν_i) (фиг. 21.1).

Очевидна е връзката между разпределението на една дискретна случайна величина и статистическото разпределение на относителните честоти. По аналогия се въвежда и функцията

Емпирична функция на разпределението:

$F^*(x)=\frac{n_x}{n}$, $x\in(-\infty,\infty)$, *където n_x - брой на вариантите, по-малки от x.*

Като се използуват:

натрупаните (кумулативни) честоти $c_i=\sum_{j=1}^{i}m_j$ *и*

относителните натрупани (кумулативни) честоти $\gamma_i=\frac{c_i}{n}=\sum_{j=1}^{i}\nu_j$

не е трудно да се види, че

$$F^*(x)=\begin{cases}0 & x\le x_1\\ \gamma_i & x_i<x\le x_{i+1},\ i=1,\ldots,k-1\\ 1 & x>x_k\end{cases}, \qquad (21.4)$$

Забележка 21.1. Както знаем, с помощта на функцията на разпределение на една случайна величина изчисляваме на квантилите на случайната величина. Приложението на емпиричната функция на разпределение е аналогично, но за да може да се използва с успех във всички случаи (виж **§22**) е по-удобно въвеждането на функция, която да бъде непрекъсната за всяко x. Тази функция ще наричаме

кумулативна функция

$$F(x)=\begin{cases}0 & x\le x_1\\ \gamma_2+\frac{\gamma_2}{x_2-x_1}(x-x_2) & x_1<x\le x_2\\ \gamma_i+\frac{\nu_i}{x_i-x_{i-1}}(x-x_i) & x_{i-1}<x\le x_i,\ i=3,\ldots,k\\ 1 & x>x_k\end{cases} \qquad (21.5)$$

Графиката на емпиричната функция на разпределението $F^*(x)$ е стъпаловидна, а графиката на кумулативната функция $F(x)$, която се нарича също *полигон на относителните кумулативни честоти*, се състои от отсечки, съединяващи точките с координати $(x_1,0)$, (x_i,γ_i) $(i=2,\ldots,k)$.

Забележка 21.2. Единствената разлика между вариационния ред (21.1) и статистическото разпределение (21.2) е в това, че данните в (21.2) са групирани по стойност..

Пример 21.1. Броят на авариите в едно предприятие през последните десет години е 5,2,3,4,5,4,0,3,4,3. Да се намери размахът и средата на извадката, статистическото разпределение на честотите и натрупаните честоти, емпиричната и кумулативната функция. Да се начертаят полигонът на относителните честоти и графиката на емпиричната функция на разпределението.

Решение. Очевидно, размахът на извадката е $5-0=5$,

средата е $\frac{5+0}{2}=2{,}5$.

Всички необходими изчисления нанасяме в таблица:

i	*варианти* x_i	*честоти* m_i	*относит. честоти* $\nu_i=\frac{m_i}{n}$	*относит. кумулат честоти.* $\gamma_i=\frac{c_i}{n}$	$\frac{\nu_i}{x_i-x_{i-1}}$
1	0	1	0,1	0,1	–
2	2	1	0,1	0,2	0,05
3	3	3	0,3	0,5	0,3
4	4	3	0,3	0,8	0,3
5	5	2	0,2	1,0	0,2

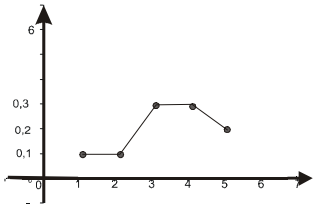
Статистическото разпределение на относителните честоти е

$$\begin{array}{c|ccccc} x_i & 0 & 2 & 3 & 4 & 5\\ \hline \nu_i & 0{,}1 & 0{,}1 & 0{,}3 & 0{,}3 & 0{,}2\end{array}, \qquad (\nu_i=\frac{m_i}{n}=\frac{m_i}{10}),$$

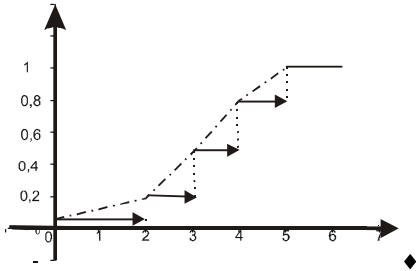
а полигонът на относителните честоти е начертан на фиг.21.1. По формули (21.4) и (21.5) получаваме емпиричната функция на разпределението $F^*(x)$ и кумулативната функция $F(x)$ (използваме четвъртата и шестата колона в таблицата)

$$F^*(x)=\begin{cases}0, & x<0\\ 0{,}1, & 0\le x<2\\ 0{,}2, & 2\le x<3\\ 0{,}5, & 3\le x<4\\ 0{,}8, & 4\le x<5\\ 1, & x\ge 5\end{cases}, \text{ и } F(x)=\begin{cases}0, & x<0\\ 0{,}2+0{,}05(x-2), & 0\le x<2\\ 0{,}5+0{,}3(x-3), & 2\le x<3\\ 0{,}8+0{,}3(x-4), & 3\le x<4\\ 1+0{,}2(x-5), & 4\le x<5\\ 1, & x\ge 5\end{cases}$$

с графики, представени на фиг. 21.2.



Фиг. 21.1



Фиг.21.2 ♦

Забележка 21.3. В интервала (0,2] функцията $F(x)$ е построена по правилото за останалите интервали, затова има прекъсване за $x=0$. За да се получи функцията (21.5), числото във втория ред на последната колона в таблицата трябва да бъде $\gamma_2/(x_2-x_1)=0{,}2:2=0{,}1$. ♦

2. Интервално статистическо разпределение.

Ако X е непрекъсната случайна величина или обемът на извадката е много голям, е целесъобразно да се постъпи по следния начин. Разделяме интервала $[x_{\min},x_{\max}]$ на k подинтервали $[x_0,x_1),\ldots,[x_{k-1},x_k)$

(препоръчително е броят на подинтервалите да е между 5 и 15). Означаваме с m_i броя на вариантите, попадащи $[x_{i-1},x_i)$. Така получаваме

интервалното статистичско разпределени на честотите (интервално честотно разпределение):

$$\begin{array}{c|cccc} [x_{i-1},x_i) & [x_0,x_1) & [x_1,x_2) & \dots & [x_{k-1},x_k) \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array}, \quad \sum_{i=1}^{k} m_i = n, \qquad (21.6)$$

интервалното статистическо разпределение на относителните честоти:

$$\begin{array}{c|cccc} [x_{i-1},x_i) & [x_0,x_1) & [x_1,x_2) & \dots & [x_{k-1},x_k) \\ \hline \nu_i & \nu_1 & \nu_2 & \dots & \nu_k \end{array}, \quad \nu_i = \frac{m_i}{n}, \quad \sum_{i=1}^{k} \nu_i = 1 \qquad (21.7)$$

Използуват се също и

натрупана (кумулативна) честота $c_i = \sum_{j=1}^{i} m_j$;

относителна натрупана (кумулативна) честота $\gamma_i = \frac{c_i}{n} = \sum_{j=1}^{i} \nu_j$.

Разпределенията (21.6) и (21.7) се представят чрез графики които се наричат хистограми (фиг.21.3).

Хистограма на честотите: фигура, състояща се от правоъгълници, с основа $[x_{i-1},x_i)$ *и лице, равно на честотата* m_i, *т.е. с височина*

$$h_i = \frac{m_i}{x_i - x_{i-1}}.$$

Хистограма на относителните честоти: фигура, състояща се от правоъгълници, с основа $[x_{i-1},x_i)$ *и лице, равно на* ν_i.

За по-нататъчната обработка на интервалното разпределение, с цел да се получи разпределение от вида (21.2), интервалът $[x_{i-1},x_i)$ се заменя със средата му $x_i^* = \frac{x_{i-1}+x_i}{2}$ ($i = 1,\dots,k$) и така се получава разпределението

$$\begin{array}{c|cccc} x_1^* & x_2^* & x_2^* & \dots & x_k^* \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array}. \qquad (21.8)$$

Сега можем да получим и други характеристики на извадката, които ще бъдат разгледани по-нататък. Например, по формула (21.4) получаваме емпиричната функция $F^*(x)$ на разпределението, но по-удобна за приложение е следната непрекъсната функция, наречена

Кумулативна функция:

$$F(x) = \begin{cases} 0 & x \le x_0 \\ \gamma_i + \frac{\nu_i}{d_i}(x - x_i) & x_{i-1} < x \le x_i,\ i = 1,\dots,k \\ 1 & x > x_k \end{cases}. \qquad (21.9)$$

където $d_i = x_i - x_{i-1}$ *е дължината на интервала* $[x_{i-1},x_i)$

Графиката на функцията $F(x)$ в интервала $(x_0\, x_k]$ се състои от отсечки, съединяващи точките $(x_0,0), (x_1,\gamma_1), \dots, (x_k,\gamma_k)$.

Забележка 21.4. Кумулативната функция както и графиката ѝ са много удобни при определяне на разгледаните в следващия параграф *квантили.*

Пример. 21.2. Дадена е извадка с обем $n = 20$:4,5; 1,2; 3,0; 4,0; 1,1; 4,2; 2,1; 3,4; 2,4; 3,9; 3,7; 4,3; 4,9; 2,7; 2,7; 3,6; 3,3; 4,2; 3,9; 2,4. Да се състави интервален статистически ред и построи хистограмата на честотите. Да се намерят емпиричната и кумулативната функции на разпределение и построят графиките им.

Решение: Подреждаме числата в групи по цялата им част:

1,2; 1,1;
2,1; 2,4; 2,7; 2,7; 2,4;
3,0; 3,4; 3,9; 3,7; 3,6; 3,3; 3,9.
4,5; 4,0; 4,2; 4,3;.4,9; 4,2,

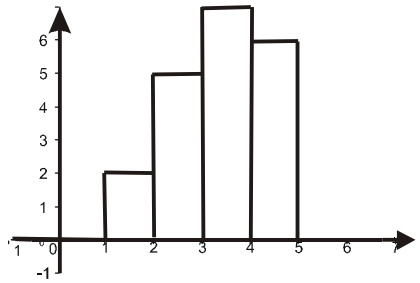
откъдето получаваме интервалното статистическо разпределение

$$\begin{array}{c|cccc} [x_i,x_{i+1}) & [1,2) & [2,3) & [3,4) & [4,5) \\ \hline m_i & 2 & 5 & 7 & 6 \end{array}.$$

По-нататък ще използуваме следната таблица, в която са отразени всички необходими пресмятания:

i	*интервал* $[x_i,x_{i+1})$	*честота* m_i	*кумулат. честота* $c_i = \sum_{j=1}^{i} m_j$	*относит. честота* $\nu_i = \frac{m_i}{n}$	*относит. кумул.честота* $\gamma_i = \frac{c_i}{n}$	*среда на интервала* $x_i^* = \frac{x_i + x_{i-1}}{2}$
1	[1,2)	2	2	0,10	0,10	1,5
2	[2,3)	5	7	0,25	0,35	2,5
3	[3,4)	7	14	0,35	0,70	3,5
4	[4,5)	6	20	0,30	1,00	4,5

Хистограмата на честотите начертаваме като използваме втората колона (понеже дължината на интервалите $[x_{i-1},x_i)$ е единица, височината на правоъгълниците съвпада с честотата)

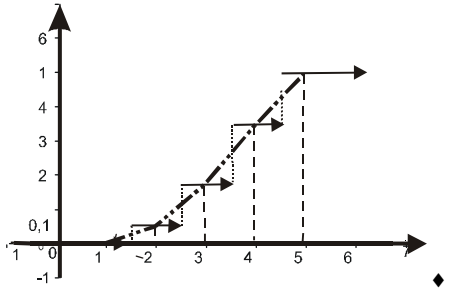


Фиг.21.3.

От тук, съгласно формули (21.4) и (21.8) получаваме емпиричната и кумулативаната функции:

$$F^*(x)=\begin{cases}0 & x<0{,}5\\ 0{,}1 & 1{,}5\le x<2{,}5\\ 0{,}35 & 2{,}5\le x<3{,}5\\ 0{,}7 & 3{,}5\le x<4{,}5\\ 1 & x>4{,}5\end{cases},\qquad F(x)=\begin{cases}0 & x<1\\ 0{,}1+0{,}1(x-2) & 1\le x<2\\ 0{,}35+0{,}25(x-3) & 2\le x<3\\ 0{,}7+0{,}35(x-4) & 3\le x<4\\ 1+0{,}3(x-5) & 4\le x<5\\ 1 & x>5\end{cases}$$

а техните графики са начертани на фиг. 21.4.



Фиг. 21.4. ♦

Упражнения

1. Извадката 5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4 представете като:
а) вариационен ред; б) неинтервално статистическо разпределение. Да се намери емпиричната функция на разпределение; в) интервално статистическо разпределение с интервали (0,4], (4,8], (8,12]. Да се намери кумулативната функция на разпределение.
2. Да се построи хистограмата на честотите на извадката

(x_{j-1},x_i)	(10,12)	(12,14)	(14,16)	(16,18)	(18,20)	(20,22)	(22,24)
m_i	2	4	8	12	16	10	3

Да се намерят кумулативната и емпиричната функция и да се начертаят графиките им.

3. Да се намери кумулативната функция на разпределение на извадката, дадена със статистическия ред

x_i	2	5	7	8
m_i	1	3	2	4

4. Записан е броят на грешките върху една страница от текст:
2, 3, 6, 1, 4, 1, 2, 0, 1, 5
1, 4, 4, 2, 3, 1, 2, 2, 1, 0.
Да се начертае полигонът на относителните честоти и намери емпиричната функция.
5. Да се намери интервално статистическо разпределение на извадката (брой на пътниците в един автобус по дадено направление):
16 35 29 30 37 20 34 40 46 10 23 48 33 45 34
50 32 12 15 13 25 25 34 10 14 39 20 25 31 54.
Да се построи хистограма на относителните честоти.

§22. Числени характеристики на статистическото разпределение. Характеристики на местоположението - мода, медиана, квантили и квартили.

Нека се изследва количествената променлива X на генералната съвкупност и нека е дадена извадка $(X_1,X_2,\ldots,X_n)$ с обем n. Освен графически, изследването на извадката се извършва с помощта на *<u>числени характеристики</u> – мерки (числа), които дават допълнителна информация за извадката.*

Числените характеристики на извадката се разделят на:

- *характеристики на местоположението;*
- *характеристики на разсейването*.

Например, размахът и средата са характеристики съответно на разсейването и на местоположението. Освен тези първи и недостатъчно прецизни мерки, от гледна точка на *местоположението на вариантите* й са въведени следните характеристики на извадката:

<u>Мода</u> - наблюдаваната стойност M_o с най-голяма честота.

Една извадка може да има повече от една мода и се нарича бимодална, ако има две моди и полимодална, ако модите са повече от две. Ако извадката е зададена с интервално разпределение, то интервалът с най-голяма честота се нарича *<u>модален интервал</u>*.

<u>Медиана</u> – такова число Q_2, за което половината от вариантите са по-малки, а другата половина – по големи от него. (друго означение M_e).

<u>Квантил</u> <u>от ред</u> p $(0<p<1)$ – число x_p, за което $100p\%$ от вариантите са по-малки от него.

Очевидно, $Q_2=x_{0,5}$.

Заедно с медианата ($p=0{,}5$), особено често се използуват квантилите от ред $p=0{,}25$ и $p=0{,}75$, които носят имената:

Долен квартил - $Q_1=x_{0.25}$. *горен квартил* - $Q_3=x_{0{,}75}$.

Друга характериситка, които се изразяват чрез квантилите са:

процентил $x_{P\%}$ *от ред* $P\%$ ($0<P<100$) *– число, съвпадащо с квантил от ред* $p=P/100$.

Всяка числена характеристика се пресмята на базата на наблюдаваните данни $(x^{(1)},x^{(2)},\ldots,x^{(n)})$, затова се явява тяхна функция.

Формулите за пресмятане на тези величини зависят от това дали е дадено статистическо или интервално статистическо разпределение на извадката.

1. Статистическо разпределение (неинтервално). За определяне на квантилите е по-удобно е да се използува вариационният ред $\hat{x}_1\le\hat{x}_2\le\ldots\le\hat{x}_n$ (който винаги може да получим от разпределението)

Квантилът x_p *(от ред* p*) се изчислява по следния начин:*

1) *Означаваме с* r *цялата част на числото* np *. Тогава* $\hat{x}_r$ е *тази варианта, за която* np *наблюдения са по-малки или равни на нея).*
2) *За квантил от ред* p *приемаме числото*

$$x_p=\begin{cases}\dfrac{\hat{x}_r+\hat{x}_{r+1}}{2}, & \text{ако } np \text{ е цяло}\\ \hat{x}_{r+1}, & \text{ако } np \text{ не е цяло}\end{cases} \qquad (22.1)$$

За $p=\dfrac{1}{2}$ се получава формулата за медианата:

$$Q_2=\begin{cases}\dfrac{\hat{x}_r+\hat{x}_{r+1}}{2}, \text{ако } n \text{ четно}\\ \hat{x}_{r+1}, \text{ако } n \text{ е нечетно}\end{cases}$$, където r цялата част на числото $\dfrac{n}{2}$

(в ляво от $\hat{x}_r$ имаме не повече от половината наблюдения).

Пример 22.1. Да се определи на медианата:

а) За вариационния ред 3, 5, 6, 7, 7. Имаме: $n=5$ - нечетно, цялата част на $\dfrac{n}{2}=2{,}5$ е 2. Следователно $r=2$ и $Q_2=\hat{x}_{2+1}=\hat{x}_3=6$.

б) За вариационния ред 3, 5, 6, 7, 7, 9. Тук $Q_2=\dfrac{\hat{x}_3+\hat{x}_4}{2}=\dfrac{6+7}{2}$ ($r=\dfrac{n}{2}=\dfrac{6}{2}=3$).

в) За вариационния ред 3, 5, 5, 6, 6, 7, 7 $Q_2=6$. ♦

Пример 22.2: За вариационния ред (с обем на извадката $n=20$)

0, 0, 0, 1, 2, 2, 3, 3, 6, 6, 6, 7, 7, 8, 8,12,12, 20, 20, 20

да се намерят:

$x_{0,2}=\dfrac{\hat{x}_4+\hat{x}_5}{2}=\dfrac{1+2}{2}=2{,}5$, тъй като $np=20.0{,}2=4$ е цяло число, а

$x_{\frac{2}{3}}=\hat{x}_{13+1}=8$, защото $np=20.\dfrac{2}{3}=13{,}3333$ не е цяло и $r=13$.

$x_{40\%}=x_{0,4}=\dfrac{\hat{x}_8+\hat{x}_9}{2}=\dfrac{3+6}{2}=4{,}5$ ($np=20.0{,}4=8$ - цяло).

$Q_1=\dfrac{\hat{x}_5+\hat{x}_6}{2}=\dfrac{2+2}{2}=2$, $Q_2=\dfrac{\hat{x}_{10}+\hat{x}_{11}}{2}=\dfrac{6+6}{2}=6$, $Q_3=\dfrac{\hat{x}_{15}+\hat{x}_{16}}{2}=\dfrac{8+12}{2}=10$. ♦

Забележка 22.1. Горните формули не са много точни, особено, когато има и други варианти, равни на x_r (какъвто е случаят за Q_1 и Q_2 в пример **22.2**) или обемът n на извадката е твърде малък. По-добре е да се използува кумулативната функция (21.5) като не е необходимо пълното й определяне, а само в интервала, в който се намира квантилът (примери **22.5, 22.6)**.

Квартилите Q_1 , Q_2 , Q_3 , най-малката $x_{\min}=\hat{x}_1$ и най-голямата $x_{\max}=\hat{x}_n$ наблюдавана стойност се използват за да се даде едно просто и нагледно представяне на данните – така наречената

диаграма от тип "кутия" (box-plot) - диаграма от вида

Тази диаграма представя най-важните елементи на извадката: половината от данните се намират в интервала (Q_1,Q_3) , т.е. те са "вътре в кутията", а медианата Q_2 като число, за което половината от данните са по-големи, а другата половина – по-малки, дава информация за симетричността на данните.

Понякога в извадката попадат данни, които не са типични за генералната съвкупност – варианти, които в сравнение с останалите са или много малки, или много големи. Едно правило за определяне на нетипичните данни е следното:

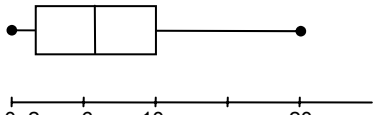
Като нетипични се приемат вариантите, които са:

по-малки от числото $l_1=Q_1-1{,}5(Q_3-Q_1)$

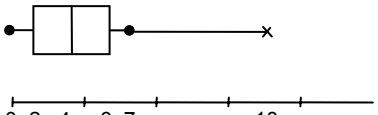
по-големи от числото $l_2=Q_3+1{,}5(Q_3-Q_1)$.

Пример 22.3. Да се начертае диаграмата от тип кутия на извадката от пример 22.2.

Решение. Имаме: $x_{\min}=0$, $x_{\max}=20$, $Q_1=2$, $Q_2=6$, $Q_3=10$. Диаграмата е начертана на фиг. 22.1.



Фиг.22.1.



Фиг.22.2

Пример 22.4. Получена е следната извадка (брой на закъснелите за работа): 0, 1, 2, 2, 3, 4, 5, 5, 6, 7, 18. Да се определи кои от данните са нетипични и начертае диаграмата от тип кутия.

Решение: Пресмятаме $Q_1=\hat{x}_3=2$, $Q_2=\hat{x}_6=4$. $Q_3=\hat{x}_9=6$,

$$l_1=2-1{,}5(6-2)=-4,\ l_2=6+1{,}5(6-2)=12.$$

Следователно, нетипична е вариантата $\hat{x}_{11}=18$, защото е по-голяма от l_2. На диаграмата тази стойност е отбелязана с кръстче (фиг.22.2). ♦

2. **Интервално статистическо разпределение**. Нека е дадена таблицата

$[x_{i-1},x_i)$	$[x_0,x_1)$	$[x_1,x_2)$	…	$[x_{k-1},x_k)$
m_i	m_1	m_2	…	m_k

.

Квантилът x_p е решение на уравнението $F(x)=x_p$, където $F(x)$ е кумулативната функция (21.9). За намирането му е добре да са изчислени относителните кумулативни честоти γ_i. Тогава интервалът, съдържащ квантила x_p, е първият интервал $[x_{r-1},x_r)$, за който $\gamma_r\ge p$, и се определя по следния начин:

Намиране на квантил x_p от ред p:

1) *Намираме най-малката относителна кумулативна честота γ_r, за която $p\le\gamma_r$. Интервалът, в който се намира квантилът, е $[x_{r-1},x_r)$.*

2) *Намираме квантила:* $x_p=x_r+\dfrac{x_r-x_{r-1}}{\nu_r}(p-\gamma_r)$ (22.2).

За частния случай на *медиана*:

Определяме медианния интервал $[x_{r-1},x_r)$, където γ_r е най-голямата честота, за която $\gamma_r\ge 0{,}5$, откъдето

$$Q_2=x_r+\frac{d_r}{\nu_r}(0{,}5-\gamma_r).,\ d_r=x_r-x_{r-1}.$$

Забележка 22.2. Освен за интервално, формули (22.2) могат да се използват и за неинтервално статистическо разпределение (в този случай, ако $r=2$, то $\nu_r=\nu_2$ трябва да се замени с γ_2).

Пример 22.5. Продължителността на разговорите по телефона в едно учреждение е представена чрез извадка с обем 100.

$(x_{i-1},x_i]$	(0,2]	(2,4]	(4,6]	(6,8]	(8,10]
m_i	10	25	20	40	5

.

Да се намерят: а) квантилът $x_{0,2}$, б) долният и горният квартил Q_1, Q_3, в) медианата ; г) 3%-персентил $x_{3\%}$.

Решение. Използваме таблицата

i	$(x_{i-1},x_i]$	m_i	$\nu_i=\dfrac{m_i}{n}$	$\gamma_i=\dfrac{c_i}{n}$
1	(0,2]	10	0,1	0,1
2	(2,4]	25	0.25	0,35
3	(4,6]	20	0,20	0,55
4	(6,8]	40	0,40	0,95
5	(8,10]	5	0,05	1

.

а) Определяме интервала на квантила: $\gamma_2=0{,}35$ е най-малката относителна кумулативна честота, за която 0,2<0,35. Следователно, $x_{0,2}\in(2,4]$ и се определя от израза (използваме втория ред от таблицата):

$$x_{0,2}=x_2+\frac{d_2}{\nu_2}(0{,}2-\gamma_2)=4+\frac{2}{0{,}25}(0{,}2-0{,}35)=4-\frac{200}{25}.0{,}15=4-1{,}2=2{,}8.$$

б) 0,25<0,35. Следователно, $Q_1=x_{0,25}\in(2,4]$ и

$$x_{0,25}=x_2+\frac{d_2}{\nu_2}(0{,}25-\gamma_2)=4+\frac{2}{0{,}25}(0{,}25-0{,}35)=4-\frac{200}{25}0{,}1=4-0{,}8=3{,}2.$$

0,75<0,95. Следователно, $Q_3=x_{0,75}\in(6,8]$ - (четвърти интервал) и

$$Q_3=x_{0,75}=x_4+\frac{d_4}{\nu_4}(0{,}75-\gamma_4)=8+\frac{2}{0{,}40}(0{,}75-0{,}95)=8-\frac{20}{4}.0{,}2=8-1=7.$$

в) Q_2 е квантилът от ред $p=0{,}5$, откъдето 0,5<0,55 $\Rightarrow Q_2\in(4,6]$,

$$Q_2=x_3+\frac{d_3}{\nu_3}(0{,}5-\gamma_3)=6+\frac{2}{0{,}2}(0{,}5-0{,}55)=6-\frac{20}{2}0{,}05=5{,}5.$$

г) $x_{3\%}=x_{0,03}$. Тогава от 0,03<0,1 следва, че $x_{0,03}\in(0,2]$ и

$$x_{0,03}=2+\frac{2}{0,1}(0,03-0,1)=2-20.0,07=2-1,4=0,6\ .\blacklozenge$$

Забележка. Всички квантили могат да се намерят графично от графиката на кумулативната функция $F^*(x)=\begin{cases}0 & x\in(-\infty,0]\\ 0,1+0,05(x-2) & x\in(0,2]\\ 0,35+0,125(x-4) & x\in(2,4]\\ 0,55+0,1(x-6) & x\in(4,6]\\ 0,95+0,2(x-8) & x\in(6,8]\\ 1+0,025(x-10) & x\in(8,10]\\ 1 & x\in(10,\infty)\end{cases}$

Пример 22.6. Ще намерим квантила $x_{0,4}$ на статистическото разпределение от пример **21.1.**

Решение. Ако приложим формула (22.2), намираме първо, че $\gamma_3=0,5>0,4$, следователно (използваме третия ред на таблицата),

$$x_{0,4}=x_3+\frac{x_3-x_2}{\nu_3}(0,4-\gamma_3)=3+\frac{3-2}{0,3}(0,4-0,5)=3-\frac{10}{3}.0,1=2\frac{2}{3}\ .$$

Ако приложим формула (22.1), то $np=10.0,4=4$, следователно, квантилът съвпада с петата варианта от вариационния ред, т.е. $x_{0,4}=3\ .\blacklozenge$

Упражнения.

1. Времето, необходимо за почистване на дома (в минути) на 100 произволно избрани домакинства е следното:

(x_{i-1},x_i)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)	(70,80)
m_i	10	12	25	42	7	4

а) Да се намерят трите квартили на извадката.
б) Да намери до колко време отделят за почистване на дома 30% от домакинствата.
в) Повече от колко минути са необходими за почистване за 45% от домакинствата.

2. Секретарка е записала броя на набраните страници в работен ден в продължение на един месец (20 работни дни):

2, 5, 0, 3, 2, 6, 6, 10, 12, 2, 5, 13, 14, 0, 1, 2, 6, 5, 5, 6.

Да се намери модата, горният квартил, медианата и квантилът от ред 0,4.

3. В продължение на 100 дни контрольор по качеството е записвал броя на дефектните изделия, които е проверил:

x_i	2	3	4	5	6	7	8	9	10
m_i	5	12	15	24	17	13	10	3	1

а) Да се начертае диаграмата от тип кутия, б) да се намерят процентилът от ред 64, и критичната точка от ред 0,2. Да се начертае диаграмата от тип кутия.

§23. Средна, дисперсия и средно квадратично отклонение, начални и централни моменти на извадката.

Друг вид числени характеристики на извадката са тези, с които се определят различни средни стойности на извадката.

Нека е дадена извадка $\begin{array}{c|cccc}X & x_1 & x_2 & \dots & x_k\\ \hline m_i & m_1 & m_2 & \dots & m_k\end{array}$ с обем n.

Средно аритметично на извадката наричаме средно аритметичното на всички получени наблюдения, т.е.

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{k}x_i m_i\ . \qquad (23.1)$$

Това е най-често използваната характеристика на извадката, затова накратко я наричаме *средна на извадката* (*извадкова средна)*. Заедно с модата M_o и медианата Q_2, средната на извадката е една от трите *характеристики на "центъра" на извадката.*

Други важни характеристики са:

Дисперсия на извадката (извадъчна дисперсия)

$$s_x^2=\frac{\sum(x_i-\overline{x})^2 m_i}{n}=\overline{x^2}-(\overline{x})^2\ , \qquad (23.2)$$

Извадъчно средно квадратично $s_x=\sqrt{s_x^2}$, (23.3)

които отразяват "разсейването" на наблюденията около средната на извадката.

Ако използуваме относителните честоти ν_i, то

$$\overline{x}=\sum x_i\nu_i\ ,\ s_x^2=\sum(x_i-\overline{x})^2\nu_i$$

Вижда се приликата с формули (8.1)-(8.3) за математическо очакване и дисперсия на дискретна случайна величина. Затова свойствата на тези характеристики са подобни на свойствата на математическото очакване и дисперсията (виж **§8**), а именно, ако $c=const$ и са дадени две независими извадки на променливите X и Y, то

$\overline{c}=c$	$s_c^2=0$.
$\overline{(cx)}=c\overline{x}$,	$s_{(cx)}^2=c^2 s_x^2$.
$\overline{(x\pm y)}=\overline{x}\pm\overline{y}$	$s_{x\pm y}^2=s_x^2+s_y^2$.

Забележка 23.1. Ако данните не са групирани в статистическо разпределение, то $\overline{x}$ е просто сумата от всички варианти, разделена на броят им:

$$\overline{x}=\frac{1}{n}(x^{(1)}+\ldots+x^{(n)})\ , \qquad (23.4)$$

където $x^{(i)}$ е наблюдаваната стойност в i-тото измерване. Във формула (23.1) повтарящите се събираеми от (23.4) са обединени в произведението $m_i x_i$ (m_i е честотата на вариантата x_i).

Аналогично, за извадъчната дисперсия на негрупирани данни имаме

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \overline{x})^2 . \qquad (23.5)$$

Разглеждат се също начални и централни моменти:

Начален момент от ред q: $\overline{x^q} = \sum_{i=1}^{k} x_i^q m_i$, $q = 1,2,...$

Централен момент от ред q: $s_x^{(q)} = \sum_{i=1}^{k}(x_i - \overline{x})^q m_i$, $q = 1,2,...$.

Между началните и централните моменти съществуват зависимостите:

$$\begin{aligned} s_x^{(2)} &= s_x^2 = \overline{x^2} - (\overline{x})^2 \\ s_x^{(3)} &= \overline{x^3} - 3\overline{x^2}\,\overline{x} + 2(\overline{x})^3 \\ s_x^{(4)} &= \overline{x^4} - 4\overline{x^3}\,\overline{x} + 6\overline{x^2}\,(\overline{x})^2 - 3(\overline{x})^4 \end{aligned} \qquad (23.6)$$

Ако величината X се мери в някакви мерни единици, то началните и централните моменти от ред q се измерват с q-тите им степени. Например, ако X се мери в метри, то средната и средно кватратичното също се мерят в метри, а дисперсията – в квадратни метри.

Въвеждат се също безразмерните характеристики

$a_x = \frac{s_x^{(3)}}{(s_x)^3}$ - *асиметрия и* $e_x = \frac{s_x^{(4)}}{(s_x)^4} - 3$ - *ексцес*.

Други средни характеристики са:

. *Средно геометрично* $G = \sqrt[n]{x_1^{m_1} x_2^{m_2} ... x_k^{m_k}}$.

Средно хармонично $\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{k}\frac{m_i}{x_i}$.

За негрупирани данни

$$G = \sqrt[n]{x^{(1)}x^{(2)}...x^{(n)}} . \qquad \frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x^{(i)}} .$$

Забележка 23.1. Ако извадката е зададена с интервален статистически ред, то всеки интервал $(x_{i-1}, x_i]$ се заменя със средата си

x_i^* и се прилагат същите формули.

За улесняване пресмятането на моментите в много случаи е удобно полагане от вида (линейна трансформация)

$u_i = ax_i - b$,

откъдето, като имаме предвид, че $x_i = \frac{1}{a}(u_i + b)$ *, получаваме*

$$\overline{x} = \frac{1}{a}[\overline{x} + b], \quad s_x^{(q)} = \frac{1}{a^q}s_u^q . \qquad (23.7)$$

Пример 23.1. За извадката

(x_{i-1}, x_i)	(1,3)	(3,5)	(5,7)	(7,9)	(9,11)
m_i	2	4	10	6	3

да се намерят началните и централните моменти до четвърти ред, асиметрията a_s и ексцесът e_x .

Решение.

1) Всеки интервал заменяме със средата му $x_i^* = \frac{x_{i+1} + x_i}{2}$:

x_i^*	2	4	6	8	10
m_i	2	4	10	6	3

.

2) Вариантата с най-голяма честота е $x_3^* = 6$, затова пресмятанията ще се опростят, ако положим $u_i = \frac{1}{2}(x_i^* - 6)$. Така получаваме таблицата

u_i	m_i	$u_i m_i$	$u_i^2 m_i$	$u_i^3 m_i$	$u_i^4 m_i$
−2	2	−4	8	−16	32
−1	4	−4	4	−4	4
0	10	0	0	0	0
1	6	6	6	6	6
2	3	6	12	24	48
$\sum =$	25	4	30	10	90

,

в последния ред на която са резултатите от сумирането по съответните стълбове.

3) Пресмятаме началните моменти:

$\overline{u} = \frac{4}{25} = 0{,}16$, $\overline{u^2} = \frac{30}{25} = 1{,}2$, $\overline{u^3} = \frac{10}{25} = 0{,}4$, $\overline{u^4} = \frac{90}{25} = 3{,}6$.

4) По формули (23.7) пресмятаме централните моменти:

$$\begin{aligned} s_u^{(2)} &= \overline{u^2} - (\overline{u})^2 = 1{,}2 - 0{,}16^2 = 1{,}1744 \\ s_u^{(3)} &= \overline{u^3} - 3\overline{u^2}.\overline{u} + 2(\overline{u})^3 = 0{,}4 - 3.0{,}16.1{,}2 + 2.0{,}16^3 = -0{,}1678 \\ s_u^{(4)} &= \overline{u^4} - 4\overline{u^3}..\overline{u} + 6(\overline{u})^2\overline{u^2} - 3(\overline{u})^4 = 3{,}5263. \end{aligned}$$

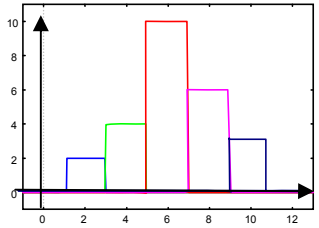
5) Пресмятаме всички моменти на изходната извадка като използваме формули (23.7). От равенството $u_i=\frac{1}{2}(x_i^*-6)$ определяме, че $x_i^*=2u_i+6$, следователно, $\bar{x}=2\bar{u}+6=2.0{,}16+6=6{,}32$.

За централните моменти имаме $s_x^{(q)}=2^q s_u^{(q)}$, откъдето

$$s_x^{(2)}=s_x^2=2^2 s_u^{(2)}=4.1{,}1744=4{,}6976 \Rightarrow s_x=\sqrt{4{,}6976}=2{,}1674,$$

$$s_x^{(3)}=2^3 s_u^{(3)}=8.(-0{,}1678)=-1{,}3424,\quad s_x^{(4)}=2^4 s_u^{(4)}=16.3{,}5263=56{,}4217.$$

3) За асиметрията и ексцеса получаваме



Фиг.23.1

$$a_x=\frac{-1{,}3424}{(s_x)^3}=-0{,}132,$$

$$e_x=\frac{s_x^{(4)}}{(s_x)^4}-3=-0{,}4432 \blacklozenge$$

Знакът на асиметрията е отрицателен, което означава, че данните са изместени на дясно (фиг. 23.1).

ОБРАБОТКА НА ДАННИ – ОСНОВНИ РЕЗУЛТАТИ

Означения: m_i - *честота*,, $\nu_i=\frac{m_i}{n}$ - *относителна честота*, $\gamma_i=\sum_{j=1}^{i}\nu_j$ - *относителна кумулативна честота* на варианатата x_i (на интервала (x_{i-1},x_i) с номер i).

***1) Неинтервално статистическо разпределение*:**

$$\begin{array}{c|cccc} X & x_1 & x_2 & \dots & x_k \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array},\quad n=\sum_{i=1}^{k}m_i$$ - *обем на извадката*.

Графично представяне – полигон.

емпирична функция: $F^*(x)=\begin{cases}0 & x\le x_1\\ \gamma_i & x_i<x\le x_{i+1},\ i=1,\dots,k-1\\ 1 & x>x_k\end{cases}$.

2) Интервално статистическо разпределение

$$\begin{array}{c|cccc} [x_{i-1},x_i) & [x_0,x_1) & [x_1,x_2) & \dots & [x_{k-1},x_k) \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array},\quad n=\sum_{i=1}^{k}m_i$$ - *обем.*

Графично представяне – хистограма.

кумулативна функция: $F(x)=\begin{cases}0 & x\le x_0\\ \gamma_i+\frac{\nu_i}{x_i-x_{i-1}}(x-x_i) & x_{i-1}<x\le x_i,\ i=1,\dots,k\\ 1 & x>x_k\end{cases}$.

***3) Основни числени характеристики на извадка*:**

Квантил x_p от ред p: $x_p=x_r+\frac{x_r-x_{r-1}}{\nu_r}(p-\gamma_r)$, *където е* γ_r - *най-малката кумулативна честота, за която* $p\le\gamma_r$.

Средно аритметично $\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i m_i$.

извадъчна дисперсия $s_x^2=\frac{\sum(x_i-\bar{x})^2 m_i}{n}=\overline{x^2}-(\bar{x})^2$, *където* $\overline{x^2}=\frac{1}{n}\sum_{i=1}^{k}x_i^2 m_i$.

Общи задачи.

1. Продължителността на 15 сърдечни трансплантации е както следва:
7,0 6,5 3,5 3,8 3,1 2,8 2,5 2,6 2,4 2,1 1,8 2,3 3,1 3,0 2,5.
а) Да се намерят медианата, долният и горният квартил и начертае бокс-плотът на извадката. Има ли случаи на изключения? б) Да се намери средната продължителност на една сърдечна трансплантация.

2. При 20 посещения на магазин, студент е записал броя на чакащите пред него: 7, 7, 4, 5, 4, 6, 6, 9, 5, 5, 6, 6, 6, 7, 7, 6, 7, 6, 7, 9.
а) Да се начертае полигонът на честотите, да се намерят горният, долният квартил и медианата и начертае бокс-плотът на извадката.
б) да се намерят извадъчните асиметрия и ексцес.

3. Дадена е извадката
107 119 99 114 120 104 88 114 124 116 101 121 152 100 125 114 95 117.
а) Да се състави интервално разпределение с дължина на интервала d=20;
б) да определят модалният и медианният интервал на разпределението;
в) да се начертае хистограма на честотите;
г) да се намерят 30%-процентил, и квантилите от редове 0,2 и 0,6;
д) да се намерят извадъчните средна и дисперсия.

4. За изследване на неизвестната величина X е получена следната извадка
2, 3, 2, 0, 3, 2, 4, 1, 0, 2, 5, 5, 1, 4, 5, 1, 2, 1, 0, 0, 3, 6, 2,1, 0
а) Да се намерят размахът и средата на извадката, медианата и горният квартил.
б) Да се начертае полигонът на относителните честоти.
в) Да се намерят извадъчните средна и дисперсия.

5. Като се извърши подходяща линейна смяна, да се намерят моментите до трети ред на извадките:

а) $\begin{array}{c|ccccc} x_i & 15 & 16 & 17 & 18 & 19 \\ \hline m_i & 4 & 5 & 4 & 2 & 1 \end{array}$;

б) $\begin{array}{c|cccccc} (x_{i-1},x_i) & (10,14) & (14,18) & (18,22) & (22,26) & (26,30) & (30,34) \\ \hline m_i & 1 & 5 & 10 & 20 & 18 & 3 \end{array}$;

в) $\begin{array}{c|cccccc} (x_{i-1},x_i) & (34,36) & (36,38) & (38,40) & (40,42) & (42,44) & (44,46) \\ \hline m_i & 2 & 3 & 30 & 40 & 20 & 5 \end{array}$.