Измерения в психологии. Шкалы измерений. Способы представления статистических данных.

лекция № 10 для студентов 1 курса, обучающихся по специальности 37.05.01– Клиническая психология

© Составитель: доцент кафедры медицинской и биологической физики, к. ф.-м. н. Романова Н.Ю.

План лекции:

- **Задачи математической статистики.**
- **Дискретные и интервальные ряды распределения. Числовые характеристики.**
- **Точечные и интервальные оценки основных числовых характеристик генеральной совокупности по выборочным данным.**

Существуют три вида лжи: ложь, наглая ложь и статистика.

Бенджамин Дизраэли

Придет время, и статистическое мышление станет таким же необходимым качеством для истинного гражданина, как умение читать и писать.

Герберт Уэллс

Нам досаждают не столько те вещи, о которых мы не знаем, сколько те, о которых мы знаем, что с ними что-то не так.

Артемус Уорд

Статистика изучает большие массивы информации и устанавливает закономерности, которым подчиняются случайные массовые явления.

Типы задач математической статистики:

Точечное оценивание – по результатам наблюдений выбрать значение параметра, которое оптимальным образом согласуется с данными.

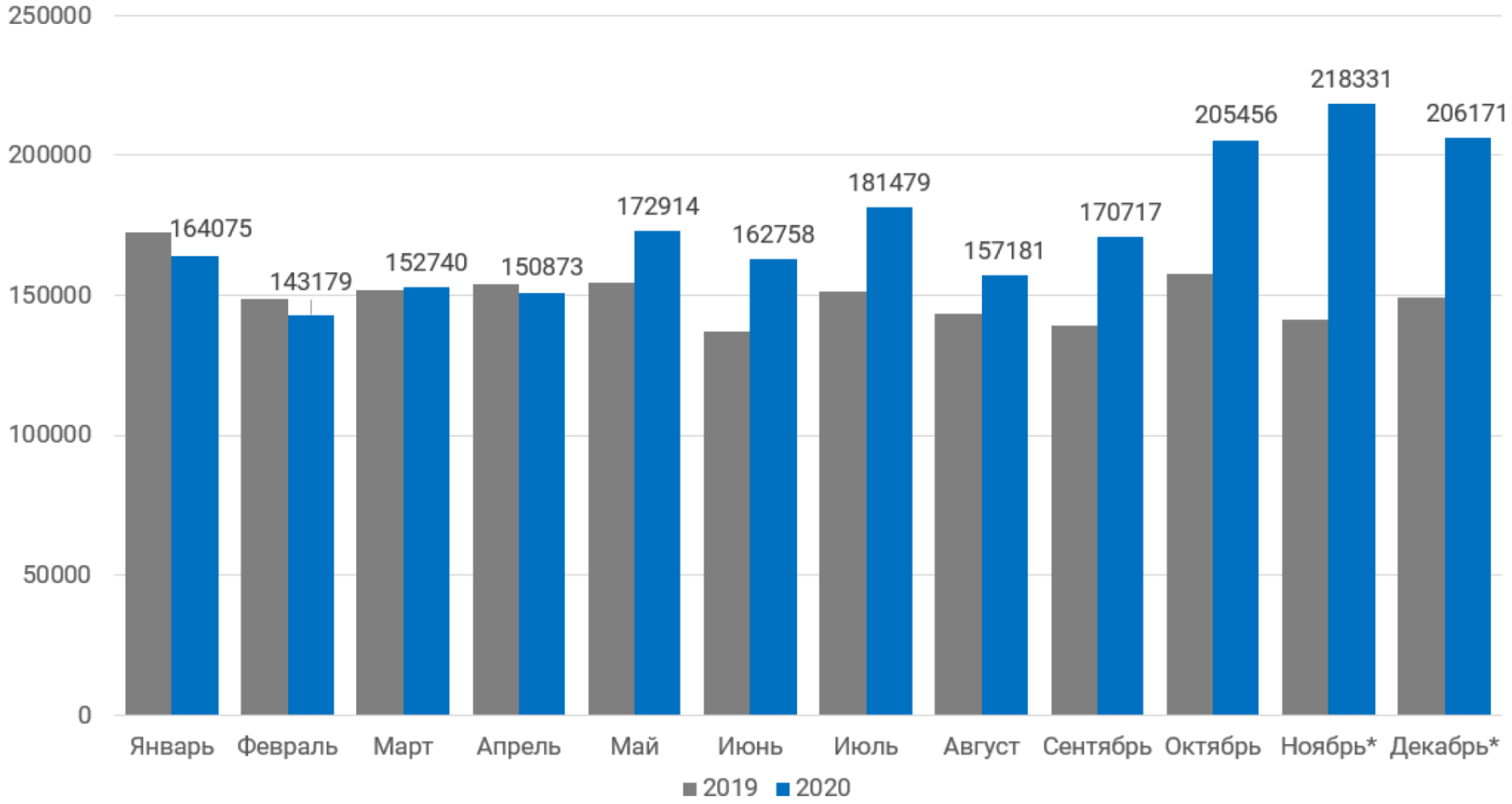
Интервальное оценивание- по результатам наблюдений выбрать область, т.е. выбор такого множества, которое накрывает теоретическое значение параметра с вероятностью не меньше (1-α).

Проверка статистических гипотез– по результатам наблюдений выбрать из нескольких взаимоисключающих гипотез H_1… H_n наиболее подходящую.

Цель математической статистики: С помощью статистических методов по свойствам выборки сделать вывод о свойствах генеральной совокупности

ОТКРЫТЫЕ МЕДИА

ЧИСЛО УМЕРШИХ В РОССИИ В 2019 И 2020 ГОДАХ



* Данные за ноябрь и декабрь 2020 года - оценка «Открытых медиа»

Источник: расчёты «Открытых медиа» и Росстат

Генеральной совокупностью (ГС) называется вся подлежащая изучению какого-либо свойства (говорят, признака) совокупность объектов.

Та часть объектов, которая отобрана для непосредственного изучения какого-либо признака ГС носит название **случайной выборки** (или просто **выборки**).

Объекты этих совокупностей обладают интересующим нас признаком - X, или признаками X, Y, Z…

Объем ГС (обозначается N) и **объем выборки (обозначается n)** – это количество элементов в них.

ГС может иметь разный объем, иногда небольшой, сопоставимый с объемом выборки. В этом случае в получаемых формулах присутствует влияние соотношений объемов в виде дроби n/N $(n \leq N)$.

Если же N велико по сравнению с n, то дробь n/N становится малой величиной и в формулах эта дробь не учитывается. Такая ситуация наиболее типическая.

КАК НАМ ИСПОЛЬЗОВАТЬ
ВЫБОРКУ...

Следует делать такую случайную выборку, чтобы на основе ее изучения можно было выносить обоснованное заключение о поведении изучаемого признака во всей ГС. Такая выборка называется **репрезентативной** (представительной) выборкой.

Репрезентативная выборка - выборка, адекватно отражающая особенности генеральной совокупности.

Для того, чтобы выборка была репрезентативной, она должна быть отобрана случайно. Случайность отбора элементов в выборку достигается соблюдением принципа **равной возможности каждого элемента ГС быть отобранным в выборку** (лотерея, датчик случайных чисел).

*По **объему** выборки делятся на*

малые - $n < 30$

средние - $30 \leq n < 100$

большие – $n \geq 100$

Выборочная совокупность может быть сформирована с помощью следующих ***методов отбора*:**

1. *случайный*;
2. *механический*;
3. *типический* (расслоенный, стратифицированный);
4. *серийный* (гнездовой);
5. комбинированный.

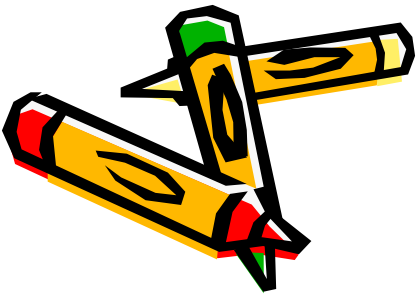
Шкалы:

- Отношений
- Интервалов
- Рангов (порядковая)
- Наименований (номинальная)

метрические (количественные)

качественные

Первичная обработка выборочных данных



В качестве ***первичной обработки статистических данных*** рассматриваются следующие операции:

- определение объема выборки - ***n***;
- ранжирование выборочных данных;
- определение максимального X_{max} и X_{min} минимального элементов, размаха выборки - ***R***;
- распределение выборочных данных по частотам (получение ***дискретного или интервального*** вариационного ряда и статистического распределения выборки)
- графическое представление данных.

Ранжированием называется упорядочивание элементов выборки по возрастанию.

n – объем выборки – количество опытных данных в выборке.

X_i – значение выборочной варианты, таких значений может быть несколько в одной выборке.

n_i – частота i-й варианты (число повторений *i*-й варианты в случайной выборке).

w_i – относительная частота i-й варианты в выборке (статистическая вероятность):

$$w_i = \frac{n_i}{n}$$

R – размах вариации – рассчитывается по максимальному и минимальному значению выборки: *Xmax - Xmin*

Дискретный вариационный ряд представляет собой таблицу соответствия n_i и X_i. Такое представление удобно для дальнейшего построения *полигона частот*:

X_i	X_1	X_2		...	X_k
n_i	n_1	n_2		...	n_k

Сумма частот Σn_i равна объему выборки n.

Статистическое распределение выборки – **соответствие вариант и их относительных частот:**

X_i	X_1	X_2		...	X_k
w_i	w_1	w_2		...	w_k

w_i иногда называют *статистической вероятностью* и обозначают p_i. Сумма частот Σw_i равна 1.

Графическое представление соответствия n_i (или w_i) от X_i называется *полигоном частот*.

Пример 1.

Пусть у нас есть выборка данных по количеству выступлений отдельных студентов на семинарах за 1 семестр:

3, 1, 3, 1, 4, 2, 7, 2, 5, 0,
2, 2, 4, 0, 3, 4, 2, 0, 2, 3,
1, 2, 0, 2, 2, 1, 0, 3, 4, 1,
2, 0, 2, 1, 4, 2, 1, 1, 2, 1,
2, 5, 1, 3, 2, 1, 1, 5, 1, 3,
3, 3, 3, 1, 4, 5, 1, 1, 2, 3.

Построить вариационный выборочный ряд и полигон частот данной выборки.

Решение.

Для начала определим объем выборки – $n = 60$ – и ранжируем выборочные данные:

0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 7.

Для удобства их можно расположить следующим образом:

0, 0, 0, 0, 0, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
4, 4, 4, 4, 4, 4,
5, 5, 5, 5,
7

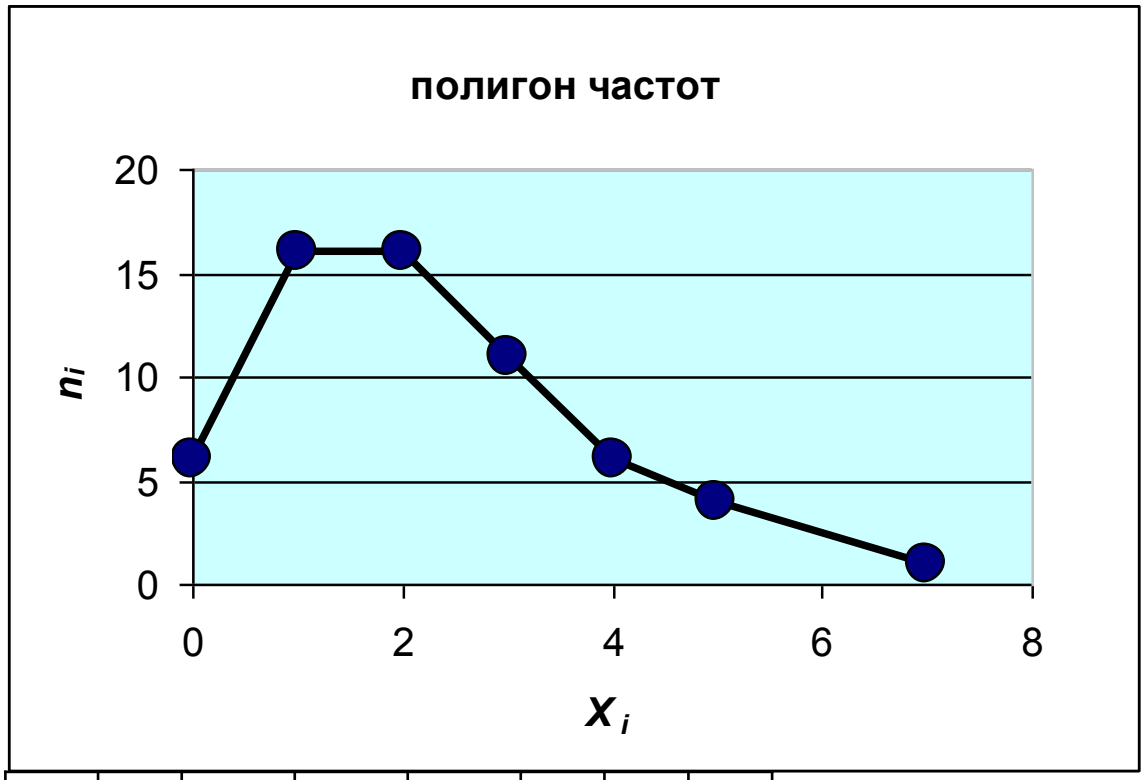
Итак, варианта может принимать всего 7 значений: *0,1,2,3,4,5,7.*

Теперь произведем группировку вариант в вариационный ряд:

X_i	0	1	2	3	4	5	7
n_i	6	16	16	11	6	4	1

$\Sigma n_i = 60$ – сумма частот.

Теперь построим полигон частот (точечную диаграмму):



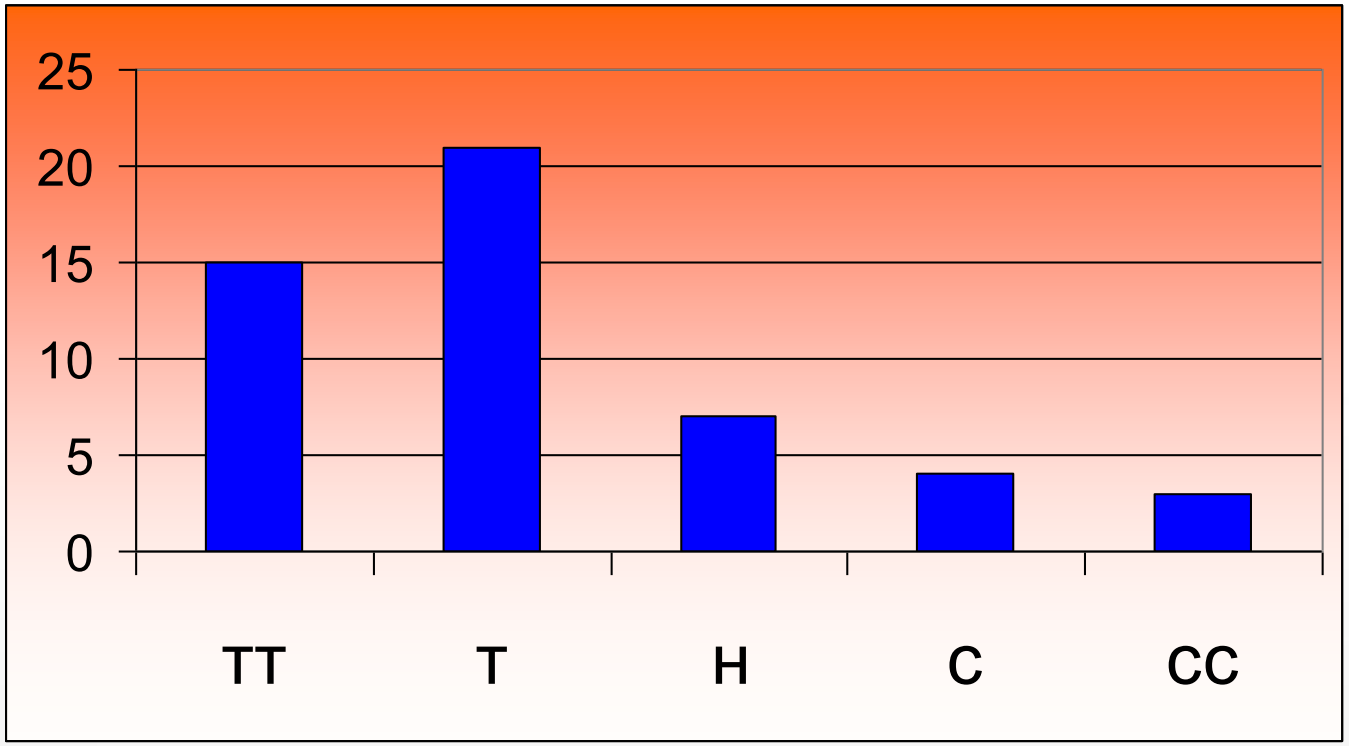
X_i	0	1	2	3	4	5	7
n_i	6	16	16	11	6	4	1

Статистические данные могут быть измерены в качественной шкале, например:

т, тт, с, с, сс, н, т, т, тт,. ....

X_i	тт	т	н	с	сс
n_i	15	21	7	4	3

по ним тоже может быть построена диаграмма:



Для случаев непрерывных статистических данных (число значений X_i велико или признак имеет непрерывную природу) более удобно представление в виде *интервального статистического ряда*, которое предполагает разбиение на классы: частоты подсчитываются не для отдельных значений вариант, а для интервалов, на которые разбивается весь интервал значений выборки. Графическое представление зависимости частоты вариант в каждом таком интервале называется ***гистограммой частот***.

Как определить количество классов?

1. По ***формуле Стерджесса***:

$$k = 1 + 3{,}322 \lg (n)$$

2. По таблице:

Объем выборки, n	Число интервалов, k
25-40	5-6
40-60	6-8
60 - 100	7-10
100-200	8-12
>200	10-15

Иногда прибегают к его оценке, как «корня квадратного»:

$$k = \sqrt{n}$$

Алгоритм составления интервального распределения выборки

1. Учитывая объем выборки *n*, **вычисляют количество классов (интервалов) – *k*.**

2. **Определяют величину класса (интервала):**

$$h_i = \frac{x_{\max} - x_{\min}}{k}$$

при необходимости его можно округлить, исходя из характера выборки.

3. **Определяют границы классов**. Границей первого класса можно выбрать:

• минимальное значение

• так, чтобы он содержал наименьшее значение, но не начинался с него, например, класс может начинаться с величины (X_{min} – $h/2$)

•крайние интервалы могут быть *полубесконечными*.

Последующие классы образуются добавлением величины интервала h_i.

4. Подсчитывают количество выборочных значений признака, попавших в каждый класс (интервал)

Если значение варианты совпадет с границей какого-либо класса:

- это значение следует отнести к тому или иному классу - по предварительной договоренности
- добавить обоим классам по ½ к частоте
- заранее выбрать границы интервалов так, чтобы они не могли совпасть с выборочными значениями.

5. Строится гистограмма частот.

Полигон и гистограмма частот являются наиболее употребимыми *аналитическими графиками математической статистики*.

Пример:

Дана выборка из 30 элементов (x_1 x_2 … x_{30}):
151 166 133 155 179 148 143 128 138 172 163 157
158 136 169 153 142 147 134 164 167 131 152 145
176 122 149 154 161 156

Объем выборки - 30

Требуется построить (вычислить):
- группированный статистический ряд
- построить гистограмму частот;

$$k = 1 + 3,322 \cdot \lg 30 = 5.9 \approx 6$$

К среднему

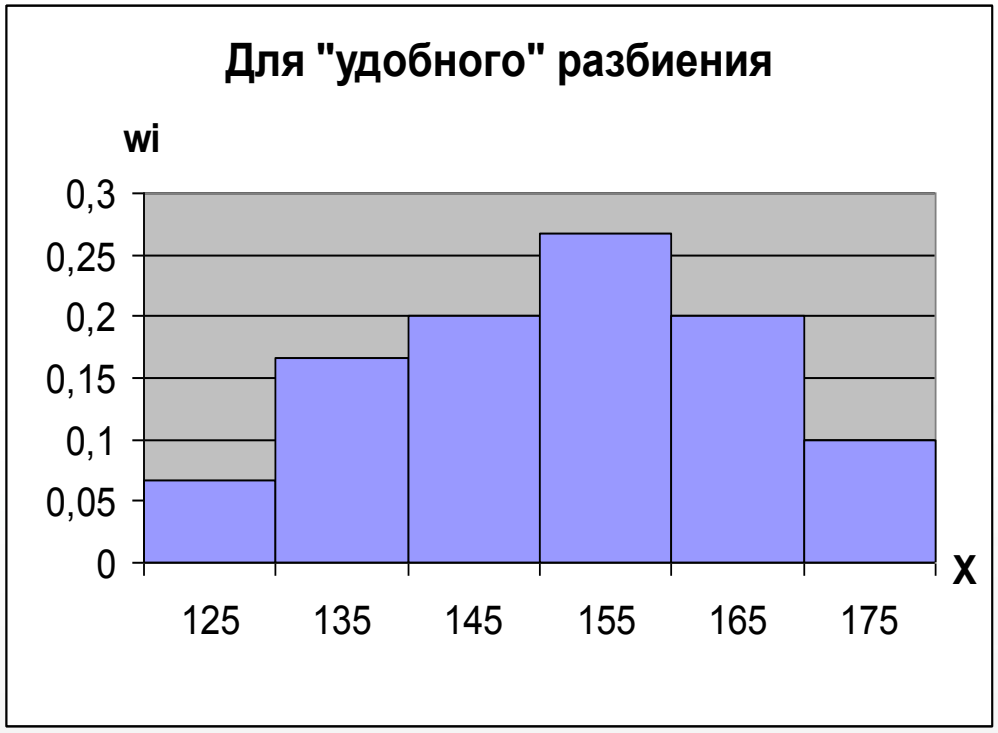
Вариант разбиения - удобный:

В данном случае округлили значение ширины интервала до 10, и выбрать соответствующие «ровные» границы классов (интервалов):

Номер интервала j	**Границы интервала**		**Центр интервала** x_j	**попало в интервал** n_j	**Относительная частота интервала** $w_i=n_j/n$
1	120	130	125	2	0,07
2	130	140	135	5	0,17
3	140	150	145	6	0,20
4	150	160	155	8	0,27
5	160	170	165	6	0,20
6	170	180	175	3	0,10

Гистограмма частот — это фигура, состоящая из прямоугольников, опирающихся на интервалы группировки. Высота ***j***-го прямоугольника полагают равной n_j - частоте соответствующего интервала, или n_j/n, - относительной частоте данного интервала - w_i.

Данные гистограммы построены для относительных частот



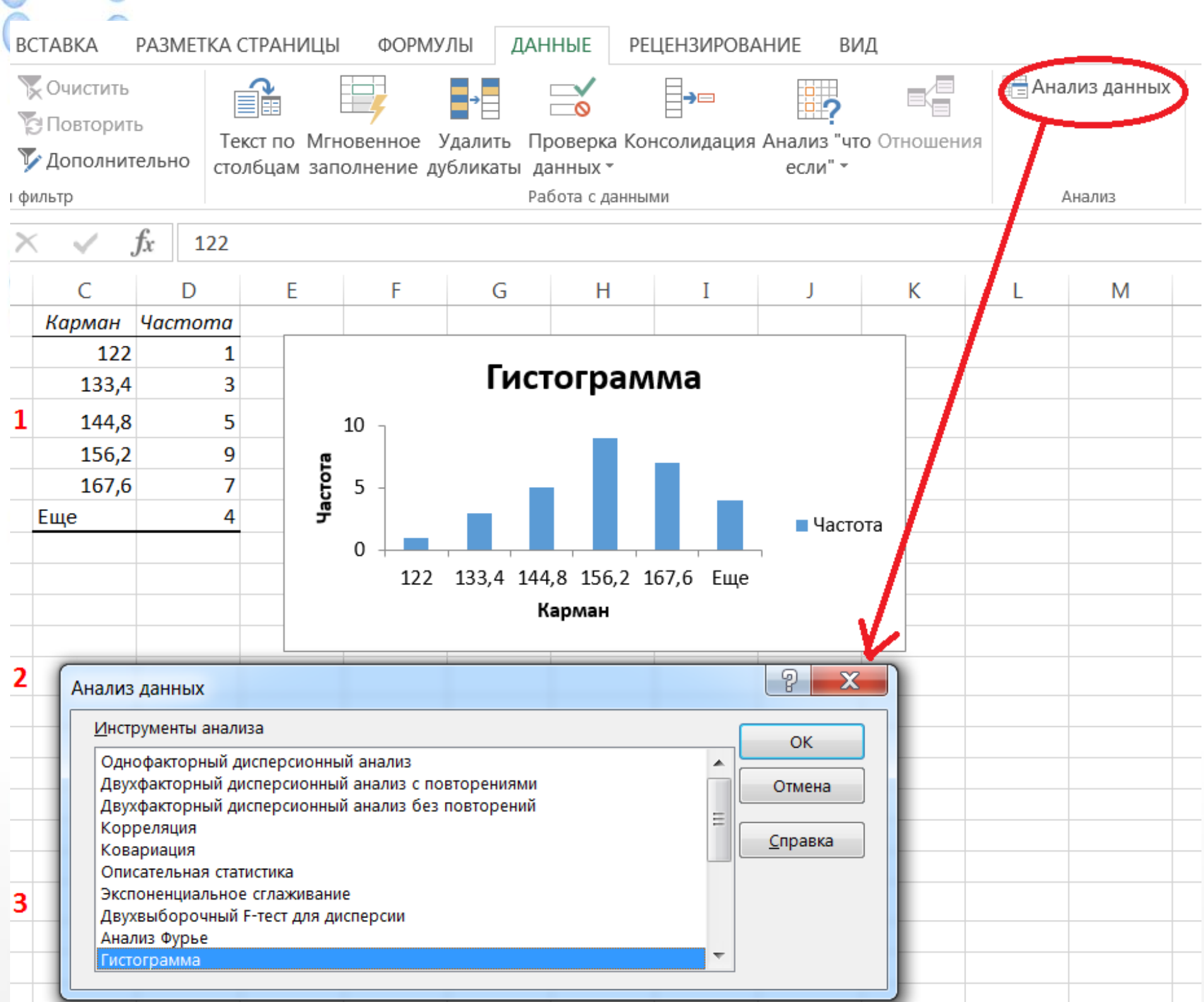
Построение гистограммы в Excel с помощью статистического пакета **Анализ данных**

Выберите из меню команду: Данные→Анализ данных→Гистограмма.

В появившемся диалоговом окне в области Входные данные укажите Входной интервал. Если в таблице данных есть «шапка», установите флажок «Метки».

В области «Параметры вывода» установите флажок *Вывод графика* и включите параметр «Выходной интервал»: для вывода гистограммы на том же листе, укажите в поле «Выходной интервал» адрес левой верхней ячейки желаемого диапазона.

Если не указывать параметр «Интервал карманов», определяющий ширину столбца, то Excel установит равные интервалы автоматически. Количество интервалов приближенно равно квадратному корню из числа значений данных.


ВСТАВКА
РАЗМЕТКА СТРАНИЦЫ
ФОРМУЛЫ
ДАННЫЕ
РЕЦЕНЗИРОВАНИЕ
ВИД
Очистить
Повторить
Дополнительно
фильтр
Текст по столбцам
Мгновенное заполнение
Удалить дубликаты
Проверка данных
Консолидация
Анализ "что если"
Отношения
Работа с данными
Анализ данных
Анализ
122
C
D
Карман
Частота
122 1
133,4 3
144,8 5
156,2 9
167,6 7
Еще 4
1
Гистограмма
10
5
0
Частота
122
133,4
144,8
156,2
167,6
Еще
Карман
Частота
2
Анализ данных
Инструменты анализа
Однофакторный дисперсионный анализ
Двухфакторный дисперсионный анализ с повторениями
Двухфакторный дисперсионный анализ без повторений
Корреляция
Ковариация
Описательная статистика
Экспоненциальное сглаживание
Двухвыборочный F-тест для дисперсии
Анализ Фурье
Гистограмма
3
OK
Отмена
Справка

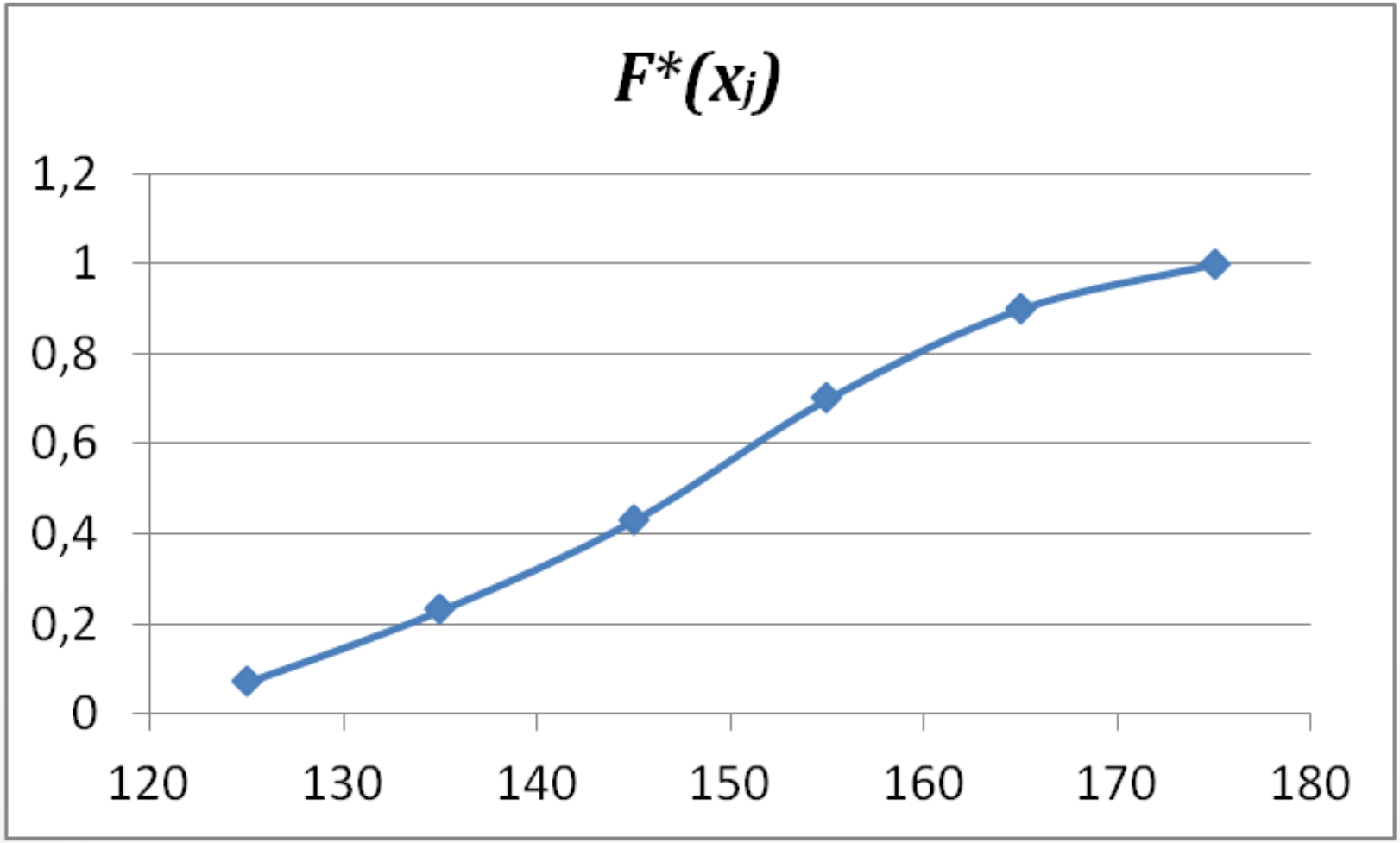
Эмпирической функцией распределения F*(x) называется относительная частота того, что случайная величина принимает значение меньше заданного (имеет смысл «накопленной частоты»):

$$F^*(x) = w(X<x) = w_x^{нак}$$

Номер интервала j	Интервал	n_j	$n_{накопл}$	$w_j=n_j/n$	$F^*(x_j)$
1	**(120; 130)**	2	2	0,0667	0,07
2	**(130; 140)**	5	2+5=7	0,1667	0,23
3	**(140; 150)**	6	7+6=13	0,20	0,43
4	**(150; 160)**	8	13+8=21	0,27	0,70
5	**(160; 170)**	6	21+6=27	0,20	0,90
6	**(170; 180)**	3	27+3=30	0,10	1

График накопленных частот называется *кумулята*.

На рисунке - интегральная функция распределения:



Конечно, представление вариационных рядов делает картину поведения статистических данных более наглядной.

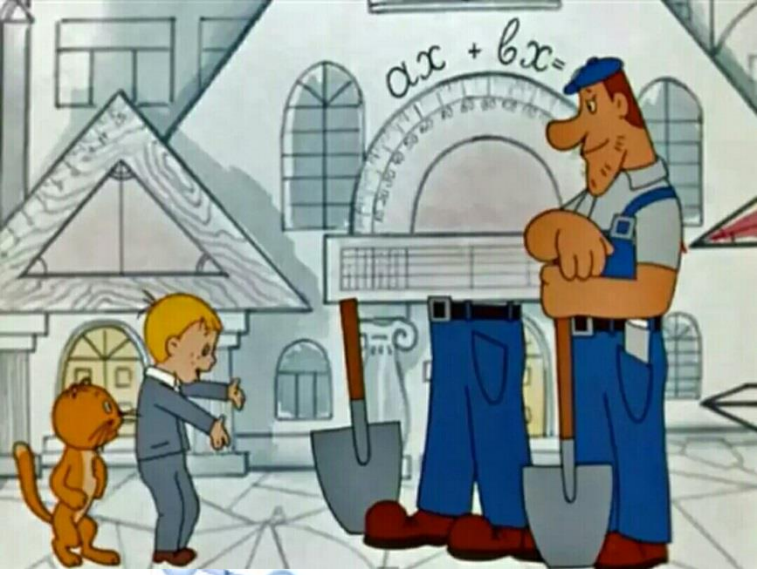
Точечные оценки

Следует дополнить вариационные ряды и их графическое изображение некоторыми сводными характеристиками вариационных рядов.

Эти обобщающие показатели в компактном виде характеризуют всю выборку (вариационного ряда) в целом. В таким обобщающим показателям относят:

1)Характеристики центральной тенденции - это средние величины, определяющие значения признака, вокруг которого концентрируются все его наблюдаемые значения;

2)Характеристики вариации (изменчивости) – это величины, определяющие колебания наблюдаемых значений признака.



1) **Наиболее распространенной из средних величин является <u>среднее арифметическое выборки:</u>**

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n}=\frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}=\sum_{i=1}^{k} x_i \cdot w_i$$

Если выборочные данные заданы в виде интервального ряда, то рекомендуется в качестве X_i выбирать середины соответствующих интервалов - $X_{i\ ср}$, а в качестве n_i - частоты интервалов.

Вычислим среднее арифметическое для примера:

$$\bar{x}=\frac{151+\ 166+\ \ldots\ +\ 156}{30}\ =\ \frac{4549}{30}\approx 151{,}6$$

Медиана – это значение признака, приходящееся на середину ранжированного ряда наблюдений.
Иначе: это то значение варианта, которое делит вариационный ряд на две равные по объему части.

Обозначение:

Теоретическое $Me(X)$;

Статистическое

$$\widetilde{Me}(X)$$

Если число вариант нечетное, т.е. n=2m+1 , то

$$\widetilde{Me}(X) = x_{m+1}$$

Если число вариант четное, т.е. n=2m , то

$$\widetilde{Me}(X) = (x_m + x_{m+1})/2$$

Для примера 1 Me=152,5. Это означает, что половина субъектов имеют оценки $\leq 152{,}5$, а другая половина имеют оценки $\geq 152{,}5$.

Простые примеры:

Найти медиану по выборкам:

А) 1,0; 1,3; 1,1; 1,2; 1,2; 1,4; 1,5; 1,2; 1,4

Ранжируем данный выборочный ряд:

1,0; 1,1; 1,2; 1,2; **1,2**; 1,3; 1,4; 1,4; 1,5

В середине ранжированного ряда находится значение **1,2** - это и есть медиана данной выборки.

Б) В выборке 66; 65; 63; 64; 68; 61; 69; 62; 66; 65 -10 значений (нечетное число), поэтому после ранжирования

61; 62; 63; 64; 64; 65; 66; 66; 68; 69

Медиану находим по формуле: (64+65)/2=64,5

В отличие от среднего арифметического, на которое оказывают влияние все значения x_i, медиана не зависит от крайних значений признака. Если в наблюдениях имеются выбросы, то медиана не реагирует на них, а среднее арифметическое может при этом очень сильно меняться.

Наглядной иллюстрацией может служить вычисление средней заработной платы. Если оценивать её как среднее арифметическое, то что реально отражает эта величина? Несомненно, на результат влияет немногочисленная группа очень высокооплачиваемых работников.
Более убедительной характеристикой реального финансового положения могла бы служить, например, медиана.

$45 000

$15 000

$10 000

← СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

$5700

$5000

$3700

← МЕДИАНА (этот парень будет посередине: дюжине сотрудникам платят больше, чем ему, а другой дюжине — меньше)

$3000

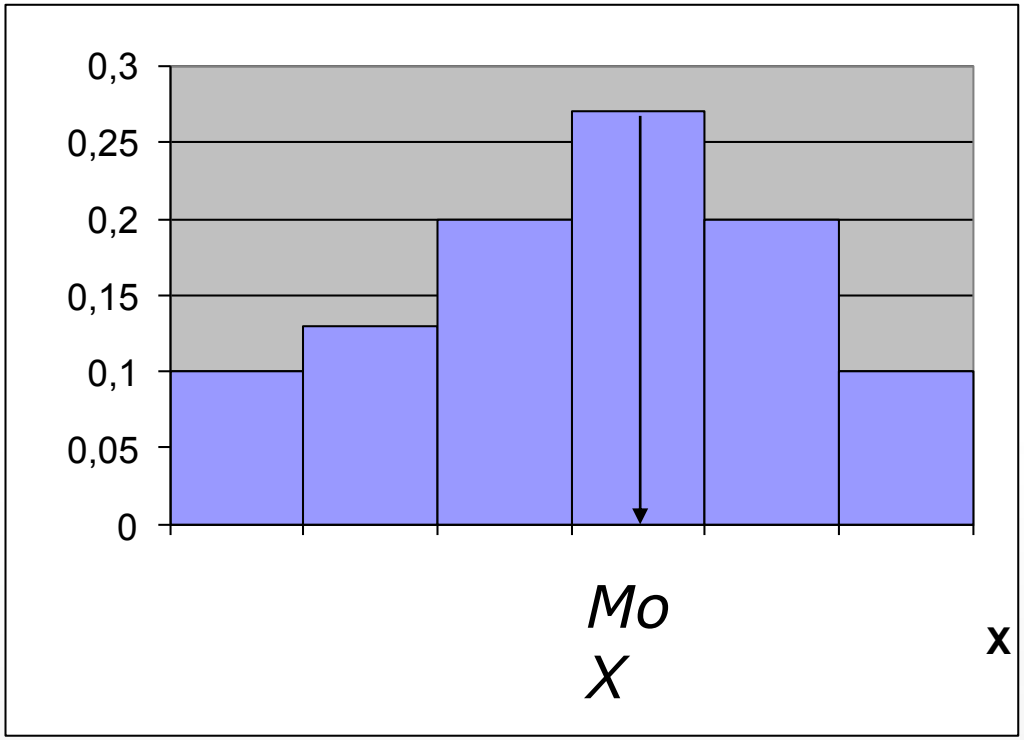
← МОДА (величина зарплаты, чаще всего встречающаяся в этой компании)

$2000

***Выборочной модой** называется значение признака, наиболее часто встречающееся в выборке.*

$\tilde{Mo}$

Иначе *мода* - то значение варианты, которому соответствует наибольшая частота.



Для того, чтобы найти значение моды в случае дискретного вариационного ряда, следует найти варианту с наибольшей частотой появления. Эту варианту и принимаем за моду.

В простом случае:

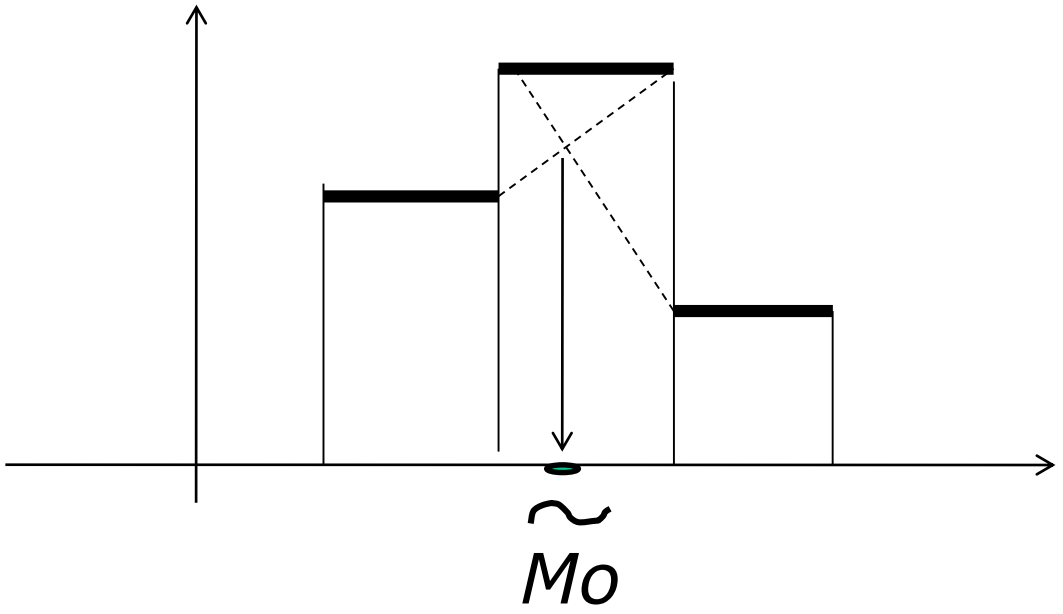
Дана выборка: 36, 35, 36, 37,5, 38, 38, 36,5, 38,5, 38, 40, 39, 39.

Статистическое распределение выборки значений размера ноги задано таблицей:

X_i	35	36	36.5	37.5	**38**	38.5	39	40
n_i	1	2	1	1	3	1	2	1

Варианта, которая наиболее часто встречается в данной выборке – это **38**, то есть *мода = 38*.

Для интервального ряда значение моды можно найти с помощью графического линейного интерполирования на гистограмме:



Если $X'_s - X''_s$ – модальный интервал, а интервалы вариационного ряда имеют постоянную ширину h, то мода изучаемого признака вычисляется:

$$Mo\ X = X'_s + h \cdot (n_s - n_{s-1}) / ((n_s - n_{s-1}) + (n_s - n_{s+1})),$$

где n_{s-1} , n_{s+1} – частоты, находящиеся в соответствии с интервалами, предшествующим модальному и следующим за ним. ***Модальный интервал выбирается по наибольшей частоте***.

Для интервального ряда из примера 2 (формальное разбиение на классы):

№ интервала	н.г.	в.г.	частота
1	122	131,5	3
2	131,5	141	4
3	141	150,5	6
4	150,5	160	8
5	160	169,5	6
6	169,5	179	3

Ф

Выбираем в качестве начала интервала 150.5, так как именно на этот интервал приходится наибольшее количество.

$$M_o = 150.59.5\frac{8-6}{(8-6)(8-6)} = 155.25$$

Таким образом мода данного стат. распределения – 155.25

Ясно, что если среднее арифметическое, мода и медиана близки друг к другу, то распределение случайной величины близко к симметричному, например, к нормальному.

Нам важно знать не только средние значения вариант, но и отличие значений вариант от среднего значения.
Для отражения изменчивости (вариации) значений признака вводят различные показатели вариации (колеблемости) ряда.

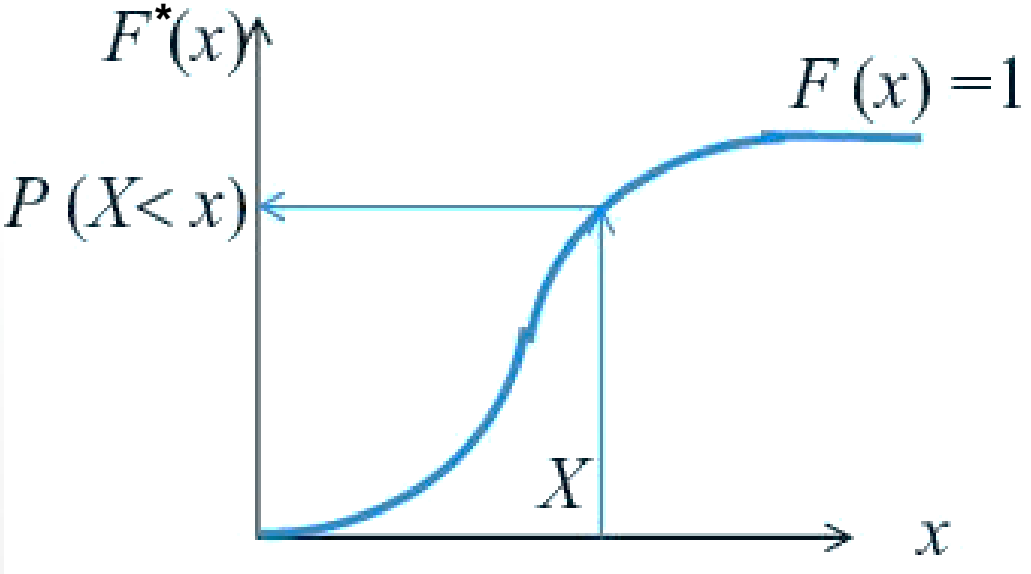
Простейшим и весьма приближенным показателем вариации является размах выборки (или размах вариации)

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

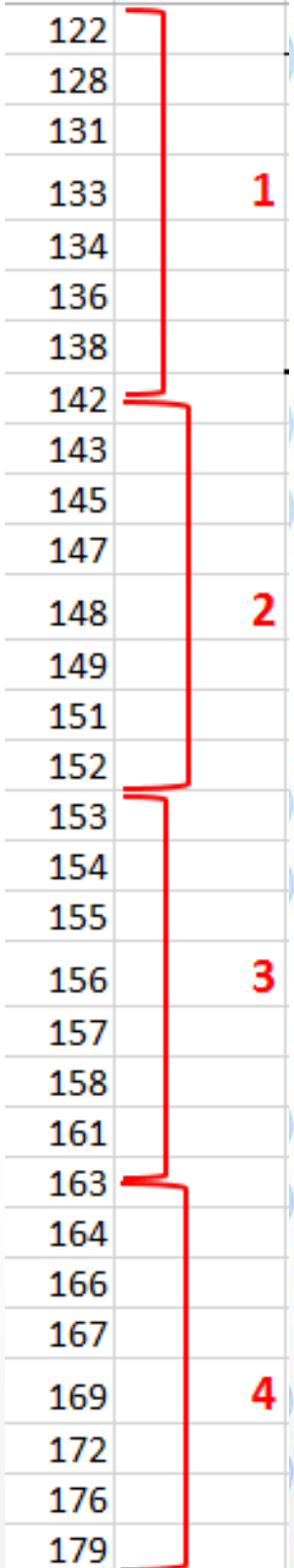
Больший интерес представляет мера рассеяния наблюдаемых значений вокруг средних величин, в частности, вокруг среднего арифметического.

- Зная эмпирическую функцию распределения, можно найти эмпирические **квантили** точно так же, как в случае обычной (теоретической) функции распределения.

- А именно: ***выборочная квантиль порядка p*** получается как ***абсцисса точки пересечения горизонтальной прямой, пересекающей ось ординат в точке p, с эмпирической функцией распределения F*(x).***



- **Если пересечение - не точка, а отрезок, то в качестве квантили можно взять абсциссу середины этого отрезка. Выборочные квантили $\tilde{X}_{0,25}, \tilde{X}_{0,5}, \tilde{X}_{0,75}$ дают соответственно выборочную нижнюю квартиль, выборочную медиану и выборочную верхнюю квартиль, а разность $\tilde{X}_{0,75} - \tilde{X}_{0,25}$ - *выборочную интерквартильную широту* Me [25%–75%].**



Для примера 2

122 128 131 133 134 136 138 142 143 145 147 148 149 151 152 153 154 155 156 157 158 161 163 164 166 167 169 172 176 179

Всего данных - 30. Резонно предположить, что 1-я квартиль находится где-то в районе 8-го данного, вторая - между 15 и 16, а 3-я - в районе 23-го.

Рассчитанные в Excel 1,2 и 3 значения квартилей (статистическая функция «КВАРТИЛЬ») имеют величины соответственно ***142,25, 152.5 и 162,5***.

Интерквартильный размах равен, соответственно, *20,25*.

Это означает, что:

25% значений выборки меньше 142,25

25% значений - больше 162,5

Выборочной дисперсией вариационного ряда называется среднее арифметическое квадратов отклонений вариантов от их среднего арифметического:

$$D_* = S_*^2 = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)^2}{n} = \sum_{i=1}^{k} \frac{\left(x_i - \overline{x}\right)^2 \cdot n_i}{n} = \sum_{i=1}^{k} \left(x_i - \overline{x}\right)^2 \cdot w_i$$

При вычислении выборочной (или эмпирической) дисперсии формулу несколько меняют. Из некоторых соображений в знаменателе этой формулы ставят не *n*, а *n-1*, и чаще применяется другая формула для вычисления дисперсии, которую называют **исправленная выборочная дисперсия**:

$$\tilde{D} = S^2 = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)^2}{n-1} = \sum_{i=1}^{k} \frac{\left(x_i - \overline{x}\right)^2 \cdot n_i}{n-1}$$

Для меры вариации, выраженной в тех же единицах измерения, что и значение признака, вычисляют **исправленное выборочное среднеквадратическое отклонение (СКО)**:

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

Вычисляем эти характеристики для примера 1, n = 30.

Получаем:

$$\tilde{D} = \frac{(151-151.6)\cdot 2 + (166-151.6)\cdot 2 + \ldots + (156-151.6)\cdot 2}{30-1} \approx$$

$$\approx 213,21$$

$$s = \sqrt{\tilde{D}} = \sqrt{213.21} = 14,6$$

Среднеквадратическая (стандартная) ошибка среднего:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{14,6}{\sqrt{30}} \approx 2,67$$

Коэффициент вариации

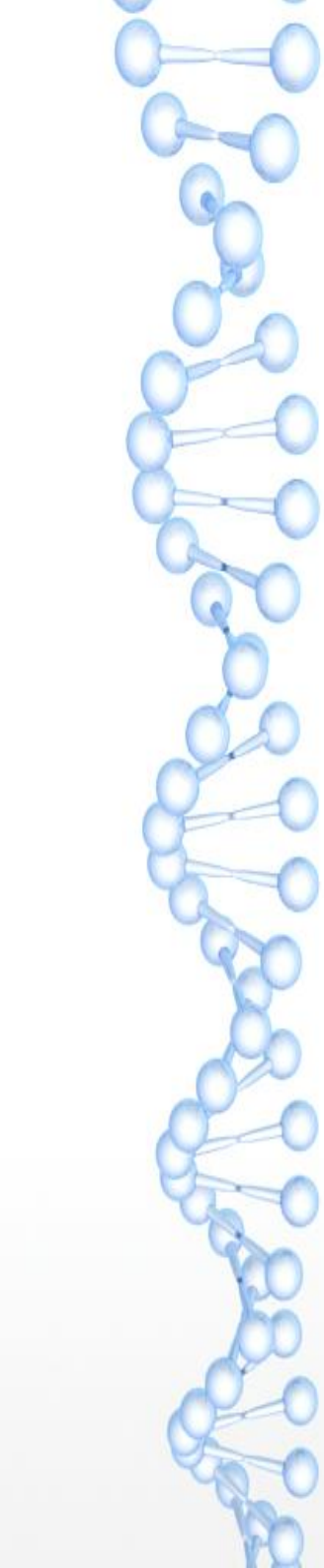
- **Характеризует относительное значение среднеквадратического отклонения и служит для сравнения разброса несоизмеримых показателей**

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

Для примера 1

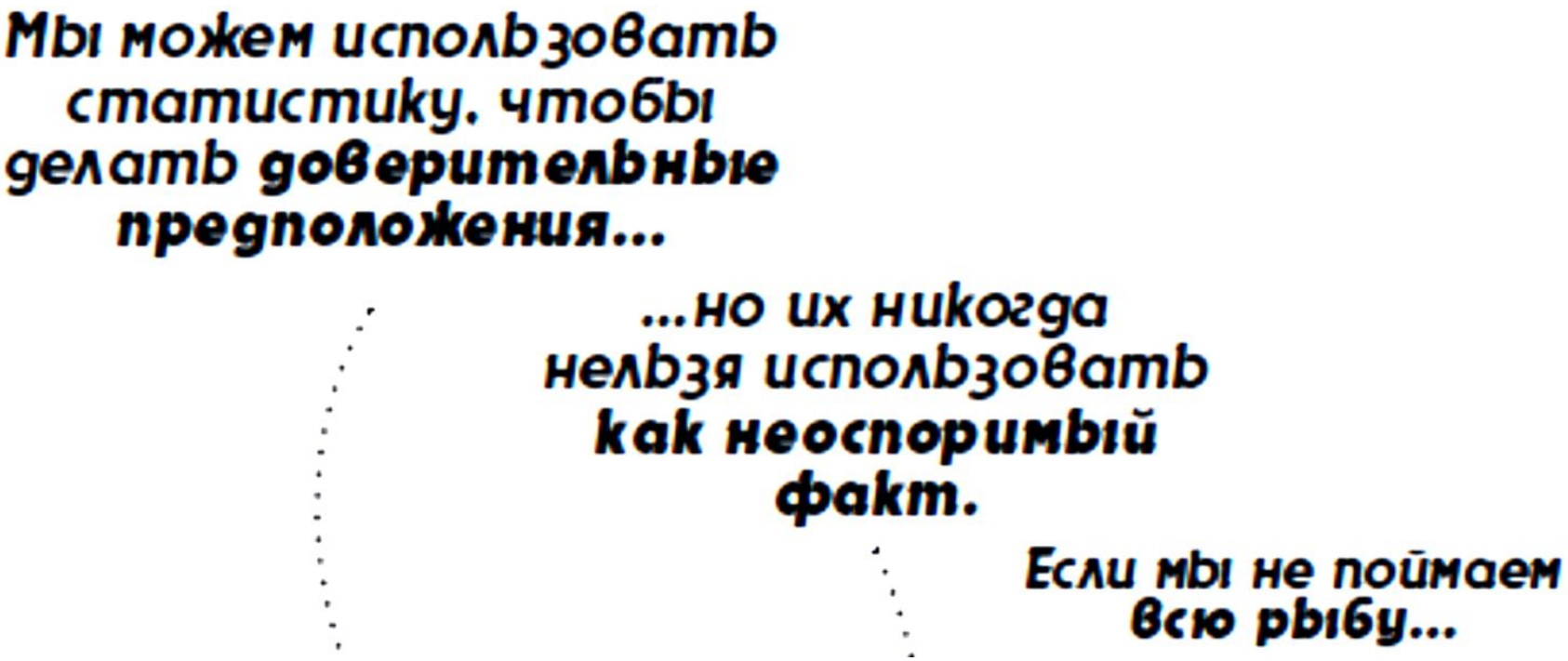
$$V = \frac{14,6}{151,6} \cdot 100\% = 9,63\%$$

- выборка однородна



Критерий хи-квадрат

Интервальные оценки



Интервальная оценка параметра ГС – это такой интервал, который с вероятностью, близкой к единице, но не равное 1 (например, 0,95; 0,99 или 0,999), накрывает истинное значение оцениваемого параметра.

Столь же часто говорят: находим «**доверительный интервал**» для оцениваемого параметра ГС вместо «интервальная оценка» (эти два термина означают одно и то же).

Значения β*=0,95; 0,975; 0,99* называется ***доверительной вероятностью*** (а также: ***надежностью, уровнем доверия***)

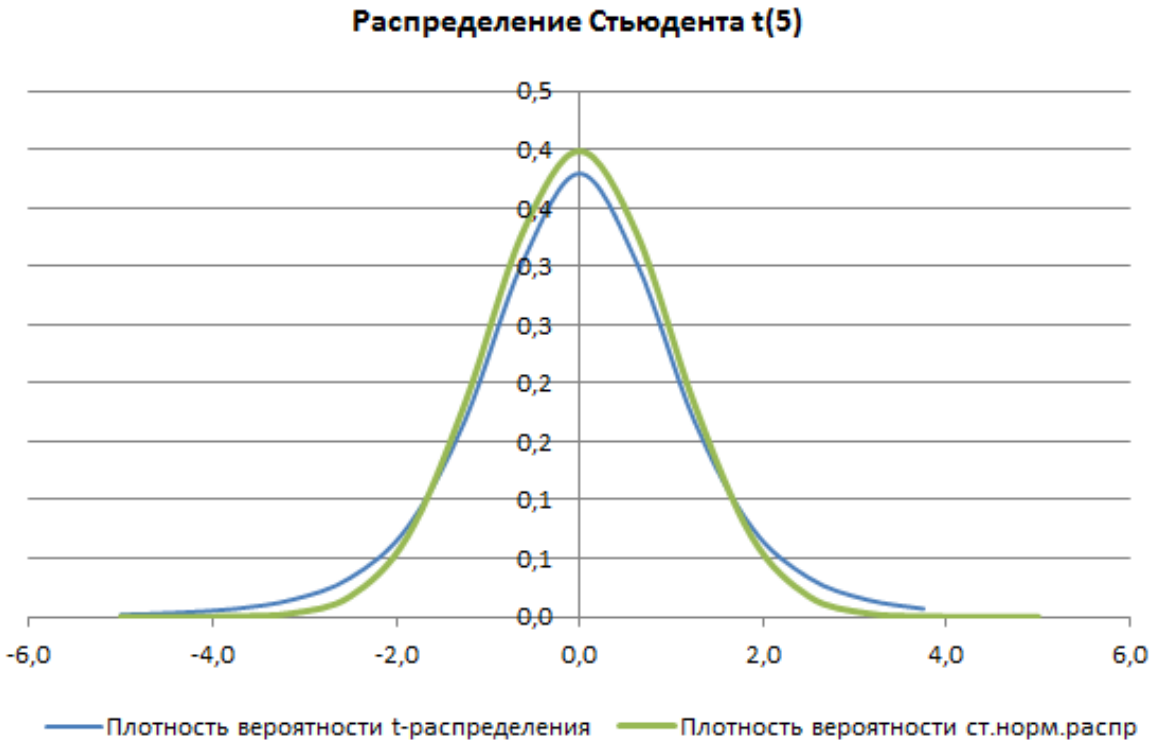
Для **малых выборок** **доверительный интервал мат. ожидания** находится по формуле:

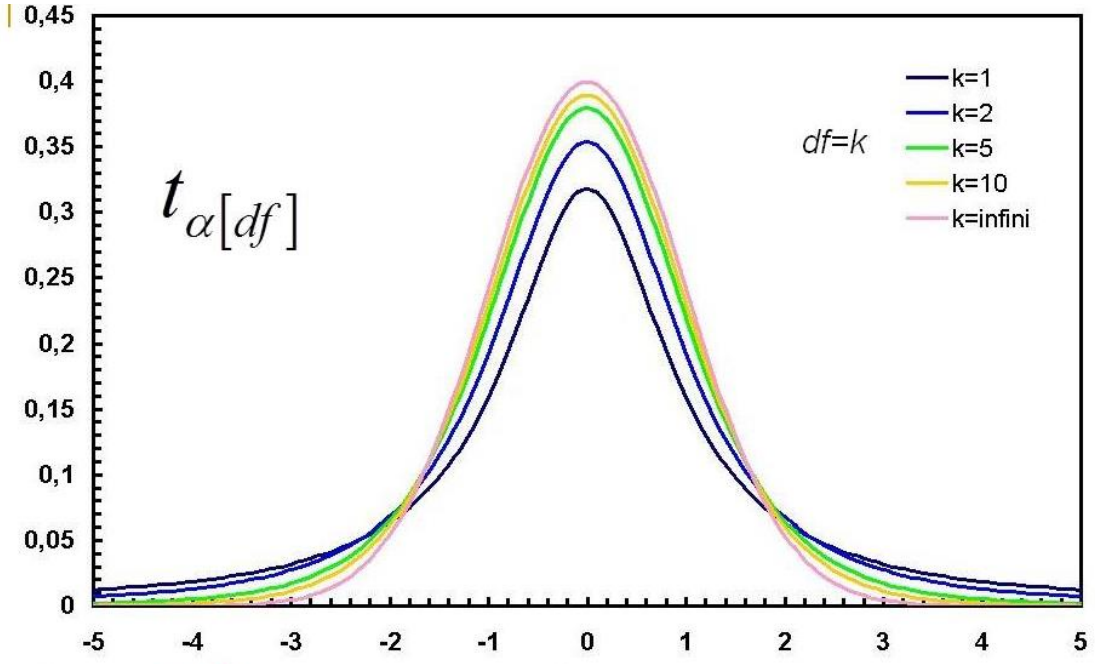
$$I_\beta = \left(\overline{x} - \frac{t_{кр} \cdot S}{\sqrt{n}}; \overline{x} + \frac{t_{кр} \cdot S}{\sqrt{n}} \right)$$

$t_{кр}$ обычно записывается как $t_{кр}(\beta, n-1)$ и находится по таблице *t*-распределения Стьюдента.

При существенном возрастании объёма выборки распределение Стьюдента переходит в стандартное нормальное распределение. При малых k замена распределения Стьюдента на нормальное распределение приводит к грубым ошибкам.

Распределение Стьюдента





При больших (>30) размерах выборок приближается к нормальному

Значения $t(\beta;k)$ для t–распределения Стьюдента

k	β				k	β			
	0.8	0.9	0.95	0.99		0.8	0.9	0.95	0.99
1	3.08	6.31	12.71	63.7	21	1.323	1.721	2.08	2.83
2	1.886	2.92	4.30	9.92	22	1.321	1.717	2.07	2.82
3	1.638	2.35	3.18	5.84	23	1.319	1.714	2.07	2.81
4	1.533	2.13	2.77	4.60	24	1.318	1.711	2.06	2.80
5	1.476	2.02	2.57	4.03	25	1.316	1.708	2.06	2.79
6	1.440	1.943	2.45	3.71	26	1.315	1.706	2.06	2.78
7	1.415	1.895	2.36	3.50	27	1.314	1.703	2.05	2.77
8	1.397	1.860	2.31	3.36	28	1.313	1.701	2.05	2.76
9	1.383	1.833	2.26	3.25	29	1.311	1.699	2.04	2.76
10	1.372	1.812	2.23	3.17	30	1.310	1.697	2.04	2.75
11	1.363	1.796	2.20	3.11	35	1.307	1.690	2.03	2.72
12	1.356	1.782	2.18	3.06	40	1.303	1.684	2.02	2.70
13	1.350	1.771	2.16	3.01	45	1.300	1.680	2.01	2.69
14	1.345	1.761	2.14	2.98	50	1.298	1.675	2.01	2.68
15	1.341	1.753	2.13	2.95	60	1.296	1.671	2.00	2.66
16	1.337	1.746	2.12	2.92	70	1.294	1.668	1.99	2.65
17	1.333	1.740	2.11	2.90	80	1.293	1.667	1.99	2.64
18	1.330	1.734	2.10	2.88	90	1.292	1.666	1.99	2.63
19	1.328	1.729	2.09	2.86	100	1.290	1.661	1.98	2.63
20	1.325	1.725	2.09	2.84	∞	1.282	1.645	1.960	2.58

Пример.

Найти доверительный интервал для оценки математического ожидания нормальной случайной величины с надежностью β = 0.95, зная выборочную среднюю – $\tilde{m} = \bar{X} = 75{,}15$, объем выборки – *n=64*, среднеквадратическое отклонение – *s=8*.

Решение.

По таблице Стьюдента определяем значение *t*: *t(0,95; 63)=2,00*. Получаем искомый доверительный интервал:

$$75{,}15 \pm 2{,}00 \cdot \frac{8}{\sqrt{64}} = 75{,}15 \pm 2{,}00$$

или I_β=(73,15; 77,15).

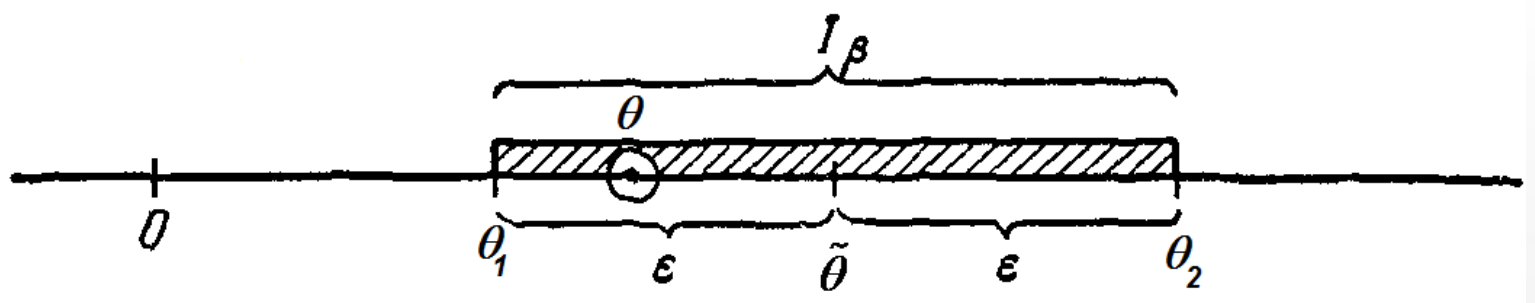
Если перейти к нормальному распределению, то *t=1,96*, а I_β=(73,19;77,11)

Интервальная оценка истинного значения параметра ГС дополняет его точечную оценку, давая более содержательную информацию об этом значении.

Можно посмотреть на полученную для ε формулу с иной точки зрения. Если задана точность оценки ε, то можно найти объем выборки, которая обеспечит эту требуемую точность:

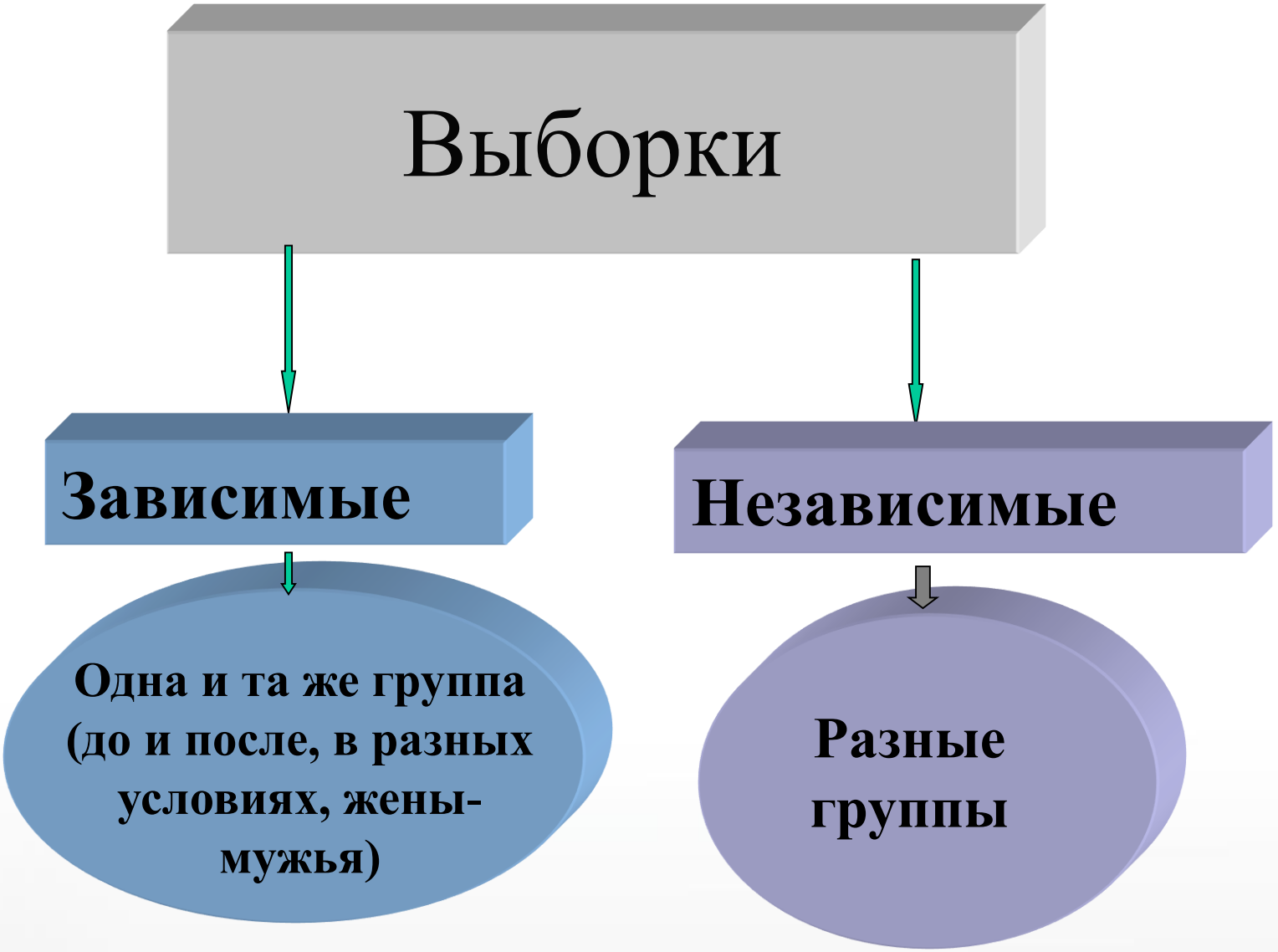
$$n_{\min} = \left(\frac{t_{кр} \cdot S}{\varepsilon} \right)^2 ; \quad при \quad n \geq n_{\min} \quad эта$$

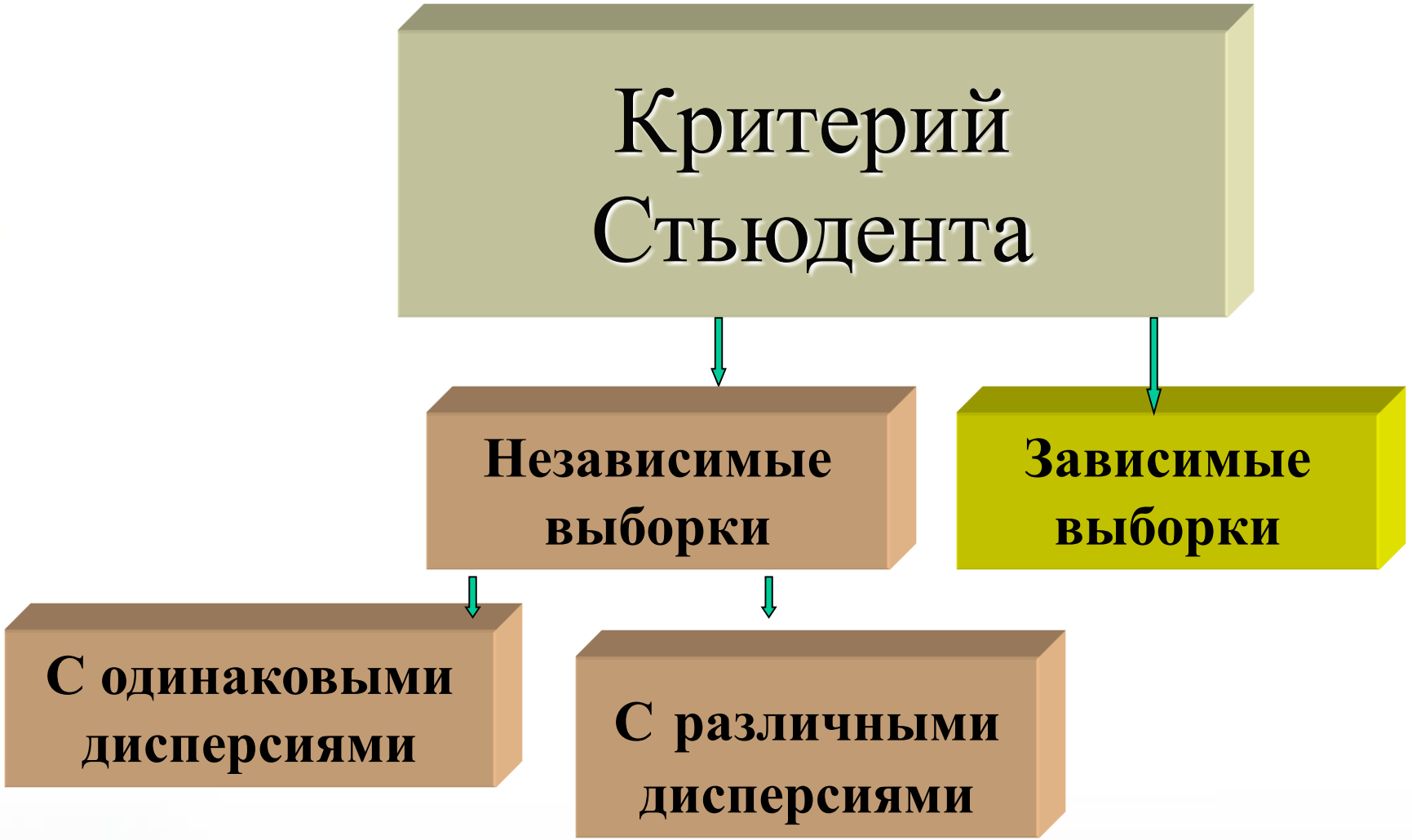
точность будет обеспечена.



На интервальных оценках мы можем проверить

ДОСТОВЕРНОСТЬ РАЗЛИЧИЙ СРЕДНИХ


Выборки
Зависимые
Независимые
Одна и та же группа (до и после, в разных условиях, жены-мужья)
Разные группы


Критерий Стьюдента
Независимые выборки
Зависимые выборки
С одинаковыми дисперсиями
С различными дисперсиями

Сравнение двух средних по *зависимым выборкам* малого объема из нормальных генеральных совокупностей

Исследовалось изменение частоты сердечных сокращений студентов до и после экзамена

N	ЧСС до	ЧСС после	d
1	90	60	-30
2	80	70	-10
3	70	70	0
4	90	70	-20
5	100	70	-30
6	110	80	-30
	90	70	-20

Нулевая гипотеза:

В генеральной совокупности нет различия между средними арифметическими выборок

Проверяем гипотезу по критерию Стьюдента t при уровне значимости $\alpha=0,05$.

1. Определяем $t_{набл}$:

$$t_{набл} = \frac{|d|}{s_d}$$

где ***d-среднее значение*** разности пульса до и после экзамена

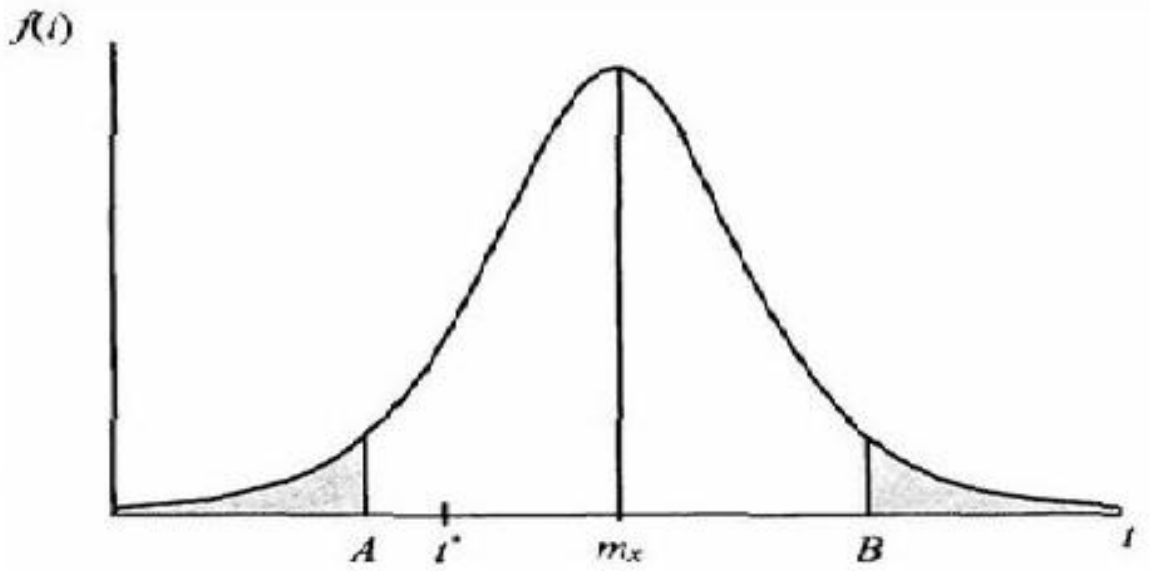
s_d-стандартная ошибка разности

Определяем критическое значение критерия

Стьюдента $t_{кр}$ для α=0,05 и df=n-1

Если $|t_{набл}| \geq t_{кр}$ *нулевая гипотеза отвергается*,
различие средних статистически значимо

Если $|t_{набл}| < t_{кр}$, *нулевая гипотеза принимается*,
различие средних статистически не значимо



Область принятия гипотезы [A,B]
и критическая область: $(-\infty, A) + (B, +\infty)$

N	$ЧСС_{до}$	$ЧСС_{после}$	d	$(d-d_{ср})^2$
1	90	60	-30	100
2	80	70	-10	100
3	70	70	0	400
4	90	70	-20	0
5	100	70	-30	100
6	110	80	-30	100
	90	70	$d_{ср}$=-20	D=160

Для разности:

$$\bar{d} = \frac{(-30)+(-10)+\ldots+(-30)}{6} = -20$$

$$D(x) = \frac{(-30-(-20))^2+(-10-(-20))^2+\ldots+(-30-(-20))^2}{5} = 160$$

$$s = \sqrt{160} = 12{,}65$$

$$s_d = \frac{12{,}65}{\sqrt{6}} = 5{,}16$$

Определим, достоверно ли определена средняя арифметическая разности:

$$t_{набл} = \frac{|d|}{s_d} = \frac{20}{5{,}16} = 3{,}87$$

$t_{кр}(0,05;5)=2,57$ $t_{набл} > t_{кр}$

Это означает, что нулевая гипотеза отвергается, снижение ЧСС статистически значимо

Группа	ЧСС уд/мин	
	До экзамена	После экзамена
МК201 (n=6)	90±14,1	70 ±6,3*

Рассчитаем относительное изменение:

$$\Delta = \frac{(70-90)}{90} * 100\% = -22\%$$

ЧСС студентов после экзамена снизилась на 22% ($\alpha<0,05$)

Сравнение генеральных средних двух групп по *независимым выборкам* из нормальных совокупностей.

Допущения:

В генеральной совокупности выборки распределены по нормальному закону

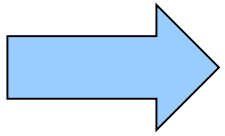
Дисперсии независимых выборок однородны (критерий Фишера)

$$F_{набл} = \frac{D_{>}}{D_{<}} < F_{крит}(\alpha; n_1, n_2)$$

Нормированное отклонение:

$$t_{набл} = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{s_{\overline{x}_1 - \overline{x}_2}}$$

1. Для n≥30, ошибка разницы s_d определяется по формуле: $\sqrt{S_{x1}^2 + S_{x2}^2}$



$$t_{набл} = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{S_{x1}^2 + S_{x2}^2}}$$

2. Критическое значение $t_{кр}$ определяется по таблице Стьюдента в зависимости от α и $df = n_1 + n_2 - 1$

Пример: $n_1=40 \quad n_2=50$

$$\bar{x}_1 = 150 \quad \bar{x}_2 = 140 \quad s_{\bar{x}_1} = 5 \quad s_{\bar{x}_2} = 6$$

Определить значимость различий при α=0,05

$$t_{набл} = \frac{150-140}{\sqrt{5^2+6^2}} = 1{,}28$$

$t_{крит}\,(0{,}05;\ 89)=1{,}99, \quad t_{набл} < t_{крит.}$

Разница средних арифметических недостоверна.

2. Для n<30, ошибка разницы определяется по формуле:

$$s_{\overline{x_1 - x_2}} = \sqrt{\frac{\sum (x_{1i} - \bar{x}_1)^2 + \sum \left(x_{2i} - \bar{x}_2\right)^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}$$

и

$$t_{набл} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Пример. Частота сердечных сокращений двух групп:

I	II	ΔX_1	$(\Delta X_1)^2$	ΔX_2	$(\Delta X_2)^2$
90	100	0	0	0	0
80	120	-10	100	20	400
70	90	-20	400	-10	100
90	70	0	0	-30	900
100	90	10	100	-10	100
110	110	20	400	10	100
	120			20	400
$\overline{X}_1$	$\overline{X}_2$	$\Sigma=$	1000		2000
90	100	$D_1=$	1000/5= 200	$D_2=$	2000/6= 333

$$F_{набл} = \frac{333}{200} = 1{,}67$$

$F_{крит}(6,5,0{,}05)=4{,}95$ - 1,67<4,95 дисперсии однородны.

$$S_d = \sqrt{\frac{(1000+2000)\cdot 13}{(5+6)\cdot 42}} = 9{,}2$$

$$t_{набл} = \frac{10}{9{,}2} = 1{,}1$$

$df=n_1+n_2-2=11$

$t_{кр}=2{,}2$

$t_{набл} < t_{кр}$, нулевая гипотеза не отвергается, различие средних арифметических статистически не значимо, выборки принадлежат одной генеральной совокупности

Excel

Функция ТТЕСТ позволяет определить, вероятность того, что две выборки взяты из генеральных совокупностей, которые имеют одно и то же среднее.

Синтаксис

ТТЕСТ(массив1;массив2;хвосты;тип)

Аргументы функции ТТЕСТ описаны ниже.

Массив1 Обязательный. Первый набор данных.

Массив2 Обязательный. Второй набор данных.

Хвосты Обязательный. Число хвостов распределения. Если значение "хвосты" = 1, функция ТТЕСТ возвращает одностороннее распределение. Если значение "хвосты" = 2, функция ТТЕСТ возвращает двустороннее распределение.

Тип Обязательный. Вид выполняемого t-теста.

Новый вариант этой функции - Функция СТЬЮДЕНТ.ТЕСТ.

Сводка основных формул

Средняя арифметическая выборки

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Дисперсия

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

Среднее квадратическое отклонение

$$s = \sqrt{D(x)}$$

Средняя квадратическая ошибка:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Критерий $t_{эксп}$ разности средних арифметических двух выборок

а) n≥30

$$t_{набл} = \frac{\left|\overline{X_1} - \overline{X_2}\right|}{\sqrt{S_{x1}^2 + S_{x2}^2}}$$

$t_{кр}$ определяется по таблице Стьюдента в зависимости от α и $df = n_1 + n_2 - 1$

б) n<30

$$t_{набл} = \frac{\left|\overline{X_1} - \overline{X_2}\right|}{\sqrt{\frac{\sum (X_{1i} - \overline{X_1})^2 + (X_{2i} - \overline{X_2})^2}{(n_1 - 1) + (n_2 - 1)} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}}$$

или

$$t_{набл} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА:

Венцель Е. Теория вероятностей, 10-е издание, стереотипное, Издательский центр „Академия” //Москва. – 2005.

Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. 3-е изд., перераб. и доп. –М.: Высшая школа, 2004. 400 с.

Шапиро Л.А., Шилина Н.Г.. Руководство к практическим занятиям по медицинской и биологической статистике. Учебное пособие для студентов медицинских вузов. – Красноярск, 2003, 94 стр.

Спасибо за внимание!