Группировка

Группировкой называется расчленение множества единиц изучаемой статистической совокупности на группы по существенным для них признакам, которые называются **группировочными признаками или основанием группировки**

Цели группировки

Группировка позволяет:

1. «Сжать» статистическую информацию.
2. Выявить структурные сдвиги в общественных явлениях
3. Выявить взаимосвязи между вариационными признаками.
4. Выявить динамику развития социально-экономических явлений.

Виды группировок

1. Типологическая
2. Структурная
3. Аналитическая

Определение типологической группировки

Типологическая группировка - это разделение исследуемой качественно разнородной совокупности на классы, социально-экономические типы, однородные группы единиц в соответствии с правилами научной группировки.

Пример типологической группировки

Группировка предприятий по формам собственности в одном из регионов (1994 г.)	Число предприятий	
	Всего единиц	% к итогу
Федеральная собственность	26 326	93, 6
Муниципальная собственность	89	0,3
Частная собственность	1366	4,9
Смешанная собственность	331	1,2
Всего	28112	100

Определение структурной группировки

Структурной называется группировка, в которой происходит разделение однородной совокупности на группы, характеризующую ее структуру по значениям какого либо варьирующегося признака.

Группировка населения России по размеру среднедушевого дохода апрель 1994 г.

Группы населения по размеру дохода, тыс. руб./ мес.	Численность населения	
	Всего, млн. человек	% к итогу
До 80	25,8	17,4
80 - 120	34,8	23,5
120 - 160	29,4	19,8
160 - 200	20,7	13,9
200 - 240	13,5	9,1
240 и более	24,2	16,3
Всего	148,4	100

Принципы группировки по количественному признаку

1. Группы не должны быть малочисленными
2. Интервалы следует выбирать таким образом, чтобы большее число единиц совокупности принадлежало интервалам средней части.
3. Не должно быть пустых интервалов, не содержащих ни одной единицы совокупности.
4. Отмеченные правила не являются строгими и носят эмпирический характер

Формальные способы определения числа интервалов

1. Формула *Стерджесса*

$$k = 1 + 3{,}322 \cdot \lg_{10} n$$

Формулой Стерджесса разумно пользоваться при $n > 30$

Виды интервалов группировки. 1

1. Открытые и закрытые интервалы

Интервалы могут быть открытыми и закрытыми.

Для открытых интервалов определена только верхняя или только нижняя границы.

Для закрытых равных интервалов величина интервала определяется формулой

$$h = \frac{X_{max} - X_{мин}}{K}.$$

Аналитическая группировка

1. Группировка, выявляющая взаимосвязи между изучаемыми явлениями называется аналитической.

2. В основу аналитической группировки кладется факторный признак.

3. Каждая группа должна характеризоваться некоторым значением результативного признака, причем должна прослеживаться связь между факторным и результативным признаками.

4. Если такой связи нет, то группировка не является аналитической

Распределение коммерческих банков России по сумме активов баланса (данные условные)

Группы банков по сумме активов баланса, тыс. руб.	Кол-во банков, единиц	В среднем на 1 банк	
		Численность занятых, чел.	Прибыль млн. руб.
До 20 000	19	184	22,5
20 000 – 30 000	8	313	31,6
30 000 – 40 000	7	374	36,0
4 0000 – 50 000	9	468	69,2
50 000 и более	7	516	205,6
Всего	50	325,1	59,4

Демонстрируется связь между суммой активов и балансовой прибылью

Сложной называется группировка, при которой формируются группы сначала по одному признаку, а затем они делятся на подгруппы по другому признаку.

Группировка семей России по месту проживания и числу детей в 1989 г. (по материалам переписи)

Группы семей по месту проживания	В том числе подгруппа семей по числу детей	Число семей, тыс.
Городское население	1 ребенок	9 605
	2 детей	6 936
	3 детей	971
	4 и более	229
Сельское население	1 ребенок	2 328
	2 детей	2 306
	3 детей	757
	4 и более	354
	Всего	23, 486

1 Относительные и абсолютные показатели

Определение статистического показателя

- Статистический показатель представляет собой количественную характеристику социально-экономических явлений и характеризует явление с качественной и количественной сторон.

- Совокупность взаимосвязанных статистических показателей, нацеленных на решение конкретной статистической задачи, называется системой статистических показателей.

Абсолютные показатели

Абсолютным показателем называется такой показатель, который отражает либо суммарное число единиц, либо суммарное свойство объекта, например, выработка электроэнергии тепловыми станциями РФ в 2001 г .

Абсолютные показатели выражаются именованными величинами в натуральных единицах измерения (штуках, килограммах, тоннах, киловаттах и т. д.)

Абсолютные показатели являются исходными при статистическом анализе, но не являются наглядными при сравнительном анализе явлений.

Относительные показатели

Относительный показатель — это показатель, который получается путем сопоставления абсолютных показателей, относящихся к разным объектам в один момент времени, разным моментам времени для одного и того же объекта, или сравнения разных свойств одного и того же объекта (относительные показатели первого порядка).

Относительные показатели, получающиеся при сопоставлении других относительных показателей, называются относительными показателями второго порядка. Если соотносятся относительные показатели второго порядка, то получаются относительные показатели третьего порядка и т. д.

Виды относительных показателей. 1

1. Относительные показатели, характеризующие структуру объекта (отношение части к целому, например число женщин к общей численности населения РФ). Размерности не имеют.

2. Относительные показатели характеризующие динамику *процесса.* Отношение показателя, характеризующего объект в текущий период, к значению аналогичного показателя в более ранний (базисный) период. Такие показатели называются темпами роста. Размерности не имеют.

Виды относительных показателей. 2

3. Относительные показатели, характеризующие соотношение разных признаков того же объекта между собой (иногда их называют показателями интенсивности). Могут быть размерными.

4. Показатели, которые являются отношением фактически наблюдаемых значений признака к его нормативным, плановым, оптимальным или максимально возможным величинам. Размерности не имеют.

Примеры статистических показателей

- *Относительный показатель динамики* $K_Д$ представляет собой отношение уровня исследуемого явления по состоянию на данный момент времени (или период времени) к уровню того же явления в прошлом:

$$K_Д = \frac{Фактический\ уровень\ за\ текущий\ период}{Фактический\ уровень\ за\ базисный\ период} \cdot 100\%.$$

- *Относительный показатель плана* $K_П$ используется для сравнения уровня планируемых и достигнутых результатов деятельности и выражается формулой

$$K_П = \frac{Уровень, планируемый\ на\ (i+1)\ период}{Уровень, достигнутый\ \ в\ i\text{ - }м\ периоде} \cdot 100\% .$$

- *Относительная величина реализации плана* $K_{ВП}$ выражает степень реализации плана

$$K_{ВП} = \frac{Фактическое\ выполнение\ плана\ за\ текущий\ период}{Планированное\ задание\ на\ текущий\ период} \cdot 100\%$$

Из определения этих показателей следует, что они связаны простым соотношением: $K_Д = K_{ВП} * K_П$.

Доказательство формулы

$$К_Д = К_{ВП} * К_П$$

Выпишем формулы для интересующих нас показателей, взяв, *i*- тый уровень за базисный, а *i*+1 - за текущий:

$$К_Д = \frac{Фактический\ уровень\ за\ текущий\ период}{Фактический\ уровень\ за\ базисный\ период} \cdot 100\%.$$

$$К_П = \frac{Уровень,\ планируемый\ на\ текущий\ период}{Уровень\ за\ базисный\ период} \cdot 100\%.$$

$$К_{ВП} = \frac{Фактический\ уровень\ за\ текущий\ период}{Уровень,\ планируемый\ на\ текущий\ период} \cdot 100\%$$

Доказательство формулы теперь очевидно.

Относительный показатель структуры K_C представляет собой отношение структурных частей изучаемого объекта и выражается в долях единицы или процентах.

$$K_C = \frac{\textit{Показатель, характеризующий часть совокупности}}{\textit{Показатель, характеризующий совокупность в целом}}.$$

Пример использования относительного показателя структуры приведен на следующем слайде.

Относительный показатель координации K_K представляет собой отношение показателей, характеризующих одну из частей совокупности, к показателю, характеризующему другую часть совокупности, выбранной в качестве базы сравнения.

Относительный показатель сравнения K_{CP} представляет собой отношение одного и того же абсолютного показателя, характеризующего разные объекты (предприятия А и В, например).

Относительный показатель интенсивности $К_И$ характеризует степень распространенности изучаемого процесса или явления и определяется отношением показателя абсолютной величины изучаемого явления к показателю, характеризующему среду распространения явления.

Относительный показатель интенсивности используется в тех случаях, когда абсолютный показатель не является достаточно наглядным.

Примером относительного показателя интенсивности может служить плотность населения, которая вычисляется как число людей, проживающих на 1 км^2 территории.

Пример использования относительного показателя структуры

	Трлн. Руб.	% к итогу
А	1	2
Внешнеторговый оборот -всего	896,7	100,0
В том числе:		
экспорт	505,6	56,4
импорт	391,1	43,6

Пример использования относительного показателя координации

Возвращаясь к таблице, характеризующей структуру внешнеторгового оборота, выберем в качестве базы сравнения структурную часть, которая имеет наибольший вес- показатель, характеризующий экспорт (506, 6 Трлн. руб.). (За базовый можно выбрать и показатель , приоритетный с экономической или социальной точек зрения)

Тогда относительный показатель координации позволяет найти, что на каждый рубль экспортируемой продукция приходится $K_K = 391,1 / 506,5 = 0,77$ рублей импортируемой продукции.

Пример использования относительного показателя сравнения

Распределение начисленной заработной платы в октябре 2001 г по отраслям хозяйства РФ по данным ГОСКОМСТАТа

Отрасль хозяйства РФ	Средняя заработная плата, руб	K_{CP}
Средняя по стране	3515	1
Нефтедобывающая	14535	4,14
Электроэнергетика	6041	1,72
Образование	1862	0,53
Финансы, кредит	8603	2,45
сельское хозяйство	1481	0,42
Машиностроение	3528	1,01

Пример использования относительного показателя интенсивности

Воспроизводство населения Рф (ГОСКОМСТАТ 2001 г)

	Январь - октябрь			
	Тысяч		На 1000 населения	
	2001	2000	2001	2000
Родившихся	1100,5	1056,6	9,2	8,7
Умерших	1871,4	1851,1	15,6	15,3
Браков	838,8	751,8	7,0	6,2
Разводов	624,0	510,4	5,2	4,2

Задача

- Торговая фирма планировала в 1997 г. по сравнению с 1996 г. увеличить оборот на 14, 5%. Выполнение установленного плана составило 102,7%. Определите относительный показатель динамики оборота.

Решение

Из условия задачи следует: $K_П = 1{,}145$;
Действительно, из определения

$$K_П = \frac{\text{Уровень, планируемый на 1997 г}}{\text{Уровень за 1996 г}} = 1{,}145 .$$

Продолжение

$$K_{ВП} = \frac{\textit{Фактический уровень за 1997 г}}{\textit{Уровень, планируемый на 1997 г}} = 1{,}027.$$

Поскольку $K_Д = K_{ВП} * K_П,$

получаем

$K_Д$ = 1,145*1,027 *100 % = 117,6 %.

2 Статистические показатели, используемые для характеристики рядов распределений. Виды средних.

Статистические показатели вариационного ряда

1. Среднее значение, мода, медиана
- характеризуют наиболее типичные значения признака

2. Среднеквадратичное отклонение, среднее линейное отклонение, размах вариации
- характеризуют разброс значений признака в статистической совокупности

Определение средней как статистического показателя

- Средняя величина – это обобщающий статистический показатель, отражающий типичный уровень изучаемого признака в расчете на единицу совокупности в конкретных условиях места и времени.

Сущность средней

- Сущность средней как статистического показателя состоит в том, в средней погашаются случайные отклонения признака элементов совокупности и лучше выявляется действие основных факторов.
- Среднее значение как статистический показатель можно использовать только для однородной совокупности.

Средняя арифметическая простая

При значении m=1 приведенная выше формула дает среднюю арифметическую простую величину, которая равна сумме отдельных значений осредняемого признака, деленной на общее число этих значений:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Эта формула применяется в тех случаях, когда имеются несгруппированные значения признака.

Средняя арифметическая взвешенная

Средняя арифметическая взвешенная - средняя сгруппированных величин x_1 , x_2 ,..., x_n вычисляется по формуле

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i},$$

Пример 1. Расчет средней статистической взвешенной

- Данные о количестве произведенной продукции рабочими цеха представлены в таблице

Производство продукции одним рабочим за смену, шт. (x_i)	8	9	10	11	12
Число рабочих, чел. (f_i)	7	10	15	12	6

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{8 \cdot 7 + 9 \cdot 10 + 10 \cdot 15 + 11 \cdot 12 + 12 \cdot 6}{7 + 10 + 15 + 12 + 6} = 10.$$

Пример 2. Расчет средней для интервального вариационного ряда

- Состав работников предприятия по стажу работы характеризуется следующими данными

Стаж работы, лет	До 5	5-10	10-15	15-20	20-25	25 и более	Всего
Число работн., чел	8	24	43	23	11	6	115

- Вычислить средний стаж работы.

Пример 2. Расчетная таблица

Интервалы	Частоты f_i	Середина интервалов x_i	$x_i f_i$
0 – 5	8	2,5	20
5 – 10	24	7,5	180
10 – 15	43	12,5	537,5
15 – 20	23	17,5	402,5
20 – 25	11	22,5	247,5
25 – 30	6	27,5	165
Всего	115	–	1552,5

Средняя гармоническая простая

Средняя гармоническая – это средняя обратных величин, что соответствует значению m = – 1 в общей формуле (см. слайд 15) .

$$\overline{X} = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{x_i}}.$$

Пример 1. Средняя урожайность

Имеются данные о валовом сборе урожая по трем районам области

Район	Валовый сбор, тыс. ц. Φ_i	Урожайность, ц /га x_i	Частоты $f_i = \Phi_i / x_i$
А	189,3	372	0,509
Б	207,8	468	0,444
В	159,6	401	0,398

$$\bar{x} = \frac{\Phi_1 + \Phi_2 + \Phi_3}{\frac{\Phi_1}{x_1} + \frac{\Phi_2}{x_2} + \frac{\Phi_3}{x_3}} = \frac{556,7}{0,509 + 0,444 + 0,398} = 412,1\ ц/га.$$

Средняя геометрическая

- Средняя геометрическая вычисляется по формуле

$$\overline{X}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

Средняя хронологическая

• Средняя хронологическая используется если вариационный признак изменяется во времени. Пусть в течение времени t_1 вариационный признак принимает значение x_1, в течение времени t_2 – x_2, в течение времени t_n – x_n. *Тогда среднее значение этого признака вычисляется по формуле*

$$\overline{x}_{\text{врем}} = \frac{x_1 t_1 + x_2 t_2 + \ldots + x_n t_n}{t_1 + t_2 + \ldots + t_n}.$$

Задача на вычисление средней хронологической

- Число работников предприятия по списку на 1 апреля составило 300 человек. 15 апреля зачислено 10 новых сотрудников, а 26 апреля двое уволилось. Определить среднее число сотрудников в апреле месяце.

Решение

Для решения задачи составим таблицу, представленную на след. слайде

Вычисление средней хронологической

Календарный период апреля	Число работников, чел. X_i	Продолжительность периода, дней t_i	Число человеко-дней $X_i t_i$
1 – 14	300	14	4200
15 – 25	310	11	3410
26 – 30	308	5	1540
Всего	—	30	9150

Вычисление средней хронологической. Ответ

$$\overline{x} = \frac{\sum x_i t_i}{\sum t_i} = \frac{9150}{30} = 305 \text{ чел.}$$

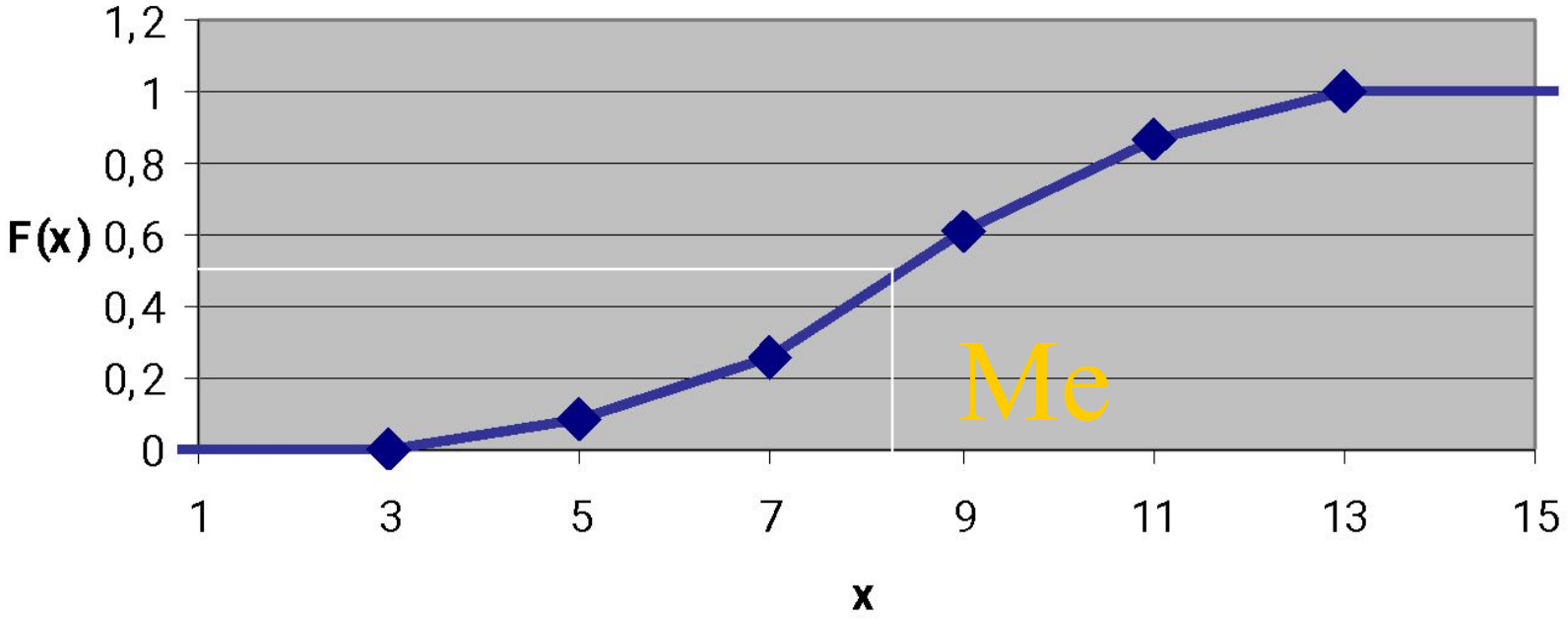
3 Медиана и мода

Медиана распределения - значение признака, которое приходится на середину ранжированной статистической совокупности

Признаку, определяющий медиану дискретного ряда (медианному интервалу непрерывного ряда)
соответствует первое значение накопленной доли, превышающее 0.5

Для интервальных рядов медиана вычисляется по специальной формуле

Функция распределения интервального ряда



$$M_e = x_0 + i \times \frac{\frac{1}{2}\sum f_i - s_{M_{e-1}}}{f_{M_e}}$$

где Xo - нижняя граница медианного интервала (медианным называется первый интервал, накопленная частота которого превышает половину общей суммы частот);

i - величина медианного интервала:

S_{me-1} - накопленная частота интервала, предшествующего медианному;

f_{Me} - частота медианного интервала.

Дискретный ряд

x_i f_i

Распределение жилого фонда городского района по типу квартир

Группы квартир по числу комнат	Число квартир, тыс. ед.
1	10
2	35
3	30
4	15
5	5
Всего	95

Середина совокупности приходится на 48 по счету квартиру (95/2=47.5). В этой квартире 3 комнаты. Медиана равна 3

Интервальный ряд

Распределение семей по размеру жилой площади, приходящейся на одного человека

Группы семей по размеру жилой площади, приходящейся на одного чел., м. кв.	Число семей, тыс.	Накопленные частоты, тыс.
3 – 5	10	10
5 - 7	20	30
7 – 9	40	70
9 – 11	30	100
11 – 13	15	115
Всего	115	—

Середина совокупности приходится на 57500-ю семью (115/2=57.5).

Медианный интервал (на котором накопленная частота впервые превышает 115/2) - интервал (7-9).

$Me=7+(57.5-30)/40 \cdot 2$

Модой распределения (Мо) называется значение признака с наибольшей частотой. В интервальном вариационном ряду модальным является интервал с наибольшей частотой.

$$Mo = x_0 + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \cdot h_{Mo}$$

где f_{Mo}, f_{Mo-1}, f_{Mo+1} - частоты в модальном, предыдущем и следующем интервалах, h_{Mo} - величина модального интервала. Из приведенной формулы видно, что если частоты f_{Mo-1}, f_{Mo+1} одинаковы, то модальный признак находится точно посередине модального интервала.

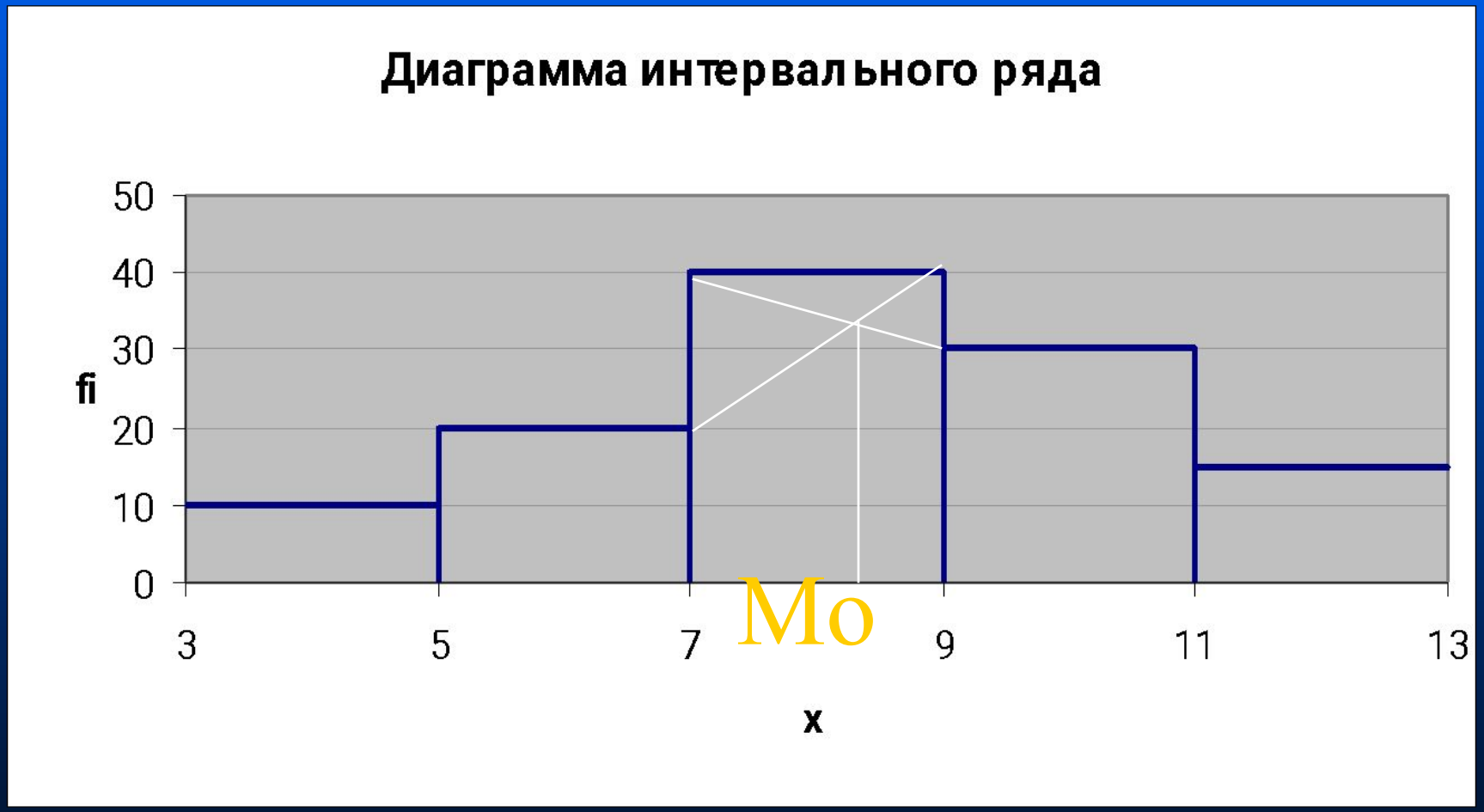


Диаграмма интервального ряда
50
40
30
20
10
0
fi
3
5
7
9
11
13
x
Mo

Распределение семей по размеру жилой площади, приходящейся на одного человека

Группы семей по размеру жилой площади, приходящейся на одного чел., м. кв.	Число семей, тыс.	Накопленные частоты, тыс.
3 – 5	10	10
5 - 7	20	30
7 – 9	40	70
9 – 11	30	100
11 – 13	15	115
Всего	115	—

Модальным является интервал (7-9)

$Mo = 7+(40-20)/(40-20+40-30) \cdot 2$

5.4. Показатели вариации

Размах вариации

Размах вариации $R = x_{max} - x_{min}$ показывает, насколько велико различие между максимальным и минимальным значением признака.

Поскольку размах вариации исчисляется только с использованием крайних значений совокупности, то он может содержать большие ошибки (из-за влияния случайных факторов крайние точки могут вообще оказаться выбросами)

Среднее линейное отклонение

Важной структурной характеристикой вариационного ряда является среднее линейное отклонение, которое вычисляется по формулам

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n}; \quad \bar{d} = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i}$$

в зависимости от формы представления вариационного ряда. В первой из этих формул суммирование производится по всем членам вариационного ряда, а во второй - по всем группам.

Дисперсия

Дисперсия характеризует степень рассеяния индивидуальных значений признака в совокупности от среднего значения и вычисляется по формулам

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

Записанное выражение называется формулой простой дисперсии. Ряд предполагается не сгруппированным и суммирование идет по всем членам ряда совокупности.

Взвешенная дисперсия

В этом случае (взвешенная дисперсия) вариационный ряд предполагается сгруппированным и суммирование ведется по всем группам. f_i - частота повторения признака в i - й группе.

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

Среднее квадратическое отклонение

Среднее квадратическое отклонение

$$\sigma = \sqrt{\sigma^2}$$

представляет собой характеристику вариационного ряда, которая отражает рассеянность членов совокупности относительно среднего значения. Чем меньше среднее квадратическое отклонение, тем лучше среднее значение характеризует всю совокупность.

Другие показатели вариации

Коэффициент осцилляции V_R

$$V_R = \frac{R}{\bar{x}} \cdot 100\ \%;$$

Линейный коэффициент вариации V_d

$$V_d = \frac{\bar{d}}{\bar{x}} \cdot 100\ \%;$$

Коэффициент вариации V_σ

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

Пример вычисления показателей вариации

Рассмотрим вычисление среднего линейного отклонения, дисперсии и среднеквадратичного отклонения для интервального ряда распределения промышленных предприятий одного из районов города по вооруженности работников промышленно – производственными основными фондами (ППОФ) представленного в табл. 24 (см. следующий слайд)

Табл.

Группы фирм по ППОФ на одного работника, тыс. руб. X_i	Число фирм в % к итогу f_i	Середина интервала X_i'
До 1	7,8	0,5
1,0 – 2,0	12,2	1,5
2,0 – 3,0	14,9	2,5
3,0 – 5,0	23,3	4
5,0 – 10,0	24,3	7,5
10,0 – 20,0	10,6	15,0
20 и более	6,9	25
Всего	100	-

Вычисление дисперсии в случае интервального ряда

В случае интервального ряда в качестве значения вариационного признака x_i берутся середины интервалов

Схема вычисления среднего линейного отклонения

f	x	xf	|(x - средн.Зн)|*f
7,80	0,50	3,90	48,08
12,20	1,50	18,30	63,00
14,90	2,50	37,25	62,04
23,30	4,00	93,20	62,07
24,30	7,50	182,25	20,31
10,60	15,00	159,00	88,36
6,90	25,00	172,50	126,52
			470,39
Средн. Зна	6,66	ср. лин откл.	4,7039

Схема вычисления дисперсии

f	x	(x - ср. знач.)^2*f
7,80	0,50	296,36
12,20	1,50	325,34
14,90	2,50	258,35
23,30	4,00	165,36
24,30	7,50	16,98
10,60	15,00	736,58
6,90	25,00	2319,84
ср. знач.=	6,664	4118,81

$$\sigma^2 = \frac{4118,81}{100} = 41,1881$$

6. Эмпирическое определение тесноты корреляционной связи. Правило сложения дисперсий.

Рассмотрим аналитическую группировку данных по двум признакам. По первому признаку (группировочный или факторный признак) мы разобьем статистическую совокупность на несколько групп, а затем исследуем в каждой группе характеристики второго признака (результативный признак). А именно, найдем для каждой группы среднее значение и дисперсию результативного признака. Для этих величин вводятся новые названия - *групповое среднее* и *групповая (внутригрупповая) дисперсия*.

№ группы	Объем группы n_i	Среднее для группы $\bar{x}_j$	Внутригруп-повая дисперсия σ_j^2
1	n_1	$\bar{x}_1$	σ_1^2
2	n_2	$\bar{x}_2$	σ_2^2
3	n_3	$\bar{x}_3$	σ_3^2
…	…	…	…

Внутригрупповой дисперсией *j* -ой группы называется обычная дисперсия, вычисленная для группы с номером *j* .

Внутригрупповая дисперсия вычисляется по формуле

$$\sigma_j^2 = \frac{\sum_i (x_{ij} - \bar{x}_j)^2 f_{ij}}{\sum_i f_{ij}},$$

где x_{ij} - значения вариант,

f_{ij} - частот,

$\bar{x}_j$ - среднее значение , а

$$n_j = \sum_i f_{ij}.$$

- объем для j -ой группы.

По имеющимся данным можно вычислить общее среднее:

$$\bar{x} = \frac{\sum_j \bar{x}_j n_j}{\sum_j n_j}$$

Межгрупповая дисперсия

Межгрупповой дисперсией называется дисперсия групповых средних, рассчитанная с учетом объема каждой группы n_j

$$\delta^2 = \frac{\sum_j (\bar{x}_j - \bar{x})^2 n_j}{\sum_j n_j}$$

Средняя из групповых дисперсий. Формула сложения дисперсий

В математической статистике показано, что между общей дисперсией, межгрупповой дисперсией и средней из групповых дисперсией, определяемой формулой

$$\overline{\sigma^2} = \sum_j \sigma_j^2 n_j / n, \quad n = \sum_j n_j.$$

существует простая связь, выражающая правило сложения дисперсий

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

Эмпирическое корреляционное отношение

- количественная характеристика тесноты связи факторного и результативного признаков - равно корню квадратному из отношения межгрупповой дисперсии к общей дисперсии

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}}$$

По величине эмпирического корреляционного отношения можно определить, насколько сильно связаны факторный и результативный признаки.

0-0.3 связь отсутствует
0.3-0.5 слабая
0.5-0.7 умеренная
0.7-1 сильная связь.

Пример решения задачи

Задача. По данным таблицы (см. след слайд) вычислить общую дисперсию, а также характеризовать степень влияния объема затрат туристических фирм на рекламу, на вариацию количества туристов, воспользовавшихся услугами этих фирм.

Таблица

Группы тур. Фирм по затратам на рекламу тыс. долл.	Число фирм в группе n_i	Среднее число туристов, восп. услугами фирм $\overline{x}_i$	Групповые дисперсии 2 σ_i
< 10	12	720	920
10 – 50	23	1850	1600
50 – 100	5	3630	2100
Всего	40	6200	—

Решение задачи

1. Вычисляем среднее значение

$$\bar{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{720 \cdot 12 + 1850 \cdot 23 + 3630 \cdot 5}{12 + 23 + 5} =$$

$$= 1733{,}5 \text{ чел.}$$

2. Найдем среднюю групповую дисперсию

$$\overline{\sigma^2} = \frac{920 \cdot 12 + 1600 \cdot 23 + 2100 \cdot 5}{12 + 23 + 5} = 1458{,}5.$$

3. Вычислим межгрупповую дисперсию

$$\delta^2 = \frac{1}{n}\sum(\bar{x}_i - \bar{x})^2 n_i = \frac{1}{40}[12\cdot(1733{,}5-720)^2 +$$

$$+23\cdot(1850-1733{,}5)^2 + 5\cdot(3630-1733{,}5)^2] =$$

$$= 766548$$

4. Общая дисперсия равна

$$\sigma^2 = \delta^2 + \overline{\sigma^2} = 766548 + 1458{,}5 = 768006{,}5.$$

Сделаем выводы

- Средняя из групповых дисперсий значительно меньше межгрупповой дисперсии. Это значит, что группы существенно отличаются одна от другой. Это в свою очередь означает, что затраты на рекламу существенно сказываются на число туристов, воспользовавшихся услугами данной фирмы . Формальным признаком этого является большое значение эмпирического корреляционного отношения η

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{\frac{766548}{768006,5}} = 0,999.$$

7. Альтернативный признак. Среднее значение и дисперсия. Эмпирическая оценка тесноты связи в случае альтернативного признака.

Рассмотрим вариационный ряд с двумя возможными значениями признака (*альтернативный признак*)

Пусть *p* - доля единиц совокупности, обладающих некоторым признаком, а q - доля единиц совокупности, не обладающих этим признаком. Тогда можно построить вариационный ряд для *альтернативного* признака *x*, принимающего всего два значения:

x_i	0	1
w_i	q	p

Вычисление среднего значения и дисперсии

Среднее значение и дисперсия такого ряда вычисляется по формулам:

$$\overline{x} = \frac{0 \cdot q + 1 \cdot p}{q + p} = p;$$

$$\sigma_p^2 = \frac{(0 - p)^2 \cdot q + (1 - p)^2 \cdot p}{p + q} = p \cdot q.$$

Внутригрупповая и межгрупповая дисперсии для альтернативного признака

Пусть имеется аналитическая группировка, включающая несколько групп, характеризуемых альтернативным признаком (с двумя возможными значениями варианты). Так же, как и в случае вариационного признака с большим количеством градаций, для этих групп можно ввести понятия внутригрупповой, межгрупповой, полной и средней из групповых дисперсий.

Внутригрупповая дисперсия и среднегрупповая дисперсии определяются по формулам:

$$\sigma_{pi}^2 = p_i \cdot q_i;$$

$$\overline{\sigma_p^2} = \frac{\sum_{i=1}^{k} p_i \cdot q_i \cdot n_i}{\sum_{i=1}^{k} n_i}, \quad i - \text{номергруппы.}$$

Формула межгрупповой дисперсии имеет вид (*p* - доля признака во всей совокупности, она же - общее среднее)

$$\delta_p^2 = \frac{\sum_{i=1}^{k}(p_i - p)^2 \cdot n_i}{\sum_{i=1}^{k} n_i}, \quad p = \frac{\sum_{i=1}^{k} p_i n_i}{\sum_{i=1}^{k} n_i}.$$

Общая дисперсия вычисляется по формуле

$$\sigma_p^2 = p(1-p).$$

Как и в случае рядов, построенных по количественному признаку, справедлива формула сложения дисперсий

$$\sigma_p^2 = \delta_p^2 + \overline{\sigma_p^2}.$$

Пример вычисления дисперсий доли

Данные об удельном весе рабочих основных специальностей в трех цехах предприятия представлены в таблице

Цех	Удельный вес рабочих основных спец. , % P_i	Численность всех рабочих n_i
1	80	100
2	75	200
3	90	150
Всего	—	450

Найдем среднюю долю основных рабочих

$$\overline{p} = \frac{0{,}8 \cdot 100 + 0{,}75 \cdot 200 + 0{,}9 \cdot 150}{100 + 200 + 150} = 0{,}81.$$

Вычислим общую дисперсию

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

Вычислим внутригрупповые дисперсии

$$\sigma_{p1}^2 = 0,8(1-0,8) = 0,154;$$

$$\sigma_{p2}^2 = 0,75(1-0,75) = 0,19;$$

$$\sigma_{p3}^2 = 0,9(1-0,9) = 0,09.$$

Средняя из групповых дисперсий

$$\overline{\sigma_p^2} = \frac{0,16 \cdot 100 + 0,19 \cdot 200 + 0.09 \cdot 150}{100 + 200 + 150} = 0,15$$

Найдем межгрупповую дисперсию

$$\delta^2 = \frac{1}{450} \cdot [(0{,}8 - 0{,}81)^2 \cdot 100 + (0{,}75 - 0{,}81)^2 \cdot 200 + (0{,}9 - 0{,}8)^2 \cdot 150] = 0{,}004$$

Проверяем вычисления, используя формулу сложения дисперсии

$$\sigma_p^2 = \delta^2 + \overline{\sigma_p^2}; \qquad 0{,}154 = 0{,}15 + 0{,}004.$$

Выводы

1. Межгрупповая дисперсия является малой. Она существенно меньше средней из внутригрупповых дисперсий. Это означает, что цеха различаются по числу основных рабочих незначимо.
2. Тот же самый вывод можно получить, вычислив эмпирический коэффициент корреляции

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{\frac{0{,}004}{0{,}154}} = 0{,}16$$