Тема 5. Показатели вариации

1. Понятие вариации. Виды показателей вариации.
2. Виды дисперсий в совокупности, разделенной на группы. Правило сложения дисперсии.
3. Характеристика закономерности рядов распределения.

1. Понятие вариации.

Различие индивидуальных значений признака внутри изучаемой совокупности называется вариацией признака. Вариация возникает в результате того, что индивидуальные значения признака складываются под совокупным влиянием разнообразных факторов.

Вариация - это колеблемость величины признака у отдельных единиц совокупности под влиянием различных факторов, как систематических, так и случайных.

Систематические факторы- действуют постоянно, являются существенными и проявляются в вариации закономерно.

Случайные факторы- вносят хаотичность в изменение значений признака.

Вариацию под влиянием случайных факторов называют *случайной вариацией*, а под влиянием систематических факторов - *систематической вариацией.*

Общая вариация учитывает влияние как систематических, так и случайных факторов.

Для изучения вариации значений признака недостаточно знать только среднюю величину признака.

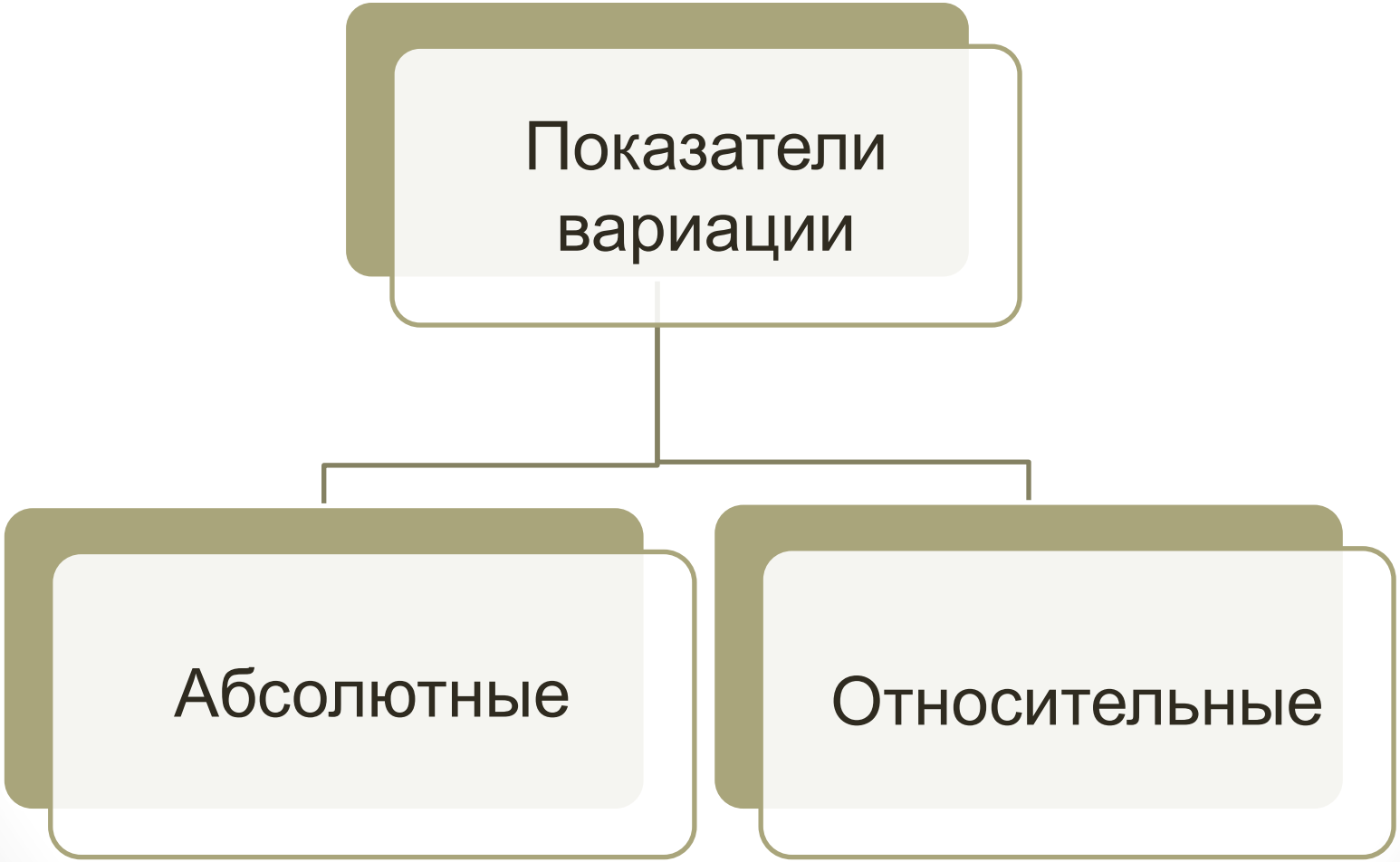
Средняя величина не показывает строения совокупности, не дает представления о том, как отдельные значения изучаемого признака группируются вокруг своей средней величины.

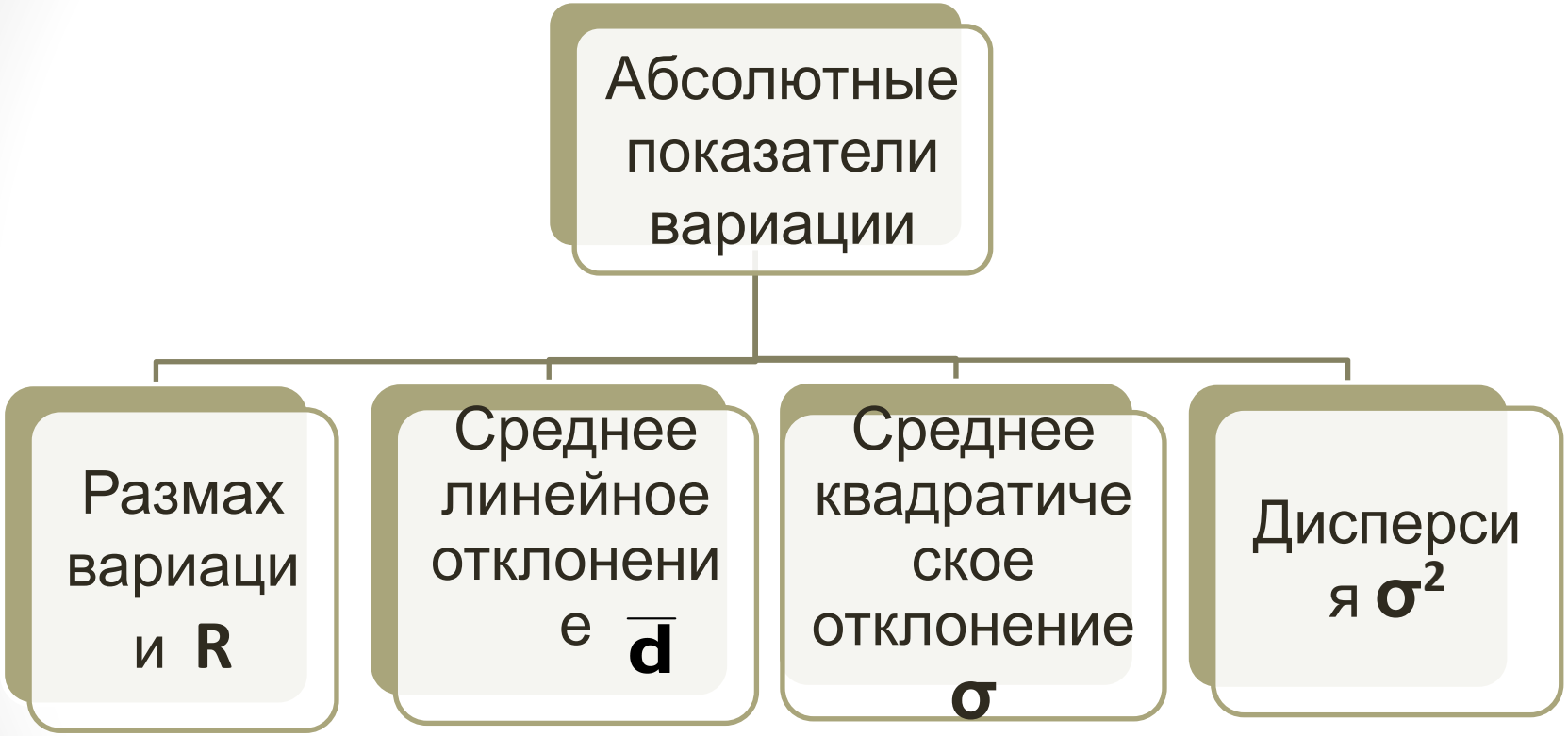
В некоторых случаях отдельные значения признака близко примыкают к средней и мало от нее отличаются. В таких случаях средняя хорошо представляет всю совокупность, т.е . будет типичной.

В других случаях, отдельные значения признака совокупности далеко отстоят от средней, тогда средняя плохо представляет всю совокупность.

Поэтому необходимо знать и разброс отдельных единиц по отношению к среднему значению.

Возникает необходимость измерять вариацию признака в совокупностях. Для этой цели вводится ряд обобщающих показателей вариации.




Абсолютные показатели вариации
Размах вариации R
Среднее линейное отклонение $\overline{d}$
Среднее квадратическое отклонение σ
Дисперсия σ^2

1. *Размах вариации* – это разность между максимальным и минимальным значением исследуемого признака в совокупности.

R= Xmax-Xmin

2. *Среднее линейное отклонение* – это средняя арифметическая абсолютных значений отклонений вариантов от их средней величины:

- простая при несгруппированных данных

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- взвешенная при сгруппированных данных

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| * f_i}{\sum_{i=1}^{n} f_i}$$

3. *Среднее квадратическое отклонение* (называется стандартным отклонением) является наиболее совершенной характеристикой вариации признака:

- простая форма

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

- взвешенная форма

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

Среднее квадратическое отклонение - это обобщающая характеристика размеров вариации признака в совокупности, оно показывает, на сколько в среднем отклоняются конкретные варианты признака от среднего значения, является абсолютной мерой колеблемости признака и выражается в тех же единицах, что и признак, поэтому экономически хорошо интерпретируется.

Сопоставление средних отклонений – квадратического **σ** и линейного $\bar{d}$ позволяет сделать вывод ***об устойчивости индивидуальных значений признака***, т.е. об отсутствии среди них «аномальных» значений вариантов.

Отношение показателей $\bar{d}$ и σ может служить ***индикатором устойчивости данных***: если $\frac{\bar{d}}{\sigma} > 0{,}8$, то значения признака неустойчивы, в них имеются «аномальные» выбросы.

Показатель вариации σ является основной абсолютной мерой вариации. Он широко используется в выборочных наблюдениях при установлении границ однородности совокупности, при установлении формы кривой распределения и др.

По значениям показателей $\bar{x}$ и σ можно определить границы диапазонов рассеяния значений признака относительно средней $\bar{x}$, т.е. установить, какая доля значений признака попадает в тот или иной диапазон отклонений от $\bar{x}$.

В нормально распределенных и близких к ним рядах вероятностные оценки диапазонов рассеяния значений признака таковы:

68,3% войдет в диапазон ($\bar{x} \pm \sigma$);

95,4% попадет в диапазон ($\bar{x} \pm 2\sigma$);

99,7% появится в диапазон ($\bar{x} \pm 3\sigma$).

Данное соотношение известно как правило «трех сигм».

По значениям $\bar{x}$ и σ, основываясь на правиле «трех сигм», можно точно оценить границы всех трех диапазонов рассеяния признака и определить, сколько значений X_i попадает в каждый из диапазонов.

4. *Дисперсия* - это квадрат среднего квадратического отклонения:

- простая

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

- взвешенная

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

.

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

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

Пример 1. Имеются данные о товарообороте магазинов райо
Необходимо рассчитать характеристики ряда распределения.

Группы магазинов x_i	Число магаз. f_i	Середина интер. x'_i	$x'_i * f_i$	$x'_i - \bar{x}$	$(x'_i - \bar{x})^2$	$(x'_i - \bar{x})^2 * f_i$	$x_i^2 * f_i$
170 – 190	10	180	1800	-36	1296	12960	324000
190 – 210	20	200	4000	-16	256	5120	800000
210 – 230	50	220	11000	4	16	800	2420000
230 – 250	20	240	4800	24	576	11520	1152000
Итого	100		21600			30400	4696000

1.

$$\bar{x} = \frac{\sum_{i=1}^{n} (x_i) * f_i}{\sum_{i=1}^{n} f_i} = \frac{21600}{100} = 216 \text{млн.руб}$$

$$D = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 * f_i}{\sum_{i=1}^{n} f_i} = \frac{30400}{100} = 304$$

2.

$$\sigma^2 = \overline{x^2} - \bar{x}^2 = \frac{4696000}{100} - 216^2 = 304$$

Дисперсия альтернативного признака

В ряде случаев возникает необходимость в измерении дисперсии альтернативных признаков, тех, которыми обладают одни единицы совокупности, и не обладают другие (брак продукции, ученая степень и др.).

Обозначим **p** –доля единиц совокупности, обладающая данным признаком и **q** – доля единиц, не обладающая данным признаком: **p+q =1**.

Альтернативный признак принимает всего два значения **0** и **1** с весами соответственно **q** и **p**.

Найдем **среднее значение** альтернативного признака:

$$\overline{x}_p = \frac{\sum x_i f_i}{\sum f_i} = \frac{1*p + 0*q}{p+q} = \frac{p}{1} = p$$

Дисперсия альтернативного признака:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} = \frac{(1-p)^2 * p + (0-p)^2 * q}{p+q} = \frac{q^2 * p + p^2 * q}{1} = pq(q+p) = pq,$$

Пример. На 10000 человек населения района приходится 4500 мужчин и 5500 женщин.

$$p = \frac{4500}{10000} = 0{,}45; \quad q = \frac{5500}{10000} = 0{,}55 \qquad \sigma_p^2 = pq = 0{,}45 * 0{,}55 = 0{,}2475$$

Среднее квадратическое отклонение альтернативного признака: $\sigma_p = \sqrt{pq} = \sqrt{p(1-p)}$

Пример. Известно, что 2% всех деталей бракованные. Найти дисперсию брака.

$$\sigma_p^2 = 0{,}02 * 0{,}98 = 0{,}0196,$$

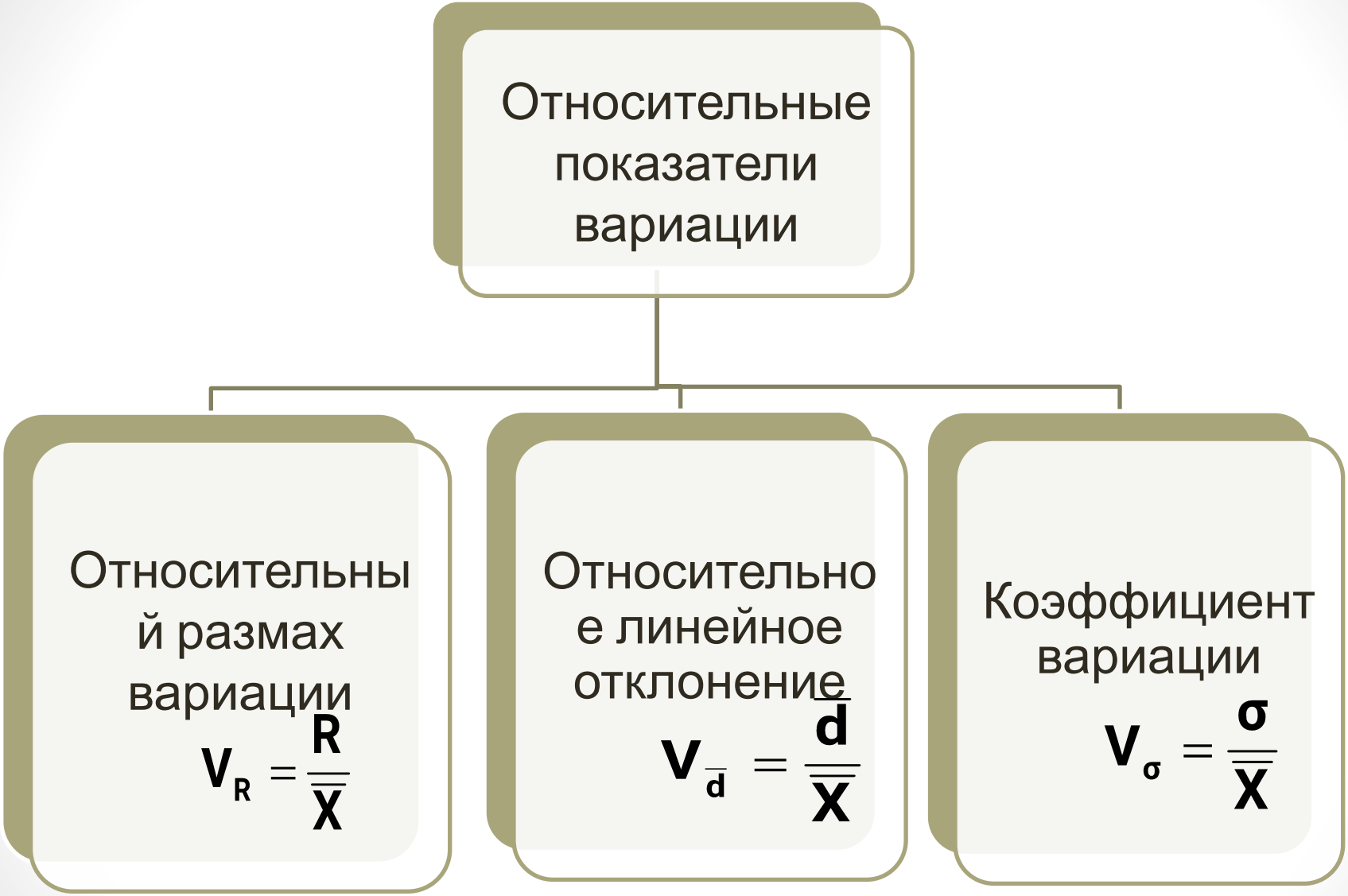
Среднеквадратическое отклонение доли брака будет

$$\sigma_p = \sqrt{0{,}0196} = 0{,}14,$$

Интенсивность вариации признака измеряется относительными показателями.

Относительные показатели вводятся для сравнительной оценки вариации совокупности по разным признакам или для сравнения вариаций нескольких совокупностей по одному и тому же признаку.

Эти показатели вычисляются как отношение абсолютных показателей вариации к средней величине.


Относительные показатели вариации
Относительный размах вариации
$V_R = \frac{R}{\overline{X}}$
Относительное линейное отклонение
$V_{\overline{d}} = \frac{\overline{d}}{\overline{X}}$
Коэффициент вариации
$V_\sigma = \frac{\sigma}{\overline{X}}$

Коэффициент вариации **V□** выражается в процентах и вычисляется по формуле:

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100$$

Величина **V□** оценивает интенсивность колебаний вариантов относительно их средней величины. Принята следующая оценочная шкала колеблемости признака:

0%<V□≤40% - колеблемость незначительная;
40%< V□≤60% - колеблемость средняя (умеренная);
V□>60% - колеблемость значительная.

Для нормальных и близких к нормальному распределений показатель **V□** служит индикатором однородности совокупности: **V□≤33%**

Пример 2. На этапе отбора претендентов для участия в проекте фирмы объявлен конкурс. Распределение претендентов по опыту работы (лет) показано в таблице :

Группы по опыту работы, лет	f, чел.	Центр интер-вала	$x_i' * f_i$	$\left(x_i' - \bar{x}\right)$	$\left(x_i' - \bar{x}\right)^2 * f_i$	$\left(x_i'\right)^2 * f_i$
А	1	2	3	4	5	6
до 4-х	10	3	30	-4.2	176.4	90
4 – 6	10	5	50	-2.2	48.4	250
6 – 8	50	7	350	-0.2	2.0	2450
8 – 10	20	9	180	1.8	64.8	1620
свыше 10	10	11	110	3.8	144.4	1210
ИТОГО	100	-	720		436.0	5620

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * f_i}{\sum_{i=1}^{n} f_i} = \frac{720}{100} = 7.2(\text{лет})$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 * f_i}{\sum_{i=1}^{n} f_i} = \frac{436}{100} = 4{,}36$$

2. Виды дисперсий в совокупности, разделенной на группы. Правило сложения дисперсии.

Вариация признака обусловлена различными факторам Поэтому, изучая вариацию по всей совокупности в целом и рассчитав общую среднюю, невозможно определить влияние отдельных факторов на колеблемость индивидуальных значений признака.

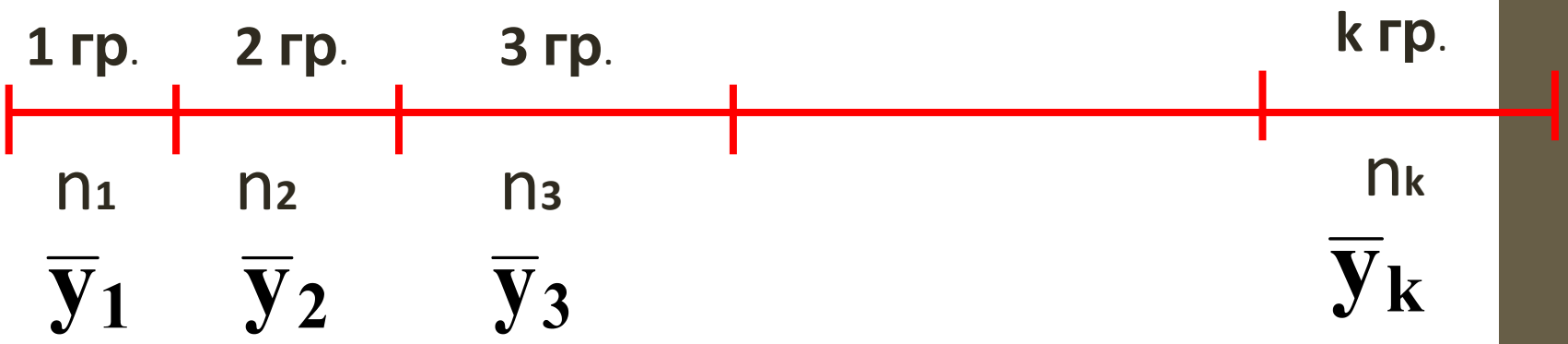
Это можно сделать, если статистическую совокупность разбить на группы по какому-либо признаку-фактору. Тогда, наряду с изучением вариации признака по всей совокупности целом, можно изучить вариацию для каждой из составляющих ее групп, а также вариацию между этими группами.

Показатели вариации могут быть использованы не только в анализе колеблемости признака, но и для оценки влияния одного признака на вариацию другого признака, т.е. в анализе взаимосвязей между показателями.

Для такого анализа совокупность должна быть разбита на группы по факторному признаку. При этом используются ***три*** вида дисперсий - это общая дисперсия, дисперсия межгрупповая и внутригрупповая (средняя из внутригрупповых дисперсий).

1 гр.
1 г

Обозначая факторный признак – **X**, результативный – **У**, дадим определение этих трех видов дисперсии.
Введем обозначения:



$n=n_1+n_2+\ldots+n_k$;
k – количество групп;

$\overline{y}_j$ – среднее значение результативного признака **У** в **j**-ой группе;

$\overline{y}_{об}$ – общая средняя по всей совокупности;

n – число единиц совокупности.

Общая дисперсия $\sigma^2_{об}$ характеризует вариацию признака во всей совокупности, сложившуюся под влиянием всех факторов (систематических и случайных), обусловивших эту вариацию.

$$\sigma^2_{об} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_{об})^2}{n}$$

Межгрупповая дисперсия δ^2_x измеряет систематическую вариацию, которая обусловлена влиянием того признака-фактора **X,** по которому произведена группировка. Такое воздействие фактора проявляется в отклонении групповых средних от общей средней.

$$\delta_x^2 = \frac{\sum_{j=1}^{k} \left(\overline{y_j} - \overline{y_{об}}\right)^2 * n_j}{\sum_{j=1}^{k} n_j}$$

$\overline{y_j}$ – групповые средние;

$\overline{y_{об}}$ – общая средняя;

n_j – численность единиц в j-ой группе;

k – количество групп.

Внутригрупповая дисперсия σ_j^2 оценивает вариацию признака, сложившуюся под влиянием других, не учитываемых в данном исследовании факторов, и не зависящую от группировочного фактора **X**.

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (y_i - \overline{y_j})^2}{n_j}$$

y_i – индивидуальные значения признака внутри группы;

$\overline{y_j}$ – среднее значение признака в группе с номером j;

n_j – численность единиц в j-ой группе.

На основании внутригрупповых дисперсий всех групп σ_j^2, вычисляется средняя из внутригрупповых дисперсий:

$$\overline{\sigma_j^2} = \frac{\sum_{j=1}^{k} \sigma_j^2 * n_j}{\sum_{j=1}^{k} n_j}$$

Правило сложения дисперсии :

$$\sigma^2_{об} = \delta^2_x + \overline{\sigma^2_j}$$

Данное правило показывает связь между различными видами дисперсий.

Это правило позволяет оценить влияние группировочного признака на образование общей вариации.

Очевидно, чем больше доля межгрупповой дисперсии в общей, тем сильнее влияние группировочного **X** признака на изучаемый результативный признак **Y**.

В статистическом анализе широко используется показатель η^2 , который называют эмпирическим коэффициентом детерминации.

Он характеризует долю межгрупповой дисперсии в общей дисперсии .

Межгрупповая дисперсия обусловлена вариацией признака, положенного в основу группировки. Она показывает силу влияния факторного признака на образования общей вариации:

$$\eta^2 = \frac{\delta_x^2}{\sigma_{об}^2}$$

Эмпирический коэффициент детерминации показывает долю вариации результативного признака **Y** под влиянием вариации факторного признака **X.**

Теснота связи между группировочным и результативным признаками оценивается показателем η, который называется эмпирическим корреляционным отношением.

Для качественной оценки тесноты связи на основ служит соотношение Чэддока:

η	0,1 – 0,3	0,3 – 0,5	0,5 – 0,7	0,7 – 0,9	0,9 – 0,99
Сила связи	слабая	умеренная	заметная	тесная	Весьма тесная

Чем значение η ближе к 1, тем теснее связь между признаками.

Пример 3.

Стоимость 1 кв.м общей площади в у.е. на рынке жилья для двух групп домов приведена в таблице 3. При этом известно, что дома 1-ой группы находятся вблизи от станции метро, а дома 2-ой группы – на значительном расстоянии от станции метро.

Необходимо установить влияет ли месторасположение домов на стоимость 1 кв.м общей площади.

Группировочный факторный признак **X** – это качественный признак (расположение дома – близость к станции метро); результативный признак **Y** – стоимость 1 кв.м общей площади.

Таблица 3

	№ п/п	Стоимость $м^2$, тыс. у.е, Y	Y^2
J=1	1	3,9	15,21
	2	3,8	14,44
	3	3,6	12,96
	4	4,1	16,81
итого	4	15,4	59,42
J=2	1	3,3	10,89
	2	2,6	6,76
	3	2,8	7,84
	4	2,2	4,84
	5	3,1	9,61
	6	2,8	7,84
итого	6	16,8	47,78
Всего	10	32,2	107,20

1. Рассчитаем среднюю стоимость одного м^2. жилья и общую дисперсию по всей совокупности в целом:

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{32,2}{10} = 3,22 \text{тыс.уе.}$$

$$\sigma^2_{об} = \overline{y^2} - \bar{y}^2 = \frac{107,20}{10} - 3,22^2 = 0,3516$$

2. Вычислим среднюю стоимость одного м^2 жилья и дисперсию для каждой группы домов.

$$\overline{y_1} = \frac{15.4}{4} = 3.85 \text{ тыс.у.е.} \qquad \sigma_1^2 = \frac{59,42}{4} - 3,85^2 = 14,855 - 14,8225 = 0,0325$$

$$\overline{y_2} = \frac{16,8}{6} = 2,8 \text{ тыс.у.е.} \qquad \sigma_2^2 = \frac{47,78}{6} - 2,8^2 = 7,9633 - 7,84 = 0,1233$$

3.Определим величину межгрупповой дисперсии

$$\delta_x^2 = \frac{\sum_{j=1}^{k} (\overline{y_j} - \overline{y})^2 * n_j}{\sum_{j=1}^{k} n_j} = \frac{(3.85 - 3.22)^2 * 4 + (2.8 - 3.22)^2 * 6}{10} = \frac{2.646}{10} = 0.2646$$

4. Найдем эмпирический коэффициент детерминации

$$\eta^2 = \frac{\delta_x^2}{\sigma_o^2} = \frac{0,2646}{0,3516} = 0.752 \quad \text{или} \quad 75,2\%$$

5. Эмпирическое корреляционное отношение

$$\eta = \sqrt{0.752} = 0.87$$

6. Определим среднею из внутригрупповых дисперсий

$$\overline{\sigma_j^2} = \frac{\sum_{j=1}^{k} \sigma_j^2 f_j}{\sum_{j=1}^{k} f_j} = \frac{0,0325 * 4 + 0,1233 * 6}{10} = \frac{0,13 + 0,7398}{10} = 0,0869$$

7. Найденные дисперсии в сумме дают общую дисперсию.

0,2646+0,0869=0,3515

Правило сложения дисперсии для доли признака.

Рассмотренное правило сложения дисперсий верно и для дисперсии доли признака.

Дисперсия альтернативного признака: $\sigma^2 = pq = p(1-p)$

Средняя величина $\overline{x}_p = p$

Тогда **внутригрупповая** дисперсия доли :

$$\sigma^2_{p_i} = p_i * (1 - p_i),$$

где p_i - доля изучаемого признака в i-ой группе.

Средняя из внутригрупповых дисперсий :

$$\overline{\sigma^2_{p_i}} = \frac{\sum_i p_i (1 - p_i) * n_i}{\sum_i n_i}$$

Формула **межгрупповой** дисперсии имеет вид:

$$\delta^2_{p_i} = \frac{\sum_i (p_i - \overline{p})^2 * n_i}{\sum_i n_i},$$

где n_i - численность единиц в отдельных группах;

$\overline{p}$ - доля изучаемого признака во всей совокупности.

Доля признака в совокупности определяется по средней арифметической взвешенной:

$$\overline{p} = \frac{\sum_i p_i n_i}{\sum_i n_i}$$

Правило сложения дисперсий доли признака выражается соотношением:

$$\sigma^2_{\overline{p}} = \overline{\sigma^2_{p_i}} + \delta^2_{p_i}$$

Пример 4. Данные удельного веса основных рабочих в трех цехах фирмы представлены в таблице.

Определить общую, внутрицеховую и межцеховую дисперсии доли основных рабочих.

Цех	Удельный вес основных рабочих, в %, p_i	Численность всех рабочих, чел, n_i
1	80	100
2	75	200
3	90	150
Итого		450

1. Определим долю основных рабочих в целом по фирме:

$$\overline{p} = \frac{\sum_i p_i n_i}{\sum_i n_i} = \frac{0{,}80 * 100 + 0{,}75 * 200 + 0{,}90 * 150}{450} = \frac{365}{450} = 0{,}81$$

2. Общая дисперсия доли основных рабочих по всей фирме в целом равна:

$$\sigma^2_{\overline{p}} = \overline{p} * (1 - \overline{p}) = 0{,}81 * (1 - 0{,}81) = 0{,}154$$

3. Внутрицеховые дисперсии равны:

$$\sigma^2_{p_i} = p_i * (1 - p_i)$$

$$\sigma^2_{p_1} = 0{,}80 * 0{,}20 = 0{,}16$$

$$\sigma^2_{p_2} = 0{,}75 * 0{,}25 = 0{,}19$$

$$\sigma^2_{p_3} = 0{,}90 * 0{,}10 = 0{,}09$$

4. Средняя из внутрицеховых дисперсий равна:

$$\overline{\sigma^2_{p_i}} = \frac{\sum_i p_i (1 - p_i) * n_i}{\sum_i n_i} = \frac{0{,}16 * 100 + 0{,}19 * 200 + 0{,}09 * 150}{450} = \frac{67{,}5}{450} = 0{,}15$$

5. Межцеховая дисперсия равна:

$$\delta^2_{p_i} = \frac{\sum_i (p_i - \bar{p})^2 * n_i}{\sum_i n_i} = \frac{(0{,}80 - 0{,}81)^2 * 100 + (0{,}75 - 0{,}81)^2 * 200 + (0{,}90 - 0{,}81)^2 * 150}{450} = \frac{1{,}945}{450} = 0{,}004$$

Проверка вычислений: 0,154 = 0,15 + 0,004.

3.Характеристика закономерности рядов распределения.

Для обобщающей характеристики особенностей формы распределения применяются кривые распределения, которые выражают графически закономерность распределения единиц совокупности по величине варьирующего признака.

Различают эмпирические и теоретические кривые распределения.

Эмпирическая кривая распределения - это фактическая кривая распределения, полученная по данным наблюдения, в которой отражаются как общие, так и случайные условия, определяющие распределение.

Теоретическая кривая распределения - это кривая, выражающая общую закономерность данного типа распределения. При этом теоретическое распределение играет роль некоторой идеализированной модели эмпирического распределения, а сам процесс анализа вариационного ряда сводится к сопоставлению эмпирического и теоретического распределений..

Кривые распределения могут быть одно-, двух- и многовершинными.

Для однородных совокупностей характерны одновершинные распределения. Многовершинность свидетельствует о неоднородности изучаемой совокупности. В этом случае необходимо сделать перегруппировку данных с целью получения однородных групп.

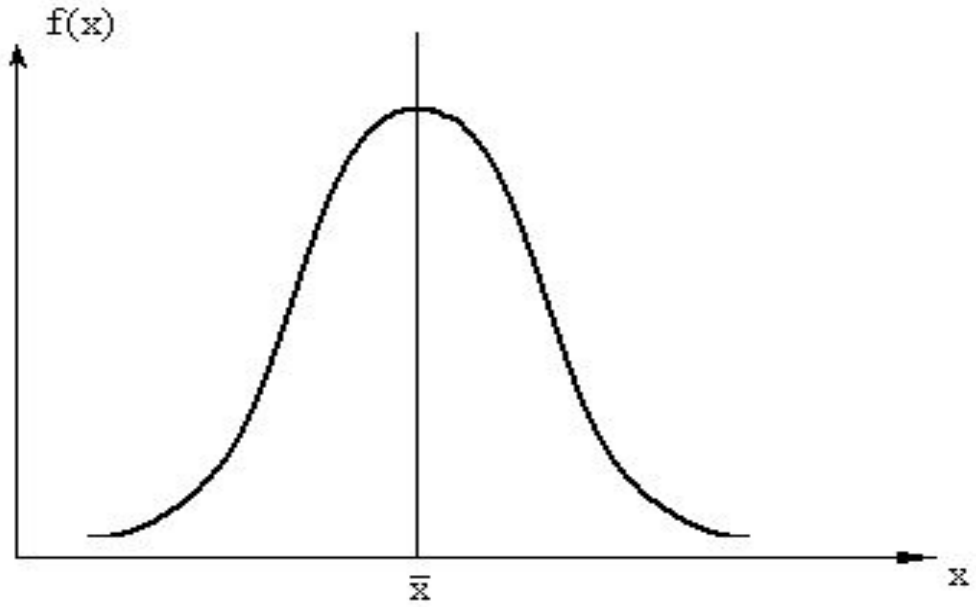
Кривые распределений бывают симметричными и асимметричными. В зависимости от того, какая ветвь кривой вытянута - правая или левая, различают правостороннюю или левостороннюю асимметрию.

Для симметричных распределений частоты любых двух вариантов, равноотстоящих от центра в обе стороны, равны между собой.

Распределение изучаемого признака характеризуется 3-мя группами показателей:

- показатели центра;
- показатели вариаций;
- показатели для изучения формы кривой.

Нормальное распределение является симметричным



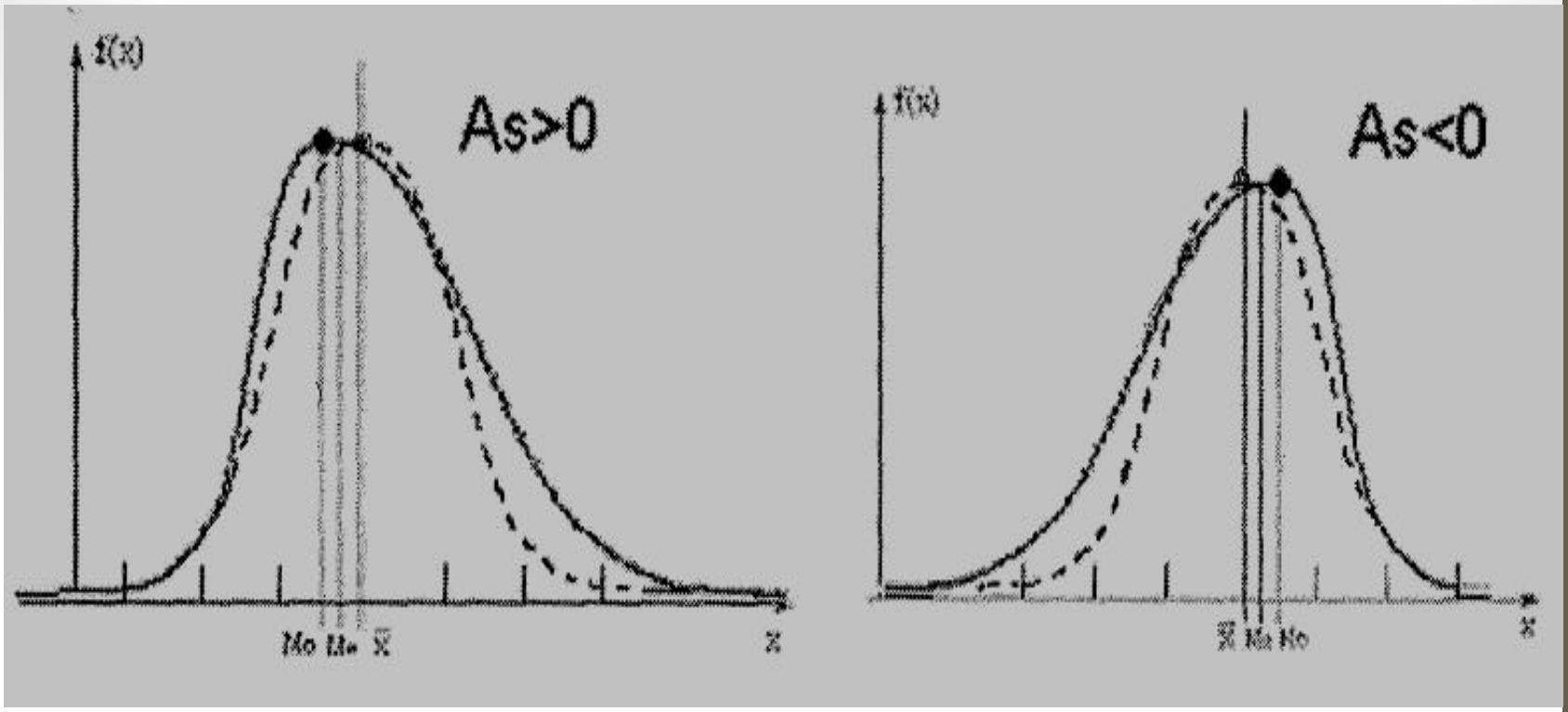
Для симметричных распределений имеют место следующие характеристики:

1. $\bar{x} = M_o = M_E$
2. $R = 6 * \sigma$
3. $\sigma = 1.25 * \bar{d}$

Если эти соотношения нарушены, то это свидетельствует о наличии асимметрии распределения.

Показатель асимметрии **As** оценивают смещение ряда распределения влево или вправо по отношению к оси симметрии нормального распределения.

В случае асимметричного распределения вершина кривой находится не в середине, а сдвинута либо влево, либо вправо.



Если вершина сдвинута влево, то правая часть кривой оказывается длиннее левой т.е. имеет место правосторонняя асимметрия, характеризующаяся неравенством $\bar{X} > Me > Mo$.

Если же вершина кривой сдвинута вправо и левая часть оказывается длиннее правой, то асимметрия левосторонняя, для которой справедливо неравенство $\bar{X} < Me < Mo$.

Установлена следующая оценочная шкала асимметричности:

|As|≤0,25 - асимметрия незначительная;

|As|>0,5 - асимметрия существенная.

0,25<|As|≤0,5 - асимметрия заметная (умеренная);

Показатель эксцесса **Ek** характеризует крутизну кривой распределения - ее заостренность или пологость по сравнению с нормальной кривой.