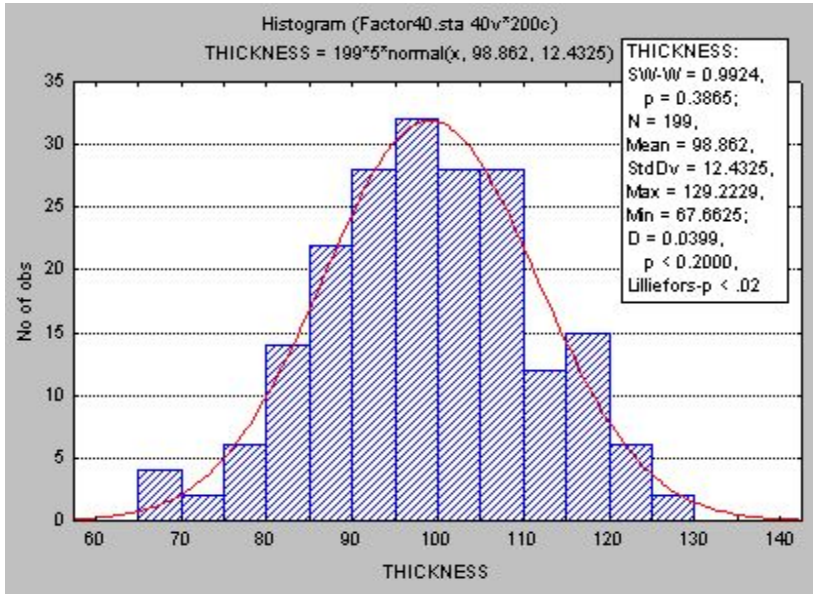
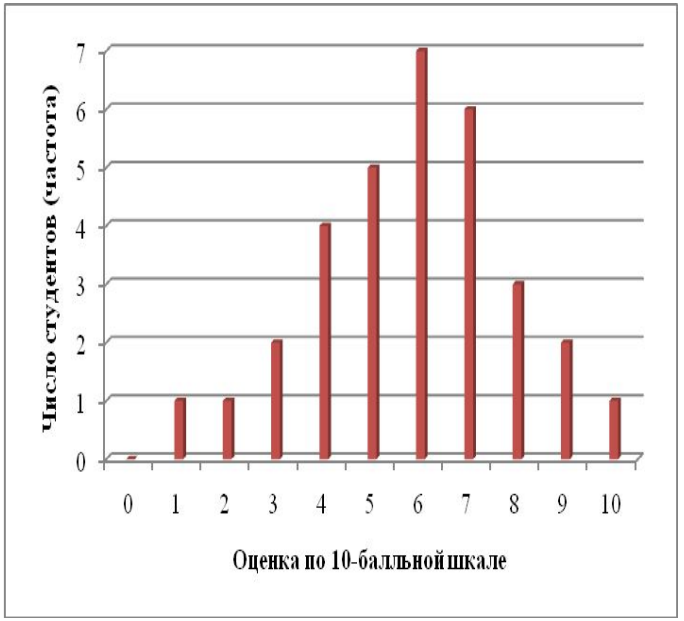
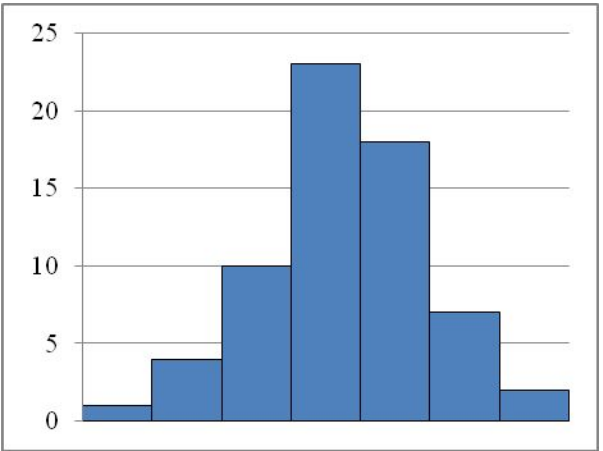
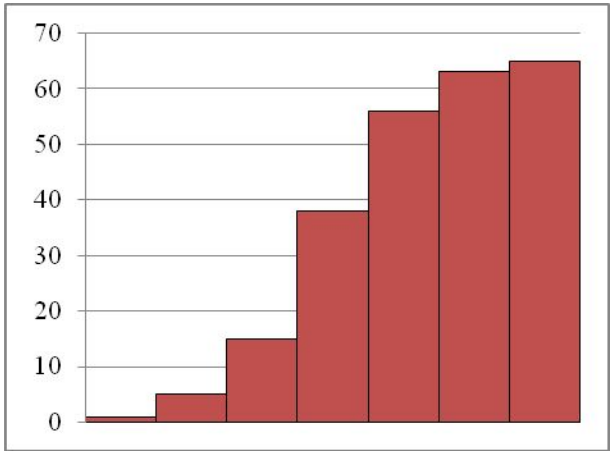
Основы статистического описания









Статистические распределения и их основные характеристики

Рассмотрим одномерную случайную величину X, принимающую n- значений

$$X : \{x_1, x_2, \ldots, x_n\}$$

Изучаемый признак

качественный

Атрибутивный ряд распределения

- Распределение рабочих по профессии
- Предприятий по форме собственности

Количественный (дискретный)

Вариационный ряд распределения

(Распределение коммерческих банков по объему активов)

- Варианты значений изучаемого признака, встречающегося в совокупности;
- Частота, соответствующая каждому варианту значения изучаемого признака

Статистические распределения и их основные характеристики

Существуют три формы вариационного ряда:

- ранжированный,
- дискретный,
- интервальный.

Статистические распределения и их основные характеристики

Ранжированный ряд — это перечень отдельных единиц совокупности в порядке возрастания (убывания) изучаемого признака.

$$x_1 \le x_2 \le \ldots \le x_i \le x_{i+1} \le \ldots \le x_n.$$

Элемент x_i называется ***i-й порядковой статистикой.***

Основные порядковые статистики:

$x_{(1)} = min\{x_{(i)}\}$ **– наименьшее значение**

$x_{(n)} = max\{x_{(i)}\}$ **– наибольшее (максимальное) значение.**

Пример Сведения о крупных банках Санкт-Петербурга, ранжированных по размерам собственного капитала на 01.10.2013 г.

Название банка	Собственный капитал, млн руб.
• Балтонэксим банк	169
• Банк «Санкт-Петербург»	237
• Петровский	268
• Балтийский	290
• Промстройбанк	1007

Статистические распределения и их основные характеристики

Если признак принимает небольшое число значений, то строится дискретный вариационный ряд.

Например, распределение футбольных матчей по числу забитых мячей.

Дискретный вариационный ряд – это таблица, состоящая из двух строк: конкретных значений варьирующего признака и числа единиц совокупности с данным значением признака (частотами).

Эти частоты называют *эмпирическими*.

Значения признака x_i	$x_{(1)}$	$x_{(2)}$	…	$x_{(i)}$	…	$x_{(k)}$
Частоты m_i	m_1	m_2	…	m_i	…	m_k

Сгруппированный дискретный вариационный ряд графически представляют в виде *гистограммы* или *полигона*.

Дискретные количественные данные

Сгруппированный кумулятивный дискретный вариационный ряд представляет собой значения признака x_i , указанные вместе с соответствующими накопленными частотами $m_{iн}$ или частостями $w_{iн}=m_{iн}/n$.

Значения признака x_i	$x_{(1)}$	$x_{(2)}$	…	$x_{(k)}$
Накопленная частота $m_{iн}$	$m_{1н}=m_1$	$m_{2н}=m_1+m_2$	…	$m_{kн}=n=\sum_{i=1}^{k} m_i$

Накопленные частоты показывают, сколько единиц совокупности имеют значения признака не больше, чем верхняя граница интервала.

Частоты и частости ряда

Частоты ряда (mi) могут быть заменены частостями ($w_i=m_i/n$) ряда, которые представляют собой частоты, выраженные в относительных числах (долях или процентах):

$$w_1 = \frac{m_1}{\sum m}; \quad w_2 = \frac{m_2}{\sum m} \ldots$$

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

Производственный стаж (лет)	Число рабочих, m	Частость, w		Накоплена я частота, S
		в долях	в %	
До 5	2	0,022 (2/90)	2,2	2
5-10	22	0,245 (22/90)	24,5	24
10-15	48	0,533	53,3	72
15-20	16	0,178	17,8	88
20 и более	2	0,022	2,2	90
Итого	90	1,000	100,0	

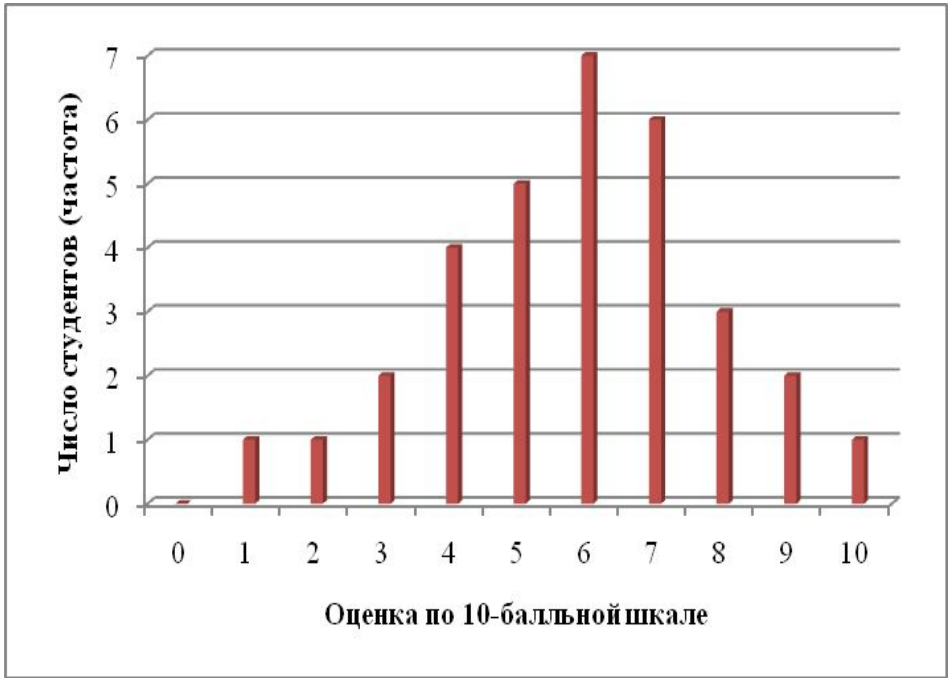
Статистические методы анализа одномерных данных

Share of distribution

Гистограмма (histogram) - диаграмма в виде столбцов, по оси абсцисс которой отображаются все возможные значения переменной,

по оси ординат – частоты встречаемости m_i каждого значения или относительные частоты – доли, частости (m_i/n).

Гистограмма была введена в статистическую практику Карлом Пирсоном в 1895 г.



Дискретные количественные данные

Полигон – графическое изображение сгруппированного дискретного вариационного ряда в виде ломаной, соединяющей точки, по оси абсцисс соответствующие всем возможным значениям признака,

а по оси ординат - значениям частот m_i или относительных частот $w_i=m_i/n$.

Полигон позволяет оценить распределение частот значений дискретной переменной, выявить *наиболее часто* ***(мода)*** *и редко встречающиеся значения признака*.

Дискретные количественные данные

Сгруппированный кумулятивный дискретный вариационный ряд графически представляют в виде *кумуляты.*

Кумулята *– графическое изображение сгруппированного кумулятивного дискретного вариационного ряда в виде столбцов, при построении которого*

по оси абсцисс откладывают все возможные значения признака,

по оси ординат - накопленные частоты или накопленные относительные частоты, относящиеся к данному значению.

Кумулята показывает количество (или долю) объектов совокупности, значения признака которых не превышают заданного значения.

Пример

Для построения кумуляты используем накопленные частоты

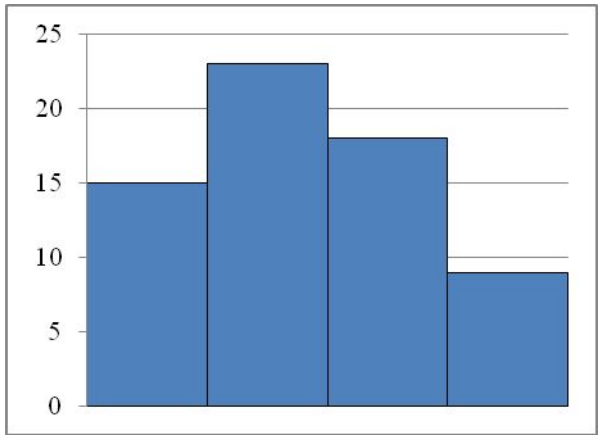
График кумуляты позволяет найти число объектов, имеющих значения признака, не превышающее заданного.

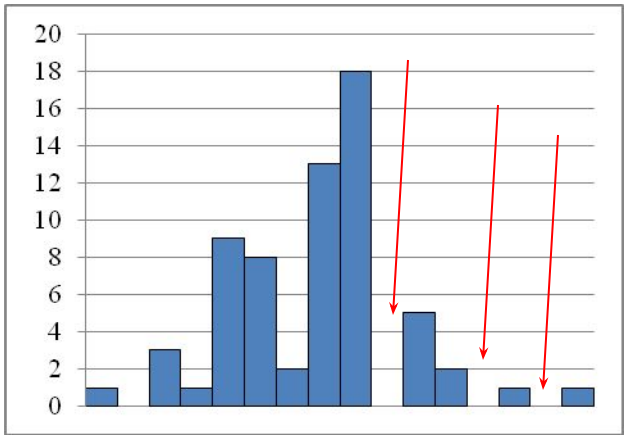
Например, 24 страницы имеют число опечаток не превышающее 5 (от 0 до 5 опечаток).

Интервальный вариационный ряд

Построение *интервального вариационного* ряда начинают с определения числа интервалов ***k***.

- Число интервалов не должно быть *слишком малым*, т.к. при этом гистограмма получается слишком сглаженной (*oversmoothed*), теряет особенности изменчивости исходных данных.
- Число интервалов не должно быть *слишком большим* – иначе мы не сможем оценить плотность распределения изучаемых данных по числовой оси – гистограмма получится «недосглаженная» (*undersmoothed*), с незаполненными интервалами, неравномерная.





Определение оптимального числа интервалов

- В 1926 г. Герберт Стерджес (*Herbert Sturges*) предложил формулу для вычисления количества интервалов, на которые необходимо разбить исходное множество значений изучаемого признака.
- Приблизительное **число интервалов s**, которое необходимо выбрать при группировке и построении гистограммы для *n* результатов измерений СВ, полученных из нормально распределенной ГС определяется по правилу Стерджеса как:

$$s \approx 1 + 3{,}322 \cdot \lg n$$

- **Ширина интервалов *h***, на которые необходимо разбить всю область возможных значений исследуемого признака по имеющимся наблюдениям $\{x_1, x_2, \ldots, x_n\}$, определяется как:

$$h \approx \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n}$$

Альтернативные подходы

✔ Метод Дэвида Скотта

Дэвид Скотт (*David W. Scott*) в 1979 г. предложил следующую формулу для вычисления оптимальной ширины интервалов *h*:

$$h^* \approx \frac{3{,}5 \cdot S}{\sqrt[3]{n}}$$

где S – среднее квадратическое отклонение.

✔ Метод квадратного корня (*Square-root choice*) – число интервалов h выбирается равным квадратному корню из числа наблюдений *n*:

$$h \approx \sqrt{n}.$$

Рекомендации

Число интервалов для небольших выборок обычно берут

- 5–6 при ***n*<50**,
- 6-8 – от 50 до 100 наблюдений;
- 8-10 классов при ***n*>100**

с расчетом, чтобы интервалы были достаточно наполнены частотами.

- Считается, что формула Стерджеса позволяет строить удовлетворительные гистограммы при числе измерений менее 200.
- Для больших массивов информации, например, порядка 10^4-10^9 наблюдений, правило Стерджеса может приводить к слишком сглаженным гистограммам.
- асимметричные распределения требуют бóльшего числа интервалов группировки.

One-Variable Data Analysis

Основные идеи при исследовании **формы распределения** (Share of distribution)

- Графическое представление исходных данных (точечное распределение (Dotplot); листовая диаграмма (Stemplot); гистограмма (Histogram).
- Характеристики положения СВ;
- Ранговые характеристики СВ;
- Характеристики разброса СВ;
- Исследование нормальности распределения (Normal Distribution)
- Диагностика выбросов (Ящичковая диаграмма Boxplot)
- Правило 68-95-99,7 (The 68-95-99,7 Rule)
- Z- преобразование.

Изучение формы распределения

- **Графическое представление исходных данных**

Для изучения формы распределения можно использовать следующие графические возможности

- **Точечное распределение (Dotplot);**
- **Диаграмма стебель-листья (Stemplot).**

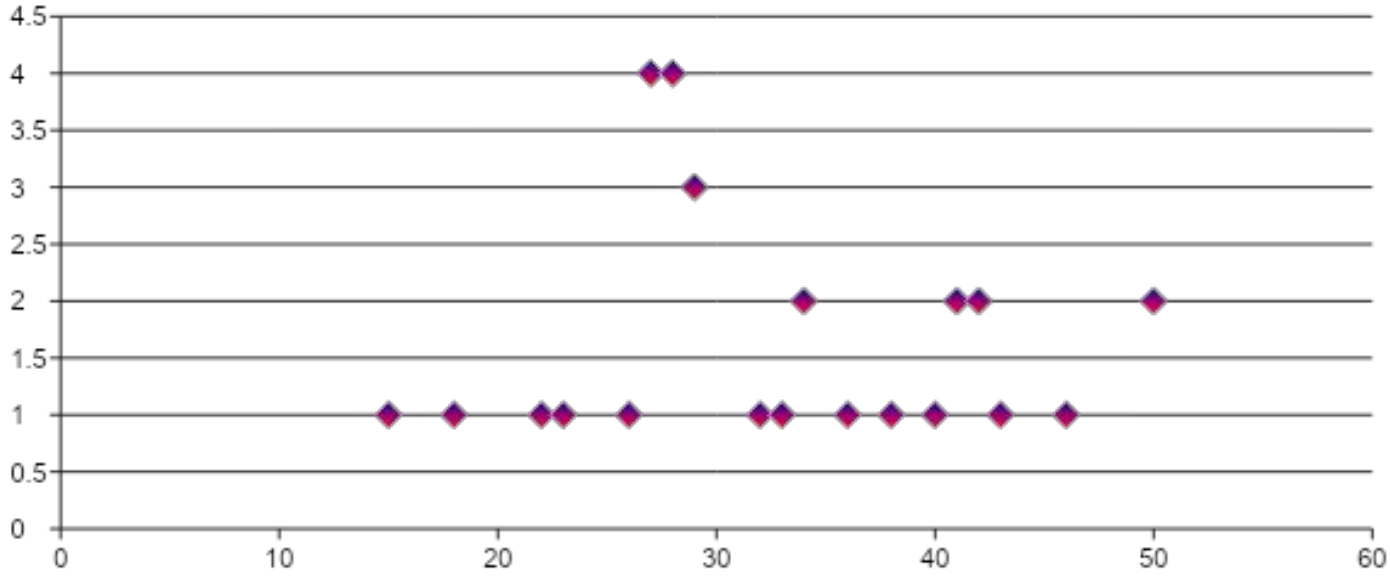
Пример

Рассмотрим 31 оценку по 50 бальной системе, которую получили студенты статистического отделения на экзамене

28	38	42	33	29	28	41	40	15	36	27	34
22	23	28	50	42	46	28	27	43	29	50	29
32	34	27	26	27	41	18					

Необходимо рассмотреть 3 типа графиков, которые помогут сделать вывод о характере распределения: Dotplot (точечное распределение), Stemplot, Histogram

Dotplot (точечное распределение)



Stemplot

28	**38**	**42**	**33**	**29**	**28**	**41**	**40**	**15**	**36**	**27**	**34**
22	**23**	**28**	**50**	**42**	**46**	**28**	**27**	**43**	**29**	**50**	**29**
32	**34**	**27**	**26**	**27**	**41**	**18**					

Xmin =15

Xmax=50

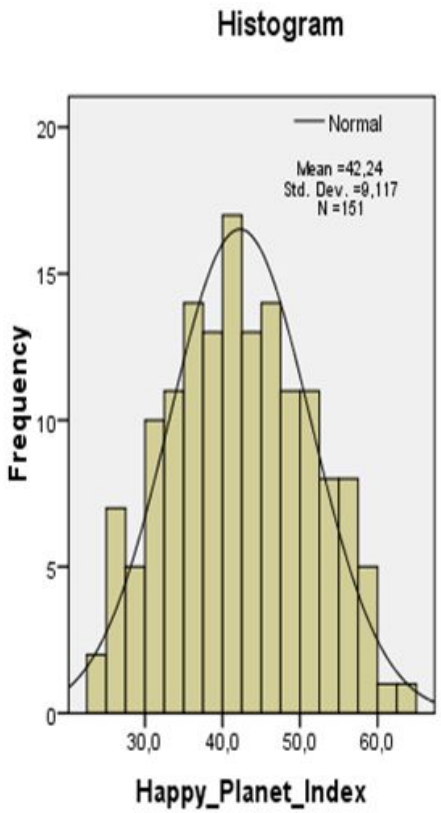
стебель	**листья**
1	58
2	23677778888999
3	234468
4	0112236
5	00

? Левосторонняя или правосторонняя асимметрия

Stemplot

28,3; 27,5; 28,1; …..

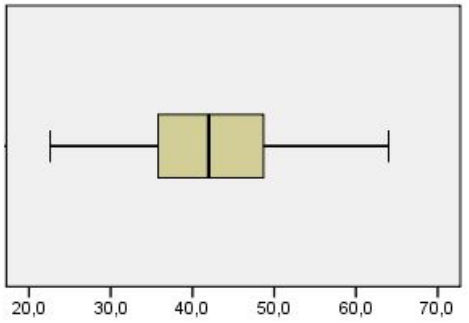
0,0018; 0.0023; 0,0021;….



Happy_Planet_Index Stem-and-Leaf Plot

Frequency	Stem & Leaf
2,00	2 . 24
12,00	2 . 556666788889
21,00	3 . 000001111233334444444
27,00	3 . 555566666777777777888999999
30,00	4 . 000000000011122222222333333444
25,00	4 . 555666666667777777889999
19,00	5 . 0001111122222233444
13,00	5 . 5566666778899
2,00	6 . 04

Stem width: 10,0
Each leaf: 1 case(s)



стебель **листья**

5	**9**
6	**00001111**
6	**2222233333333333**
6	**444444455555555555555555**
6	**66666666666667777777777777777**
6	**888888888899999999**
7	**00**
7	**2**

Happy_Planet_Index Stem-and-Leaf Plot

Frequency	Stem & Leaf
2,00	2 . 24
12,00	2 . 556666788889
21,00	3 . 000001111233334444444
27,00	3 . 555566666777777777888999999
30,00	4 . 000000000011122222222333333444
25,00	4 . 555666666667777777889999
19,00	5 . 0001111122222233444
13,00	5 . 5566666778899
2,00	6 . 04

Stem width: 10,0
Each leaf: 1 case(s)

One-Variable Data Analysis

Исследование формы распределения (Shape of the data)

- Нахождение характеристик положения случайной величины (Center of the data)

средней, моды и медианы (mean, median, mode);

Характеристики положения

Погода в определенном пункте земного шара в один и тот же день в разные годы может быть очень различной.

Например, в Санкт-Петербурге 31 марта температура воздуха за сто с лишним лет наблюдений колебалась
от **-20,1°** в 1883 г. до **+12,24°** в 1920 г.

Примерно такие же колебания наблюдаются и в другие дни года.

По таким индивидуальным данным о погоде в какой-то произвольно взятый год нельзя составить представление о климате Санкт-Петербурга.

Характеристики климата - это средние значения за длительный период времени.

Характеристики положения

	Простая выборка	Выборка по сгруппированным данным
Выборочная средняя	$\overline{x} = v_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i m_i$
Выборочная мода x_{mod}	наиболее часто встречающееся в выборочных наблюдениях значение переменной (модное значение).	

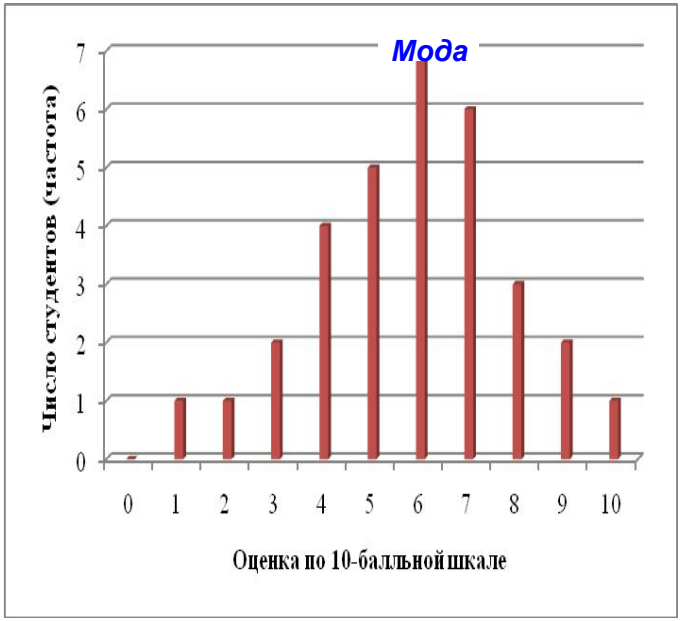
Характеристики положения

Мода может быть <u>не единственной.</u>

Если два или несколько значений переменной обладают одинаковой максимальной частотой, то в этом случае распределения называются *бимодальными и полимодальными.*

! **Для описания категориальных переменных не используются никакие числовые характеристики (например, «средний пол»).**

Единственной полезной характеристикой является мода.



Характеристики положения

Медиана (median) – значение признака, приходящееся на середину ранжированного ряда наблюдений.

Положение медианы определяется ее номером.

(нечетный и четный ряд)

Характеристики положения

Хотя среднее и медиана характеристики центра, которые используются для описания характера распределения, медиана является наиболее устойчивой оценкой

(менее подвержена влиянию экстремальных наблюдений).

Характеристики положения

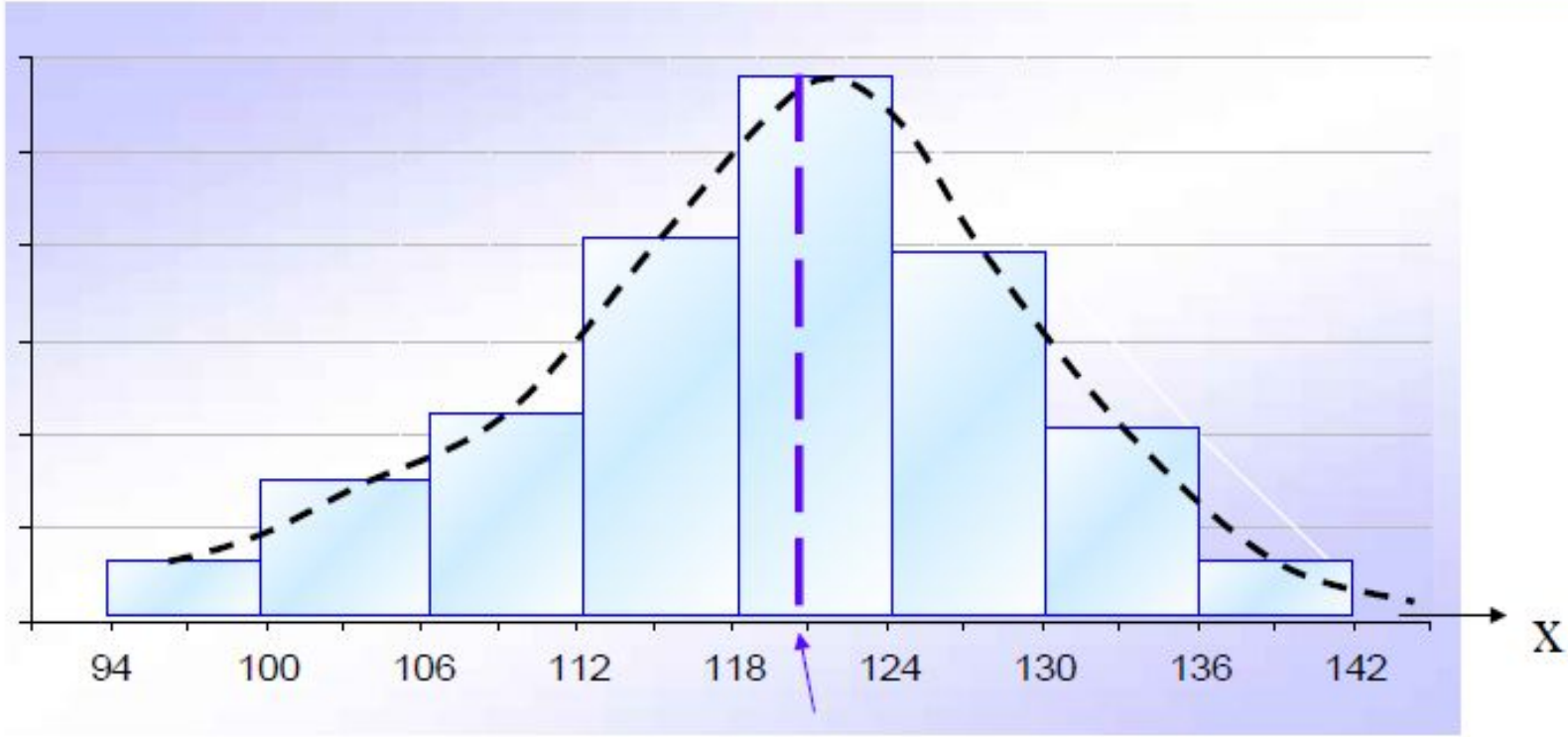
Пример

Зарплата 5 школьных учителей в колледже США составила $32,700; $32,700; $38,500; $41,600; $44,500.

Среднее значение и медиана составляют $38,160; $38,500.

Преподаватель более высокой квалификации заменил коллегу во время болезни. Его зарплата составляет $174,300.

В этом случае медиана не изменится и составит $38,500, а среднее значение увечится до $64,120



Средняя арифметическая = 119,2
Мода = 121, медиана = 121

Вывод: значение выборочных показателей свидетельствует в пользу выбора симметричного закона распределения для анализируемой генеральной совокупности.

Изучение формы распределения

- ***Ранговые характеристики –*** варианты, занимающие в ранжированном вариационном ряду определенное место.

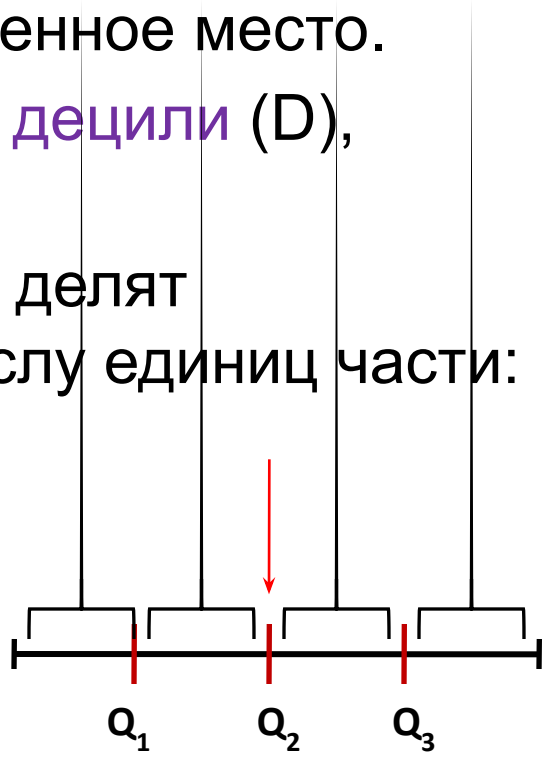
  К их числу относятся квартили (Q), квинтили, децили (D), перцентили (P).

  Квартили (Q) – значения признака, которые делят ранжированный ряд на четыре равные по числу единиц части:

  первая квартиль Q1, вторая Q2 и третья Q3.

  Вторая квартиль является медианой.

  Определение положения квартили



$$N_{Q_1} = \frac{n+1}{4}; \quad N_{Q_2} = \frac{n+1}{4} \cdot 2 = \frac{n+1}{2}; \quad N_{Q_3} = \frac{n+1}{4} \cdot 3$$

n- общее число единиц совокупности.

Ранговые характеристики

Децили – значения признака, которые делят ранжированный ряд на десять равных по численности частей (всего 9).

Расчет децилей аналогичен расчету квартилей.

При растете децилей определяют сначала порядковые номера каждой из девяти децилей:

$$N_{D_1} = \frac{n+1}{10}; \quad N_{D_2} = \frac{2(n+1)}{10} = \frac{n+1}{5}; \ ...; \ N_{D_9} = \frac{9(n+1)}{10}$$

По накопленным частотам в ДР определяют местоположение децилей и их значения.

Ранговые характеристики

Перцентили – значения признака, которые делят ранжированный ряд на 100 равных по числу единиц частей. (всего 99).

One-Variable Data Analysis

Алгоритм описания данных:

- Исследование характеристик разброса (рассеяния) случайной величины

- *Вариация (размах вариации и коэффициент вариации)*
- *Межквартильная разница (interquartile Range),*
- *Квартильное отклонение ,*
- *Относительный показатель квартильной вариации;*
- *Относительное линейное отклонение.*
- *Дисперсия, стандартное отклонение.*

Исследование характеристик разброса (рассеяния) случайной величины

Вариация признака – различие индивидуальных значений признака у единиц совокупности в один и тот же период или момент времени.

Разность наибольшего и наименьшего значений признака называется ***размахом вариации:***

$$R = x_n - x_1 = x_{max} - x_{min}.$$

Размах служит самостоятельной характеристикой разброса значений изучаемого признака. Используется не часто, т.к. хотим знать как точки распределяются вокруг центра.

Группировка данных

Относительные показатели вариации:

- ***Коэффициент вариации*** является безразмерной величиной и вычисляется по формуле

$$V = \frac{s}{\bar{x}} \cdot 100\%$$

Наиболее распространенный коэффициент (часто используется на практике).

Совокупность считается однородной, если коэффициент вариации не превышает 33%.

Характеристики рассеяния

Межквартильная разница (interquartile Range)- IQR

IQR=Q3-Q1

Me=Q2

IQR может не включать в себя 50 % наблюдений.

Пример: Определить *Q3 и Q1 для следующего ряда:*

5 5 6 7 8 9 11 13 17

Медиана ? $M_e = \frac{n+1}{2} = 5$ *позиция* ⇒ ***Me=8***

Левая часть 5 5 6 7 Q1=5,5

Правая часть 9 11 13 17 Q3=12

IQR*=Q3-Q1= 12-5,5=6,5*

Характеристики рассеяния

Квартильное отклонение - d_k

Применяется вместо размаха вариации, чтобы избежать недостатков, связанных с использованием крайних значений.

$$d_k = \frac{Q_3 - Q_1}{2} = \frac{6{,}5}{2} = 3{,}25$$

Характеристики рассеяния

Квартильное отклонение - d_k

Применяется вместо размаха вариации, чтобы избежать недостатков, связанных с использованием крайних значений.

$$d_k = \frac{Q_3 - Q_1}{2} = \frac{6{,}5}{2} = 3{,}25$$

- ***Относительный показатель квартильной вариации***

$$K_{d_k} = \frac{Q_3 - Q_1}{2Q_2} \cdot 100\% = \frac{6{,}5}{16} \cdot 100\% = 40{,}6\%$$

или

$$K_{d_k} = \frac{d_k}{M_e} \cdot 100\%$$

Относительные показатели вариации

- ***Относительное линейное отклонение***

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

где $\bar{d}$ - среднее линейное отклонение

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

Характеристики рассеяния

	Простая выборка	Выборка по сгруппированным данным
Выборочная дисперсия	$S^2 = \mu_2^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 m_i = \frac{1}{n}\sum_{i=1}^{n} x_i^2 m_i - (\bar{x})^2$
Выборочное среднее квадратическое отклонение	$S = \sqrt{S^2}$	

Исследование формы распределения

Нормальный закон - это один из многих типов распределений, имеющихся в природе, с относительно большим удельным весом практической применимости.

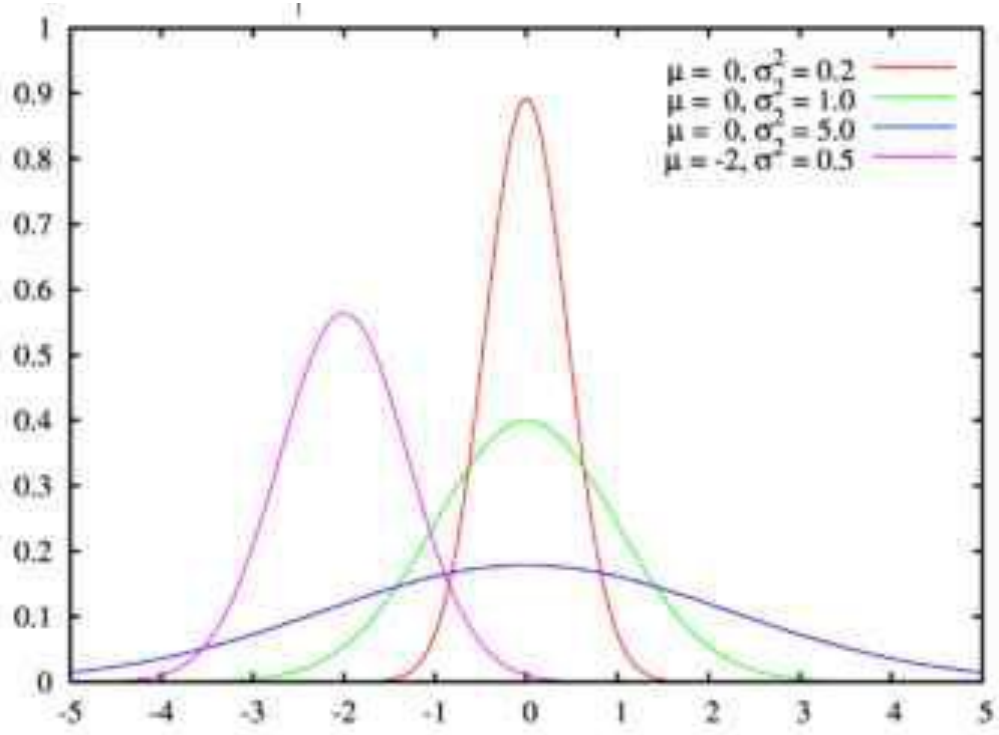
В случае отклонения исследуемых экспериментальных данных от нормального закона существуют два пути его использования:

а) использовать его в качестве *первого приближения*; при этом оказывается, что подобное допущение дает достаточно точные с точки зрения конкретных целей исследования результаты;

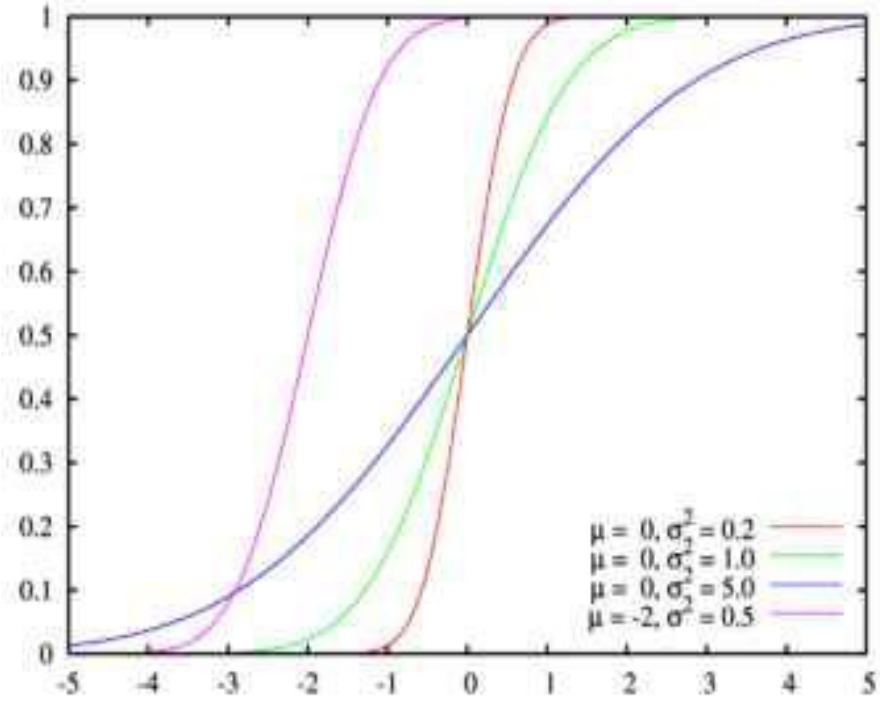
б) подобрать такое преобразование исследуемой случайной величины X, которое видоизменяет исходный «ненормальный» закон распределения, превращая его в нормальный.

Область применения:

Функция плотности



Функция распределения



Основные законы распределения случайных величин

Нормальный закон распределения

Наиболее распространённый

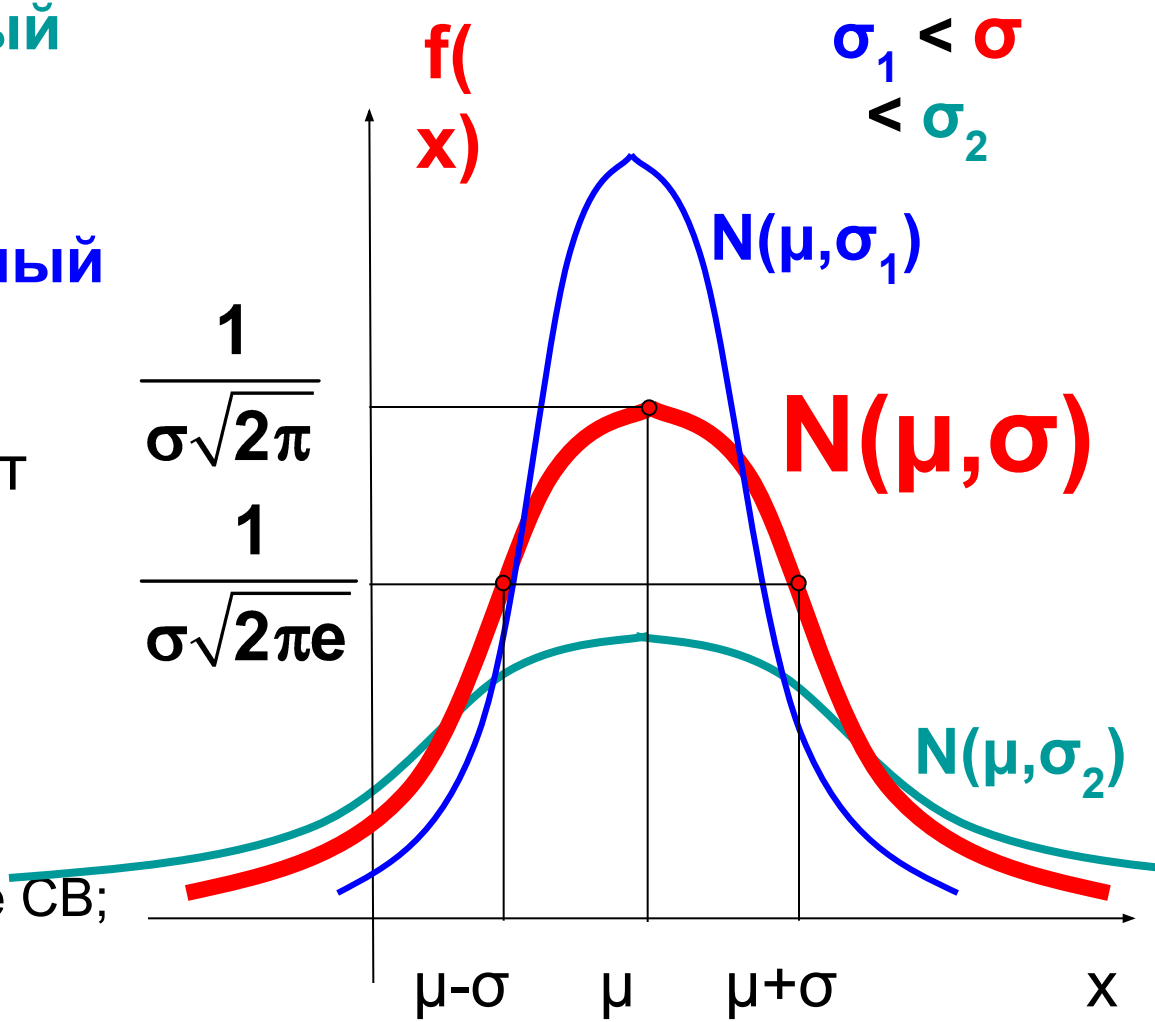
Предельный

- Непрерывная случайная величина X имеет **нормальный закон распределения** с параметрами **μ** и **σ**, если её плотность вероятности имеет вид:

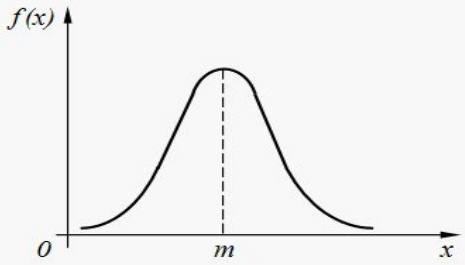
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

где **μ** – математическое ожидание СВ;

σ^2 – дисперсия, **σ** – среднее квадратическое отклонение



Нормальный закон распределения



Свойства нормального распределения

1. Кривая нормального распределения расположена над осью ОХ, $f(x) > 0$
2. При $x \longrightarrow \pm\infty$ плотность распределения стремится к 0. Кривая распределения асимптотически приближается к оси ОХ
3. В точке $x = \mu$ плотность нормального распределения имеет максимум $$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$
4. Кривая нормального распределения симметричная относительно точки $x = \mu(m)$

Математическое ожидание, мода и медиана совпадают

Нормальный закон распределения

Свойства нормального распределения:

5. Кривая распределения имеет две точки перегиба с координатами

$$(\mu-\sigma;\ \frac{1}{\sigma\sqrt{2\pi e}})\ и\ (\mu+\sigma;\ \frac{1}{\sigma\sqrt{2\pi e}})$$

6. Форма нормальной кривой не изменяется при изменении математического ожидания (кривая сдвигается вдоль оси OX)

   При изменении σ меняется форма кривой

7. При $\mu=0\ и\ \sigma=1$ плотность распределения вероятности называется *нормированной плотностью*,

а ее график – *нормированной нормальной кривой распределения*

Нормальный закон распределения

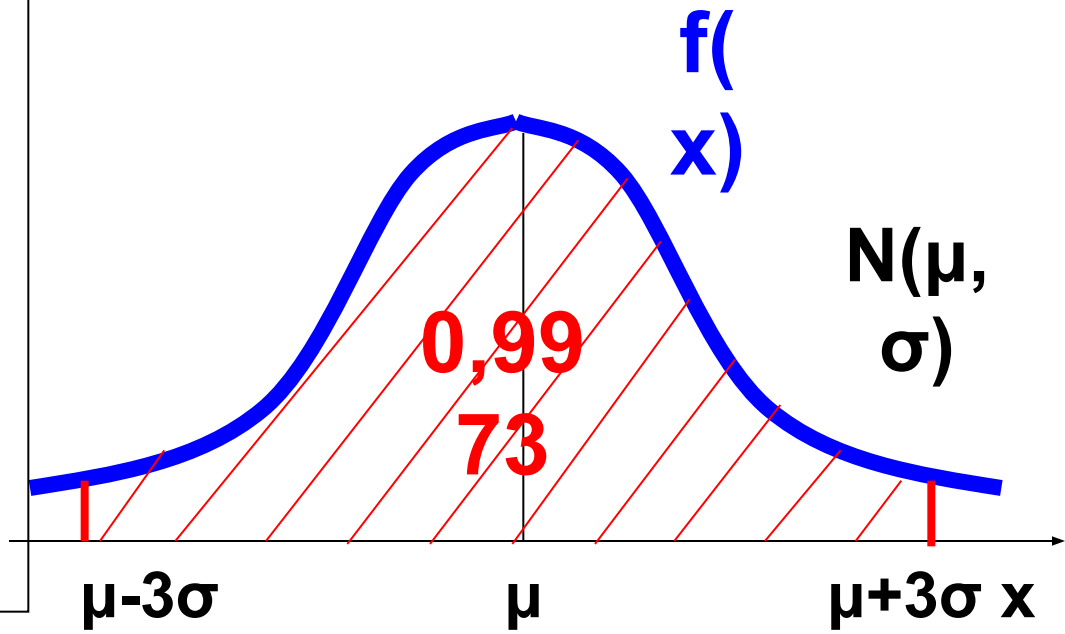
Правило «68-95-99,7»

«Правило одной сигмы»

«Правило двух сигм»

«Правило трёх сигм»

Если случайная величина X имеет нормальный закон распределения X є N(μ,σ), то практически достоверно, что её значения заключены в интервале (μ-3σ; μ+3σ)
(Вероятность «выброса» составляет 0,0027)



•Кривая плотности распределения

Для изучения формы распределения необходимо рассчитать **коэффициенты асимметрии и эксцесса**

? Симметричное ли распределение (форма распределения, холмообразная или нет)

- **Скос**
- **Ассиметрия**
- **Бимодальность**
- **Однородность.**

Выборочные коэффициенты асимметрии и эксцесса

Для характеристики особенностей формы распределения применяются показатели **асимметрии и эксцесса**.

Асимметрия

$$A_s = \frac{\mu_3}{S^3} \qquad A_s = \frac{\bar{x} - M_o}{s}$$

Относительный показатель асимметрии

Эксцесс

$$E_k = \frac{\mu_4}{S^4} - 3$$

μ_3 – центральный момент третьего порядка;
μ_4 – центральный момент четвертого порядка.

Исследование формы распределения

Асимметрия (skewness) показывает, в какую сторону относительно среднего сдвинуто большинство значений распределения.

Нулевое значение асимметрии означает симметричность распределения относительно среднего значения, что соответствует нормальному закону распределения.

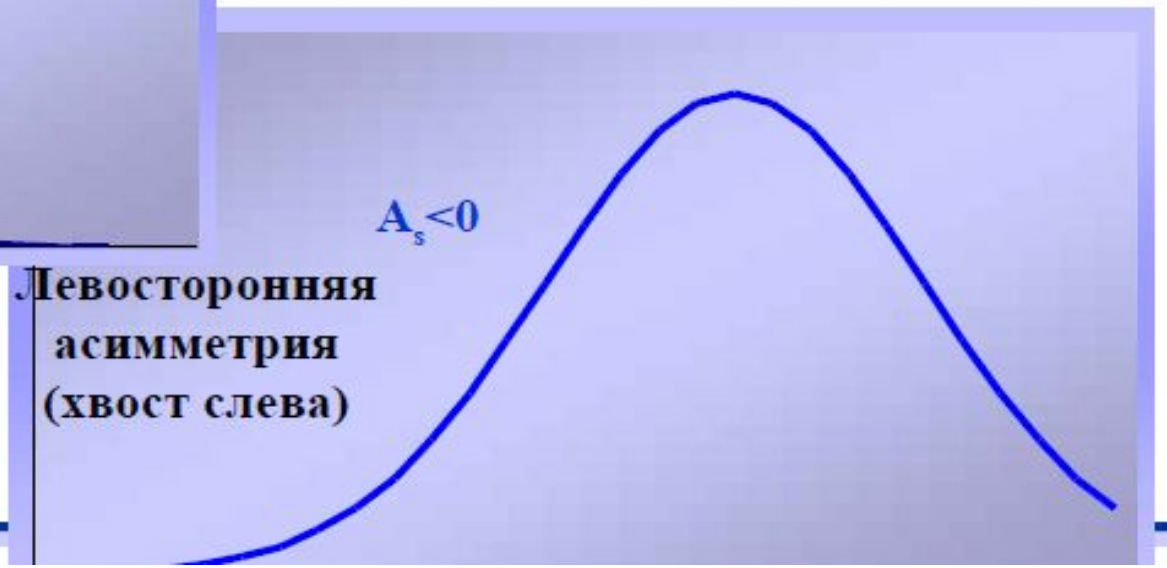
Чем больше абсолютная величина коэффициента, тем больше степень скошенности.

Выборочные коэффициенты асимметрии и эксцесса

Коэффициент асимметрии A_s – показатель асимметричности распределения, определяющий степень скошенности кривой по сравнению с нормальным распределением.

$\|A_s\|>0,5$	Значительная асимметрия

Относительный показатель асимметрии



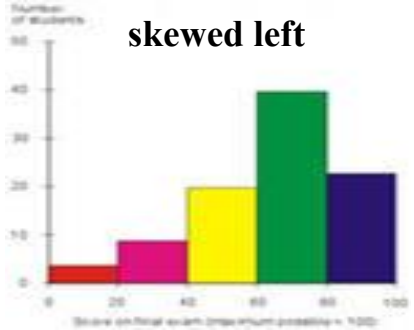
Исследование формы распределения

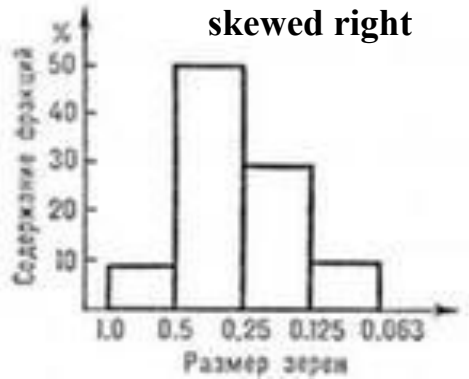
Оценка степени существенности асимметрии осуществляется с помощью средней квадратической ошибки:

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

Если $\frac{|A_s|}{\sigma_{A_S}} > 3$, асимметрия существенна и распределения признака в ГС не является ***симметричным***.

Если $\frac{|A_s|}{\sigma_{A_S}} < 3$, асимметрия несущественна, ее наличие объясняется влиянием случайных факторов.





Исследование формы распределения

ДЛЯ СИММЕТРИЧНЫХ РАСПРЕДЕЛЕНИЙ РАССЧИТЫВАЮТ ПОКАЗАТЕЛЬ **ЭКСЦЕССА** ***(kurtosis),*** характеризующего **крутизну** вершины (островершинность).

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

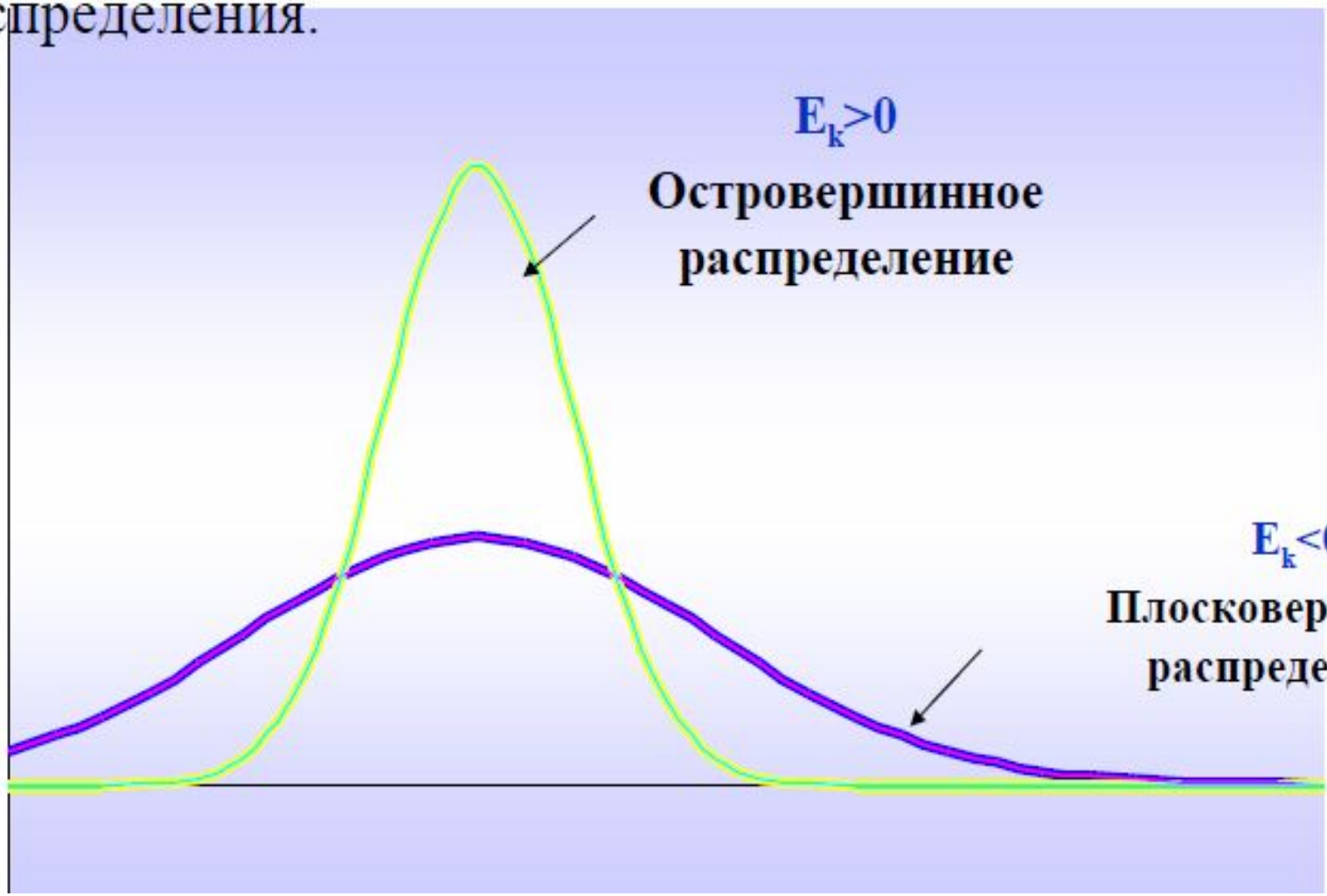
Для симметричных распределений $E_k=0$

(в нормальном распределении крутизна вершины, равная нулю, взята за эталон).

в случае островершинности распределения $E_k>0$,

в случае плосковершинности распределения $E_k<0$.

Коэффициент эксцесса E_k – показатель, служащий мерой крутости (плосковершинности или островершинности) графика вариационного ряда в сравнении с кривой нормального распределения.



Исследование формы распределения

Средняя относительная ошибка эксцесса вычисляется по формуле:

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

Характеристики положения

Считается, что распределение с эксцессом и асимметрией в диапазоне от -1 до +1 приблизительно соответствует нормальному распределению.

В большинстве случаев вполне допустимо считать нормальным распределение с асимметрией и эксцессом по модулю не превосходящими 3 (более мягкое правило).

Закон распределения

Закон распределения	*Характеристики*
Симметричное	$\bar{x} = Mo = Me;$ $A_s=0$
Нормальное	$\bar{x} = Mo = Me;$ $A_s= 0; E_k = 0.$
Правосторонняя асимметрия	$Mo > Me > \bar{x}$
Левосторонняя асимметрия	$\bar{x} > Me > Mo$
Равномерное	$A_s=0, E_k=-1,2$
Экспоненциальное	$A_s=2, E_k=9$

Диагностика выбросов (outliers)

Анализ выбросов очень важен, так как позволяет увидеть , что какой-то объект является нетипичным, необычным. Когда мы контролируем какой-то процесс, то такая информация является сигнальной.

Нахождение выбросов базируется на

- **среднем значении**
- **медиане**.

Диагностика выбросов (outliers)

- **Диагностика с использованием среднего значения**

  Определяют сколько стандартный отклонений от точки до среднего значения.

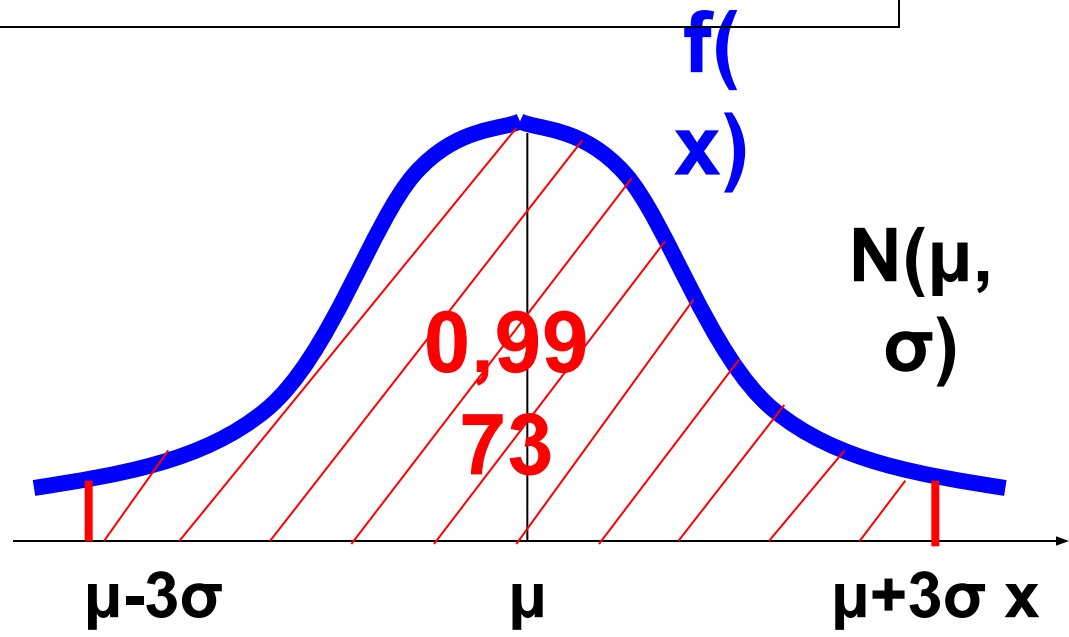
  Часто определяют, что выброс – это точка, которая отстоит от среднего значения белее, чем на 2σ или 3σ.

В случае симметричного распределения (НЗР) только 5% точек (2σ) и 0,3 % точек (3σ) имеют вероятность попасть в выбросы.

Нормальный закон распределения

Правило «68-95-99,7»

Если случайная величина X имеет нормальный закон распределения X є N(μ,σ), то практически достоверно, что её значения заключены в интервале (μ-3σ; μ+3σ) (Вероятность «выброса» составляет 0,0027)



Диагностика выбросов (outliers)

- **Диагностика выбросов с использованием медианы**

Правило 1,5 IQR *(1,5 IQR rule) - «мягкое правило»*

- *IQR* (**IQR=Q3-Q1)**
- Multiply *IQR by 1,5*
- *Find Q1-1,5 (IQR) and Q3+1,5(IQR)*
- *Any value below Q1-1,5 (IQR) or above Q3+1,5(IQR) is an outlier*

Диагностика выбросов (outliers)

Правило 1,5 IQR *(1,5 IQR rule)*

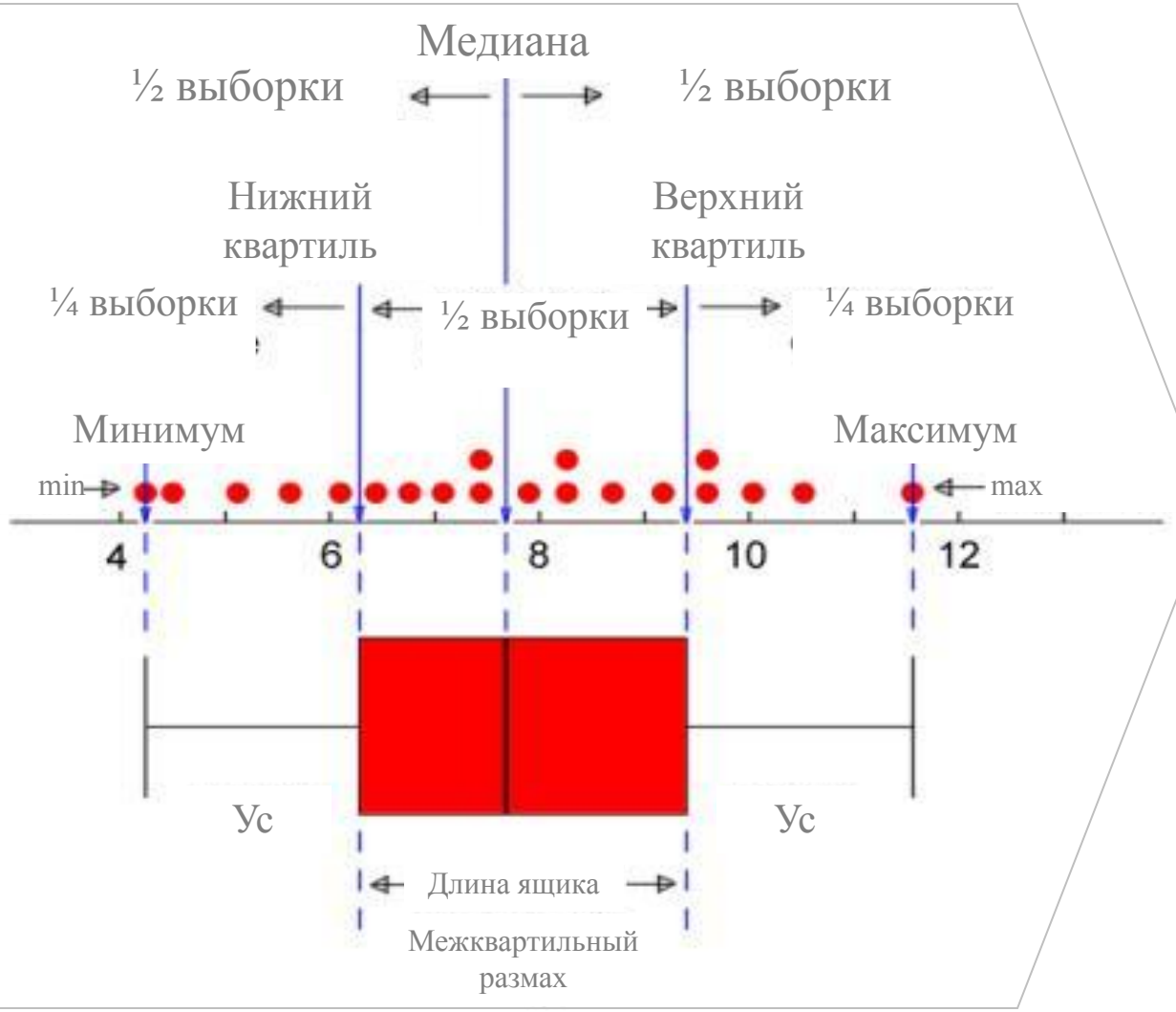
- *IQR* (**IQR=Q3-Q1)**
- Multiply *IQR by 1,5*
- *Find Q1-1,5 (IQR) and Q3+1,5(IQR)*
- *Any value below Q1-1,5 (IQR) or above Q3+1,5(IQR) is an outlier*

Правило ***3 IQR*** *(3 IQR rule)* :

Выброс или экстремальное значение в том случае, если наблюдение отличается от *Q1 и Q3 более, чем на* ***три IQR***.

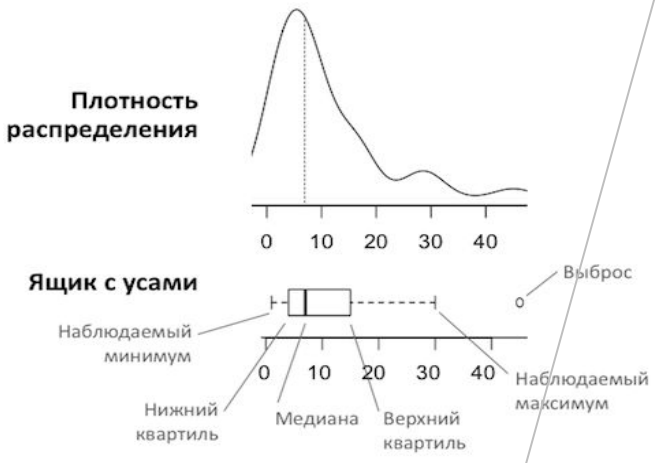
«Ящик с усами» или box-plot используется в описательной статистике и показывает 5 статистик выборки

Минимум 1	Нижний квартиль 2	Медиана 3	Верхний квартиль 4	Максимум 5



«Ящик с усами» может быть построен в любой ориентации! Большинство стат. пакетов по умолчанию используют **вертикальную**

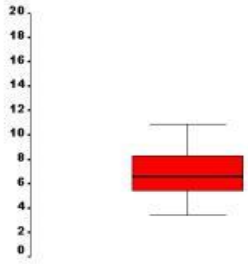
Связь с плотностью распределения



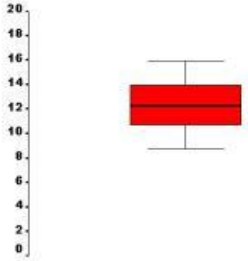
«Ящик с усами» выступает как индикатор 4-х характеристик выборки

Центрированность | Разброс | Симметричность | Размер хвоста

Центрированность

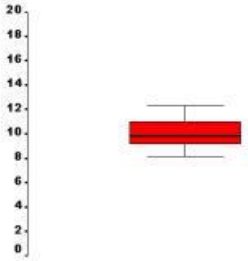


Бокс-плот выборки из 20 наблюдений с серединой – 7

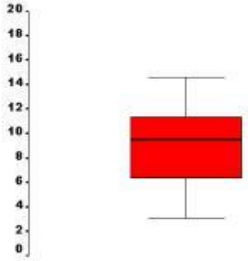


Бокс-плот выборки из 20 наблюдений с серединой –12

Разброс

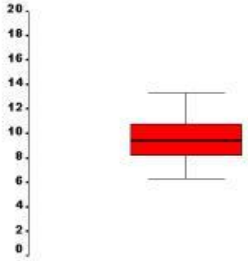


Бокс-плот выборки из 20 наблюдений с серединой в 10 и станд.отклон 1

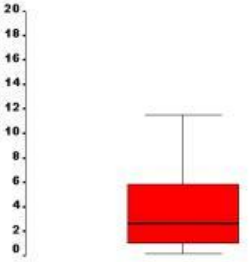


Бокс-плот выборки из 20 наблюдений с серединой в 10 и станд.отклон 3

Симметричность

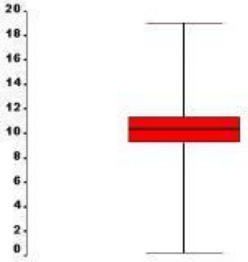


Бокс-плот выборки из 20 наблюдений с симметричным распределением

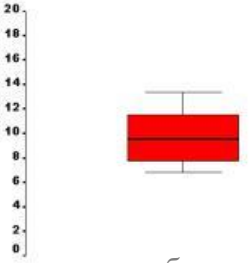


Бокс-плот выборки из 20 наблюдений с распределением скошенным направо

Размер хвоста

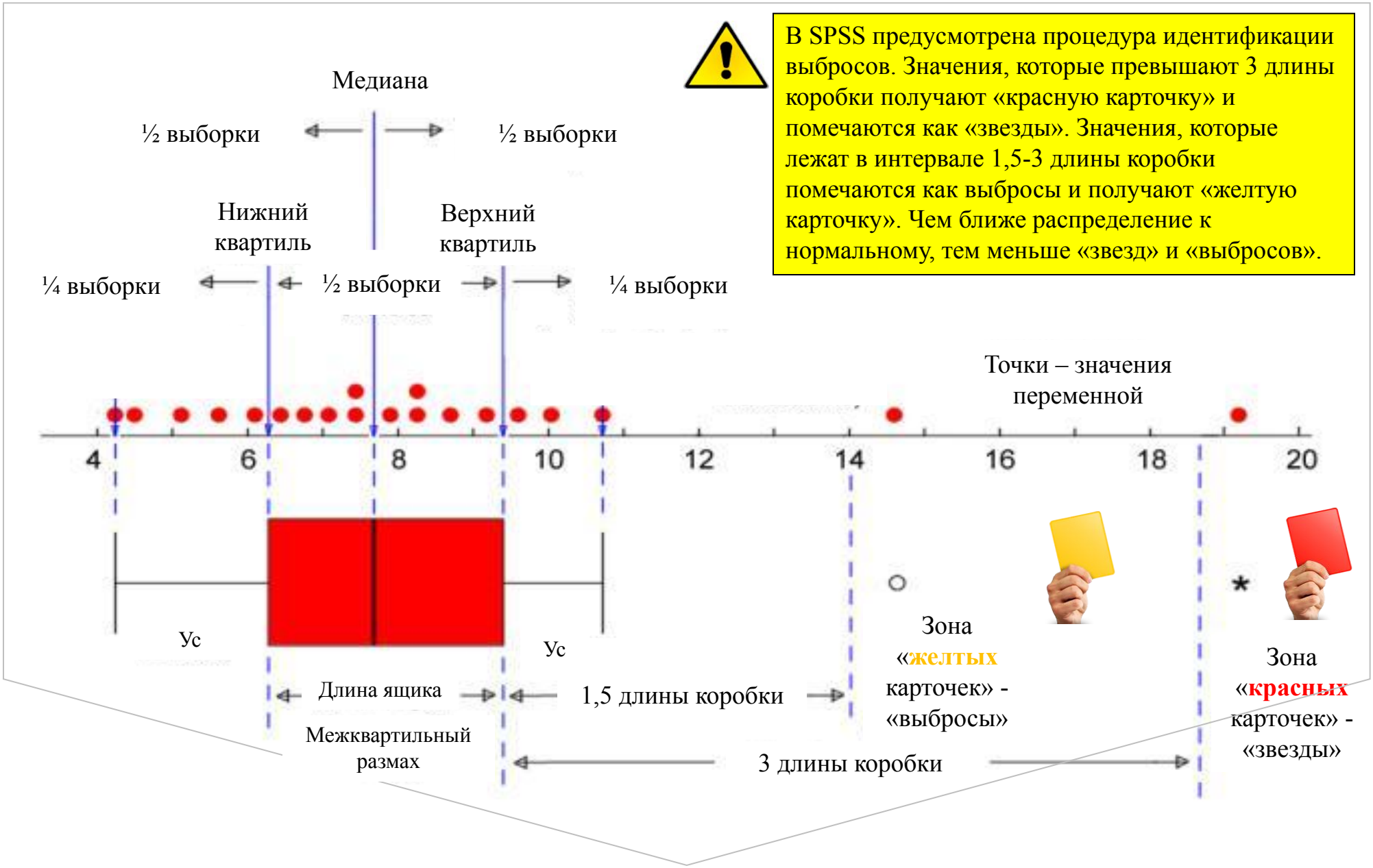


Бокс-плот выборки из 20 наблюдений с длинным хвостом



Бокс-плот выборки из 20 наблюдений с коротким хвостом

«Ящик с усами» также позволяет диагностировать наличие выбросов



В SPSS предусмотрена процедура идентификации выбросов. Значения, которые превышают 3 длины коробки получают «красную карточку» и помечаются как «звезды». Значения, которые лежат в интервале 1,5-3 длины коробки помечаются как выбросы и получают «желтую карточку». Чем ближе распределение к нормальному, тем меньше «звезд» и «выбросов».

Построение графика в Excel происходит в 3 этапа

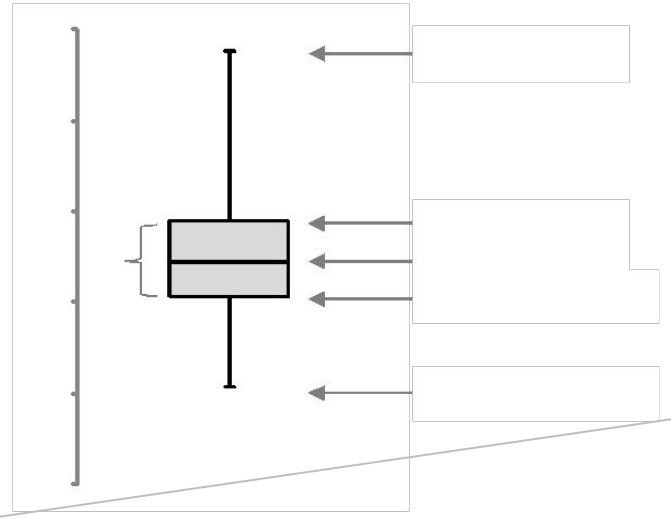
1. Вычисление необходимых параметров для графика
2. Выбор подходящей диаграммы
3. Редактирование диаграммы

Налоговое бремя в различных странах 2014

Страна	Total tax rate (% of commercial profits)
World	40,9
European Union	41,9
Russian Federation	48,9
United States	43,8
Germany	48,8
Italy	65,4
United Kingdom	33,7
Japan	51,3
China	64,6
Macedonia, FYR	21,4
Comoros	95,2

Five-Number Summary

Промежуточные вычисления	
Минимум	21,4
1-ый квартиль	41,4
Медиана	48,8
3-ий квартиль	58,0
Максимум	95,2
Блок 1	41,4
Блок 2	7,4
Блок 3	9,2
Ус 1	37,3
Ус 2	20,0



Z-преобразование

Определение позиции точки в распределении → на сколько стандартных отклонений она выше или ниже среднего значения.

Это позволяет сделать Z-преобразование (z-score).

$$z_{x_i} = \frac{x_i - \bar{x}}{s}$$

$$z_{x_i} > 0, \ если \ x_i > \bar{x}; \quad z_{x_i} < 0, \ если \ x_i < \bar{x}$$

Например: если z_3=1,5- это означает,

что 3 на $1{,}5s > \bar{x}; \quad z_3 = -2, \ то \ это \ означает, \ что \ 3 \ на \ 2s < \bar{x}$

Пример Петр сдал тест на 68. при этом средняя оценка для группы составляет 73, при s=3. Определить Z-преобразование для Петра

$$z_{68} = \frac{68 - 73}{3} = -1{,}67$$

Оценка Петра на 1,67s меньше средней оценки в группе.

Пример: данные о прибыли 100 фирм города Москвы

i	x_i	m_i	$m_i^н$
1	97	3	3
2	103	7	10
3	109	11	21
4	115	20	41
5	121	28	69
6	127	19	88
7	133	10	98
8	139	2	100

- **Мода:** x = 121
- **Медиана**: x = 121

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i m_i =$$

$$= \frac{1}{100}\left(97 \cdot 3 + 103 \cdot 7 + \ldots + 139 \cdot 2\right) = 119{,}2$$

Пример: данные о прибыли 100 фирм города Москвы

i	x_i	n_i
1	97	3
2	103	7
3	109	11
4	115	20
5	121	28
6	127	19
7	133	10
8	139	2

Дисперсия:

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 =$$

$$= \frac{1}{100}\left((97-119{,}2)^2 \cdot 3 + (103-119{,}2)^2 \cdot 7 + \ldots + (139-119{,}2)^2 \cdot 2\right) =$$

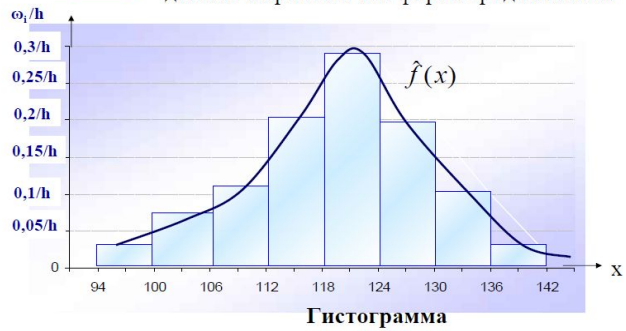
$$= 87{,}48$$

$$S = \sqrt{87{,}48} = 9{,}35$$

Коэффициент вариации:

$$\tilde{v} = \frac{9{,}35}{119{,}2} \cdot 100\% = 7{,}84(\%)$$

Пример данные о прибыли 100 фирм города Москвы



Гистограмма

i	x_i	n_i
1	97	3
2	103	7
3	109	11
4	115	20
5	121	28
6	127	19
7	133	10
8	139	2

$$\bar{x} = 119{,}2; \quad S = 9{,}35$$

$$A_s = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{S^3} =$$

$$= \frac{\frac{1}{100}\left((97-119{,}2)^3 \cdot 3 + (103-119{,}2)^3 \cdot 7 + \ldots + (139-119{,}2)^3 \cdot 2\right)}{9{,}35^3} =$$

$$= -0{,}3$$

$$E_k = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^4}{S^4} - 3 =$$

$$= \frac{\frac{1}{100}\left((97-119{,}2)^4 \cdot 3 + (103-119{,}2)^4 \cdot 7 + \ldots + (139-119{,}2)^4 \cdot 2\right)}{9{,}35^4} - 3 =$$

$$= -2{,}999$$

Вывод: незначительная левосторонняя асимметрия, плосковершинное распределение

Непрерывные количественные данные

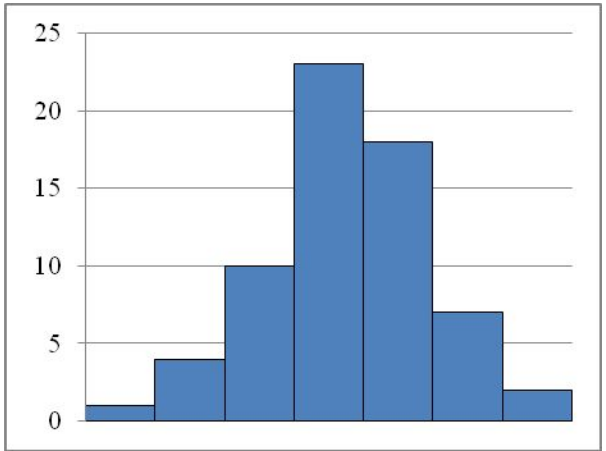
- Если исследуемый признак имеет ***непрерывный*** характер, то необходимо выбрать оптимальное число ***интервалов группировки*** признака.
- Для группировки *непрерывных* случайных величин весь вариационный размах признака $R=x_{(n)}-x_{(1)}$ разбивают на некоторое количество интервалов $\boldsymbol{k}$.
- ***Сгруппированным интервальным (непрерывным) вариационным рядом*** *называют ранжированные по значению признака интервалы* $(a_i \le x < b_i)$, *где* $i=1,2,\ldots k$, *указанные вместе с соответствующими частотами* (m_i) *числа наблюдений, попавших в i-й интервал, или относительными частотами* (m_i/n).

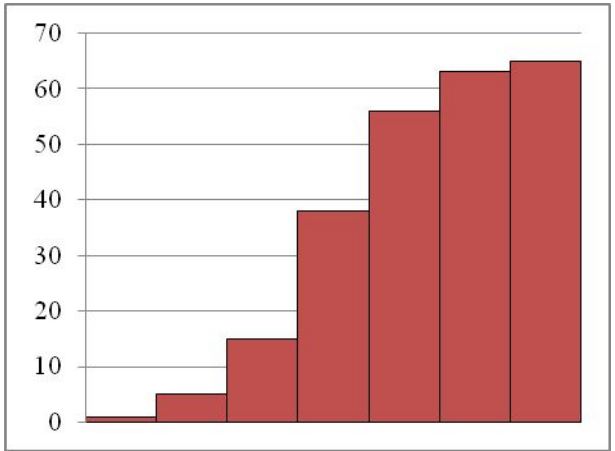
Интервалы значений признака $a_i \div b_i$	$a_1 \div b_1$	$a_2 \div b_2$	…	$a_i \div b_i$	…	$a_k \div b_k$
Частота m_i	m_1	m_2	…	m_i	…	m_k

Непрерывные количественные данные

Гистограмма и *кумулята (огива)* строятся для непрерывных данных так же, как и для дискретных, только с учетом того, что непрерывные данные сплошь заполняют область своих возможных значений, принимая любые значения.

- Высота столбика соответствует частоте m_i – числу наблюдений, попавших в данный интервал, или относительной частоте m_i/n – доле наблюдений. Интервалы не должны пересекаться, и должны, как правило, иметь одинаковую ширину.
- Гистограмма и кумулята являются эмпирическими оценками функций плотности вероятности и функции распределения СВ.





СПАСИБО ЗА ВНИМАНИЕ !

Основные выборочные характеристики

- выборочная (эмпирическая) функция распределения $\hat{F}_n(x)$
- выборочная (эмпирическая) функция плотности $\hat{f}_n(x)$
- выборочная (эмпирическая) относительная частота появления i-го возможного значения дискретной случайной величины $\hat{w}_i$
- выборочные начальные и центральные моменты анализируемой случайной величины: $\hat{\nu}_i;\ \hat{\mu}_i$
  - выборочное среднее значение $\bar{x} = \hat{\nu}_1$
  - выборочная дисперсия $s^2 = \hat{\mu}_2$
- Показатели формы распределения (ассиметрия, эксцесс)

Основные выборочные характеристики

Эмпирическая (или выборочная, т. е. построенная по выборке объема n) **функция распределения:**

$$\overline{F}_n(x) = \frac{m_x}{n}, \qquad \overline{F}_n(x) = \frac{m_1 + m_2 + \ldots + m_{ix}}{n}$$

По сгруппированным данным

где m_x - число наблюдаемых значений исследуемой случайной величины в выборке $x_1, x_2, \ldots, x_n$, меньших x;

m_i - число наблюдаемых значений в выборке, попавших в i-й интервал группирования,

ix - номер самого правого из интервалов группирования, правый конец которых не превосходит x.

Основные выборочные характеристики

Выборочная (эмпирическая) **относительная частота:**

$$\overline{w}_i = \frac{m_{x_i^0}}{n},$$

которая определяется как отношение числа $m_{x_i^o}$ наблюдений в выборке, равных x_i^0, к общему объему выборки n.

Накопленная частота m_i^H - сумма частот i-го и всех предшествующих интервалов.

Основные выборочные характеристики

Для построения эмпирической (выборочной) **функции плотности** на всей области ее определения (т,е, для всех возможных значений исследуемой величины) используют предварительно сгруппированные данные и полагают

$$\hat{f}(x) = \frac{m_{k(x)}}{n \cdot \Delta_{k(x)}},$$

где к(х) - порядковый номер интервала группирования, который накрывает точку х;

$m_{k(x)}$ - число наблюдений, попавших в этот интервал,

$\Delta_{k(x)}$ - длина интервала.

Геометрическое изображение эмпирической функции плотности наз. **гистограммой.**

ХАРАКТЕРИСТИКИ РАСПРЕДЕЛЕНИЯ

Расчет описанных характеристик является первым этапом анализа собранных статистических данных и позволяет

- Обосновать некоторые закономерности исследуемого процесса
- Выбрать статистический инструментарий

Задание

Распределение предприятий по региона по величине розничного товарооборота в текущем году.

№ п/п	Розничный товарооборот, тыс.руб.	№ п/п	Розничный товарооборот, тыс.руб.
1.	151331	16.	21253
2.	56440	17.	47248
3.	99212	18.	92955
4.	34088	19.	178291
5.	43520	20.	68865
6.	38196	21.	9767
7.	208492	22.	60674
8.	104518	23.	23944
9.	82972	24.	127725
10.	45561	25.	24559
11.	137445	26.	21946
12.	28970	27.	44876
13.	51387	28.	117021
14.	156775	29.	33775
15.	65680	30.	36637