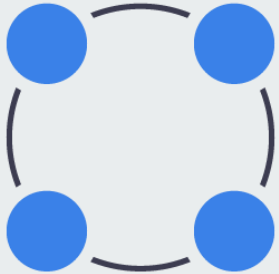
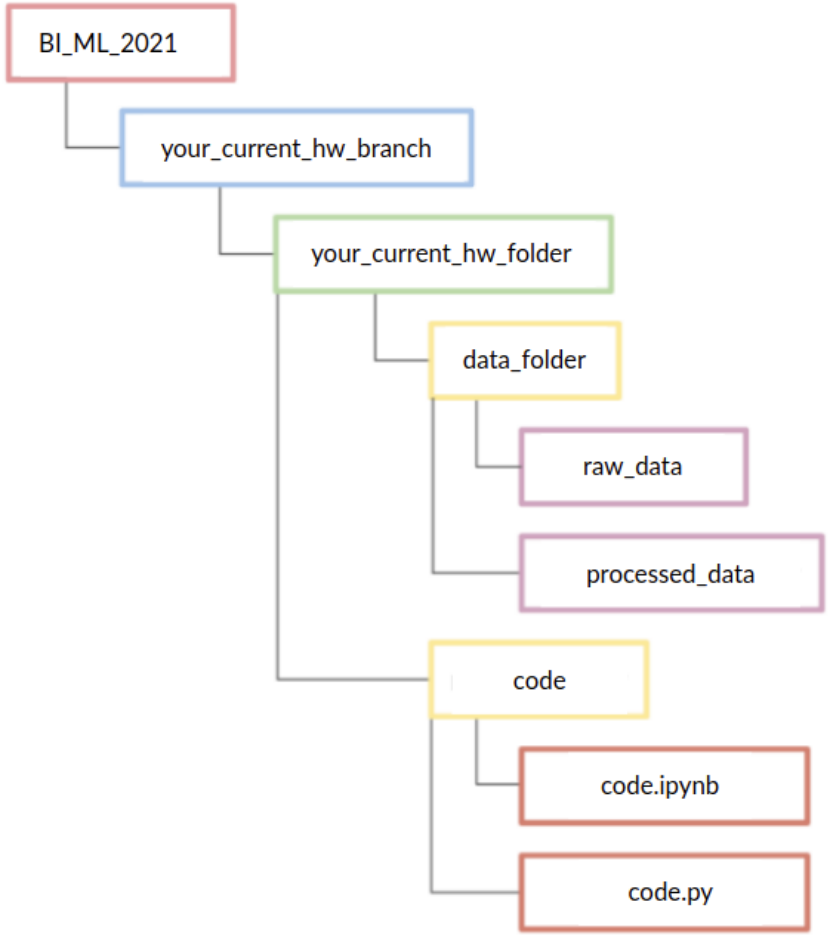
Статистика и анализ данных

Лекция 4. Введение в статистику

(08.10.2022)




BI_ML_2021
your_current_hw_branch
your_current_hw_folder
data_folder
raw_data
processed_data
code
code.ipynb
code.py

Recap

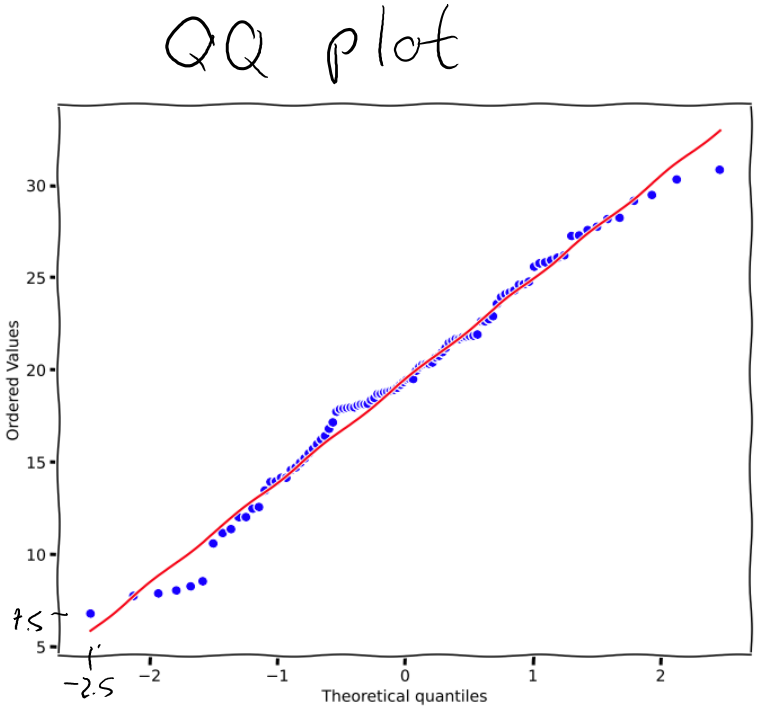
Квантили

Квантили — это значения, которые делят ряд наблюдений на N равных частей.

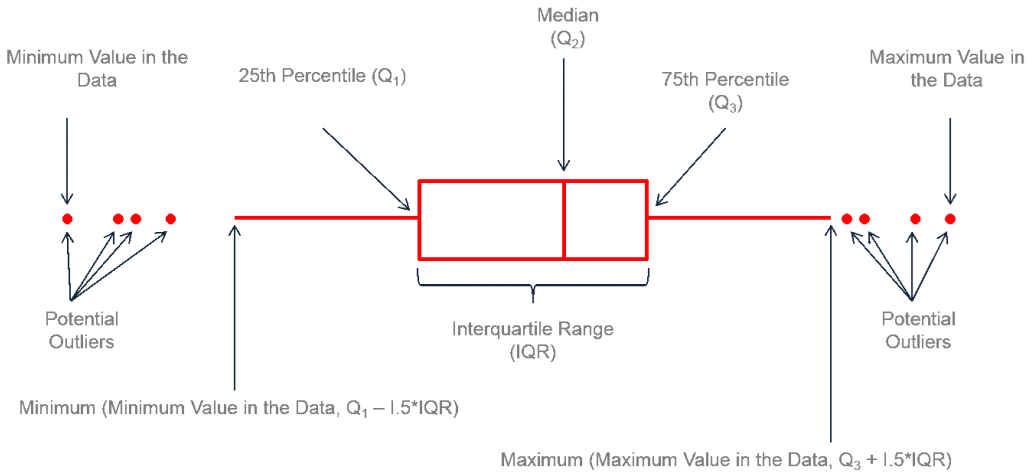
Возможные варианты квантилей:

- 2-квантиль — медиана
- 4-квантиль — квартиль
- 100-квантиль — перцентиль

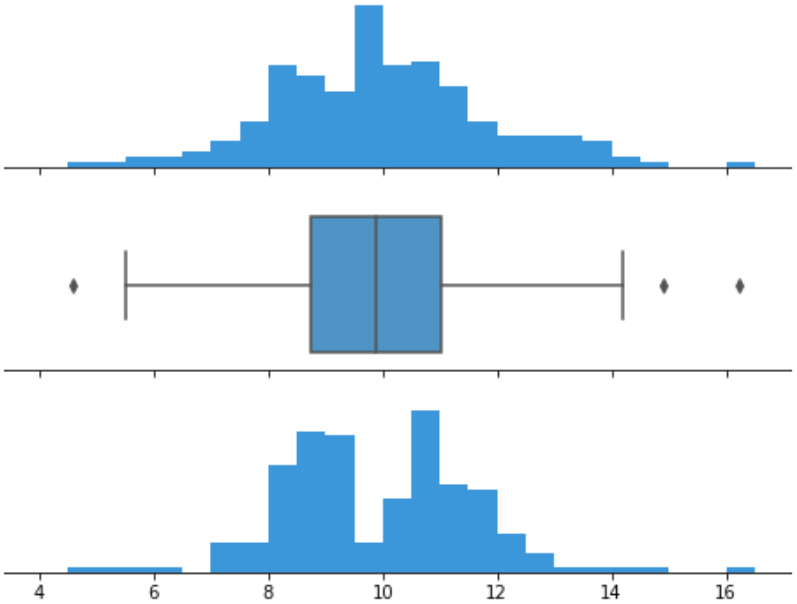
[300, 2500, 2800, 3000, 3100, 3400, 3600, 7000]



Боксплот (график)



Боксплот (недостатки)

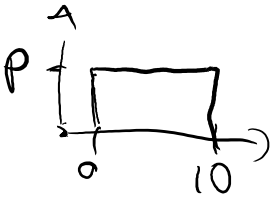


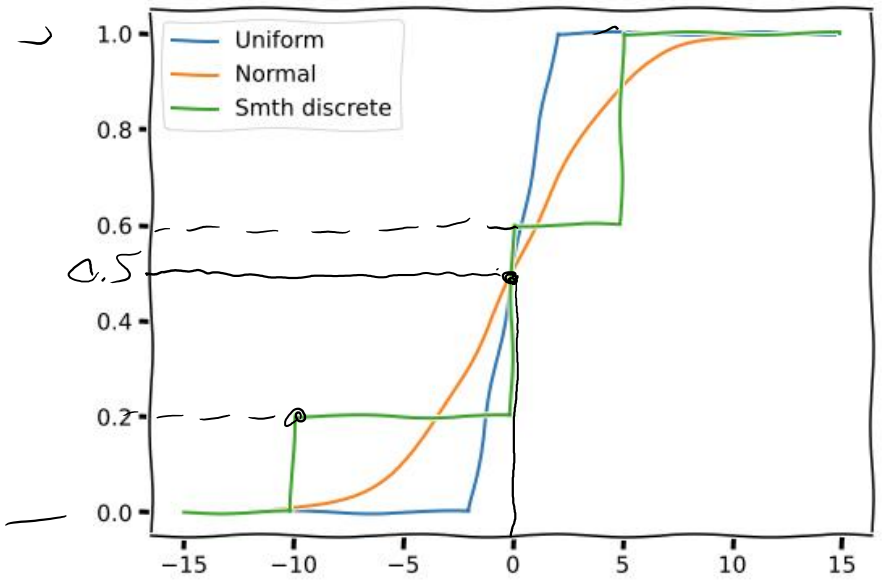
Распределения

Функция распределения

$$F(X) = P(\xi \leq X)$$

Функция распределения — это такая функция, которая для значения X, равна вероятности получить значение меньшее или равное этому X.
Например, для роста F(180 см) = ?



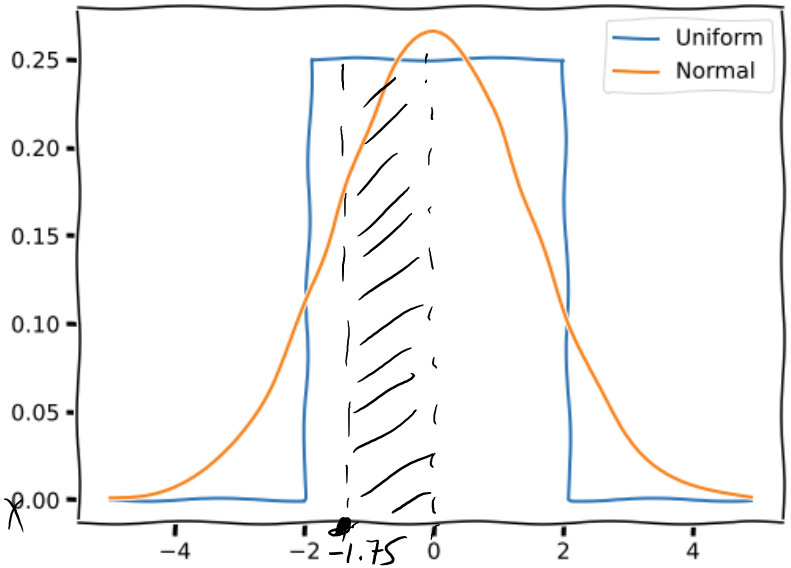


Функция плотности

Пришло время наконец узнать, что же такое *непрерывная СВ*.
Непрерывной СВ называется такая СВ, для которой существует функция *f(x)*, называемая **функцией плотности**, такая что:

$$F(X) = \int_{-\infty}^{X} f(x)\,dx$$

$$\int_{-1.75}^{\infty} f(x)\,dx$$



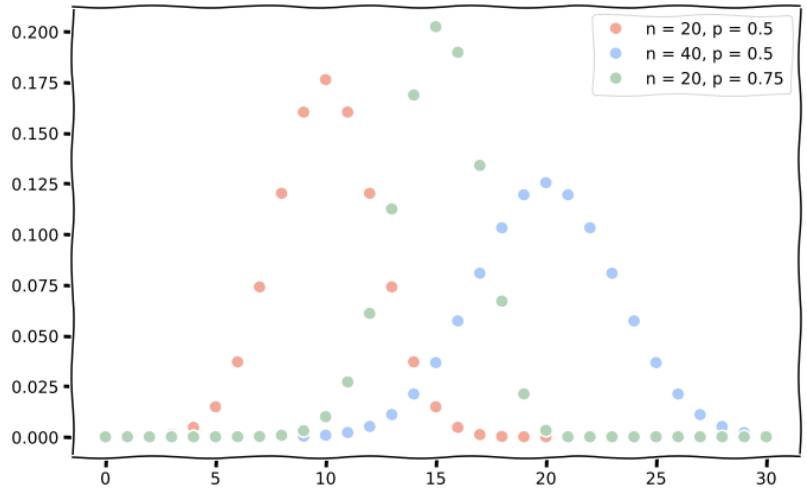
$$f(-1.75) = 0.25$$

Биномиальное распределение

Биномиальное распределение в теории вероятностей — распределение количества «успехов» в последовательности из n независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна p.

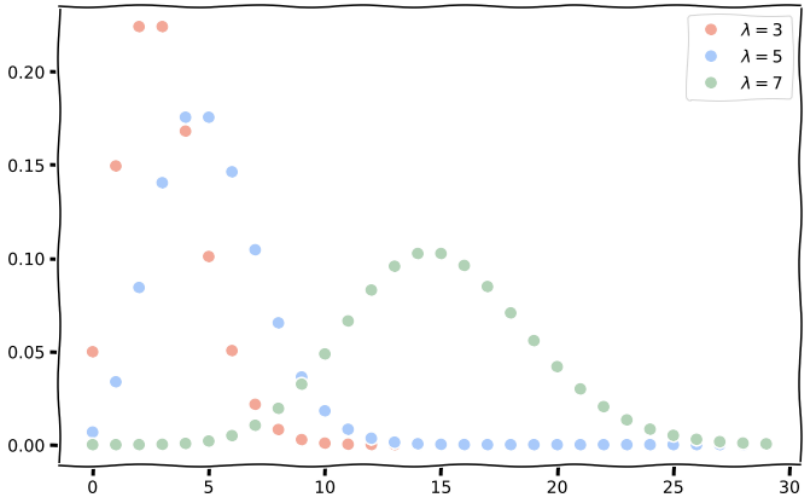
.

$$p(k) \equiv \mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Распределение Пуассона

Распределение Пуассона — распределение дискретного типа случайной величины, представляющей собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга.



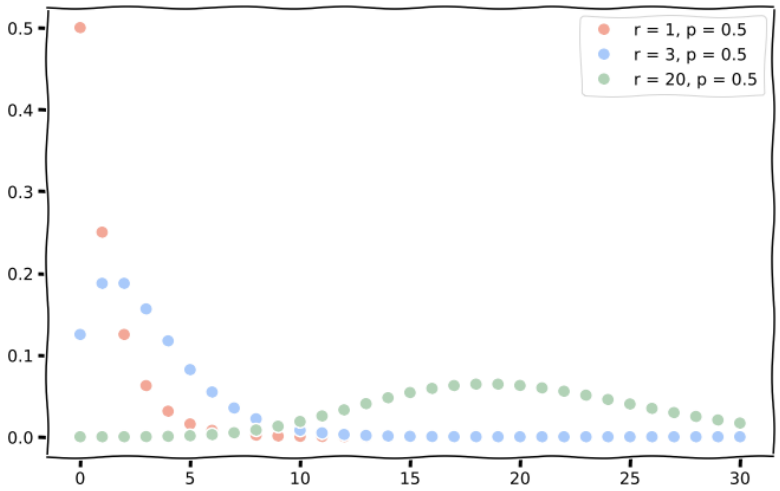
$$p(k) \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\lim_{n \to \infty} \binom{n}{k} \cdot p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} \cdot e^{-\lambda} \qquad \lambda = n \cdot p$$

Отрицательное биномиальное

Отрицательное биномиальное распределение, также называемое распределением Паскаля — это распределение дискретной случайной величины, равной числу произошедших неудач в последовательности испытаний Бернулли с вероятностью успеха p, проводимых до r-го успеха.

$$p(k) \equiv \mathbb{P}(Y = k) = \binom{k+r-1}{k} p^r (1-p)^k$$



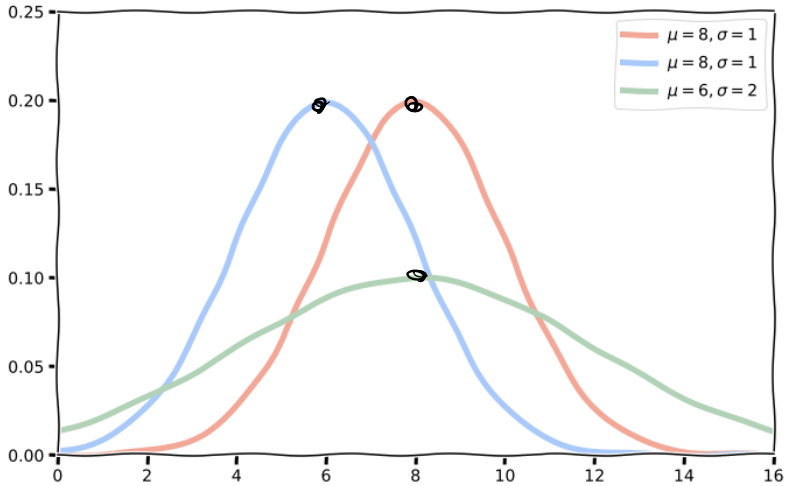
Нормальное распределение

Параметры:

μ — математическое ожидание

σ — стандартное отклонение

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



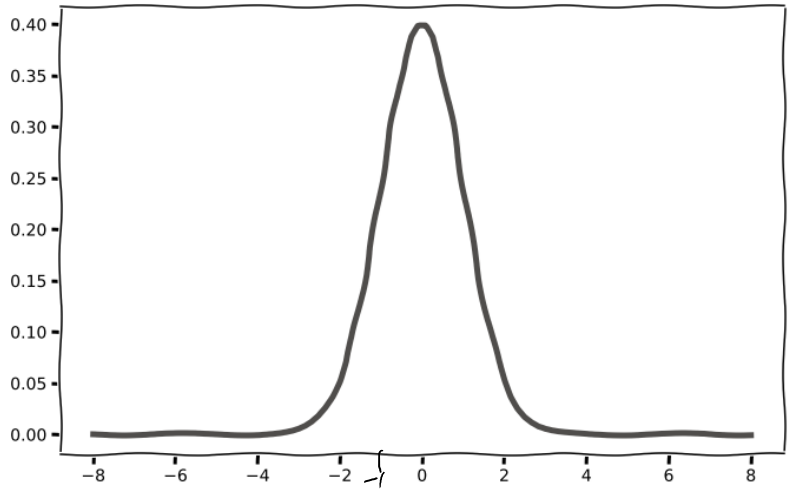
Стандартное нормальное распределение

Параметры:

μ = 0

σ = 1

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



$$z = \frac{x - \bar{x}}{\sigma}$$

Стандартизация

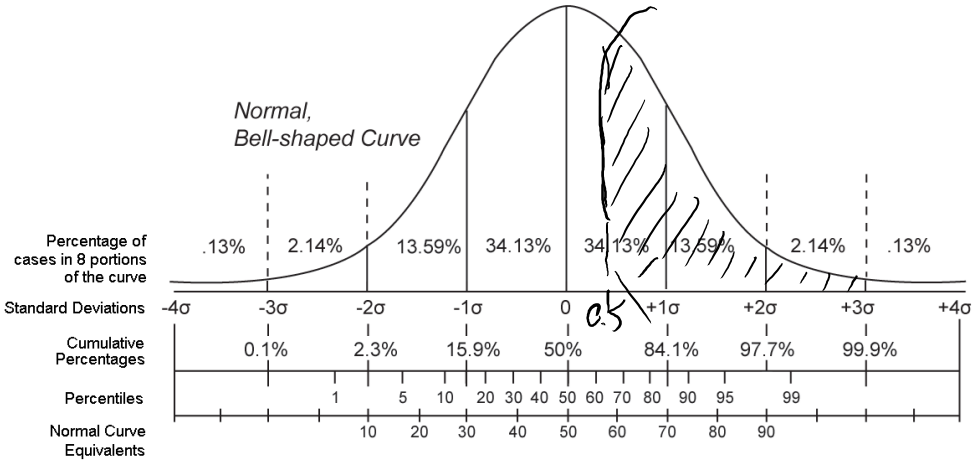
Нормальное распределение

1. Унимодально
2. Симметрично
3. Отклонения подчиняются закону

Например:

- В диапазоне от среднего до 1σ (одного стандартного отклонения) будет находиться примерно 34.1% всех наблюдений
- В диапазоне от 1σ до 2σ – примерно 13.6%
- Очень маловероятно встретить наблюдение, которое бы превосходило среднее значение больше чем на 3 стандартных отклонения (3σ)

Отклонение от среднего равновероятно как в большую, так и в меньшую стороны.



Правило "двух" и "трех" сигм

- $M_x \pm \sigma \approx 68\%$ наблюдений находятся в этом интервале
- $M_x \pm 2\sigma \approx 95\%$ наблюдений находятся в этом интервале
- $M_x \pm 3\sigma \approx 100\%$ наблюдений находятся в этом интервале

Пример: Среднее значение равняется 150, а стандартное отклонение равно 8. Какой процент наблюдений превосходит значение, равное 154?

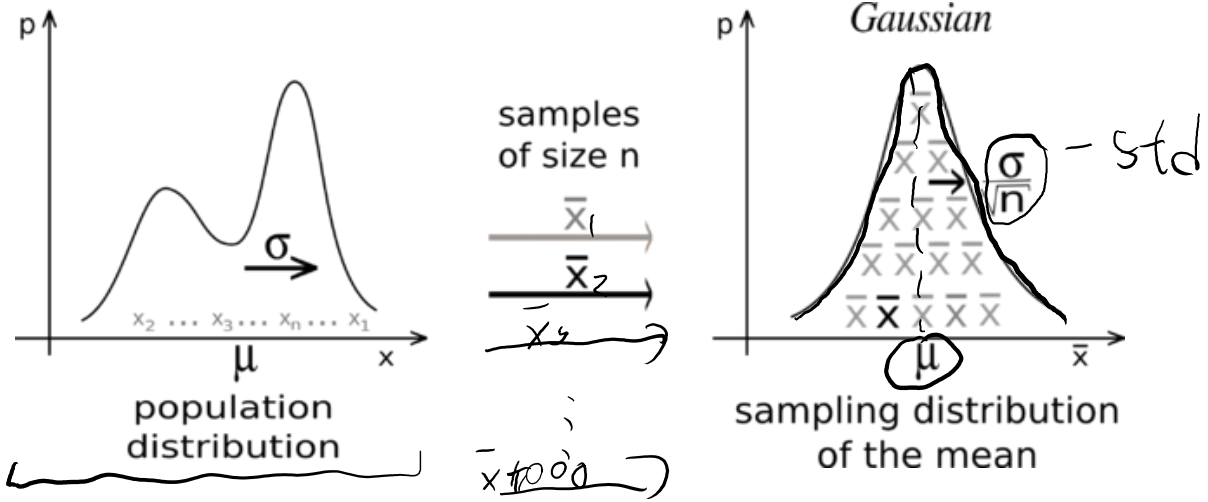
Для этого нужно сделать Z-преобразование. Как найти интересующее нас Z-значение? Из 154 нужно вычесть среднее значение по нашей выборке и разделить на стандартное отклонение. В результате:

$$z = \frac{154 - 150}{8} = \frac{4}{8} = 0.5$$

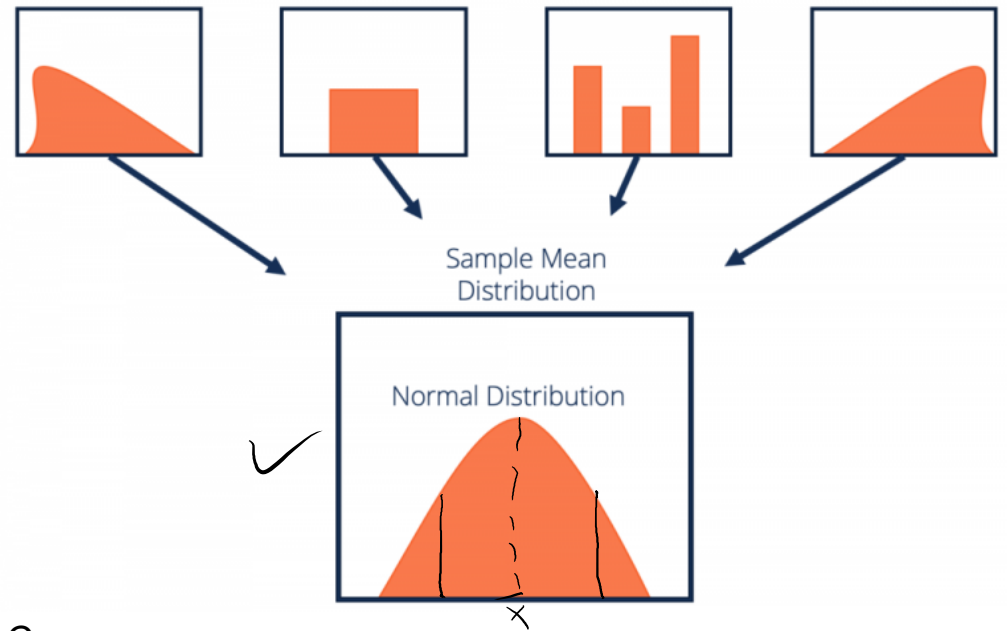
$$z = \frac{x - \bar{x}}{\sigma}$$

Центральная предельная теорема

Класс теорем в теории вероятностей, утверждающих, что сумма достаточно большого количества слабо зависимых одинаково распределенных случайных величин, имеющих примерно одинаковые масштабы, имеет распределение, близкое к нормальному (wikipedia).



Центральная предельная теорема



Стандартное отклонение этого распределения называется стандартной ошибкой среднего. Она показывает, насколько выборочные средние отклоняются от среднего ГС.

Пример

Проверим

colab

Стандартная ошибка среднего

Стандартная ошибка среднего (SE) показывает, насколько выборочные средние "разбросаны" вокруг среднего генеральной совокупности. SE при увеличении размера выборки будет стремиться к нулю.

1) $$se = \frac{\sigma}{\sqrt{n}}$$

- std of population
- sample size

Если выборка репрезентативна и число наблюдений достаточно велико, то в качестве стандартного отклонения ГС мы можем использовать стандартное отклонение нашей выборки:

2) $$se = \frac{sd_x}{\sqrt{n}}$$

Стандартная ошибка среднего

числа в выборке

1. $T = (x_1 + x_2 + \cdots + x_n)$

2. $\mathrm{Var}(T) = \big(\mathrm{Var}(x_1) + \mathrm{Var}(x_2) + \cdots + \mathrm{Var}(x_n)\big) = n\sigma^2.$

3. $\bar{x} = T/n.$

4. $\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{T}{n}\right) = \frac{1}{n^2}\mathrm{Var}(T) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$

5. $\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$

$\frac{std\,x}{\sqrt{n}}$

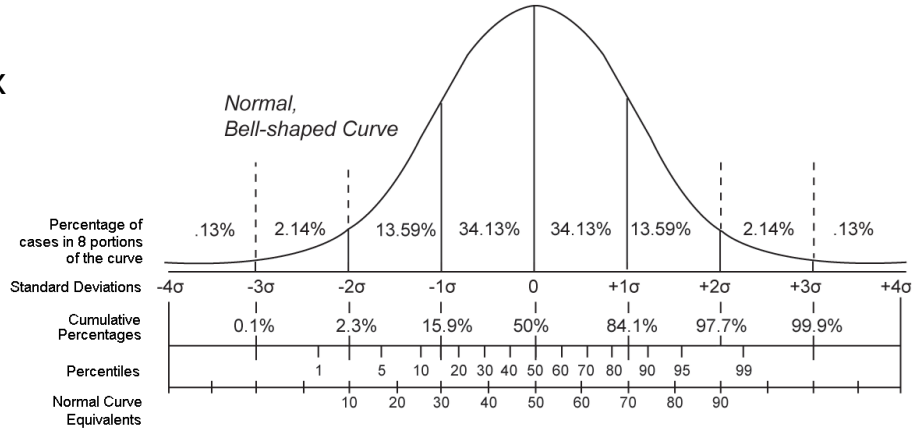
Доверительный интервал при известной дисперсии

Интервал такой ширины, что при многократном повторении эксперимента в 95% из полученных интервалов будет среднее ГС:

$$\overline{x} \pm 1.96 \cdot se$$

И в 99%:

$$\overline{x} \pm 2.58 \cdot se$$



sample $[x_1, x_2, \ldots x_n]$

$\overline{x}$

$se = \frac{6}{\sqrt{n}}$

$\overline{x} \pm 1.96 \cdot se$

$$se = \frac{\sigma}{\sqrt{n}}$$

[175 – 185]

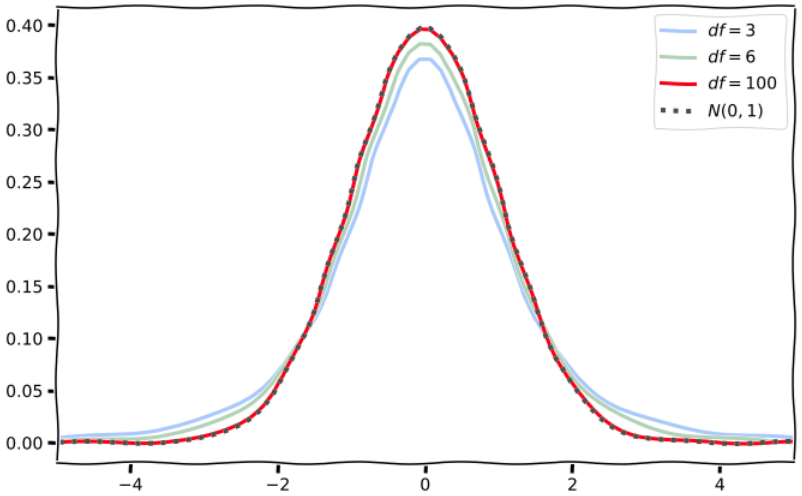
Тренируемся

colab

Доверительный интервал при неизвестной дисперсии

Если число наблюдений в выборке невелико и σ (стандартное отклонение генеральной совокупности) неизвестно (почти всегда), используется распределение Стьюдента (T-distribution), чтобы описать, как будут себя вести все выборочные средние.

1. Унимодально
2. Симметрично
3. Но: наблюдения с большей вероятностью попадают за пределы ±2σ от M



$$\overline{x} \pm t_{0.95} \cdot se$$

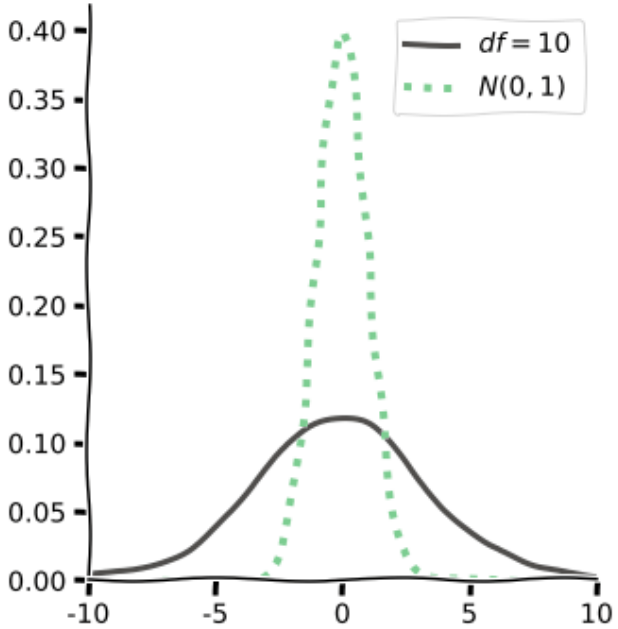
$$\overline{x} \pm t_{0.99} \cdot se$$

$$se = \frac{sd_x}{\sqrt{n}}$$

t-распределение

n — число степеней свободы. На деле это означает, сколько Y_i мы суммируем в знаменателе. По сути мы получаем число (Y_0) из распределения $N(0, 1)$, после чего получаем еще n чисел (Y_i) из такого же распределения. В конце остается лишь подставить их в данную формулу, и вы получите какое-то значение из t-распределения.

$$t = \frac{Y_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}$$



Тестирование гипотез

Тестирование гипотез

- Гипотезы H0 и HA должны быть взаимоисключающими
- Нулевая гипотеза H0 — описание ситуации отсутствия различий
- Альтернативная гипотеза HA — вопрос исследователя и формулируется до начала эксперимента
- Двусторонняя/односторонняя альтернативная гипотеза

Примеры гипотез

...

Проверяем гипотезы

colab

Итоги

1. Узнали, что такое функция распределения, и несколько распределений
2. Поняли, как работает ЦПТ
3. Узнали про стандартную ошибку среднего
4. Научились тестировать гипотезы с помощью доверительных интервалов