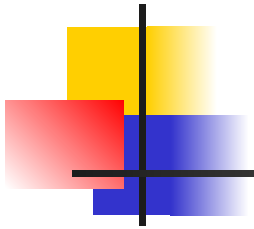


ОСНОВЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

Понятие выборочного наблюдения

Выборочное наблюдение — это такой вид статистического наблюдения, при котором обследованию подвергается не вся изучаемая совокупность, а лишь часть ее единиц, отобранных в определенном порядке.

При этом вся исследуемая совокупность называется ***генеральной***, а единицы, подлежащие наблюдению, составляют ***выборочную совокупность***, или ***выборку***.

Цель выборочного наблюдения

Цель выборочного наблюдения - определение параметров генеральной совокупности (генеральной средней — **x** и генеральной доли **p**) на основе параметров выборочной совокупности (выборочной средней $\tilde{x}$ и выборочной доли **ω**).

Разница между генеральными и выборочными параметрами называется **ошибкой выборки** или **ошибкой репрезентативности**.

Классификация выборок

Выборкой называют часть изделий, отобранных из общей их совокупности для получения информации о всей массе изделий, называемой ***общей*** или ***генеральной совокупностью***.

Генеральная совокупность подразумевает однородную совокупность параметров качества контролируемых изделий.

Если выборка достаточно хорошо представляет соответствующие характеристики генеральной совокупности, то такую выборку называют **представительной** или **репрезентативной**.

Статистический ряд и его характеристики

При проведении выборочного наблюдения возможны три способа отбора:

- **случайный**,
- **отбор** единиц **по** определенной **схеме**,
- сочетание первого и второго способов.

Классификация выборок

При анализе и контроле технологических процессов выборку классифицируют по ряду признаков:

1) **по способу образования**:
   - повторные;
   - бесповторные;
2) **по преднамеренности отбора**:
   - пристрастные;
   - случайные;
3) **по отношению ко времени образования**:
   - единовременные;
   - текущие;
4) **по целевому назначению**:
   - общепроизводственные;
   - одноагрегатные и т. д.

Классификация отборов

- **повторный** - соответствует схеме «возвращенного шара»: после отбора какой-либо единицы она возвращается в генеральную совокупность и снова может быть выбранной. Таким образом, вероятность попадания каждой отдельной единицы в выборку остается постоянной на всем протяжении отбора
- **бесповторный** - отобранная единица не возвращается в генеральную совокупность, и тем самым вероятность попасть в выборку для оставшихся единиц увеличивается с каждым шагом отбора.

Классификация выборок

Повторная выборка образуется путем извлечения изделий из генеральной совокупности с последующим возвращением в последнюю после измерения параметров качества. Такое извлечение и возвращение может быть проведено многократно.

При **бесповторной** выборке извлеченные изделия не возвращаются в генеральную совокупность, при этом дается гарантия, что ни одно изделие не попадет дважды в выборку.

Классификация выборок

Если при отборе изделий из генеральной совокупности одним отдается ***предпочтение*** по отношению к другим, например отбор изделий с заранее оговоренным признаком, то такую выборку называют **пристрастной**.

Случайная выборка образуется при отборе изделий из генеральной совокупности, если возможность попадания в выборку каждого из них ***равновероятна***. Например, изделие отбирается наугад из разных источников (разные поточные линии, разные единицы оборудования и т. д.).

Классификация выборок

Единовременная выборка образуется из партии изделий после их изготовления ***независимо*** от того, в какой момент времени изготовлено каждое из них.

В отличие от единовременной **текущая** выборка состоит из изделий, ***последовательно*** изготовленных за определенный промежуток времени.

Классификация выборок

Общепроизводственные выборки преследуют цель получения общей оценки технологического процесса независимо от того, сколько поточных линий, единиц оборудования и т. п. занято в производстве продукции.

Одноагрегатная выборка образуется из изделий, изготовленных на определенном оборудовании (агрегате).

Понятие статистического ряда

Значение параметров качества изделий выборки представляет собой ***первичный статистический материал***, подлежащий обработке, осмыслению и научному анализу.

Такая однородная совокупность называется «***простой статистической совокупностью***» или «***простым статистическим рядом***». Обычно простая статистическая совокупность оформляется в виде таблицы с одним входом, в первом столбце которой стоит номер опыта, а во втором— замеренное значение параметра.

Понятие статистического ряда

Если расположить замеренные значения параметра в ***возрастающем*** или ***убывающем*** порядке, то получится так называемый **упорядоченный (ранжированный) ряд** или упорядоченное распределение различных значений одного и того же параметра качества.

Для группирования одинаковых значений параметра статистический материал должен быть подвергнут дополнительной обработке—строится так называемый «**статистический ряд**».

Статистический ряд и его характеристики

Ошибки выборки:

- **средняя (стандартная);**
- **предельная;**
- **относительная.**

Ошибки выборки:

средняя ошибка выборки для средней величины ($\mu_{\bar{x}}$) (при случайном и механическом отборах)

- при повторном отборе:

$$\mu_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}},$$

- при бесповторном отборе:

$$\mu_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$$

σ^2 — дисперсия признака в генеральной совокупности;

n — численность выборки;

N— численность генеральной совокупности

Ошибки выборки:

Соотношение дисперсии признака в генеральной совокупности $\boldsymbol{\sigma^2}$ и выборочной дисперсии $\boldsymbol{S^2}$

$$\sigma^2 = S^2 \frac{n}{n-1}$$

$$n \to \infty \quad \frac{n}{n-1} \to 1 \quad \Rightarrow \quad \sigma^2 \to S^2$$

Ошибки выборки:

Величина **дисперсии доли** в генеральной совокупности

$$\sigma^2_{доли} = p(1-p)$$

p — доля единиц, обладающих каким-либо значением признака в генеральной совокупности

Ошибки выборки:

Дисперсия доли в генеральной совокупности заменяется дисперсией доли в выборочной совокупности

$$S^2_{доли} = \omega(1-\omega)$$

ω — доля единиц, обладающих каким-либо значением признака в выборочной совокупности

Ошибки выборки:

Средняя ошибка выборочной доли для

- повторного отбора

$$\mu_\omega = \sqrt{\frac{\omega(1-\omega)}{n}}$$

- бесповторного отбора

$$\mu_\omega = \sqrt{\frac{\omega(1-\omega)}{n} \cdot \left(1 - \frac{n}{N}\right)}$$

Ошибки выборки:

Предельная ошибка выборки

$$\Delta = t \cdot \mu$$

t — коэффициент доверия, который определяется по таблице значений интегральной функции Лапласа при заданной доверительной вероятности

Наиболее часто употребляемые уровни доверительной вероятности и соответствующие им значения t

p(t)	**0.683**	**0.950**	**0.954**	**0,990**	**0,997**
t	**1,00**	**1.96**	**2,00**	**2,58**	**3.00**

Ошибки выборки:

Зная величину выборочной средней ($\tilde{x}$) или доли (ω), а также предельную ошибку выборки (Δ), можно определить доверительные интервалы, в которых находятся значения генеральных параметров

$$\tilde{x} - \Delta \le \bar{x} \le \tilde{x} + \Delta$$

$$\omega - \Delta \le P \le \omega - \Delta$$

Задача:

Для определения среднего срока пользования краткосрочным кредитом в банке была произведена 5%-я механическая выборка, в которую попали 200 счетов.

По результатам выборки установлено, что средний срок пользования кредитом составляет 60 дней при среднеквадратическом отклонении 20 дней.

В 8 счетах срок пользования кредитом превышал 6 месяцев.

Необходимо с вероятностью 0,99 определить пределы, в которых находятся срок пользования краткосрочными кредитами банка и доля краткосрочных кредитов со сроком пользования более полугода.

Определение объема выборки:

Расчет объема выборки для повторного отбора

$$n_{повт} = \frac{t^2 \sigma^2}{\Delta^2}$$

Если полученный объем выборки превышает 5% численности генеральной совокупности, расчеты корректируют «на бесповторность»

$$n_{бесп} = \frac{t^2 \sigma^2 N}{t^2 \sigma^2 + \Delta^2 N}$$

Для оценки величины генеральной дисперсии можно использовать:

1) выборочную дисперсию по данным прошлых или пробных обследований;
2) дисперсию, найденную из соотношения для среднего квадратического отклонения:

$$\sigma = \frac{1}{3}\bar{x}$$

$\bar{x}$ — среднее значение признака в генеральной совокупности;

3) дисперсию, определенную из соотношения для асимметричного распределения

$$\sigma = \frac{1}{5}(x_{max} - x_{min})$$

x_{max}, x_{min} — соответственно максимальное и минимальное значения признака в генеральной совокупности

4) дисперсию, вычисленную из соотношения для нормального распределения

$$\sigma = \frac{1}{6}(x_{max} - x_{min})$$

Относительная ошибка выборки - отношение предельной ошибки выборки к среднему значению признака, характеризует относительную погрешность выборочного наблюдения

$$\Delta_{относ} = \frac{\Delta}{\bar{x}} \cdot 100\%$$

Тогда объем выборки:

$$n_{повт} = \frac{t^2 \nu^2}{\Delta^2_{относ}} \qquad n_{бесповт} = \frac{t^2 \nu^2 N}{t^2 \nu^2 + \Delta^2_{относ} N}$$

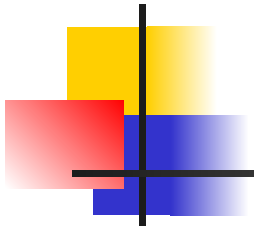
ν - коэффициент вариации $\nu = \frac{\sigma}{\bar{x}} \cdot 100\%$

Задача:

В городе зарегистрировано 30 тыс. безработных. Для определения средней продолжительности безработицы организуется выборочное обследование.

По данным прошлых лет известно, что коэффициент вариации продолжительности безработицы составляет 40%.

Какое число безработных необходимо охватить выборочным наблюдением, чтобы с вероятностью 0,997 утверждать, что полученная предельная ошибка выборки не превышает 5% средней продолжительности безработицы?



СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ

Статистическая гипотеза

Статистической гипотезой называется любое предположение о **виде** *неизвестного* **закона распределения** или о **параметрах** *известных* распределений.

Предположим, что на основании имеющихся данных есть основания выдвинуть предположения о законе распределения или о параметре закона распределения случайной величины (или генеральной совокупности, на множестве объектов которой определена эта случайная величина).

Задача проверки статистической гипотезы заключается в **подтверждении** или **опровержении** этого предположения на основании выборочных (экспериментальных) данных.

Статистическая гипотеза

Проверка статистической гипотезы означает проверку *соответствия выборочных данных* выдвинутой *гипотезе*.

Параллельно с выдвигаемой **основной** гипотезой, рассматривают и противоречащую ей гипотезу, которая называется **конкурирующей** или **альтернативной**.

Альтернативная гипотеза считается **справедливой**, если **основная** выдвинутая **гипотеза отвергается**.

Статистическая гипотеза

Параметрической гипотезой называется гипотеза о значениях **параметров распределения** или о сравнительной величине параметров двух распределений.

*Примером **параметрической** статистической гипотезы является гипотеза о равенстве математических ожиданий двух нормальных совокупностей.*

Непараметрическими гипотезами называются гипотезы о **виде распределения** случайной величины.

Статистическая гипотеза

Нулевой, основной или проверяемой гипотезой называется *первоначально выдвинутая* гипотеза, которая обозначается $\mathbf{H_0}$.

Конкурирующей или **альтернативной** гипотезой называется гипотеза, которая *противоречит основной гипотезе* $\mathbf{H_0}$ и обозначается $\mathbf{H_1}$.

Например, ***основная гипотеза*** $\boldsymbol{H_0}$ *состоит в том, что математическое ожидание* μ *равно какому-то значению* μ_0. *В этом случае конкурирующая гипотеза* $\boldsymbol{H_1}$ *может состоять в предположении, что математическое ожидание* μ *не равно (больше или меньше) значения* μ_0:

$H_0: \mu=\mu_0$; $H_1: \mu\neq\mu_0$, или $H_1: \mu>\mu_0$, $H_1: \mu<\mu_0$.

Статистическая гипотеза

При проверке статистических гипотез существует **вероятность допустить ошибку**, приняв или опровергнув верную гипотезу.

Уровнем значимости (α) называется *вероятность совершения ошибки первого рода*.

Значение уровня значимости а обычно задается близким к нулю (например, 0,05; 0,01; 0,02 и т. д.) - чем меньше значение уровня значимости, тем меньше вероятность совершить **ошибку первого рода**, состоящую в **опровержении** верной гипотезы $\mathbf{H_0}$.

Статистические критерии

Проверка справедливости статистических гипотез осуществляется с помощью различных **статистических критериев**.

В статистике чаще всего пользуются тремя уровнями значимости:

α=0,10, тогда P=0,90 (в 10 случаях из 100)

α=0,05, тогда P=0,95 (в 5 случаях из 100)

α=0,01, тогда P=0,99 (в 1 случае из 100) может быть отвергнута правильная гипотеза

Статистические критерии

Статистическим критерием называется **случайная величина**, которая используется с целью проверки **нулевой** гипотезы.

Статистические критерии называются соответственно по тому закону распределения, которому они подчиняются:

- **F-критерий** подчиняется распределению Фишера-Снедекора,
- **χ^2-критерий** подчиняется χ^2-распределению,
- **t-критерий** подчиняется распределению Стьюдента,
- **U-критерий** подчиняется нормальному распределению.

Статистические критерии

Областью принятия гипотезы или **областью допустимых значений** называется *множество возможных значений* статистического *критерия*, при которых **основная** гипотеза **принимается**.

Если наблюдаемое значение статистического критерия, рассчитанное по данным выборочной совокупности, **принадлежит критической области**, то **основная** гипотеза **отвергается**.

Если наблюдаемое значение статистического критерия **принадлежит области принятия гипотезы**, то **основная** гипотеза **принимается**.

Теоретические и эмпирические частоты

При анализе вариационных рядов распределения большое значение имеет, насколько **эмпирическое распределение** признака соответствует **нормальному**.

Для этого частоты фактического распределения нужно сравнить с теоретическими, которые характерны для нормального распределения. Значит, нужно по фактическим данным вычислить теоретические частоты кривой нормального распределения, являющиеся функцией нормированных отклонений.

Иначе говоря, **эмпирическую кривую** распределения нужно **выровнять кривой нормального распределения**.

Критерии нормальности

Объективная характеристика соответствия теоретических и эмпирических частот может быть получена при помощи специальных статистических показателей, которые называют **критериями согласия**.

Критерием согласия называют критерий, который позволяет установить, *является ли расхождение эмпирического и теоретического распределений* **случайным** или **значимым**, т. е. согласуются ли данные наблюдений с выдвинутой статистической гипотезой или не согласуются.

Распределение генеральной совокупности, которое она имеет в силу выдвинутой гипотезы, называют **теоретическим**.

Критерии нормальности

Обычно эмпирические и теоретические частоты различаются в силу того, что:

- расхождение **случайно** и связано с ограниченным количеством наблюдений;
- расхождение **неслучайно** и объясняется тем, что *статистическая гипотеза* о том, что генеральная совокупность распределена нормально — *ошибочна*.

Таким образом, **критерии согласия** позволяют **отвергнуть** или **подтвердить** правильность выдвинутой при выравнивании ряда *гипотезы о характере распределения* в эмпирическом ряду.

Эмпирические частоты получают в результате **наблюдения**.

Теоретические частоты рассчитывают **по формулам**.

Критерий согласия Пирсона χ^2

Критерий согласия Пирсона χ^2 – один из основных, который можно представить как сумму отношений квадратов расхождений между теоретическими (fT) и эмпирическими (f) частотами к теоретическим частотам:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_t)^2}{f_t}$$

k –число групп, на которые разбито эмпирическое распределение,

fi –наблюдаемая частота признака в i-й группе,

f_t –теоретическая частота.

Критерий согласия Пирсона χ^2

Для распределения χ^2 составлены таблицы, где указано **критическое значение** критерия согласия χ^2 для выбранного уровня значимости α и степеней свободы df (или ν). Уровень значимости α – вероятность ошибочного отклонения выдвинутой гипотезы, т.е. вероятность того, что будет отвергнута правильная гипотеза. P — статистическая достоверность принятия верной гипотезы.

Число степеней свободы df определяется как число групп в ряду распределения минус число связей: $df = m - n$. Под числом связей понимается число показателей эмпирического ряда, использованных при вычислении теоретических частот, т.е. показателей, связывающих эмпирические и теоретические частоты. Например, при выравнивании по кривой нормального распределения имеется три связи. Поэтому при выравнивании по кривой нормального распределения число степеней свободы определяется как $df = k - 3$.

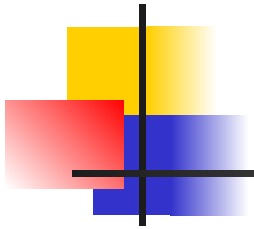
Критерий согласия Пирсона χ^2

Для оценки существенности, **расчетное** значение сравнивается с **табличным** $\chi^2_{табл}$

При полном совпадении теоретического и эмпирического распределений $\chi^2=0$, в противном случае $\chi^2>0$. Если $\chi^2_{расч} > \chi^2_{табл}$, то при заданном уровне значимости и числе степеней свободы **гипотезу о несущественности** (случайности) расхождений **отклоняем**.

В случае, если $\chi^2_{расч} < \chi^2_{табл}$ то **гипотезу принимаем** и *с вероятностью P=(1-α) можно утверждать, что расхождение между теоретическими и эмпирическими частотами* ***случайно***. Следовательно, есть основания утверждать, что **эмпирическое распределение подчиняется нормальному** распределению.

Критерий согласия Пирсона используется, если объем совокупности достаточно велик (N>50), при этом, частота каждой группы должна быть не менее 5



ИЗУЧЕНИЕ ВАРИАЦИИ

Ряды распределения

Статистические ряды распределения представляют собой упорядоченное распределение единиц совокупности по группам и группировкам.

Ряды распределения изучают **структуру** совокупности, позволяют изучить ее **однородность**, **размах** и **границы**.

Ряды распределения, образованные по **качественным** признакам, называют **атрибутивными**.

При группировке по **количественному** признаку выделяются **вариационные** ряды.

Вариационные ряды – ряды распределения единиц совокупности по признакам, имеющим **количественное** выражение, т. е. образованы **численными** значениями.

Вариационные ряды

Дискретные (прерывные) – основаны на прерывных вариациях признака. Это такие ряды, где значения **вариант** имеют **значения целых чисел** (т. е. не могут принимать дробные значения).

Дискретные признаки отличаются друг от друга на некоторую конкретную величину.

Интервальные (непрерывные) – имеют любые, в том числе и дробные количественные выражения и представлены в виде **интервалов**.

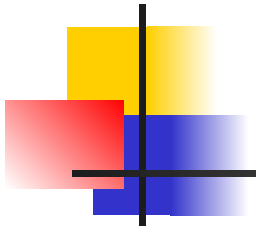
Непрерывные признаки могут отличаться один от другого на сколь угодно малую величину.

Изучение вариации

Вариацией признака называют отличие в численных значениях признаков единиц совокупности и их колебания около средней величины.

Чем меньше вариация, тем более однородна совокупность и более надежна (типична) средняя величина.

Этапы анализа вариации



1. Построение вариационного ряда
2. Графическое изображение вариационного ряда
3. Расчет показателей центра распределения и структурных характеристик вариационного ряда
4. Расчет показателей размера и интенсивности вариации
5. Оценка вариационного ряда на асимметрию и эксцесс

Этапы анализа вариации

1. Построение вариационного ряда

Построение вариационного ряда (ряда распределения) – это упорядоченное распределение единиц совокупности по возрастающим или убывающим значениям признака и подсчета числа единиц с тем или иным значением.

Статистический ряд случайных величин

- значения параметра в ранжированном порядке и соответствующие частоты

X	X1	X2	X3	...	Xi	. . .	Xk
m	***m1***	***m2***	***m3***	. . .	***mi***	. . .	***mk***

X- обозначение i-го интервала;
mi—соответствующая частота;
k—число интервалов

Этапы анализа вариации

1. Построение вариационного ряда

Исходные данные

Среднедушевой денежный доход в среднем за месяц, тыс. руб.	Число жителей		Накопленные частости (S)	Середина интервала (x)	xf	xw
	чел. (f)	в % к итогу (w)				
До 0,5	26	0,9	0,9	0,25	6,5	0,2
0,5-1,0	463	16,5	17,4	0,75	347,25	12,4
1,0-1,5	690	24,6	42,0	1,25	862,5	30,7
1,5-2,0	528	18,8	60,8	1,75	924	32,9
2,0-2,5	434	15,5	76,2	2,25	976,5	34,8
2,5-3,0	350	12,5	88,7	2,75	962,5	34,3
3,0 и более	318	11,3	100,0	3,25	1033,5	36,8
Итого	**2809**	**100**			**5112,75**	**182,0**

Этапы анализа вариации

1. Построение вариационного ряда

В составе любого вариационного ряда можно выделить три основных элемента:

- **варианты** – это значения, которые принимает исследуемый признак совокупности; если варианты представлены в виде целочисленных величин, вариационный ряд называют дискретным, а если в виде интервалов - интервальным,
- **частоты** вариационного ряда – абсолютная численность отдельных групп с различными значениями признака,
- **частости** вариационного ряда – удельные веса (доли) отдельных групп в общей численности совокупности.

Этапы анализа вариации

2. Графическое изображение вариационного ряда

Графическое изображение вариационного ряда облегчает его анализ и позволяет судить о форме распределения.

Способы графического представления вариационного ряда:

- **гистограмма;**
- **полигон частот;**
- **кумулята распределения**

Этапы анализа вариации

2. Графическое изображение вариационного ряда

Гистограмма – столбиковая диаграмма, для построения которой на оси абсцисс откладывают отрезки, равные величине интервалов вариационного ряда. На отрезках строят прямоугольники, высота которых в принятом масштабе по оси ординат соответствует частотам (или частостям).

Гистограмма

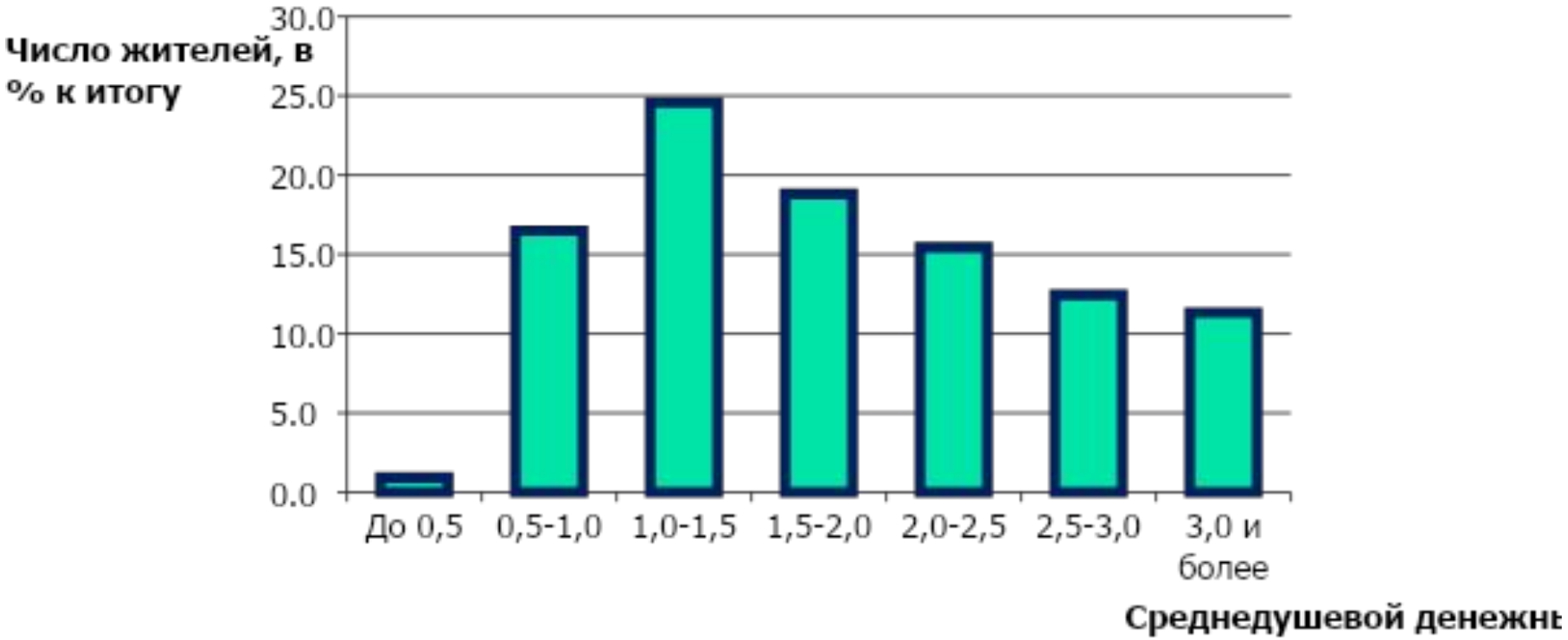
служит для графического анализа распределения

Алгоритм построения гистограммы:

1. Определяются минимальное min(X) и максимальное max(X) значения выборки (наибольшее и наименьшее значения).
2. Определяется размах выборки как разность между наибольшим и наименьшим значениями указанного показателя D = max(X)- min(X),
3. Рассчитывается число интервалов гистограммы: К=1,5+3,3 lg n, где К - число интервалов; n - число значений случайной величины).
4. Определяется ширина интервала h гистограммы путем деления диапазона гистограммы на число интервалов h = D/K.
   В случае, когда ширина интервала не превышает двукратной цены деления измерительного средства, необходимо уменьшить число интервалов К, чтобы не получить полигон частот вместо гистограммы распределения.
5. Диапазон гистограммы разбивается на интервалы. Подсчитывается число попаданий результатов в каждый j-тый интервал p(hj).
6. Определяется частота попаданий в интервал wj путем деления числа попаданий на размер выборки. wj = p(hj)/n.
7. Строится столбчатая диаграмма.

2. Графическое изображение вариационного ряда

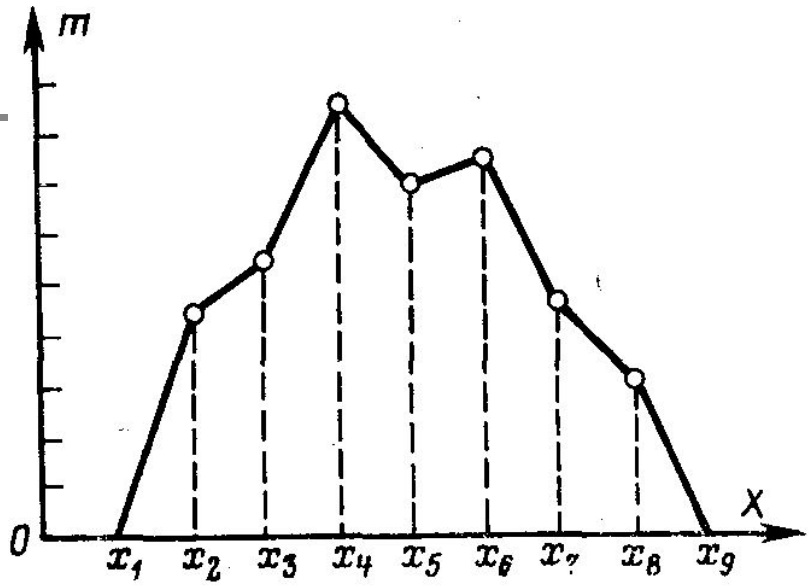
Гистограмма



Гистограмма

2. Графическое изображение вариационного ряда

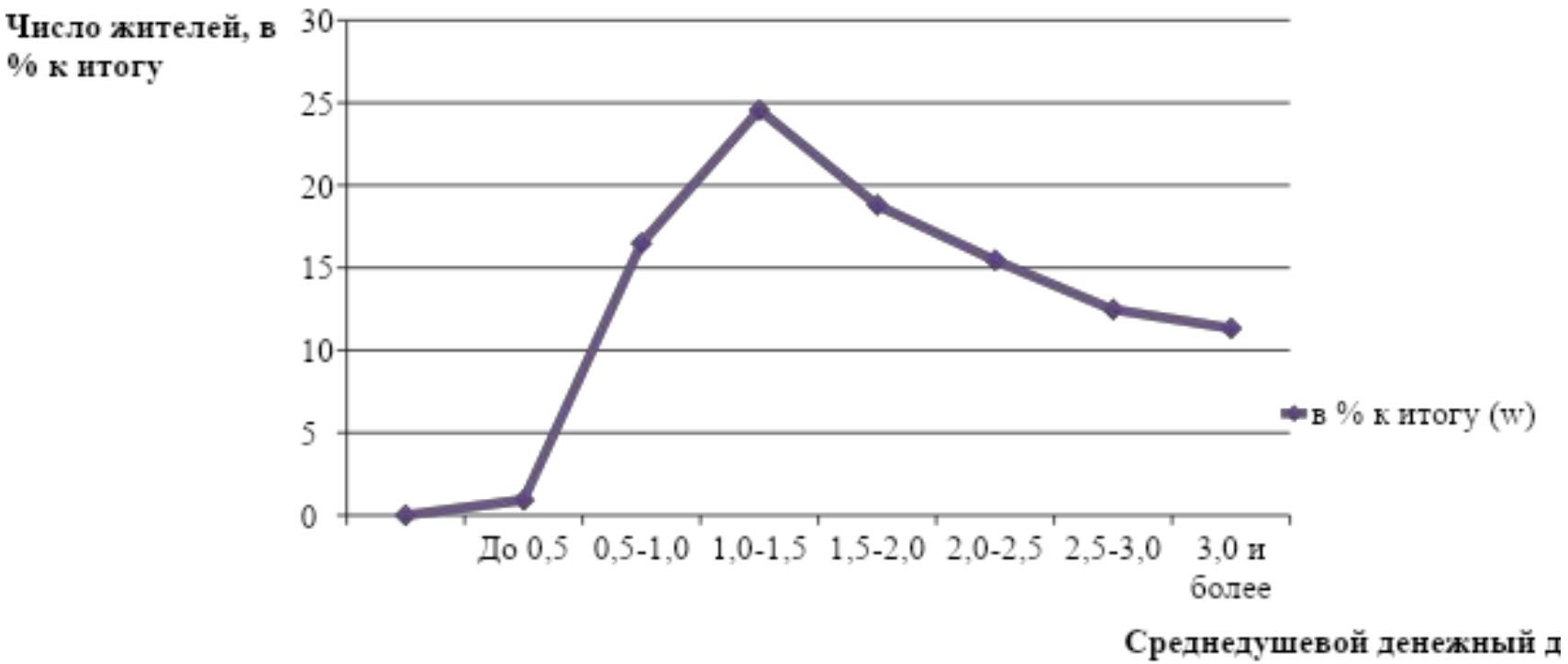
Полигон частот



- полигон распределения строится в прямоугольной системе координат;
- по оси абсцисс откладываются значения параметра, а по оси ординат—соответствующие им частоты;
- вершины ординат соединяются прямыми линиями.

2. Графическое изображение вариационного ряда

Полигон частот



2. Графическое изображение вариационного ряда

Кумулята

- Кумулята распределения строится по накопленным частотам (частостям). Накопленные частоты (частости) определяют последовательным суммированием частот (частостей). Они показывают, сколько единиц совокупности имеют значение признака не больше, чем рассматриваемое значение

Этапы анализа вариации

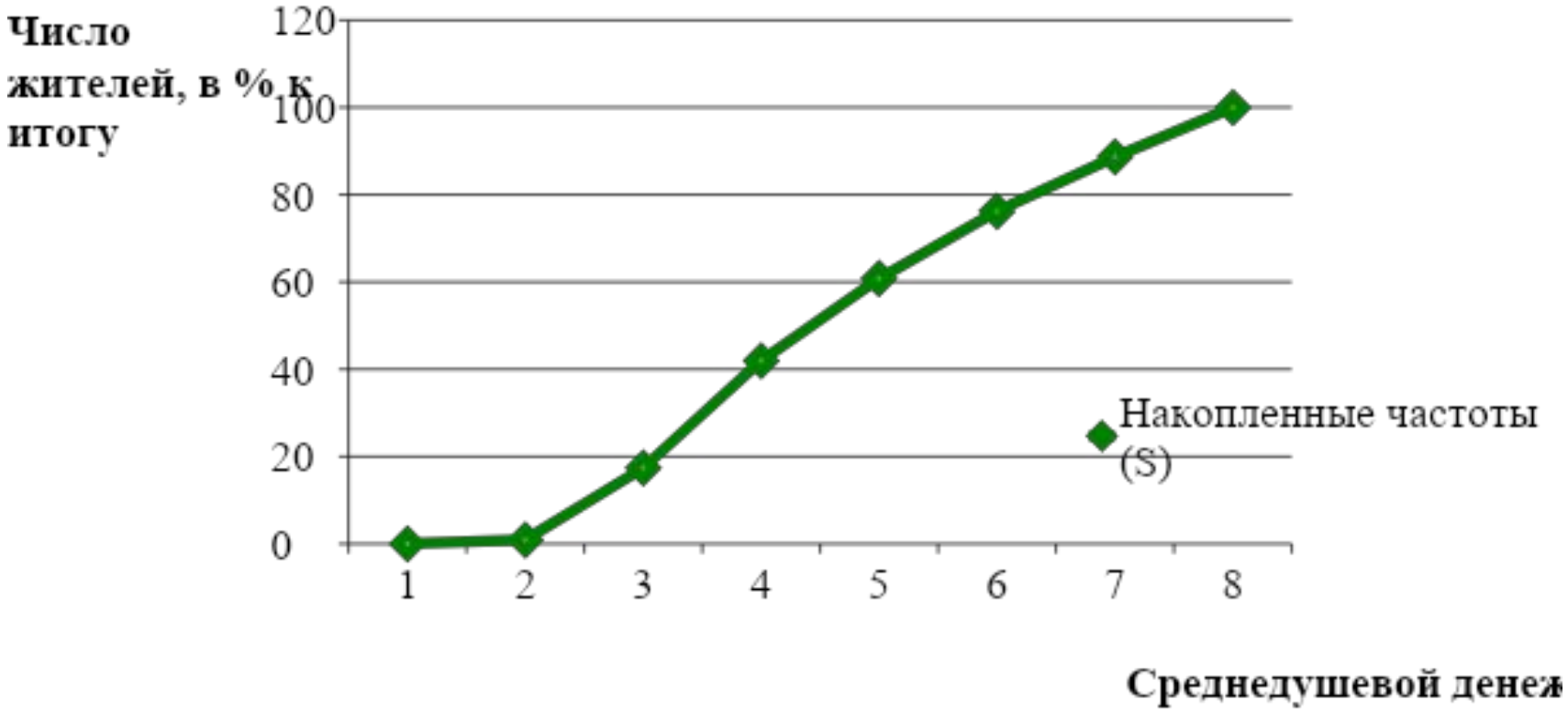
1. Построение вариационного ряда

Исходные данные

Среднедушевой денежный доход в среднем за месяц, тыс. руб.	Число жителей		Накопленные частости (S)	Середина интервала (x)	xf	xw
	чел. (f)	в % к итогу (w)				
До 0,5	26	0,9	0,9	0,25	6,5	0,2
0,5-1,0	463	16,5	17,4	0,75	347,25	12,4
1,0-1,5	690	24,6	42,0	1,25	862,5	30,7
1,5-2,0	528	18,8	60,8	1,75	924	32,9
2,0-2,5	434	15,5	76,2	2,25	976,5	34,8
2,5-3,0	350	12,5	88,7	2,75	962,5	34,3
3,0 и более	318	11,3	100,0	3,25	1033,5	36,8
Итого	**2809**	**100**			**5112,75**	**182,0**

2. Графическое изображение вариационного ряда

Кумулята



3. Показатели центра распределения и структурные характеристики вариационного ряда

Для характеристики среднего значения признака в вариационном ряду используются показатели центра распределения. К ним относятся:

- **средняя величина признака**
- **мода**
- **медиана**

3. Показатели центра распределения и структурные характеристики вариационного ряда

Средняя величина признака

Рассчитывается по формуле средней арифметической взвешенной:

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

x – варианты признака
f – частоты (частости)

3. Показатели центра распределения и структурные характеристики вариационного ряда

Средняя величина признака

Среднедушевой денежный доход в среднем за месяц, тыс. руб.	Число жителей		Накопленные частости (S)	Середина интервала (x)	xf	xw
	чел. (f)	в % к итогу (w)				
До 0,5	26	0,9	0,9	0,25	6,5	0,2
0,5-1,0	463	16,5	17,4	0,75	347,25	12,4
1,0-1,5	690	24,6	42,0	1,25	862,5	30,7
1,5-2,0	528	18,8	60,8	1,75	924	32,9
2,0-2,5	434	15,5	76,2	2,25	976,5	34,8
2,5-3,0	350	12,5	88,7	2,75	962,5	34,3
3,0 и более	318	11,3	100,0	3,25	1033,5	36,8
Итого	**2809**	**100**			**5112,75**	**182,0**

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{5112,75}{2809} = 1,82$$ тыс. руб.

Месячный среднедушевой доход составляет 1820 руб.

3. Показатели центра распределения и структурные характеристики вариационного ряда

Средняя величина признака

Среднедушевой денежный доход в среднем за месяц, тыс. руб.	Число жителей		Накопленные частости (S)	Середина интервала (x)	xf	xw
	чел. (f)	в % к итогу (w)				
До 0,5	26	0,9	0,9	0,25	6,5	0,2
0,5-1,0	463	16,5	17,4	0,75	347,25	12,4
1,0-1,5	690	24,6	42,0	1,25	862,5	30,7
1,5-2,0	528	18,8	60,8	1,75	924	32,9
2,0-2,5	434	15,5	76,2	2,25	976,5	34,8
2,5-3,0	350	12,5	88,7	2,75	962,5	34,3
3,0 и более	318	11,3	100,0	3,25	1033,5	36,8
Итого	**2809**	**100**			**5112,75**	**182,0**

$$\bar{x} = \frac{\sum x\omega}{\sum \omega} = \frac{182,0}{100} = 1,82$$ тыс. руб.

3. Показатели центра распределения и структурные характеристики вариационного ряда

Мода

Мода - значение признака, наиболее часто встречающееся в изучаемой совокупности. в· дискретном ряду модой является вариант с наибольшей частотой (частостью). В интервальном вариационном ряду мода рассчитывается по формуле:

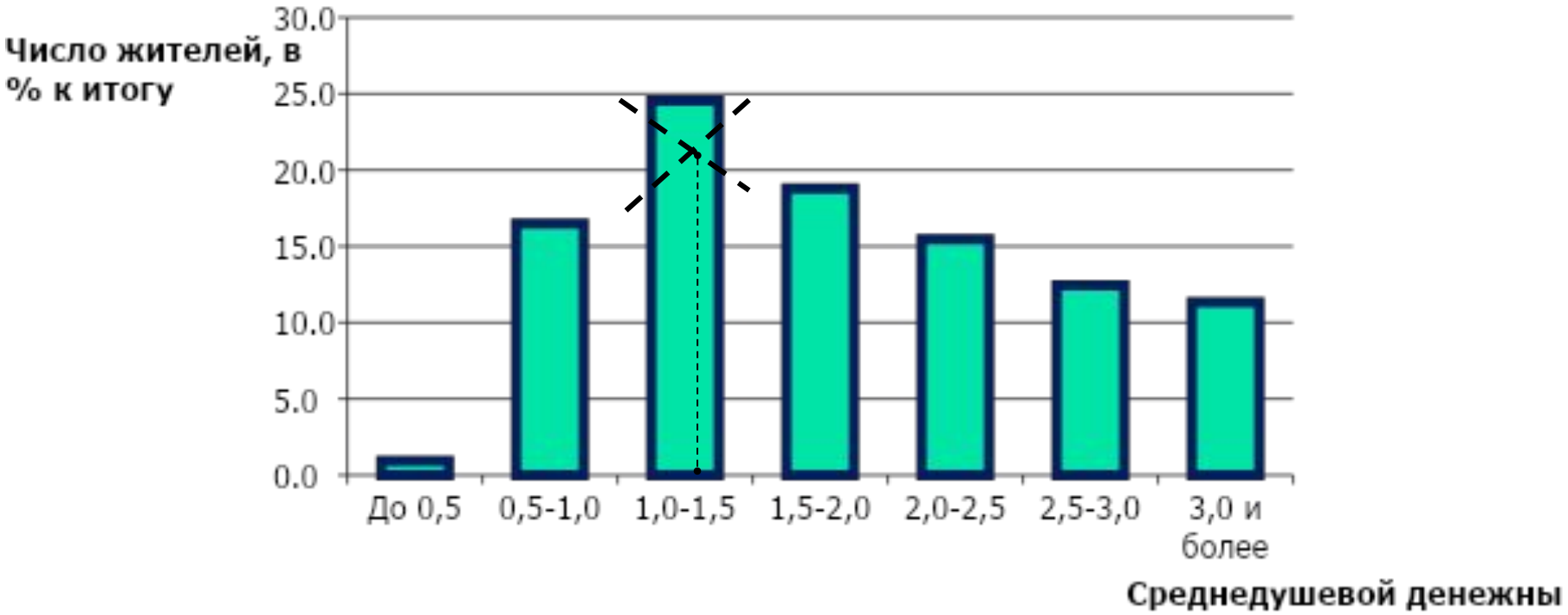
$$Mo = x_{Mo} + i_{Mo} \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})},$$

где x_{Mo} – нижняя граница модального интервала; i_{Mo} – величина модального интервала; f_{Mo}, f_{Mo-1}, f_{Mo+1} – частоты (частости) соответственно модального, домодального и послемодального интервалов.

Модальный интервал - это интервал, имеющий наибольшую частоту (частость).

2. Графическое изображение вариационного ряда

Мода



3. Показатели центра распределения и структурные характеристики вариационного ряда

Мода

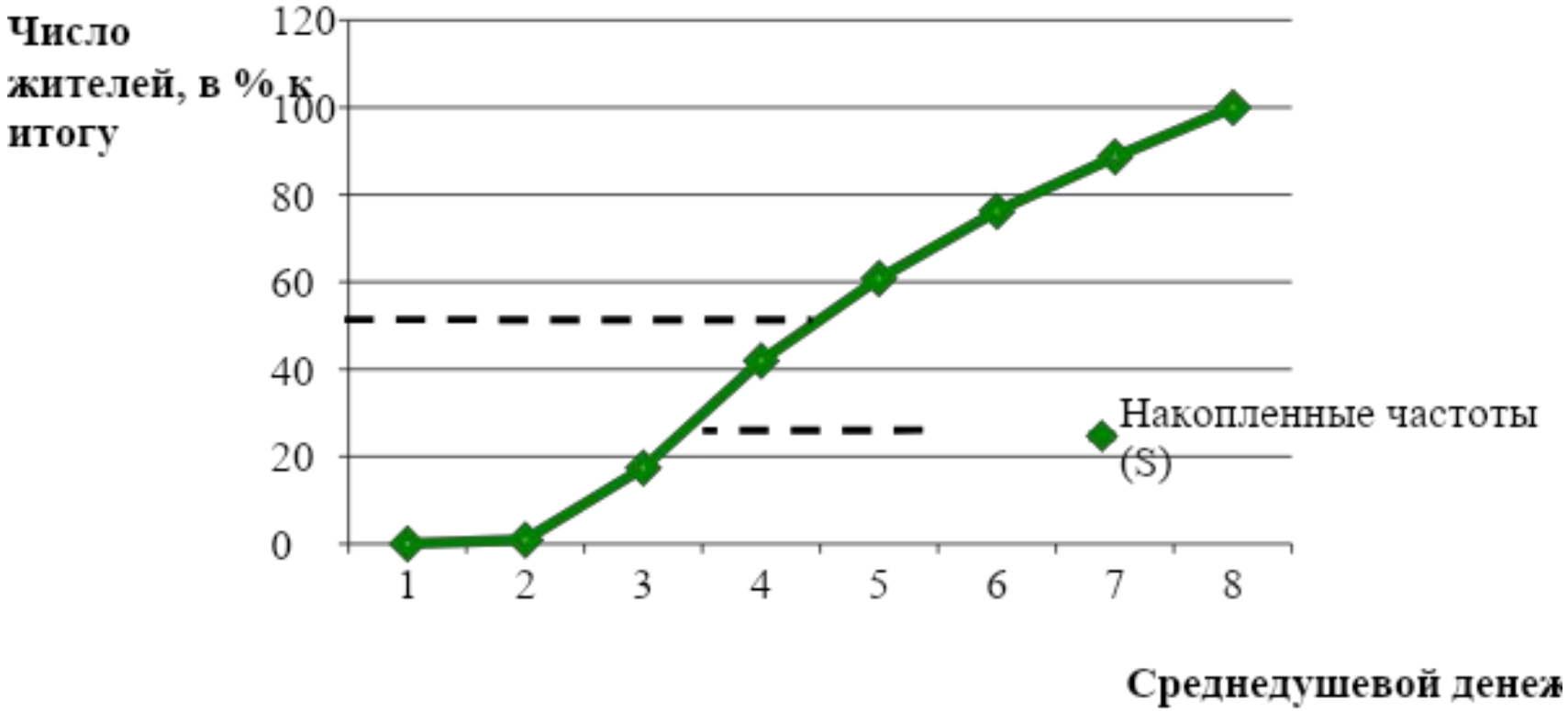
Медиана- вариант, расположенный в середине упорядоченного вариационного ряда, делящий его на две равные части, таким образом, что половина единиц совокупности имеют значения признака меньше, чем медиана, а половина -больше, чем медиана. В интервальном ряду медиана определяется по формуле:

$$Me = x_{Me} + i_{Me} \cdot \frac{\frac{1}{2}\sum f + 1 - S_{Me-1}}{f_{Me}},$$

где x_{Me} – начало медианного интервала; i_{Me} – величина медианного интервала; $\sum f$ – сумма частот (частостей) вариационного ряда; f_{Me} – частота (частость) медианного интервала; S_{Me-1} – сумма накопленных частот (частостей) в домедианном интервале.

3. Показатели центра распределения и структурные характеристики вариационного ряда

Мода



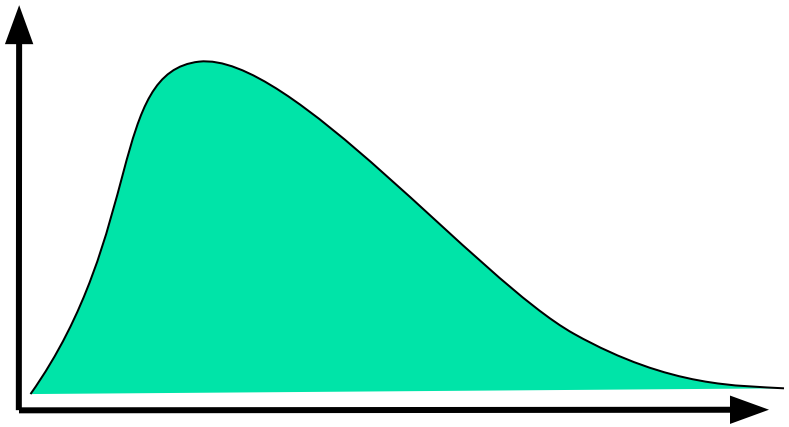
3. Показатели центра распределения и структурные характеристики вариационного ряда

- По соотношению характеристик центра распределения (средней величины, моды и медианы) можно судить о симметричности эмпирического ряда распределения.
- **Симметричным** является распределение, в котором частоты двух вариантов, равностоящих в обе стороны от центра распределения, равны между собой.
- В симметричном распределении средняя величина, медиана и мода равны между собой:

$$\bar{x} = Me = Mo.$$

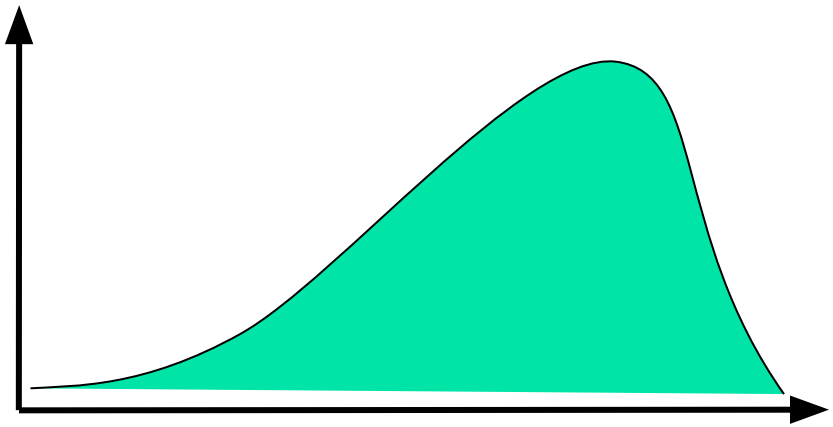
3. Показатели центра распределения и структурные характеристики вариационного ряда

- Если $\bar{x} > Me > Mo$, ет место **правосторонняя асимметрия**, т. е. бОльшая часть единиц совокупности имеет значения изучаемого признака, превышающие модальное значение.
- На графике распределения правая ветвь относительно максимальной ординаты вытянута больше, чем левая.

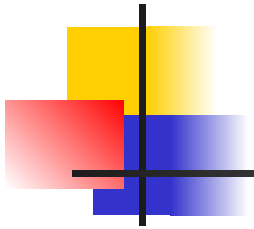


3. Показатели центра распределения и структурные характеристики вариационного ряда

- Соотношение $\bar{x} < Me < Mo$ характерно для левосторонней асимметрии, при которой бОльшая часть единиц совокупности имеет значения признака ниже модального.
- На графике распределения левая ветвь вытянута больше, чем правая.



4. Показатели размера и интенсивности вариации



Обязательным этапом в изучении вариационных рядов является расчет показателей размера и интенсивности вариации.

Для характеристики размера вариации в статистике применяются

- абсолютные показатели вариации:
- **размах** вариации,
- **среднее линейное отклонение**,
- **среднее квадратическое отклонение**,
- **дисперсия**.

4. Показатели размера и интенсивности вариации

Размах вариации (размах колебаний) представляет собой разность между максимальным и минимальным значениями признака в совокупности:

$$R = X_{max} - X_{min}.$$

4. Показатели размера и интенсивности вариации

Размах вариации зависит от величины только крайних значений признака. Более точно характеризуют вариацию признака показатели, основанные на учете колеблемости всех значений признака,- среднее линейное отклонение (d) и среднее квадратическое отклонение (σ)

$$d = \frac{\sum_i |x_i - \bar{x}| f_i}{\sum f_i},$$

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

где x_i – значение признака в i-й группе (для интервальных вариационных рядов – середина i-го интервала); $\bar{x}$ – средняя величина признака в совокупности; f_i – частота (частость) i-го интервала.

4. Показатели размера и интенсивности вариации

Квадрат среднего квадратического отклонения называется дисперсией (σ^2):

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

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

где $\overline{x_i^2}$ – средний квадрат значений признака в совокупности:

$$\overline{x_i^2} = \frac{\sum_i x_i^2 f}{\sum f};$$

$\bar{x}^2$ – квадрат среднего значения признака в совокупности.

4. Показатели размера и интенсивности вариации

- Квадрат среднего квадратического отклонения называется дисперсией (σ^2):

Среднедушевой денежный доход в среднем за месяц, тыс. руб.	Число жителей, в % к итогу (f_i)	Середина интервала (x_i)	$\lvert x_i - \bar{x} \rvert$ ($\bar{x} = 1{,}82$)	$\lvert x_i - \bar{x} \rvert f_i$	$(x_i - \bar{x})^2 f_i$
До 0,5	0,9	0,25	1,57	1,413	2,218
0,5–1,0	16,5	0,75	1,07	17,655	18,891
1,0–1,5	24,6	1,25	0,57	14,022	7,993
1,5–2,0	18,8	1,75	0,07	1,316	0,092
2,0–2,5	15,4	2,25	0,43	6,622	2,847
2,5–3,0	12,5	2,75	0,93	11,625	10,811
3,0 и более	11,3	3,25	1,43	16,159	23,107
Итого	100,0	–	–	68,812	65,959

$$d = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i} = \frac{68{,}812}{100} = 0{,}688 \text{ тыс. руб.};$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} = \frac{65{,}959}{100} = 0{,}660;$$

$$\sigma = \sqrt{0{,}660} = 0{,}812 \text{ тыс. руб.}$$

4. Показатели размера и интенсивности вариации

Для оценки интенсивности вариации, а также для сравнения ее величины в разных совокупностях или по разным признакам используют относительные показатели вариации, которые рассчитываются как отношение абсолютных показателей вариации к средней величине признака:
·относительный размах вариации (коэффициент осцилляции), относительное линейное отклонение и др.

Наиболее часто на практике применяют **коэффициент вариации** (ν), который представляет собой относительное квадратическое отклонение:

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

4. Показатели размера и интенсивности вариации

По величине коэффициента вариации можно судить об интенсивности вариации признака, а следовательно, и об однородности состава изучаемой совокупности. Чем больше величина коэффициента вариации, тем больше разброс значений признака вокруг средней, тем больше неоднородность совокупности. Существует Шкала определения степени однородности совокупности в зависимости от значений коэффициента вариации.

Коэффициент вариации (%)	Степень однородности совокупности
До 30	Однородная
30–60	Средняя
60 и более	Неоднородная

5.5. Оценка вариационного ряда на асимметрию и эксцесс

- **Асимметрия** и **эксцесс** являются важнейшими характеристиками формы распределения.
- Ряды распределения могут иметь один и тот же центр группирования (показатели центра распределения) и одинаковые пределы варьирования признака (показатели вариации), однако при этом отличаться характером распределения единиц совокупности вокруг центра.
- Если большая часть совокупности расположена **левее центра**, имеет место **левосторонняя** асимметрия, если **правее - правосторонняя**.

5.5. Оценка вариационного ряда на асимметрию и эксцесс

Для оценки степени асимметричности применяют моментный и структурный коэффициенты асимметрии.

Моментный коэффициент асимметрии (стандартизованный момент третьего порядка) определяется по формуле:

$$As = \frac{M_3}{\sigma^3},$$

где M_3 – центральный момент третьего порядка.

$$M_3 = \frac{\sum (x_i - \bar{x})^3 f}{\sum f}.$$

5.5. Оценка вариационного ряда на асимметрию и эксцесс

Степень существенности асимметрии можно оценить с помощью средней квадратической ошибки коэффициента асимметрии, которая зависит от объема изучаемой совокупности и рассчитывается по формуле:

$$\sigma_{As} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}},$$

где n — число единиц совокупности.

Если отношение $|As| : \sigma_{As} > 3$, асимметрия считается существенной, если $|As| : \sigma_{As} < 3$, то асимметрия признается несущественной, вызванной влиянием случайных обстоятельств.

5.5. Оценка вариационного ряда на асимметрию и эксцесс

Структурные показатели (коэффициенты) **асимметрии** характеризуют асимметричность только в центральной части распределения, т. е. основной массы единиц, и в отличие от моментного коэффициента не зависят от крайних значений признака.

Наиболее часто применяют структурный коэффициент асимметрии, предложенный английским статистиком К. Пирсоном:

$$As_{II} = \frac{\bar{x} - Mo}{\sigma}.$$

Учитывая, что в умеренно асимметричном распределении расстояния между показателями центра распределения характеризуются следующим равенством $|\bar{x} - Me| = |\bar{x} - Mo| \cdot 3$, формула К. Пирсона может быть записана следующим образом:

$$As = \frac{3(\bar{x} - Me)}{\sigma}.$$

5.5. Оценка вариационного ряда на асимметрию и эксцесс

- Другим свойством рядов распределения является эксцесс
- Под **эксцессом** понимают островершинность или плосковершинность распределения по сравнению с нормальным распределением при той же силе вариации.
- **Эксцесс** -это отклонение вершины эмпирического распределения вверх или вниз от вершины кривой нормального распределения. При этом эксцесс определяется только для симметричных и умеренно асимметричных распределений.

5.5. Оценка вариационного ряда на асимметрию и эксцесс

- Чаще всего на практике эксцесс оценивается с помощью следующего показателя:

$$Ex = \frac{M_4}{\sigma^4} - 3,$$

где M_4 – центральный момент четвертого порядка.

$$M_4 = \frac{\sum (x_i - \bar{x})^4 f}{\sum f}.$$

5.5. Оценка вариационного ряда на асимметрию и эксцесс

Формула эксцесса основана на отклонении от нормального распределения (в нормальном распределении отношение $M_4 : \sigma^4 = 3$).

Распределения более островершинные, чем нормальные, обладают положительным эксцессом ($Ex > 0$), более плосковершинные – отрицательным ($Ex < 0$).

Положительный эксцесс свидетельствует о том, что в совокупности есть слабоварьирующее по данному признаку «ядро», а в плосковершинных распределениях такого «ядра» нет и единицы рассеяны по всем значениям признака более равномерно.

Чтобы оценить существенность эксцесса распределения, рассчитывают среднюю квадратическую ошибку эксцесса:

$$\sigma_{Ex} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}.$$

Если отношение $|Ex| : \sigma_{Ex} > 3$, то отклонение от нормального можно считать существенным.

Числовые характеристики статистического ряда

- средние

Выборочная средняя:	а) характеризует типичное для выборки значение признака X; б) приближенно характеризует (оценивает) типичное для генеральной совокупности значение признака X (см. п. 3.2);
$\bar{x}_в = \frac{1}{n}\sum_{j=1}^{n} x_j$	– *средняя арифметическая*; применяется к вариационному ряду (данные наблюдения не сгруппированы);
$\bar{x}_в = \frac{\sum_{i=1}^{k} x_i \cdot m_i}{\sum_{i=1}^{k} m_i} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot m_i$ $\bar{x}_в = \sum_{i=1}^{k} x_i \cdot w_i$	– *взвешенная средняя арифметическая* (частоты m_i, и частости w_i называют весами); используется, если данные сгруппированы; непосредственно применима только к статистическому распределению дискретного признака (дискретному ряду).
Структурные (*порядковые*) средние.	Если $\bar{x}_в = x_{мо} = x_{ме}$, то распределение симметричное. При нарушении симметрии равенство нарушается (хотя бы одно).
$x_{ме} = \frac{x_j + x_{j+1}}{2}$, если $n = 2j$ – четное; $x_{ме} = x_{j+1}$, если $n = 2j+1$ – нечетное.	*Медиана* – это серединное значение признака X; по определению: $F^*(x_{ме}) = \frac{1}{2}$.
$x_{мо} = x_i$, если $m_i = m_{max}$ (справедливо только для дискретного ряда).	*Мода* – наиболее часто встречающееся значение признака X.

Числовые характеристики статистического ряда

- характеристики вариации (рассеяния)

$D_в = \overline{(x - \bar{x}_в)^2}$ $D_в = \overline{x^2} - \bar{x}_в^2$	– *выборочная дисперсия* есть выборочная средняя арифметическая квад95 ратов отклонений значений признака X от выборочной средней $\bar{x}_в$ (равна "среднему квадрату без квадрата средней"):
$D_в = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x}_в)^2$	– *выборочная дисперсия*; применяется к вариационному ряду (данные наблюдения не сгруппированы);
$D_в = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot m_i}{\sum_{i=1}^{k} m_i} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot m_i$ $D_в = \sum_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot w_i$	– *выборочная взвешенная дисперсия*; используется, если данные сгруппированы; непосредственно применима только к статистическому распределению дискретного признака (дискретному ряду);
$\overline{x^2} = \frac{1}{n}\sum_{j=1}^{n} x_j^2$ $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot m_i$	– *средний квадрат* есть выборочная средняя арифметическая квадратов значений признака X (для вариационного ряда и для дискретного распределения соответственно).
$\sigma_в = \sqrt{D_в}$	– *выборочное среднее квадратическое отклонение* есть арифметическое значение корня квадратного из дисперсии; оно показывает, на сколько в среднем отклоняются значения x_i признака X от выборочной средней $\bar{x}_в$.
$R = x_{max} - x_{min}$	– *размах вариации*.
$v = \frac{\sigma_в}{\bar{x}_в} \cdot 100\,\%$	– *коэффициент вариации*; применяют для сравнения вариации признаков сильно отличающихся по величине, или имеющих разные единицы измерения (разные наименования).

Проверка гипотезы о нормальном распределении генеральной совокупности

Во многих практических задачах точный закон распределения исследуемого признака X генеральной совокупности неизвестен. В этом случае необходимо проверить *гипотезу* о предполагаемом законе распределения.

Выдвигаются *нулевая* гипотеза $H0$ и ей *конкурирующая* $H1$.

$H0$: признак X имеет нормальный закон распределения.

$H1$: признак X имеет закон распределения, отличный от нормального.

Нулевая гипотеза проверяется с помощью *критерия согласия*.

Проверка гипотезы о нормальном распределении генеральной совокупности

Критерий χ^2 ("хи-квадрат") *Пирсона* – наиболее часто употребляемый критерий, может применяться для проверки гипотезы о любом законе распределения. Независимо от того, какое распределение имеет X, распределение случайной величины χ^2:

$$\chi^2 = \sum_{i=1}^{s} \frac{(m_i^э - m_i^т)^2}{m_i^т},$$

где $m_i^э$ – эмпирические частоты, $m_i^т$ – теоретические частоты; при $n \to \infty$ стремится к χ^2 – *распределению с k степенями свободы.*

Теоретические частоты определяются, исходя из предположения о законе распределения генеральной совокупности, в данном случае о нормальном законе.

Так как $p_i = \frac{m_i}{n}$, где p_i – теоретическая вероятность, то $m_i^т = n \cdot p_i$.

Для дискретного ряда:

$p_i = \frac{h}{\sigma_в} \cdot f(u_i)$, где $u_i = \frac{x_i - \bar{x}_в}{\sigma_в}$, $f(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}$ –дифференциальная функция нормированного нормального распределения, шаг $h = x_i - x_{i-1}$, $\bar{x}_в$ – выборочная средняя, $\sigma_в$ – выборочное среднее квадратическое отклонение.

Для интервального ряда:

$p_i = P(x_{i-1} < X < x_i) = \Phi\left(\frac{x_i - \bar{x}_в}{\sigma_в}\right) - \Phi\left(\frac{x_{i-1} - \bar{x}_в}{\sigma_в}\right)$, где $\Phi(t)$ – функция Лапласа.