Статистические распределения и их основные характеристики

Различия индивидуальных значений признака у единиц совокупности называются ***вариацией признака.***

Она возникает в результате того, что индивидуальные значения складываются под совместным влиянием разнообразных условий (факторов), по разному сочетающихся в каждом отдельном случае.

Вариация, которая не зависит от факторов, положенных в основу выделения групп, называется ***случайной вариацией***.

Приемы изучения вариации в пределах одной группы:

- простроение вариационного ряда (ряда распределения);
- графическое изображение;
- исчисление основных характеристик распределения: показателей центра распределения; показателей вариации.

Вариационный ряд -

групповая таблица, построенная по количественному признаку, в сказуемом которой показывается число единиц в каждой группе.

Форма построения вариационного ряда зависит от характера изменения изучаемого признака.

Он может быть построен в форме дискретного ряда или в форме интервального ряда.

Распределение рабочих по тарифному разряду

Тарифный разряд рабочего, x	Число рабочих, имеющих этот разряд, f	Частость W	Накопленная (кумулятивная) частота,S
2	1	0,05	1
3	5	0,25	5+1=6
4	8	0,4	6+8=14
5	4	0,2	14+4=18
6	2	0,1	18+2=20
итого	20	1	

Частость расчитывается по формуле

$$W_i = \frac{f_i}{\sum f_i}$$

Замена частот частостями позволяет сопоставить вариационные ряды с различным числом наблюдений.

Средняя квалификация работников

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{2*1+3*5+4*8+5*4+6*2}{1+5+8+4+2} \approx 4$$

- Т.е в среднем рабочие имеют 4 тарифный разряд

Для признака, имеющего непрерывное изменение строится ***интервальный вариационный ряд распределения.***

Определение величины интервала производится

$$i = \frac{x_{\max} - x_{\min}}{m}$$

$$\text{нижняя граница} = \text{x}_{\min}$$

$$\text{верхняя граница} = \text{x}_{\min} + i$$

Показатели центра распределения.

Средняя арифметическая для дискретного ряда расчитывается по формуле средней арифметической взвешенной:

$$\overline{x} = \frac{\sum xf}{\sum f}$$

В интервальном ряду расчет производится по этой же формуле, но в качестве x берется середина интервала. Она определяется так

$$\frac{\text{нижняя граница} + \text{верхняя граница}}{2}$$

Распределение банков по размеру прибыли.

Размер прибыли, млн. крон,x	Середина интервала, x'	Число банков, f	Накопленная частота, S
3,7 - 4,6	4,15	3	3
4,6 - 5,5	5,05	4	3+4=7
5,5 - 6,4	5,95	5	7+5=12
6,4 - 7,3	6,85	6	12+6=18
7,3 - 8,1	7,7	2	18+2=20
итого	-	20	

Средний размер прибыли

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{4{,}15 * 3 + 5{,}05 * 4 + 5{,}95 * 5 + 6{,}85 * 6 + 7{,}7 * 2}{3 + 4 + 5 + 6 + 2} = 5{,}945$$

Структурные средние

Медиана

Мода

Квартиль

Медиана (Me)

- соответствует варианту, стоящему в середине ранжированного ряда. Положение медианы определяется ее номером:

$$N_{Me} = \frac{n+1}{2}$$

- где n - число единиц в совокупности.

Медиана в дискретном ряду

- По накопленным частотам определяют ее численное значение в дискретном вариационном ряду.

Вставленная фукция в EXCEL

MEDIAN()

Расчет медианы в дискретном ряду

- Медиана тарифного разряда рабочих будет найдена следующим образом:

$$N_{Me} = \frac{n+1}{2} = \frac{20+1}{2} = 10{,}5$$

- Следовательно, среднее значение 10-го и 11-го признаков будут соответствовать медиане. По накопленным частотам находим 10-й и 11-й признаки. Их значение соответствует 4-му тарифному разряду, следовательно медиана в данном ряду равна 4.

Медиана в интервальном ряду

- В интервальном ряду распределения по номеру медианы указывают интервал, в котором находится медиана.
- Численное значение определяется по формуле:

$$Me = X_{Me} + i_{Me} * \frac{\frac{n+1}{2} - S_{Me-1}}{f_{Me}}$$

Расчет медианы в интервальном ряду

- По накопленным частотам определяем, что медиана находится в интервале

5,5 - 6,4 так как номер медианы

$$N_{Me} = \frac{n+1}{2} = \frac{20+1}{2} = 10,5$$

а это значение включает кумулятивная частота 12.

Расчет медианы в интервальном ряду

- Тогда медиана

$$M_e = 5{,}5 + (6{,}4 - 5{,}5) * \frac{\frac{20+1}{2} - 7}{5} = 6{,}13$$

- Таким образом, 50% банков имеют прибыль менее 6,13 млн. крон, а другие 50% - более 6,13.

Мода (Mo)

- наиболее часто встречающееся значение признака.
- В дискретном ряду - это варианта с наибольшей частотой.

Вставленная фукция в EXCEL

MODE()

Значение моды в интервальном ряду

- В интервальном ряду сначала определяется модальный интервал, т.е. тот, который имеет наибольшую частоту, а затем расчитывают моду по формуле:

$$Mo = X_{Mo} + i_{Mo} \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}$$

Определение значения моды в приведенных выше дискретном и интервальном рядах

- В примере 1 наибольшую частоту - 8 имеет четвертый тарифный разряд, следовательно значение моды равно 4 тарифному разряду
- В примере 2 модальный интервал 6,4 -7,3 так как такой уровень прибыли имеют наибольшее число банков.

$$Mo = 6{,}4 + (7{,}3 - 6{,}4) * \frac{(6-5)}{(6-5)+(6-2)} = 6{,}58$$

Квартиль

- это значения признака, которые делят ранжированный ряд на четыре равные по численности части.

- Таких величин будет три:

  первая квартиль(Q1),

  вторая квартиль (Q2),

  третья квартиль (Q3).

- Вторая квартиль является медианой.

Сначала определяется положение или место квартили:

$$N_{Q1} = \frac{n+1}{4}$$

$$N_{Q2} = \frac{n+1}{4} * 2 = \frac{n+1}{2}$$

$$N_{Q3} = \frac{n+1}{4} * 3$$

Квартиль в дискретном ряду

- В дискретном ряду численное значение квартили определяют по накопленным частотам.

Вставленная фукция в EXCEL

QUARTILE()

Квартиль в интервальном ряду

- В интервальном ряду распределения сначала указывают интервал, в котором лежит квартиль, затем определяют ее численное значение по формуле:

$$Q = x_Q + i\frac{N_Q - S_{(Q-1)}}{f_Q}$$

Показатели вариации (колеблемости) признака.

К абсолютным показателям относят:

- Размах колебаний;
- Среднее линейное отклонение;
- Дисперсию;
- Среднее квадратическое отклонение;
- Квартильное отклонение.

Размах колебаний (размах вариации)

- представляет собой разность между максимальным и минимальным значениями признака изучаемой совокупности:

$$R = x_{\max} - x_{\min}$$

- Размах вариации зависит только от крайних значений признака, поэтому область его применения ограничена достаточно однородными совокупностями.

Точнее характеризуют вариацию признака показатели, основанные на учете колеблемости всех значений признака.

К таким показателям относят:

- среднее линейное отклонение,
- дисперсию,
- среднее квадратическое отклонение.

Среднее линейное отклонение d

для несгруппированных данных расчитывается по формуле

$$d = \frac{\sum \left| x - \overline{x} \right|}{n}$$

Вставленная фукция в EXCEL

AVEDEV( )

Для n вариационного ряда:

$$d = \frac{\sum \left| x - \overline{x} \right| \cdot f}{\sum f}$$

Расчет среднего линейного отклонения

Произведено продукции одним рабочим за смену, шт, x	Число рабочих f	xf	$x - \bar{x}$	$\lvert x - \bar{x} \rvert f$
8	7	56	8 - 10 = -2	|8 - 10|*7 = 14
9	10	90	9 - 10 = -1	|9 -10|*10 =10
10	15	150	10 - 10 = 0	|10-10|*15 = 0
11	12	132	11 - 10 = 1	|11-10|*12=12
12	6	72	12 - 10 = 2	|12-10|*6 = 12
	50	500		48

Дисперсия

- - это средняя арифметическая квадратов отклонений каждого значения признака от общей средней.
- Дисперсия обычно называется средним квадратом отклоненй.
- В зависимости от исходных данных дисперсия может вычисляться по средней арифметической простой или взвешенной:

Дисперсия простая

$$\sigma^2 = \frac{\sum \left(x - \bar{x} \right)^2}{n}$$

Вставленная фукция в EXCEL

VARP ( )

Дисперсия взвешенная

$$\sigma^2 = \frac{\sum \left(x - \bar{x}\right)^2 \cdot f}{\sum f}$$

Среднее квадратическое отклонение

- ***стандартное отклонение (Standard Deviation)***

представляет собой корень квадратный из дисперсии

Среднее квадратическое отклонение невзвешенное

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

Вставленная фукция в EXCEL

STDEVP ( )

Среднее квадратическое отклонение взвешенное

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f}}$$

Данные о производительности труда рабочих

Произведено продукции одним рабочим, шт. (x)	Число рабочих f	xf	x - □x	(x - □x)^2	(x - □x)^2*f
8	7	56	-2	4	28
9	10	90	-1	1	10
10	15	150	0	0	0
11	12	132	1	1	12
12	6	72	2	4	24
итого	50	500			74

Расчет показателей дисперсии и среднего квадратического отклонения

1. Исчислим среднюю арифметическую взвешенную:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{500}{50} = 10$$

Расчет показателей дисперсии и среднего квадратического отклонения

2. Определим дисперсию.

$$\sigma^2 = \frac{\begin{array}{l}(8-10)^2 * 7 + (9-10)^2 * 10 + (10-10)^2 * 15 + \\ (11-10)^2 * 12 + (12-10)^2 * 6\end{array}}{7+10+15+12+6} 1{,}48$$

Расчет показателей дисперсии и среднего квадратического отклонения

3. среднее квадратическое отклонение будет равно

$$\sigma = \sqrt{\sigma^2} = \sqrt{1,48} = 1,22$$

- Это означает, что отклонение от средней производительности составило 1,2 шт.

Другой метод расчета дисперсии

- Дисперсия равна разности средней из квадратов признака и квадрата средней.

$$\sigma^2 = \overline{x^2} - \bar{x}^2$$

Относительные показатели вариации

Применяются для оценки интенсивности вариации и для сравнения ее в разных совокупностях.

- относительный размах вариации (коэффициент осцилляции)

$$K_o = \frac{R}{\bar{x}} * 100\%$$

Относительные показатели вариации

- Относительное линейное отклонение (отклонение по модулю)

$$K_o = \frac{d}{\overline{x}} * 100\%$$

- Коэффициент вариации

$$V = \frac{\sigma}{\overline{x}} \cdot 100\%$$

Относительные показатели вариации

- Относительный показатель квартильной вариации (относительное квартильное расстояние)

$$K_{dk} = \frac{d_k}{M_e} * 100\%$$

$$K_Q = \frac{Q_3 - Q_1}{2Q_2} * 100\%$$

- Оценка степени интенсивности вариации возможна только для каждого отдельного признака и совокупности определенного состава.

Предположим вариация производительности труда на предприятиях Эстонии v<10% рассматривается как слабая,10%<v<25% - умеренная, сильная при v>25%. Однако, если рассматривается вариация роста взрослых людей, то при v=4% следует говорить об очень сильной интенсивности

Моменты распределения и показатели его формы.

- Центральные моменты распределения порядка – это средние значения разных степеней отклонений отдельных величин признака от его средней арифметической величины.
- Момент первого порядка равен нулю.
- Второй центральный момент представляет собой дисперсию.
- Третий момент используется для оценки асимметрии
- Четвертый – для оценки эксцесса.

Моменты распределения

Порядок момента	Формула	
	по несгруппированным данным	по сгруппированным данным
Первый μ_1	$\frac{\sum_{(i)}\left(x_i-\bar{x}\right)}{n}$	$\frac{\sum_{(j)}\left(x_j-\bar{x}\right)f_j}{\sum_{(j)} f_j}$
Второй μ_2	$\frac{\sum_{(i)}\left(x_i-\bar{x}\right)^2}{n}$	$\frac{\sum_{(j)}\left(x_j-\bar{x}\right)^2 f_j}{\sum_{(j)} f_j}$

Моменты распределения

Порядок момента	Формула	
	по несгруппированным данным	по сгруппированным данным
Третий μ_3	$\dfrac{\sum\limits_{(i)}\left(x_i-\bar{x}\right)^3}{n}$	$\dfrac{\sum\limits_{(j)}\left(x_j-\bar{x}\right)^3 f_j}{\sum\limits_{(j)} f_j}$
Четвертый μ_4	$\dfrac{\sum\limits_{(i)}\left(x_i-\bar{x}\right)^4}{n}$	$\dfrac{\sum\limits_{(j)}\left(x_j-\bar{x}\right)^4 f_j}{\sum\limits_{(j)} f_j}$

Показатели асимметрии

На основе момента третьего порядка можно построить коэффициент асимметрии

$$A_S = \frac{\mu_3}{\sigma^3}$$

или показатель Пирсона

$$A_{Mo} = \frac{\bar{x} - Mo}{\sigma}$$

Показатели асимметрии

- Если $A > 0$, то асимметрия правосторонняя, а если $A < 0$, то асимметрия левосторонняя, в симметричном распределении – $A=0$.

- В EXCEL используется функция

SKEW ( ).

Характеристика эксцесса распределения

$$E = \frac{\mu_4}{\sigma^4} - 3$$

- В нормальном распределении $E = 0$, поэтому, если $E > 0$, то эксцесс выше нормального (островершинная кривая),

$E < 0$, эксцесс ниже нормального (плосковершинная кривая).

- В EXCEL используется функция

KURT ( ).

Характеристика эксцесса распределения

- По значению показателей асимметрии и эксцесса можно судить о близости распределения к нормальному.
- Если $\frac{As}{\sigma_{as}} \le 2$ и $\frac{Ex}{\sigma_{ex}} \le 2$

то распределение можно считать нормальным

Оценка диапазона изменения статистической переменной

По теореме Чебышева:

- в интервале (μ - 2σ, μ +2σ) находится 75 % значений,
- в интервале (μ - 3σ, μ +3σ) находится 89 % значений.

Оценка диапазона изменения статистической переменной

«Правило трех сигм» справедливо для нормального распределения

- в интервале (μ - σ, μ + σ) находится 68% значений,
- в интервале (μ - 2σ, μ +2σ) находится 95.4% значений,
- в интервале (μ - 3σ, μ +3σ) находится 99.7% значений.

Закон (правило) сложения дисперсий.

$$\sigma_o^2 = \delta^2 + \overline{\sigma}^2$$

- σ_o^2 **- величина общей дисперсии**
- δ^2 **- межгрупповая дисперсия**
- $\overline{\sigma}^2$ **- средняя внутригрупповая дисперсия**

Межгрупповая дисперсия

$$\delta^2 = \frac{\sum \left(\overline{x}_i - \overline{x}\right)^2}{n}; \qquad \delta^2 = \frac{\sum \left(\overline{x}_i - \overline{x}\right)^2 \cdot f}{\sum f}$$

Средняя внутригрупповая дисперсия

$$\overline{\sigma}^2 = \frac{\sum \sigma_i^2}{n}; \qquad \overline{\sigma}^2 = \frac{\sum \sigma_i^2 \cdot f}{\sum f}$$

Имеются следующие данные о времени простоя автомобиля под разгрузкой:

№ пункта разгрузки	1	2	3	4	5	6	7	8	9	10
Число грузчиков	3	4	4	3	3	4	4	4	3	4
Время простоя мин.	12	10	8	15	19	12	8	10	18	8

Вспомогательная таблица для расчета общей дисперсии.

Время простоя под разгрузкой мин., x	Число выполненных разгрузок, f	x*f	$x - \square x_0$	$(x-\square x_0)^2$	$(x-\square x_0)^2 f$
8	3	24	-4	16	48
10	2	20	-2	4	8
12	2	24	0	0	0
15	1	15	3	9	9
18	1	18	6	36	36
19	1	19	7	49	49
итого	10	120	-	-	150

- Среднее время простоя

$$\bar{x} = \frac{120}{10} = 12 мин$$

- Общая дисперсия

$$\sigma_o^2 = \frac{150}{10} = 15$$

Расчет внутригрупповой дисперсии по первой группе (число грузчиков, участвующих в разгрузке, 3 чел)

Время простоя под разгрузкой, мин., x	Число выполнен-ных разгрузок, f	x*f	$x - \square x_1$	$(x - \square x_1)^2 f$
12	1	12	-4	16
15	1	15	-1	1
18	1	18	2	4
19	1	19	3	9
итого	4	64	-	30

Дисперсия первой группы

$$\bar{x}_1 = \frac{64}{4} = 16 \text{ мин}$$

$$\sigma_1^2 = \frac{30}{4} = 7{,}5$$

Расчет внутригрупповой дисперсии по второй группе (число грузчиков, участвующих в разгрузке, - 4)

Время простоя под разгрузкой, мин., x	Число выполненных разгрузок, f	x*f	$x - \square x_2$	$(x - \square x_2)^2 f$
8	3	24	-1,33	5,31
10	2	20	0,67	0,90
12	1	12	2,67	7,13
итого	6	56	-	13,37

Дисперсия второй группы

$$\bar{x}_2 = \frac{56}{6} = 9{,}33 \text{ мин}$$

$$\sigma_2^2 = \frac{13{,}37}{6} = 2{,}23$$

Средняя из внутригрупповых дисперсий

$$\bar{\sigma}^2 = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{7{,}5 * 4 + 2{,}23 * 6}{4 + 6} = 4{,}3$$

Межгрупповая дисперсия

$$\delta^2 = \frac{\sum \left(\bar{x}_i - \bar{\bar{x}}\right)^2 \cdot f}{\sum f} = \frac{(16-12)^2 * 4 + (9{,}33-12)^2 * 6}{4+6} = 10{,}7$$

Общая дисперсия

$$\sigma_o^2 = 4{,}3 + 10{,}7 = 15{,}0$$