Лекция 2

Статистическая обработка данных

Ростов-на-Дону
2012

Содержание лекции №2

- Генеральная совокупность и выборка.
- Статистическое распределение. Гистограмма.
- Характеристики положения и рассеяния.
- Оценка параметров генеральной совокупности по выборке.
- Доверительный интервал и доверительная вероятность.
- Сравнение средних.

Математическая статистика (МС) – это наука, изучающая методы обработки результатов наблюдений **массовых** случайных явлений, обладающих статистической устойчивостью, закономерностью с целью выявления этой закономерности по исследованию **части** этого массива данных.

*Возможности **МС***

1. Выявляет закономерности ***массовых*** явлений (т.е.**царица** в области ***больших чисел*).**

2. **Предсказывает** наличие внешних влияний.

Два основных направления МС:

или

Задачи математической статистики

1. Оценка неизвестных параметров.

2. Проверка статистических гипотез.

Основные понятия МС:

- **генеральная совокупность**
- **выборка**

Генеральная совокупность и выборка

Генеральная совокупность – это множество всех **мыслимых** значений наблюдений, однородных относительно некоторого признака, которые могли быть сделаны.

Объем *генеральной совокупности*

N

Пример: число единиц товара, произведенных фирмой за год.

Рост студентов I курса всей Ростовской области

Выборка – совокупность ***случайно*** отобранных наблюдений.

Выборка – это множество случаев, с помощью **определенной процедуры** выбранных из генеральной совокупности для участия в исследовании.

Выборка характеризуется:

x_i - **варианта**

n_i - **частота** встречаемости

ВОПРОС:

Зачем используют выборку?

- Объект исследования **очень большой.**
- Существует необходимость **сбора первичной** информации

Объем выборки. Репрезентативность

Объем выборки – это ***количественная*** характеристика выборки.

Это ***количество вариант в выборке.*** Это число случаев, включенных в выборочную совокупность.

n

ВОПРОС:

А есть качественная характеристика выборки?

Да. Кого или **Что** именно выбирают. Какие способы построения выборки для этого используют.

Репрезентативность (*фр. representation – представление)* – это **соответствие** характеристик выборки характеристикам генеральной совокупности.

Репрезентативность – это **свойство** выборки представлять параметры генеральной совокупности.

Выборка *должна быть* ***репрезентативной,*** то есть свойства выборки должны **отражать** свойства генеральной совокупности.

Статистическое распределение (вариационный ряд)

Статистическое распределение – это совокупность вариант и соответствующих им частот.

x_i -*варианта*

n_i - *частота встречаемости*

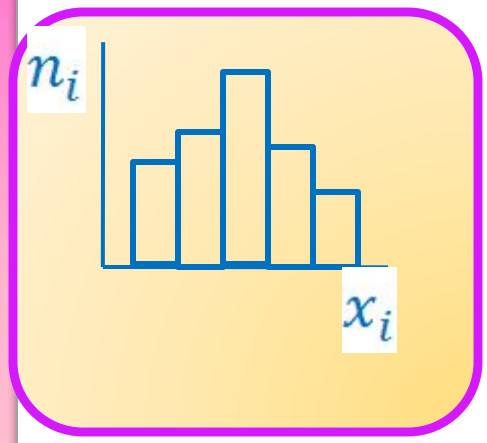
Пример: Рост 175 см встретился 5 раз; рост 168 см – 7 раз; 180 см – 8 раз.

Вариационный ряд - это та же самая выборка, но расположенная в порядке ***возрастания*** элементов.

Пример: 168 см – 7 раз; 175 см – 5 раз; 180 см – 8 раз.

Гистограмма

Гистограмма – это ***ступенчатая*** фигура, состоящая из смежных ***прямоугольников***, построенных на одной прямой, ***основания*** которых одинаковы и равны ширине класса, а ***высоты*** равны относительной частоте.



вариационный размах

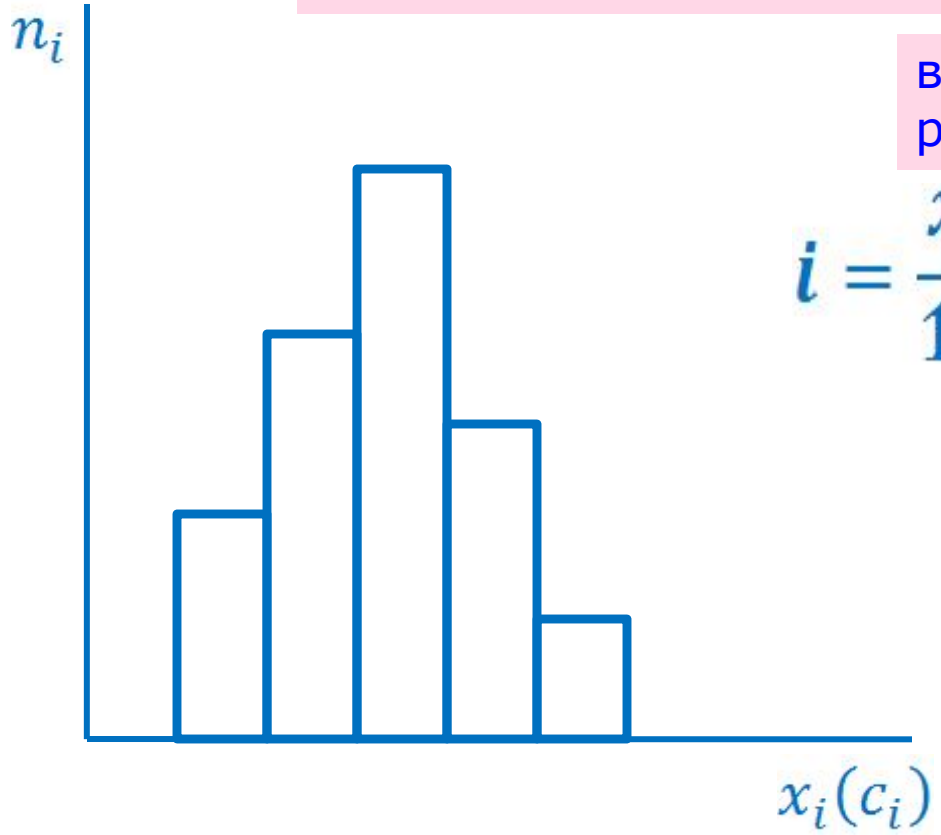
Формула Стерджеса

Ширина класса

$$i = \frac{x_{max} - x_{min}}{1 + 3,32lgn}$$

Измеряют рост.
Объем выборки n=10

168; 155; 168; 177; 189; 192; 196; 184; 189; 165



вариационный размах

$$i = \frac{x_{max} - x_{min}}{1 + 3,32lgn}$$

Гистограмма распределения

Характеристики положения (мода, медиана, выборочное среднее) и рассеяния (выборочная дисперсия и выборочное среднее квадратическое отклонение).

Характеристики положения:

- **Мода** (Mo) – наиболее ***часто*** встречающаяся ***варианта*** в данной совокупности.

Пример:

x_i	1	4	7	25
n_i	3	2	19	6

$\text{Mo} = \mathbf{7}$

172, 168, 172, 175, 187, 172, 164 $\Rightarrow \text{Mo} = 172$

Мода – это такое значение ***варианты***, что предшествующие и следующие за ней значения имеют ***меньшие*** частоты встречаемости.

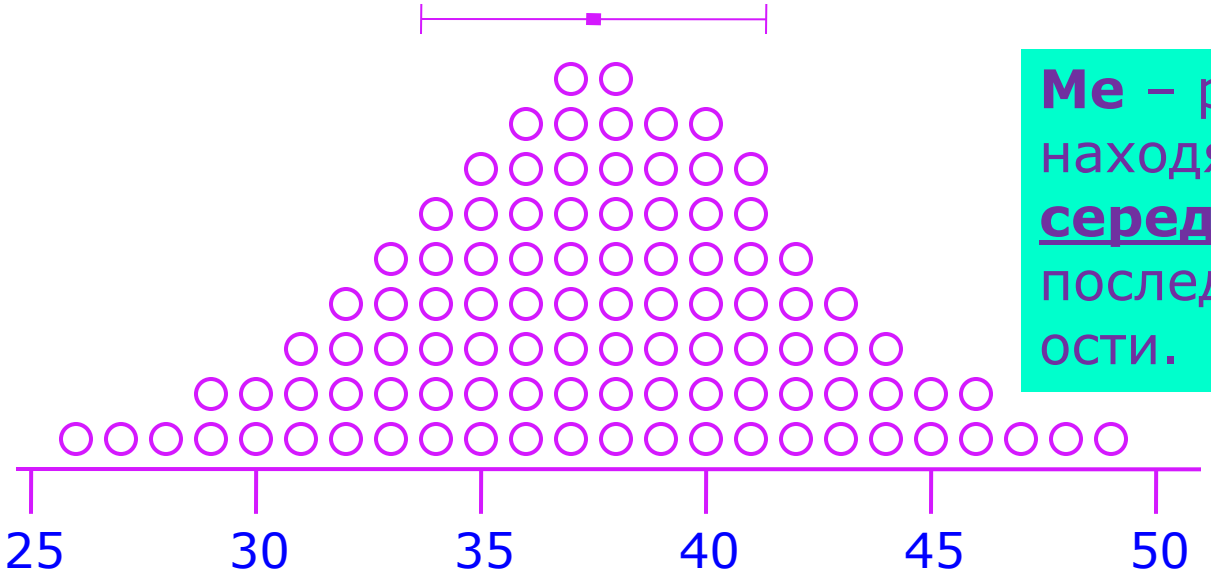
x_i	4	8	10	17	20
n_i	5	3	12	9	6

$\Rightarrow \text{Mo} =$ **10**

- **Медиана** (Me) – это структурная средняя признака, относительно которой вариационный ряд делится на **две равные части**.

Пример:

- 2 4 6 8 10 12 14 $\quad \mathrm{Me} = 8,$
- 2 4 6 8 10 12 14 16 $\quad \mathrm{Me} = \dfrac{8+10}{2} = 9.$



Me – результат, находящийся **в середине** последовательности.

- ***Выборочная средняя*** – это среднее арифметическое значение вариант статистического ряда.

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$$

n- объем выборки

n_i *- частота встречаемости*

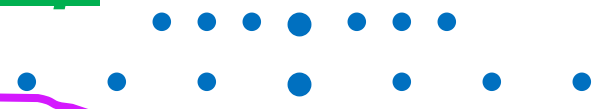
x_i *-варианта*

Пример: *Гемоглобин (*Не) в крови одной группы мужчин (n_1=30) равен **70%**, а для другой группы мужчин того же возраста (n_2= 20) – **50%.** Найти среднюю арифметическую этих двух средних.

$$\bar{x} = \frac{30 \cdot 70 + 20 \cdot 50}{30 + 20} = 62\%.$$

Характеристики рассеяния определяют отклонение каждой варианты от средней арифметической.

Пример:



Но “+” компенсируют “-” ∑=0. Поэтому ***возводим в квадрат*** и находим среднее.

$x_i - \overline{x}_в$ - ***отклонение***

- ***Выборочная дисперсия***

$$S_в^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x}_в)^2 \cdot n_i,$$

где n - объем выборки,
n_i - частота встречаемости,
x_i - варианта,
$\overline{x}_в$ - выборочное среднее.

- ***Среднее квадратическое отклонение*** = **стандартное отклонение**

$$S_в = \sqrt{S_в^2}.$$

Пример: Дана выборка

x_i	1	2	3	4
n_i	20	15	10	5

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$$

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

$$\bar{x}_в = \frac{1 \cdot 20 + 2 \cdot 15 + 3 \cdot 10 + 4 \cdot 5}{20 + 15 + 10 + 5} = \frac{100}{50} = 2,$$

$$S_в^2 = \frac{(1-2)^2 \cdot 20 + (2-2)^2 \cdot 15 + (3-2)^2 \cdot 10 + (4-2)^2 \cdot 5}{50} =$$

$$= \frac{50}{50} = 1,$$

$$S_в = \sqrt{1} = 1.$$

Пример.

Дана выборка

3, 4, 5

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$$

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

$$\bar{x}_в = \frac{3+4+5}{3} = 4,$$

$$S_в^2 = \frac{(3-4)^2 + 0 + (5-4)^2}{3} = \frac{2}{3} = 0{,}66,$$

$$S_в = \sqrt{0{,}66} = 0{,}8.$$

Оценка параметров генеральной! совокупности по характеристикам ее выборки! (точечная и интервальная)

Оценка параметра – это любая функция от значений выборки.

Требования

- **несмещенная**
- состоятельная
- эффективная

Генеральная совокупность –это гипотетическое множество элементов, объединенных **общей характеристикой.**

Выборка - множество испытуемых из генеральной совокупности.

Выборка

ПАРАМЕТРЫ

1. Выборочное среднее $\overline{x}_в$
2. Выборочная дисперсия $S_в^2$

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot n_i$$

$$S_в^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \overline{x}_в)^2 \cdot n_i,$$

Генеральная совокупность

ПАРАМЕТРЫ

1. Генеральное среднее $\overline{x}_{ген} = \mu$
2. Генеральная дисперсия $\sigma_{ген}^2$

I. Точечная оценка

Точечная оценка – это **выборочная характеристика**, используемая в качестве **приближенного** значения неизвестной генеральной характеристики.

- Определяется одним числом (точкой на числовой оси).
- Выборка должна быть **большого** объема.
- Дает лишь некоторое приближенное значение параметра.

- ***Генеральное среднее***

$$\overline{x}_{ген} = \mu = M(\overline{x}_в)$$

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot n_i$$

- это **несмещенная** оценка математического ожидания

Генеральное среднее *равно математическому ожиданию выборочной средней*

- ***Генеральная дисперсия***

Генеральная дисперсия **не равна** математическому ожиданию **выборочной** дисперсии

$$\sigma^2_{ген} \neq M(S^2_в)$$

- это ***смещенная*** оценка дисперсии

- **Исправленная** дисперсия (более точная)

$$S^2 = \frac{n}{n-1} S_в^2, \qquad s^2 = \frac{n}{n-1} \sum_{i=1}^{k} (x_i - \bar{x}_в)^2 \cdot n_i,$$

$$\sigma^2_{ген} = M(S^2)$$

Генеральная дисперсия *равна математическому ожиданию исправленной дисперсии.*

- $m_{\bar{x}} = \frac{S}{\sqrt{n}}$,

$m_{\bar{x}}$ - средняя ошибка выборочной средней,

S - исправленное среднее квадратическое отклонение,

n - объем выборки,

- $CV = \frac{S}{\bar{x}_в} \cdot 100\%$,

CV - коэффициент вариации. Характеризует изменчивость признака в единых единицах %

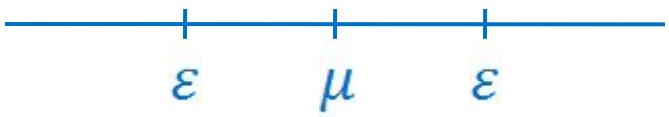
II. Интервальная оценка

– это числовой интервал, содержащий неизвестный параметр генеральной совокупности с заданной вероятностью.

- Определяется **двумя** числами –границами интервала.
- Более точная, надежная и информативная, так как дает информацию о степени близости к соответствующему теоретическому параметру.
- Используется, если выборка **малого** объема.

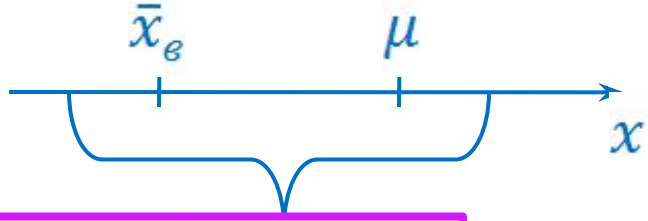
$$|\overline{x}_в - \mu| < \varepsilon,$$

$$\overline{x}_в - \varepsilon < \mu < \overline{x}_в + \varepsilon.$$



Доверительный интервал и доверительная вероятность

Доверительный интервал – это интервал, в котором с той или иной **заранее!** заданной вероятностью! находится генеральный параметр.



$$\mu = \bar{x}_в \pm mt, \quad (P \geq 0{,}95)$$

$$\bar{x}_в - mt \leq \mu \leq \bar{x}_в + mt \; (P \geq 0{,}95).$$

$\bar{x}_в$ - выборочное среднее,

m - средняя ошибка выборочной средней.

$$m_{\bar{x}} = \frac{S}{\sqrt{n}},$$

$t_{\alpha,f=n-1}$ - нормированный показатель распределения Стьюдента, с *(n-1)* степенями свободы

$t_{\alpha,f=n-1}$ - ***нормированный показатель распределения Стьюдента,*** с *(n-1)* степенями свободы, который определяется вероятностью попадания генерального параметра в этот интервал.

Стьюдент
(Уильям Д. Госсет)
1876-1937гг.

1899г.
Дублин, Ирландия,
Пивоваренный завод Гиннеса

f/α	0.1	0.05
1.	6,314	12,706
2.	2,920	4,303
3.	2,353	3,182
4.	2,132	2,776
5.	2,015	2,571
6.	1,943	2,447
7.	1,895	2,365
8.	1,860	2,306
9.	1,833	2,262
10.	1,812	2,228
11.	1,796	2,201
12.	1,782	2,179
13.	1,771	2,160
14.	1,761	2,145
15.	1,753	2,131

Доверительная вероятность *P* – это такая вероятность, что событие *1-P* – можно считать невозможным.

Признана достаточной для уверенного суждения о генеральных параметрах на основании известных выборочных показателей.

Обычно в качестве доверительных используют вероятности, близкие **к 1**. Тогда событие, что генеральный параметр попадет в этот интервал будет практически достоверным.

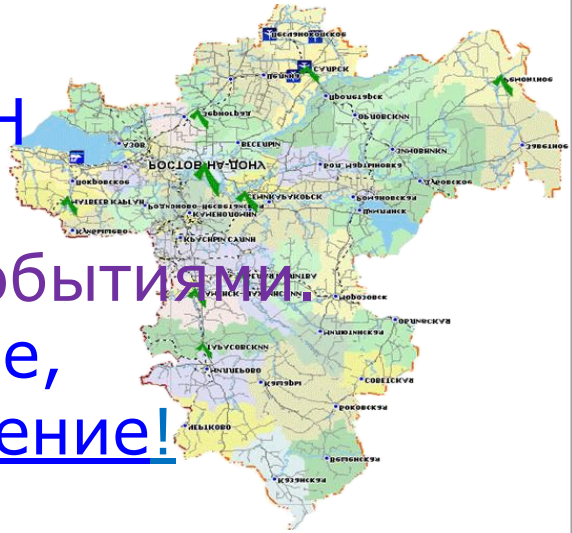
95%	99%	99,9%
$P \geq 0,95$	$P \geq 0,99$	$P \geq 0,999$

Уровень значимости = уровень ошибки α

$$\alpha = 1 - P.$$

$\alpha \leq 0,05$	$\alpha \leq 0,01$	$\alpha \leq 0,001$

Статистическая проверка гипотез.



- ***В жизни***: **Гипотеза** (hypothesis) Н – предположение, описывающее возможную **взаимосвязь** между событиями.

***В науке*: Гипотеза** – предположение, вызывающее сомнение!

- ***В математической статистике:***

Гипотеза – предположение, которое вызывает сомнение, и которое мы собираемся проверять!

Статистическая гипотеза – это всякое высказывание о **генеральной**! (всегда!) совокупности, *проверяемое по выборке*!

***Например*: Статистическая гипотеза – это предположение о виде неизвестного распределения или о параметрах известного распределения.**

Тест: Какая гипотеза, из нижеприведенных, является статистической?

1. Генеральная совокупность распределена по нормальному закону.
2. Зимой на экзамене я, может быть, получу “4”.
3. Генеральные дисперсии равны $\sigma_1^2 = \sigma_2^2$.
4. Летом, может быть, я поеду на море.

Ответ: 1, 3.

Общая постановка задачи проверки гипотез

Проверка гипотезы – это процедура сопоставления высказанной гипотезы о генеральной совокупности с выборочными данными.

Этапы проверки гипотезы (общая схема)

① H_0: Выдвигают **нулевую** гипотезу H_0. Это основная гипотеза.

Сущность H_0: разница между сравниваемыми генеральными параметрами = 0, и различия, наблюдаемые между выборочными данными носят **случайный!** характер.

② H_1: Формулируют **альтернативную** гипотезу H_1, конкурирующую с H_0. Это логическое отрицание H_0.

3 α: Задаются уровнем значимости критерия α.

Уровень значимости критерия α – это вероятность ошибки отвергнуть H_0, если на самом деле она верна.

ВОПРОС: Откуда ошибка?

Решение о справедливости H_0 принимается по выборочным данным, т.е. по ограниченному ряду наблюдений. ➡ оно может быть ошибочным.

α задается **заранее!** малым числом.

ВОПРОС: Почему малым?

Потому что это вероятность ошибочного заключения.

ВОПРОС: Каким малым числом?

Обычно это стандартное значение $\alpha \leq 0,05.$

Но можно выбрать более ограничивающее

$\alpha \leq 0,01$ $\alpha \leq 0,001.$

4 $К_{набл}$ (*из <u>выборочных</u> данных*). Для проверки H_0 вычисляют величину **критерия К**, отвечающего H_0.

<u>Статистический критерий</u> – это правило, позволяющее основываясь только на <u>выборке</u> принять или отвергнуть $\boldsymbol{H_0}$.

<u>Критерий</u> – это случайная величина, которая служит для проверки $\boldsymbol{H_0}$. Эти функции распределения <u>табулированы</u> и приводятся в специальных таблицах для различных степеней свободы f (или объема выборки n) и разных α.

5 $К^{крит}_{\alpha,f}$ или $К_{крит}(\alpha, f)$ (*<u>из таблиц).</u>*

По таблице известного распределения вероятности определяют **<u>критическое</u>** значение, превышение которого при справедливости H_0 маловероятно.

6 **Сравнение** $К_{набл}$ и $К_{крит}(\alpha, f)$

$К_{набл} < К^{крит}_{\alpha,f} \Rightarrow H_0$

$К_{набл} > К^{крит}_{\alpha,f} \Rightarrow \not{H_0} \Rightarrow H_1$

7 Интерпретация или **Выводы**

Различие незначимо

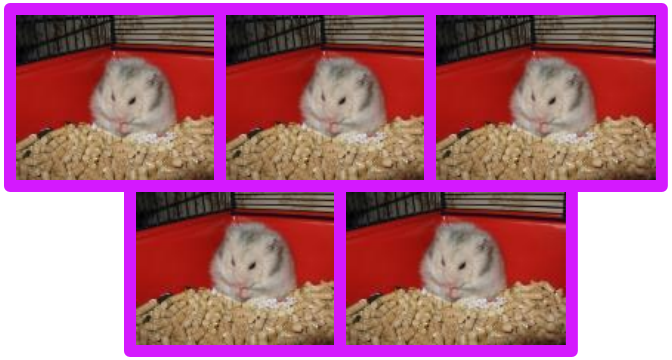
Различие значимо $\alpha \leq 0,05$

Это в случае использования параметрических! критериев.
Если непараметрический критерий, то **наоборот**.

ВОПРОС: Как понимать термин “параметрический критерий”?

Проверка гипотез относительно средних

Одна серия экспериментов



Другая серия, например, **контроль**

Средний результат $\overline{x}_1$ *отличается* $\overline{x}_2$

Возникает вопрос:

это расхождение ***случайно*** или оно вызвано ***некоторыми закономерностями***?

Общая схема проверки гипотезы:

1. Выдвигаем H_0:
2. Выдвигаем H_1:
3. Задаем α
4. Рассчитываем $K_{набл}$ (по выборке)
5. Находим $K_{крит}(\alpha, f)$ (из таблиц)
6. Сравниваем $K_{набл}$ и $K^{крит}_{\alpha,f}$
7. Выводы

1. $H_0: \mu_1 = \mu_2$ или $\bar{X}_{1_{ген}} = \bar{X}_{2_{ген}}$.
2. $H_1: \mu_1 \neq \mu_2$.
3. $\alpha \leq 0,05$.
4. Для проверки H_0 ***можно*** использовать параметрический **критерий Стьюдента**, если выполняются следующие

Требования к критерию Стьюдента (***t*-критерий**)

1. НЗР
2. $\sigma_1^2 = \sigma_2^2$.

По выборочным данным рассчитываем $t_{набл}$, отвечающее H_0.

1908г. $$t_{набл} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\cdot S_1^2 + (n_2-1)\cdot S_2^2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 \cdot n_2}}}.$$

Это отношение имеет *t*-распределение Стьюдента с $f = n_1 + n_2 - 2$ степенями свободы.

5 По таблице известного распределения находим $t^{крит}_{\alpha, f=n_1+n_2-2}$

6 Сравниваем

Выводы:

Если $|t_{набл}| < t^{крит}_{\alpha,f} \Rightarrow H_0$ Различие ***недостоверно***.

Если $|t_{набл}| > t^{крит}_{\alpha,f} \Rightarrow \not{H}_0 \Rightarrow H_1$,

Различие ***достоверно*** $(\alpha \leq 0,05)$.

Различие ***значимо***

1876-1937

t-критерий Стьюдента

f/α	0.5	0.2	0.1	0.05
1.	1,000	3,078	6,314	12,706
2.	0,816	1,886	2,920	4,303
3.	0,765	1,638	2,353	3,182
4.	0,741	1,533	2,132	2,776
5.	0,727	1,476	2,015	2,571
6.	0,718	1,440	1,943	2,447
7.	0,711	1,415	1,895	2,365
8.	0,706	1,397	1,860	2,306
9.	0,703	1,383	1,833	2,262
10.	0,700	1,372	1,812	2,228
11.	0,697	1,363	1,796	2,201
12.	0,695	1,356	1,782	2,179
13.	0,689	1,350	1,771	2,160
14.	0,692	1,345	1,761	2,145
15.	0,691	1,341	1,753	2,131