Математические методы в биологии

Блок 3. Математическая статистика

Лекция 5

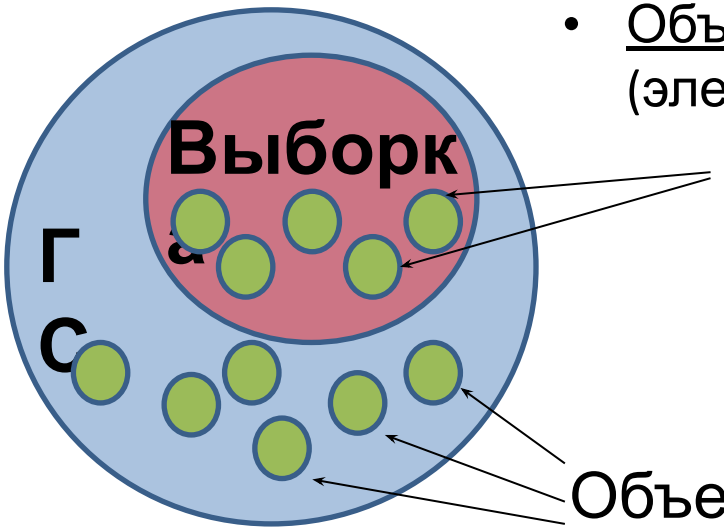
Козлова Ольга Сергеевна
89276755130, olga-sphinx@yandex.ru

Основные определения

- **Генеральная совокупность** – всё то множество объектов, относительно которого исследователь хотел бы делать выводы в рамках определённого исследования

Примеры ГС: все совершеннолетние жители Казани; все люди с заболеванием Альцгеймера

- **Выборка** – некая часть ГС, её модель, на основе изучения которой исследователь делает выводы о всей ГС.
- **Репрезентативность** выборки – её способность отражать существенные для исследования характеристики ГС
  - У объектов изучаются признаки (колич. либо кач.)
  - Объём выборки (или ГС) – количество объектов (элементов), содержащихся в выборке (или ГС)



Объекты (наблюдения)

Объекты

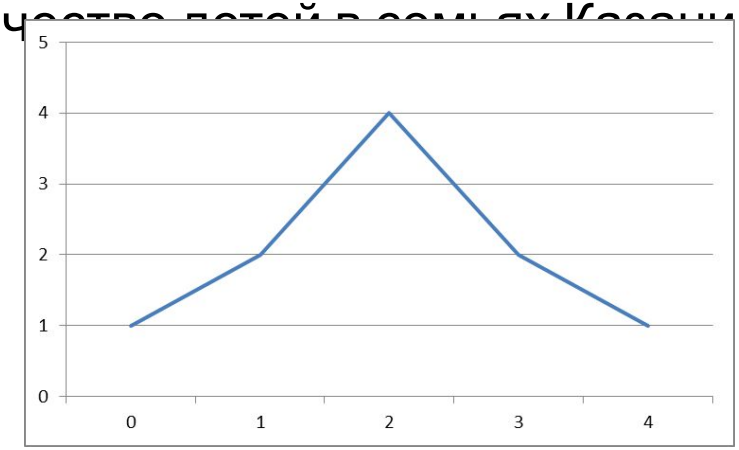
Задача статистики – делать выводы о распределении признака в ГС на основе изучения этого распределения в выборке!!!

Визуализация выборок

- Полигон – график, сопоставляющий варианты значений признака с их частотами (абсолютными или относительными) *(для дискретных признаков)*

Пример. Изучаем количество детей в семьях Казани. Объём выборки – 10 семей.

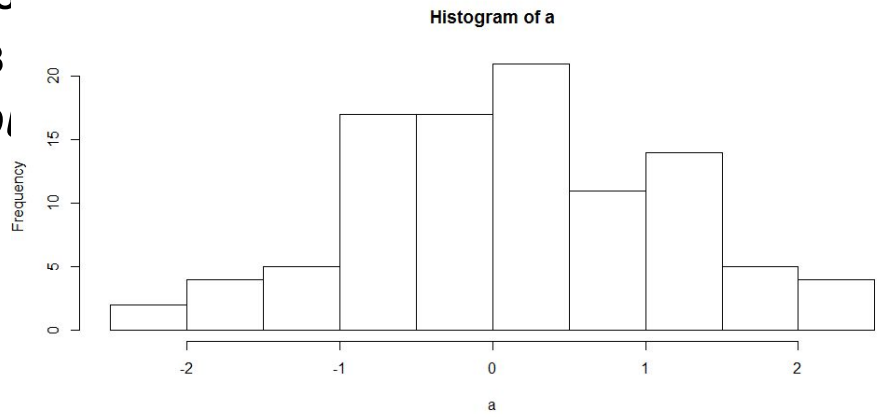
Кол-во семей	Кол-во детей
1	4
2	3
4	2
2	1
1	0



x – кол-во детей в семье;
y – кол-во семей с таким кол-вом детей
Значения по y абсолютные.

- Гистограмма – ступенчатая фигура из прямоугольников с основанием, равным ширине интервала по оси x (значения признака) и высотой, равной частоте значений признака из интервала (абсолютной, относительной) *(для непрерывных признаков)*

Пример. Гистограмма абсолютных частот нормально распределённого признака с параметрами $\mu=0$ и $\sigma=1$ (объём выборки 100)



Описательные статистики

- Их цель – описать, охарактеризовать выборку x безотносительно ГС

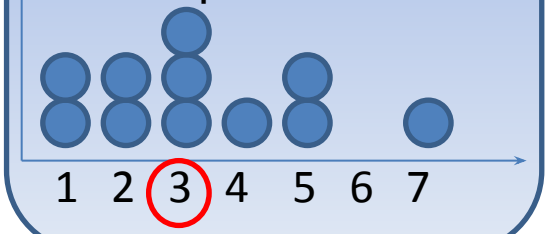
Меры центральной тенденции

- Мода
- Медиана
- Выборочная средняя $\bar{x}$

Меры изменчивости

- Размах
- Выборочная дисперсия D_x
- Стандартное отклонение $sd = \sqrt{D_x}$

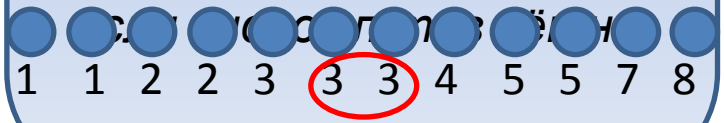
Мода – самое часто встречаемое значение признака в выборке



Мода – единственная описат.статистика для качест.признака;
Мод м.б. несколько и 0

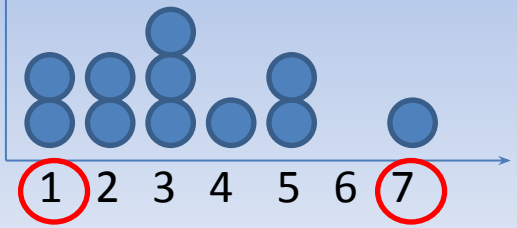
Медиана – то значение признака, которое делит упорядоченную выборку пополам:

если число эл-тов нечётно:



(3+3)/2=3 (ср.арифм.е)

Размах – расстояние между x_{min} и x_{max}



7-1=6

Выборочная средняя и выборочная дисперсия

- Выборочная средняя - среднее арифметическое всех значений признака в выборке:

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

- Выборочная средняя – случайная величина (изменяется от выборки к выборке для одной и той же ГС)
- Мат.ожидание выборочной средней как случайной величины есть **истинная средняя** – средняя для всей ГС: $M(\bar{x}) = \mu$ (т.е. $\bar{x}$ - ***несмещённая*** оценка μ)
- Сумма отклонений значений признака от выборочной средней равна 0:

$$\sum_{i=1}^{n} x_i - \bar{x} = 0$$

- Выборочная дисперсия – сумма квадратов отклонений значений признака от выборочной средней, делённая на n-1 (n – объём выборки):

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

- Выборочную дисперсию можно вычислять так:

$$D_x = \overline{x^2} - \bar{x}^2$$

(Все правила для M(X) и D(X) (лекции №№3,4) переводятся на язык выборок)

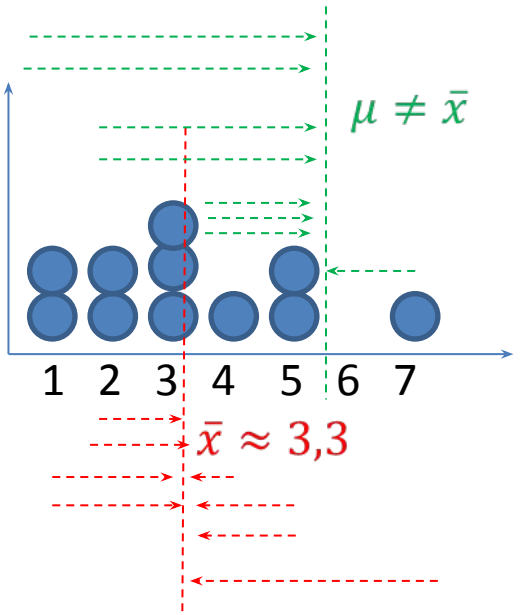
И всё же, откуда в формуле выборочной дисперсии (n-1)?

Рассмотрим «очевидное» выражение для выборочной дисперсии:

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

Это случ.величина, зависящая от выборки дважды (так как включает в себя случайную величину $\bar{x}$, а не неизвестную константу μ).

Сумма квадратов отклонений значений признака от $\bar{x}$ меньше, чем сумма квадратов отклонений значений признака от любого другого числа (в т.ч. от постоянной μ),



поэтому $D_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ - это всегда ***заниженная, смещённая оценка*** генеральной (истинной) дисперсии.

Чтобы «приподнять» её, используют поправочный коэффициент $\frac{n}{n-1}$ $(\text{стремится к } 1 \text{ при } n \to \infty)$

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

Использование поправочного коэффициента имеет обязательный характер при n<30 и практически не влияет на значение дисперсии и станд.отклонения при n≥100.

Стандартная ошибка среднего (SE)

- Выборочная средняя $\bar{x}$ - случайная величина с мат.ожиданием, равным μ

ВОПРОС: чему равна дисперсия $\bar{x}$? Пусть из ГС извлечено много выборок одинакового объёма n. Тогда

$$D(\bar{x}) = D\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) = \frac{1}{n^2} D(x_1 + x_2 + \cdots + x_n)$$

$$= \frac{1}{n^2}\left(D(x_1) + \cdots + D(x_n)\right) = \frac{1}{n^2} * n * D(x) = \frac{D(x)}{n}$$

Истинная дисперсия

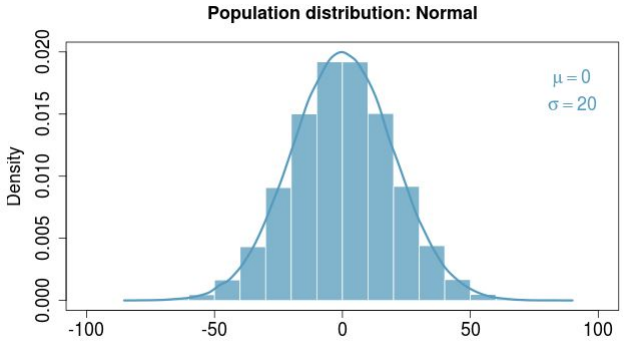
Все наблюдения – из одной ГС, значит, и изменчивость одинакова

Используем выб.дисперсию D_x как оценку $D(x)$:

Работает для n≥30!

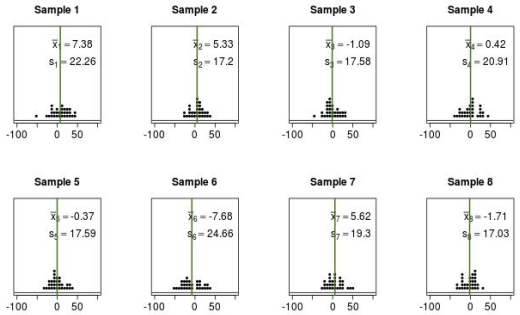
$$D(\bar{x}) = \frac{D_x}{n} \Rightarrow sd(\bar{x}) = \frac{sd(x)}{\sqrt{n}}$$

Формула стандартной ошибки

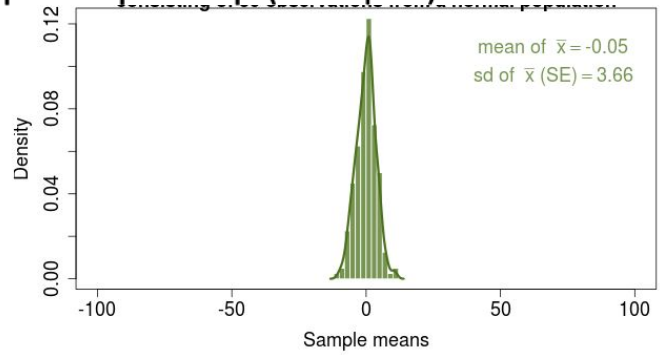


Генеральная совокупность

Т.к. $\bar{x}$ - это, по сути, сумма взаимно независ.сл.вел, то распределение $\bar{x}$ имеет норм.характер (по ЦПТ)



Выборки объёма



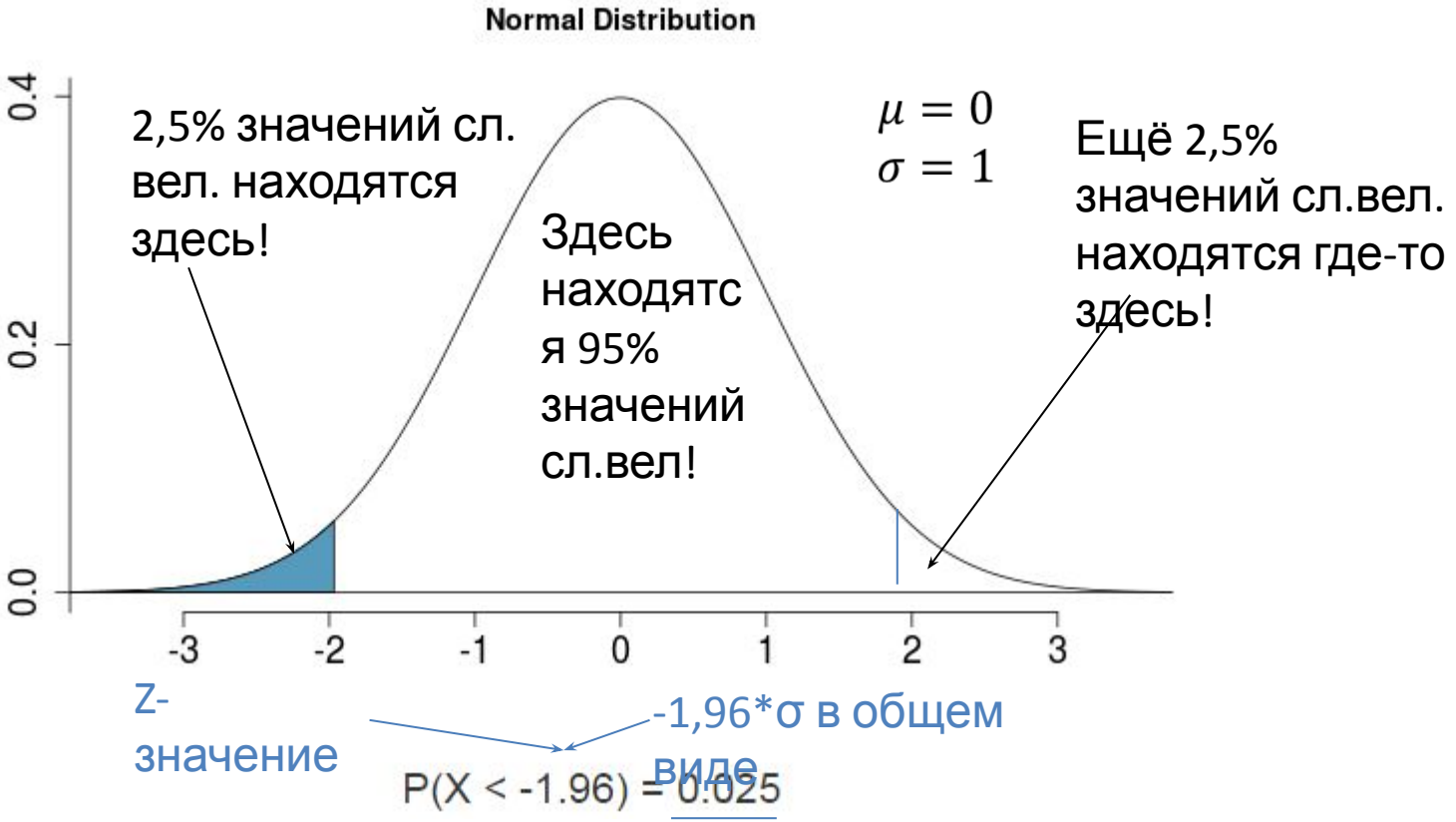
Распределение выб.

Построение доверительного интервала для среднего

- Пусть у нас есть выборка, и мы знаем $\bar{x}$ и *sd*

ВОПРОС: Можем ли мы найти μ (истинное среднее)?

ОТВЕТ: И да, и нет!! Точное значение μ мы не узнаем, но можем указать численный интервал, в котором μ находится с определённой вероятностью (этот интервал называется ***доверительным***).



Дов. инт-л – <u>симметричный</u> относительно мат.ожидания интервал, насчёт которого мы можем на сколько-то уверенно сказать, что там находится

Построение доверительного интервала для среднего. Пример

Пусть есть некая выборка из 64 наблюдений с выборочным средним, равным 100, и стандартным отклонением, равным 4. Построить 95%-ный доверительный интервал для истинного среднего.

Решение. n=64, $\bar{x}$=100, sd=4.

Рассчитаем стандартную ошибку среднего: $SE = sd(\bar{x}) = \frac{sd}{\sqrt{n}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0{,}5$

Истинное среднее имеет нормальное распределение с $\mu = \bar{x} = 100$ и σ=0,5.

95% значений истинного среднего расположены в интервале

от $\bar{x} - 1.96 * 0{,}5$ до $\bar{x} + 1.96 * 0{,}5$, значит, *мы можем на 95% быть уверенны в том, что мат.ожидание (истинное среднее) находится где-то на отрезке [99,02;100,98].*

А как же другие интервалы?

%	Z-значение (то, на что умножать SE)
90	1,645
95	1,96
99	2,575

Гипотезы и их проверка

Понятие статистической гипотезы

Статистическая гипотеза - некое предположение о виде неизвестного распределения или о его параметрах.

Примеры статистических гипотез:

1) *Распределение роста студентов нормально*
2) *Средняя продолжительность жизни в России – 67 лет*

Нулевая гипотеза (H_0)– основное предположение, выдвинутое в статистическом исследовании (обычно пессимистична).

Альтернативная гипотеза (H_1) – гипотеза, противоречащая нулевой.

Гипотезы проверяются **статистическими тестами**.

Результат статистического теста – отклонение (☺) или не отклонение нулевой гипотезы (☹)

Отклонение нулевой гипотезы означает принятие альтернативной (☺)

НО: Не отклонение нулевой гипотезы – это ещё не отклонение альтернативной!

Типовой пример статистической задачи на проверку гипотез

Средний срок выздоровления от некоторого заболевания – 20 дней. Для борьбы с заболеванием было разработано новое лекарство. Данные по его применению: n=64, $\bar{x} = 18{,}5, sd = 4$.

ВОПРОС: Действительно ли новое лекарство влияет на срок выздоровления или эти различия случайны (попалась «везучая» выборка)?

ПОСТАНОВКА ГИПОТЕЗ:

H_0: $\mu = 20$ (мат.ожидание случайной величины «средний срок выздоровления после приёма нового лекарства» не отличается от 20, т.е. наблюдаемые различия носят случайный характер)

H_1: $\mu \neq 20$ (различия не случайны, лекарство влияет на срок выздоровления)

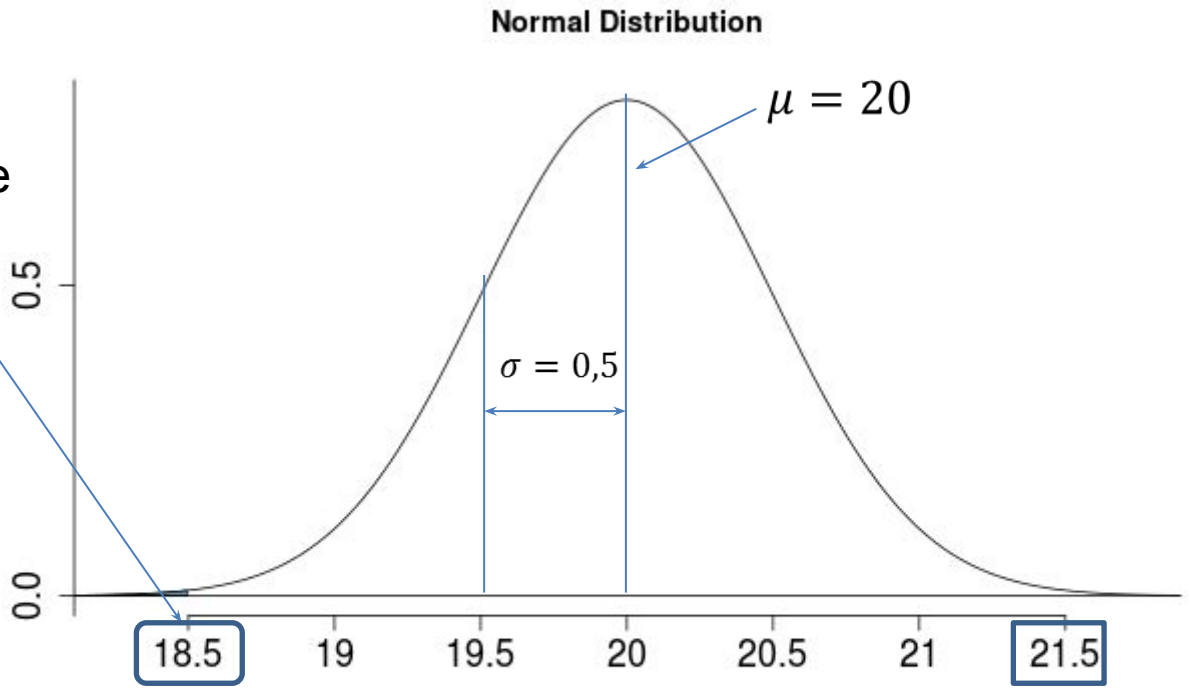
ДОПУСТИМ, ВЕРНА НУЛЕВАЯ ГИПОТЕЗА (в среднем, выборка из 64 человек выздоравливает за 20 дней даже после нового лекарства)

Рассчитаем стандартную ошибку среднего: $SE = \frac{4}{\sqrt{64}} = 0{,}5$

При условии соблюдения H_0 случайная величина «средний срок выздоровления после приёма нового лекарства» имеет нормальное распределение с мат.ожиданием 20 и стандартным отклонением 0,5.

Распределение сл.вел. «средний срок выздоровления после приёма нового лекарства» при условии принятия H_0

А вот где наше выборочное среднее!!



Если мы принимаем H_0, то наше выборочное среднее отклоняется от 20 аж на 3 стандартных отклонения! Вероятность наблюдать такие и более серьёзные отклонения составляет 0,00135*2=**0,0027 (p-value, или уровень значимости)** .

P-value (уровень значимости)

- Это вероятность наблюдения заданных отклонений (различий) при условии, что верна H_0 (вероятность случайности заданного выборочного значения)
- Чем меньше, тем большее право имеем на отклонение H_0
- «Золотой стандарт» порогового уровня p-value – 0,05 (<0,05 – отклоняем H_0 и принимаем H_1, если ≥0,05 – оснований для отклонения H_0 недостаточно!)
- Обычно двусторонний (вычисляем вероятность отклонения как в одну, так и в другую сторону)

Статистические ошибки

- Ошибка первого рода – отклонили H_0, хотя она была верна (выборочные данные были получены случайно)

Последствия – получили ложно статистически значимый вывод.

Возможный способ борьбы – уменьшить пороговое p-value (до 0,001, например). P-value – вероятность совершить ошибку первого рода.

- Ошибка второго рода – не отклонили H_0, хотя она не была верна (верна H_1).

Последствия – не получили статистического вывода.

Возможный способ борьбы – увеличить объём выборки.

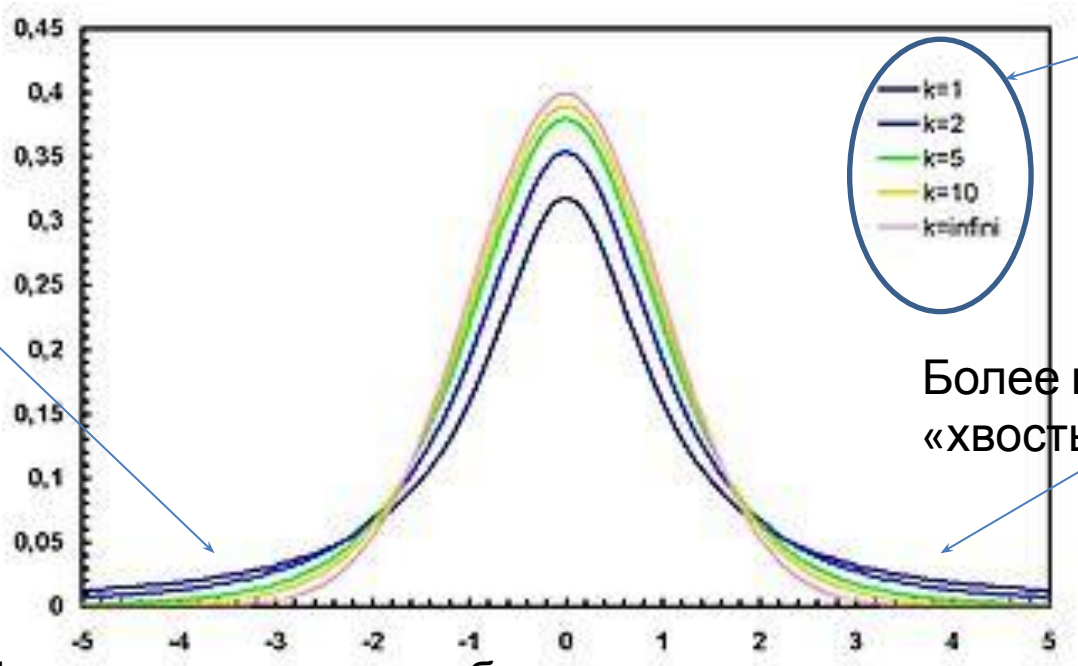
Чем чреваты маленькие выборки (n<30)

- Выборочные средние сильнее отклоняются от μ => нарушаются условия ЦПТ, т.е. нормальность распределения $\bar{x}$
- Выборочные стандартные отклонения хуже описывают истинные => не имеем права заменить истинное ст.отклонение на выборочное в формуле для вычисления стандартной ошибки среднего

Что же делать?

Заменить нормальное распределение для $\bar{x}$распределением Стьюдента (t-рас-е)

Увеличивается вер-ть попадания с.в. в крайние интервалы (±2σ)



Пар-р k (число степеней свободы) k=n-1 (n-объём выборки)

Более высокие «хвосты»

N(0,1)

$$T = \frac{Z}{\sqrt{V/k}}$$

$\chi^2(k)$

Бледно-фиолетовая линия – обыкновенное нормальное

Нормальное распределение vs распределение Стьюдента

- Вероятностные характеристики N постоянны – для t они зависят от k (k=n-1, так как, зная выборочное среднее, последнее значение тоже известно)

Пример. Есть выборка с параметрами: $\bar{x} = 10{,}8; \quad sd = 2; \quad n = 25$

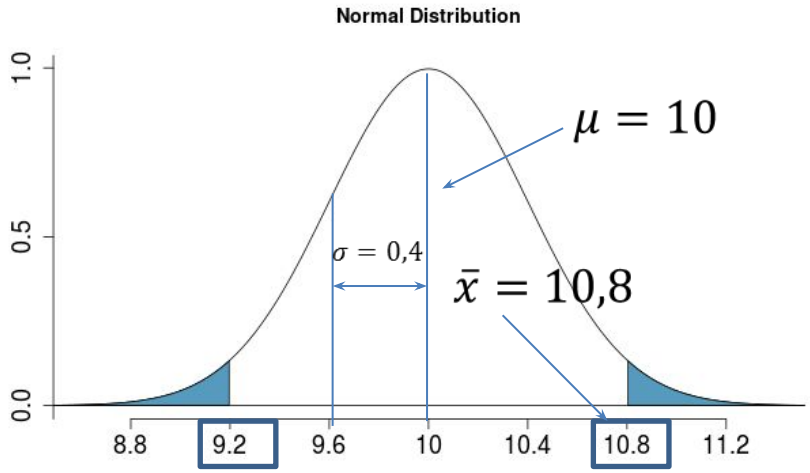
H_0: $\mu = 10$

По нормальному распределению:

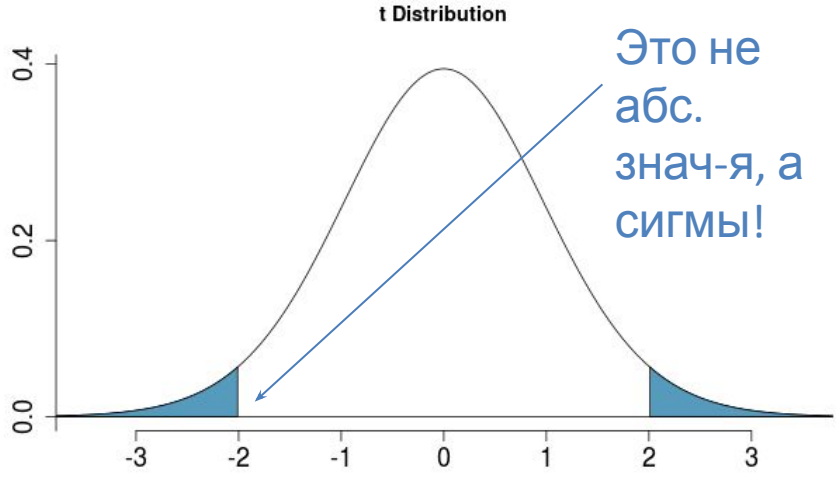
$$SE = \frac{2}{\sqrt{25}} = 0{,}4$$

По распределению Стьюдента:

k (число степ.свободы)=25-1=24



P(X < 9.2 or X > 10.8) = 0.0455



P(X < -2 or X > 2) = 0.0569

Вер-ть наблюдать случайное отклонение на $\geq 2\sigma$

p-value=0,0455 <0,05=> H_0

p-value=0,0569 > 0,05=> H_0 не

Сравнение средних – парный t-тест

Постановка задачи.

Есть две выборки:

$$\overline{x_1} = x_1$$
$$sd = sd_1$$
$$n = n_1$$

$$\overline{x_2} = x_2$$
$$sd = sd_2$$
$$n = n_2$$

$H_0: \mu_1 = \mu_2$ — Выборки принадлежат одной ГС

$H_1: \mu_1 \neq \mu_2$ — Выборки принадлежат разным ГС

$\overline{x_1}$ и $\overline{x_2}$ - случайные величины => $\overline{x_1} - \overline{x_2}$ - тоже случайная величина.

ЕСЛИ ВЕРНА НУЛЕВАЯ ГИПОТЕЗА, то она распределена с $\mu(\overline{x_1} - \overline{x_2}) = 0$ и

$sd(\overline{x_1} - \overline{x_2}) = \sqrt{\frac{{sd_1}^2}{n_1} + \frac{{sd_2}^2}{n_2}}$ (квадратный корень суммы квадратов стандартных ошибок средних) (т.к. $D(X - Y) = D(X) + D(Y)$, см. пред.презентацию)

Это t-распределение с числом степеней свободы k = n_1-1+n_2-1=n_1+n_2-2

$t = \frac{\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{{sd_1}^2}{n_1} + \frac{{sd_2}^2}{n_2}}}$ ($\mu_1 - \mu_2 = 0$) - величина, показывающая, сколько станд.отклонений укладывается в отклонение между с.в. $\overline{x_1} - \overline{x_2}$ и конст. $\mu_1 - \mu_2$

p-value

Пример на сравнение средних

Процесс денатурации ДНК (разрушения водородных связей между её цепями) зависит от температуры, которая может различаться у разных видов.

В исследовании сравнивали температуру денатурации ДНК у двух биологических видов.

		sd	n
Вид №1	89,9	11,3	20
Вид №2	80,7	11,7	20

ВОПРОС: Правда ли, что у вида 1 и вида 2 разная температура денатурации (**<u>являются ли различия статистически значимыми</u>**)?

ФОРМУЛИРОВКА ГИПОТЕЗ:

H_0: $\mu_1 = \mu_2$

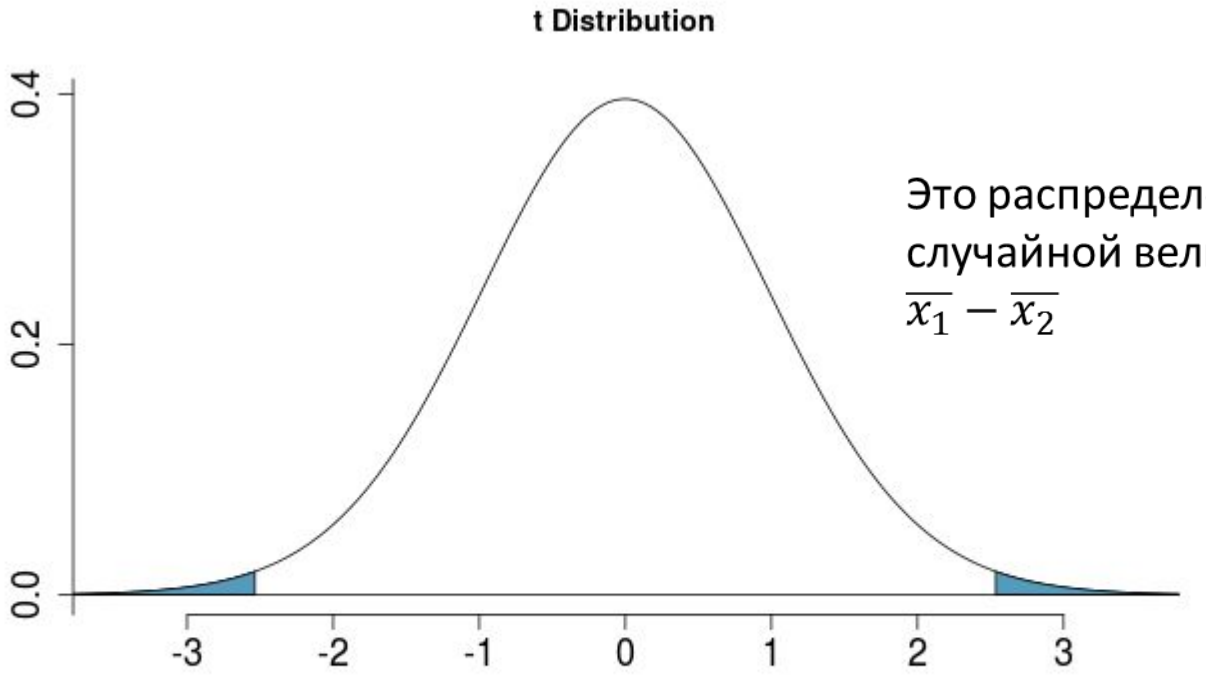
H_1: $\mu_1 \neq \mu_2$

Рассчитаем t-значение: $t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{sd_1}^2}{n_1} + \frac{{sd_2}^2}{n_2}}} = \frac{89{,}9 - 80{,}7}{\sqrt{\frac{11{,}3^2}{20} + \frac{11{,}7^2}{20}}} = \frac{9{,}2}{\sqrt{\frac{127{,}69 + 136{,}89}{20}}} \approx 2{,}53$

(разность между выб.средними отклонилась от разности между мат.ожиданиями на 2,53 стандартных отклонений)

t-распределение с 38 степенями свободы

(38=20+20-2)



Это распределение случайной величины $\overline{x_1} - \overline{x_2}$

При условии, что верна нулевая гипотеза, вероятность наблюдать отклонение разности выборочных средних от 0 в более чем 2,53 стандартных отклонения равна 0,0157 (<0,05) => отклоняем нулевую гипотезу (разница между температурами денатурации статистически значима).

Резюме (или что мы умеем делать из статистики)

1. Строить доверительный интервал для среднего с исп-ем ЦПТ (например, $(\bar{x} - 1.96 * SE;\ \bar{x} + 1.96 * SE)$ для 95%-ного интервала)
2. Проверять гипотезу о соответствии мат.ожидания выборочного среднего числу с исп-ем свойств норм.распр-я и t-распр-я
3. Проводить t-тест и сравнивать средние двух выборок