ML на Python

Неделя 3. День 9. Статистика

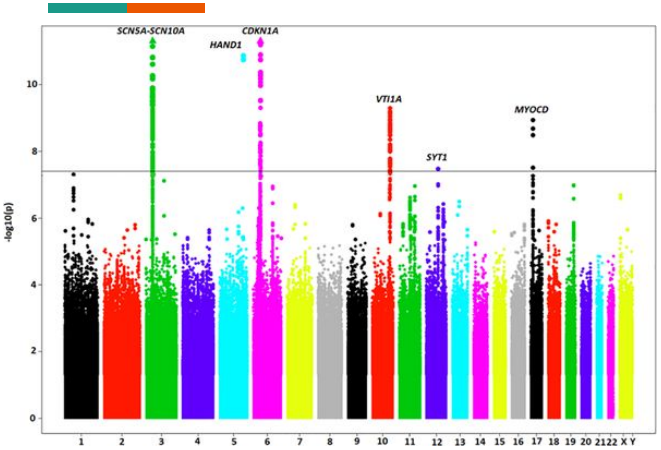
(02.03.2024)

Кто я?

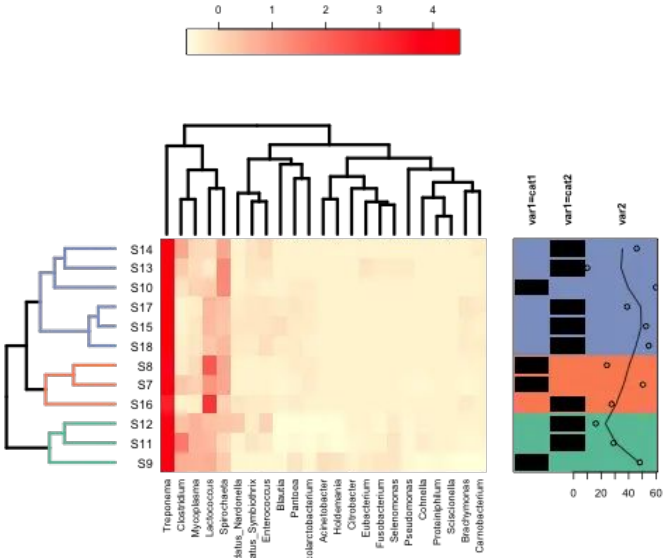
- BSc Biology (MSU)
- MSc Structural Biology (MSU)
- Junior Bioinformatician (ImmunoMind)
- Bioinformatician (BostonGene)
- PhD student (University of Basel)

Дани(и)л Литвинов
(Даня)
@Danil_litvinov

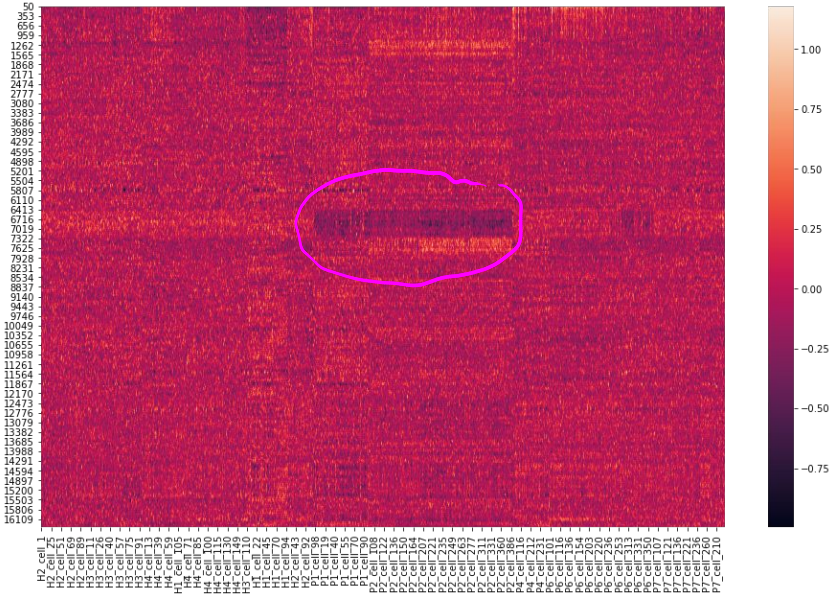
GWAS



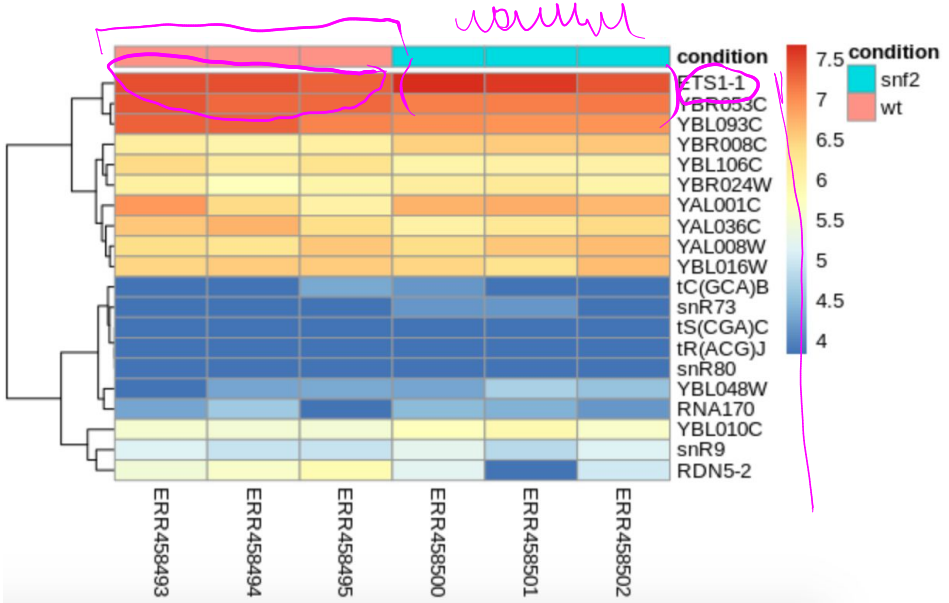
Метагеномика



Детекция анеуплоидных пациентов

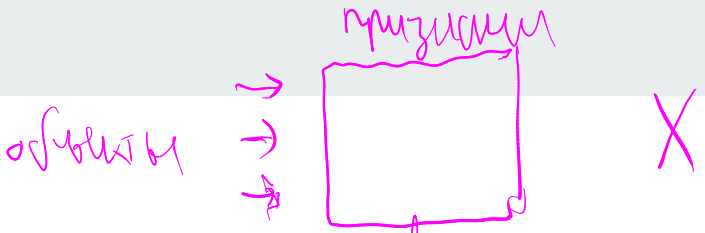


Дифференциальная экспрессия генов



Введение в статистику

Типы переменных



- Данные (признаки)
  - Количественные
    - непрерывные
    - дискретные
  - Качественные
    - категориальные (между собой сравнить нельзя)
    - порядковые

Генеральная совокупность. Выборка

Генеральная совокупность — совокупность всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи.

Выборка или выборочная совокупность — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

VS

Способы создания выборок

Вероятностные выборки

Невероятностные выборки (детерминированные) связаны с субъективными критериями

Вероятностные выборки

Вероятностные выборки – при создании таких выборок мы предполагаем, что генеральная совокупность достаточно однородна и все её элементы одинаково доступны.

Простая случайная выборка (simple random sample) – случайный набор объектов из генеральной совокупности. Пример: 100 мужчин, участвующих в Олимпийских играх

Стратифицированная выборка (stratified sample) – перед тем, как случайным образом отобрать объекты из генеральной совокупности, мы разбиваем её на несколько страт (групп).

Пример: мужчины 18-25 лет, 36-31, 32-36 и так далее.

Потом уже из этих групп случайно набираем по N человек.

Групповая выборка (cluster sample) – также сначала делим генеральную совокупность на кластеры, только считаем, что они между собой схожи.

Пример: рост жителей Санкт-Петербурга. Мы делим их на районы (Адмиралтейский, Василеостровский и т.д.), а потом случайно набираем людей из нескольких случайно выбранных районов для исследования.

Невероятностные выборки

Невероятностные выборки – отбор в такой выборке осуществляется не по принципам случайности, а по субъективным критериям – доступности объектов, типичности или равного представительства. Такие выборки часто встречаются в социологических исследованиях, однако данные, полученные на них обладают меньшей достоверностью и лучше их обходить стороной.

Метод снежного кома - у каждого респондента, начиная с первого, просятся контакты его друзей, коллег, знакомых, которые подходили бы под условия отбора и могли бы принять участие в исследовании. Основная проблема такой выборки - то, что затрагивается не случайная группа лиц, а лица, связанные общими интересами, хобби и т.д.

Стихийная выборка - производится опрос наиболее доступных респондентов. Размер и состав стихийных выборок заранее не известен и определяется только одним параметром - активностью респондентов.

Пример: опрос, проведенный в газете или журнале, большинство интернет-опросов

Выборка типичных случаев происходит отбор отдельных единиц генеральной совокупности, которые обладают типичным значением признака (часто это среднее значение). При этом возникает проблема выбора признака и определения его типичного значения.

Мера центральной тенденции

- Отражает типичное наблюдение в выборке
- Существует много вариантов

Арифметическое среднее

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Время (мин), которое студенты тратят на ДЗ: [2500, 3100, 3600, 2800, 3000]

$$\overline{T} = \frac{2500 + 3100 + 3600 + 2800 + 3000}{5} = 3000$$

А если так...

$$\overline{X} = \frac{X_1+X_2+...+X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

[2500, 3100, 3600, 2800, 3000, 17000]

$$\overline{T} = \frac{2500+3100+3600+2800+3000+17000}{6} = 5333$$

Усеченное среднее

Мы просто убираем по n % наименьших и наибольших значений в нашей выборке. Таким образом усеченное среднее лучше подходит, когда в выборке могут быть выбросы.

[2500, 3100, 3600, 2800, 3000, 1700]

$$\overline{T_{cut}} = \frac{3100+3600+2800+3000}{4} = 3125$$

Медиана

Значение, которое делит отсортированную выборку пополам.

[2500, 2800, 3000, 3100, 3400, 3600, 17000]

А если четное число наблюдений

Значение, которое делит отсортированную выборку пополам.

[2500, 2800, 3000, 3400, 3600, 17000]

3000+3400

↓

3200

Мода

Значение, которое встречается в выборке наиболее часто.

Значение	Частота
3000	4
2750	**10**
2700	7
17000	1

А какие есть варианты?

1) mean < med: [-1000, 1, 2, 3, 4]

2) mean < med = mode: [0, 1, 1]

Меры разброса

1) $x_{max} - x_{min}$ – размах

2) $\sum_{i=1}^{n} |x_i - \bar{x}| \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$

3) $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ – дисперсия (смещенная)

4) $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ – дисперсия (несмещенная)

5) $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ – стандартное отклонение.

Меры разброса

$$SS_{total} = \sum_{i=1}^{n}(X_i - \overline{X})^2 \; - \; \text{сумма квадратов отклонений}$$

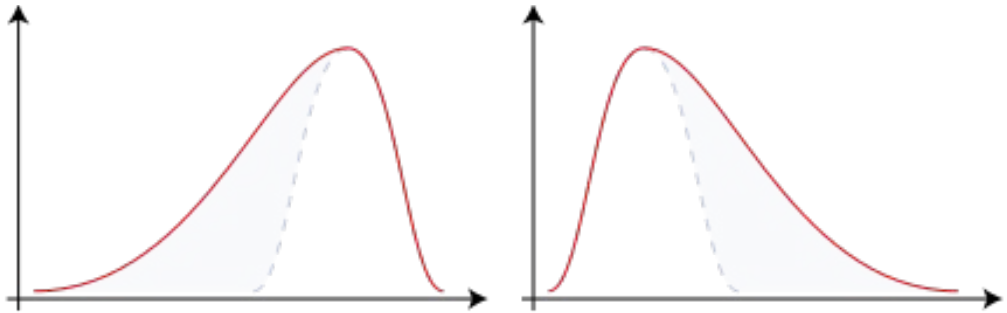
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X))^2 \; - \; \text{дисперсия при известном } E(X)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \; - \; \text{дисперсия при неизвестном } E(X)$$

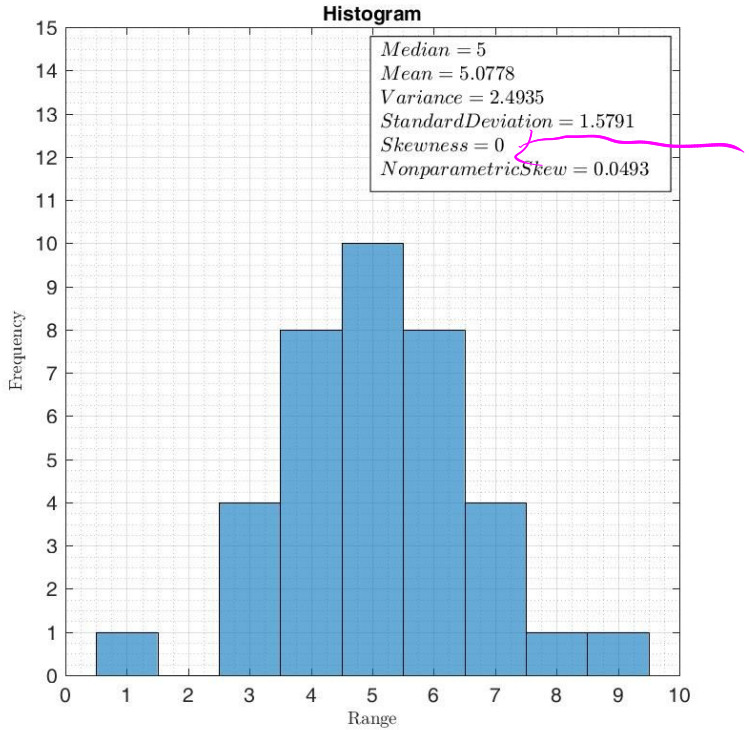
$$\overline{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \; - \; \text{стандартное отклонение}$$

Асимметрия

$$\overline{\gamma_3} = \frac{1}{n\overline{\sigma}^3}\sum_{i=1}^{n}(X_i - \overline{X})^3$$



Асимметрия. Контрпример



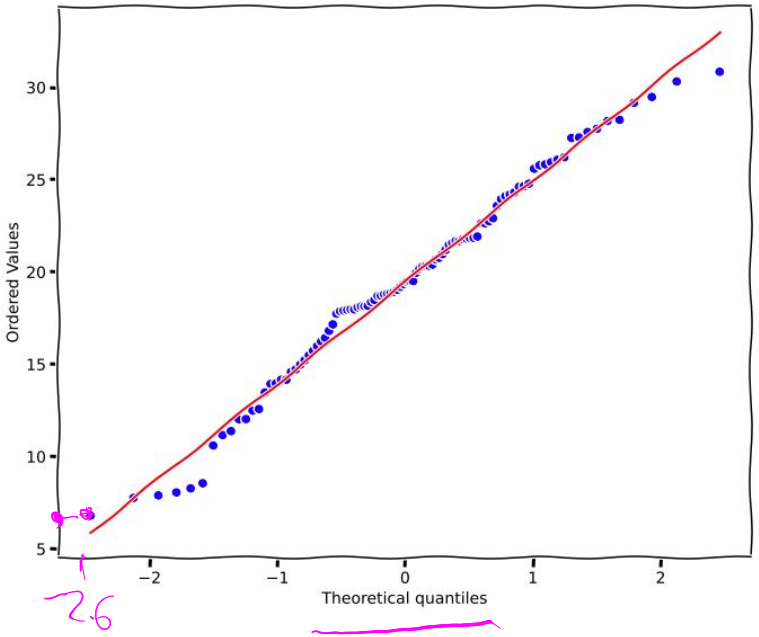
Квантили

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

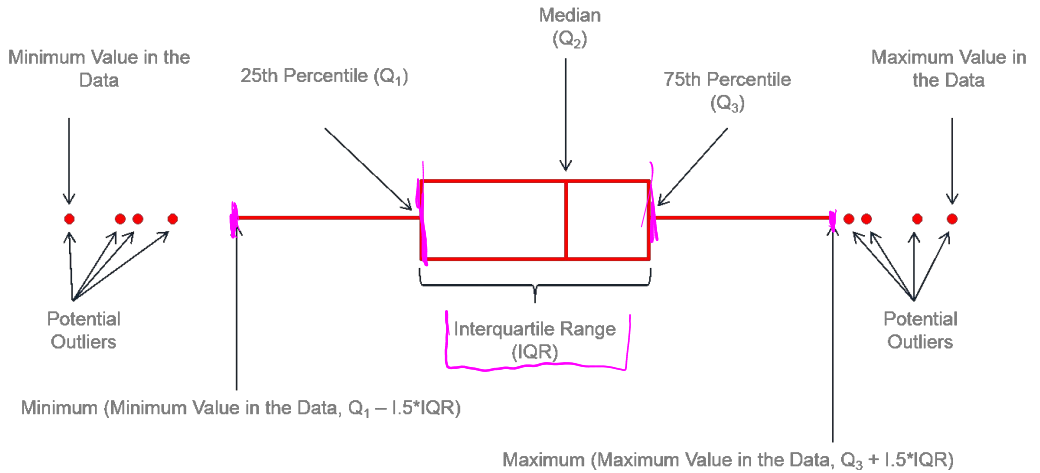
Возможные варианты квантилей:

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

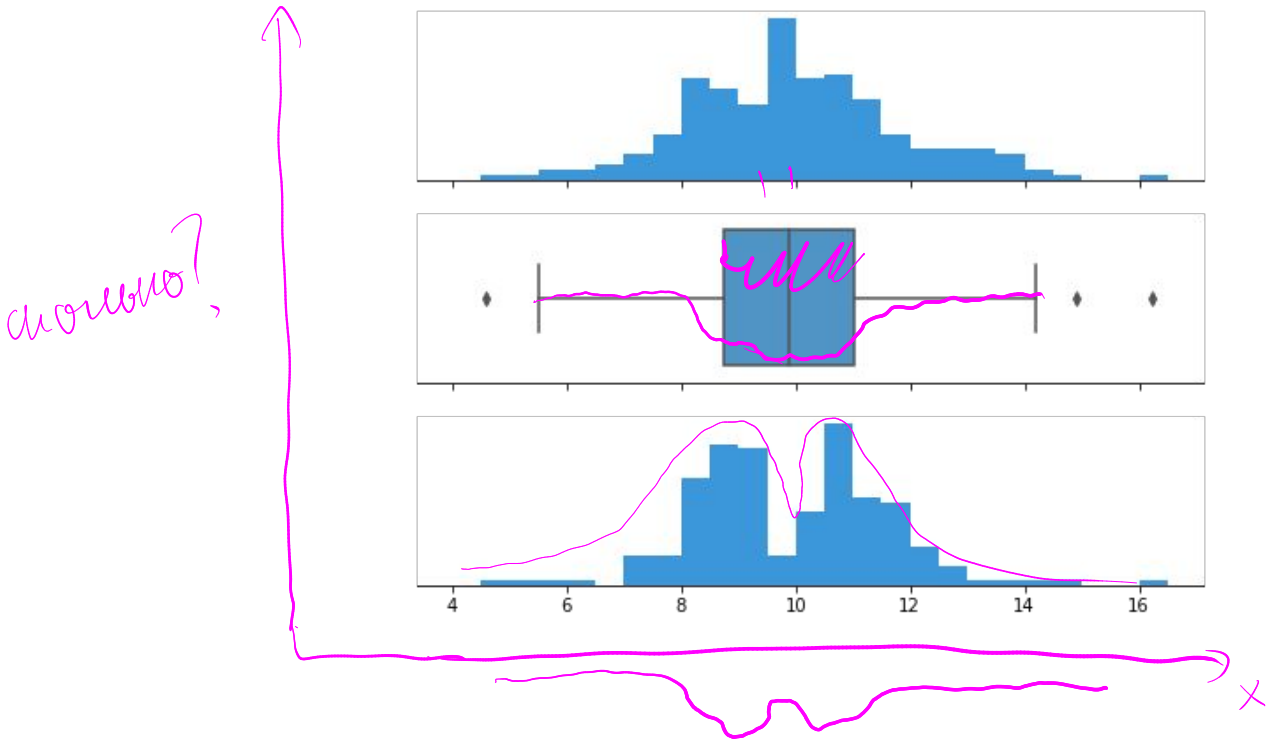
[300, 2500, 2800, 3000, 3100, 3400, 3600, 7000]



Боксплот (график)



Боксплот (недостатки)

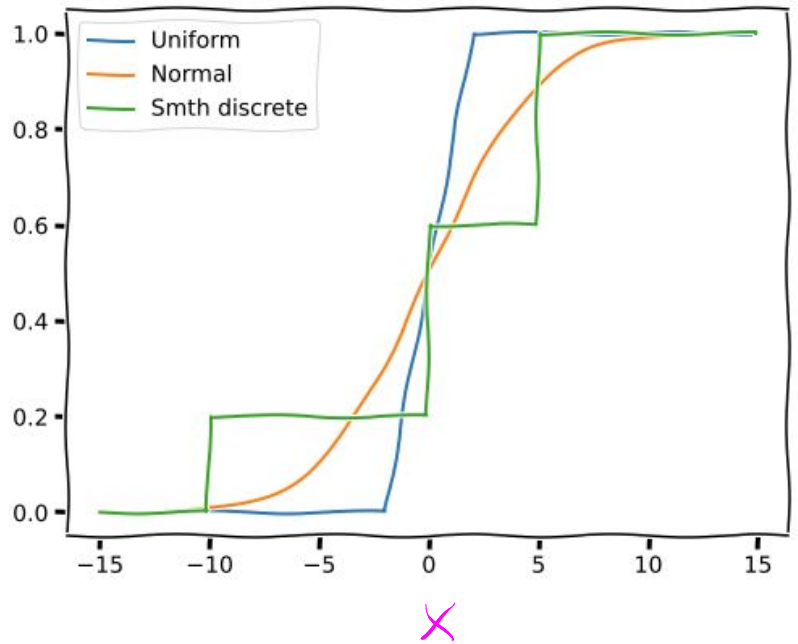


Распределения

Функция распределения

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

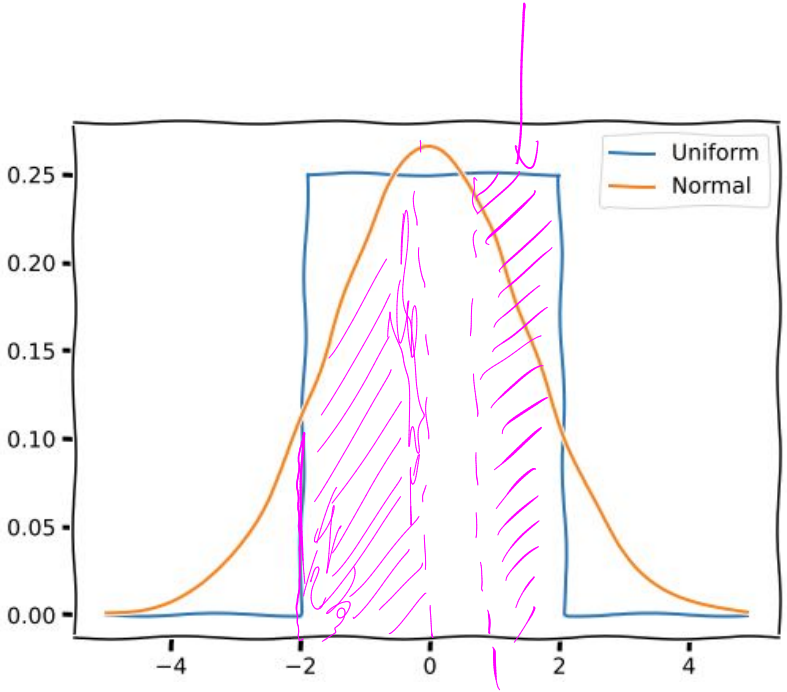
Функция распределения — это такая функция, которая для значения X, равна вероятности получить значение меньшее или равное этому X.
Например, для роста F(180 см) = ?



Функция плотности

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

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

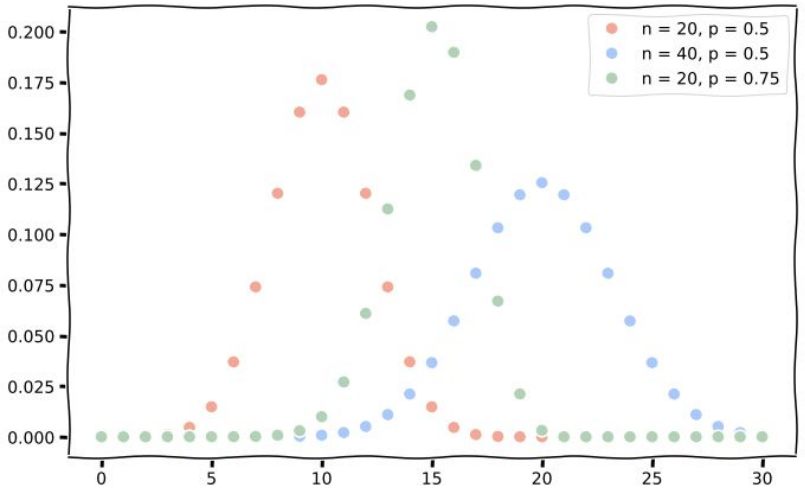


Биномиальное распределение

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.

.

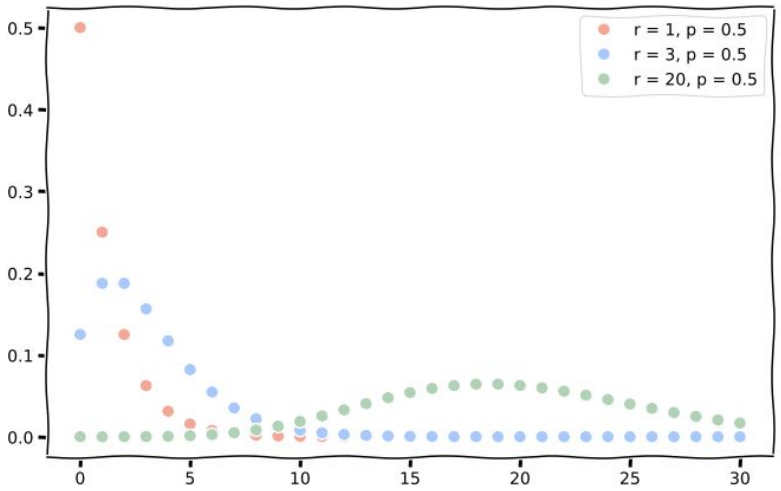
$$p(k) \equiv \mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



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.

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



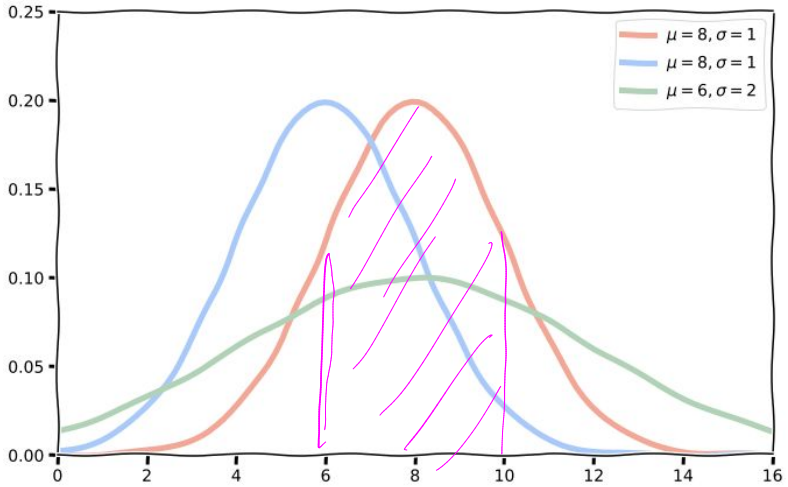
Нормальное распределение

Параметры:

μ — математическое ожидание

σ — стандартное отклонение

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



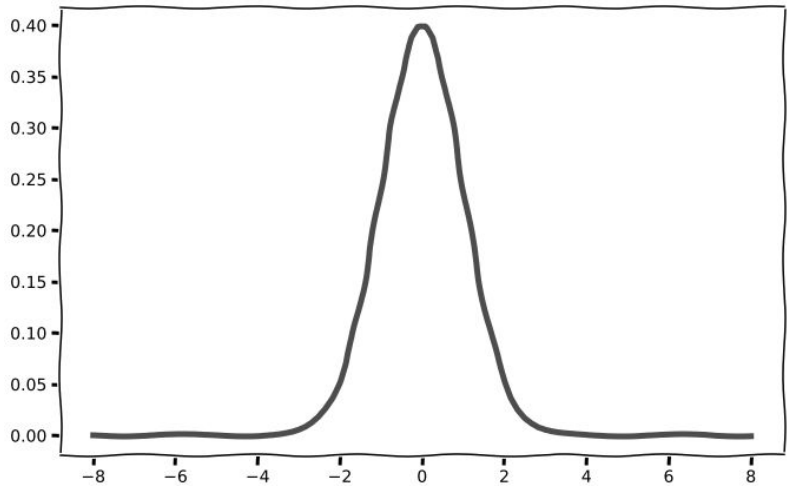
Стандартное нормальное распределение

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

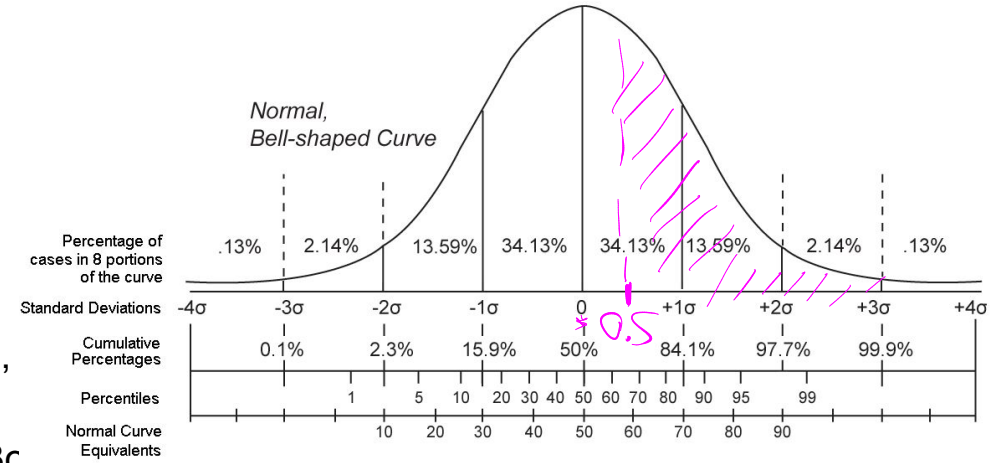
Нормальное распределение

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

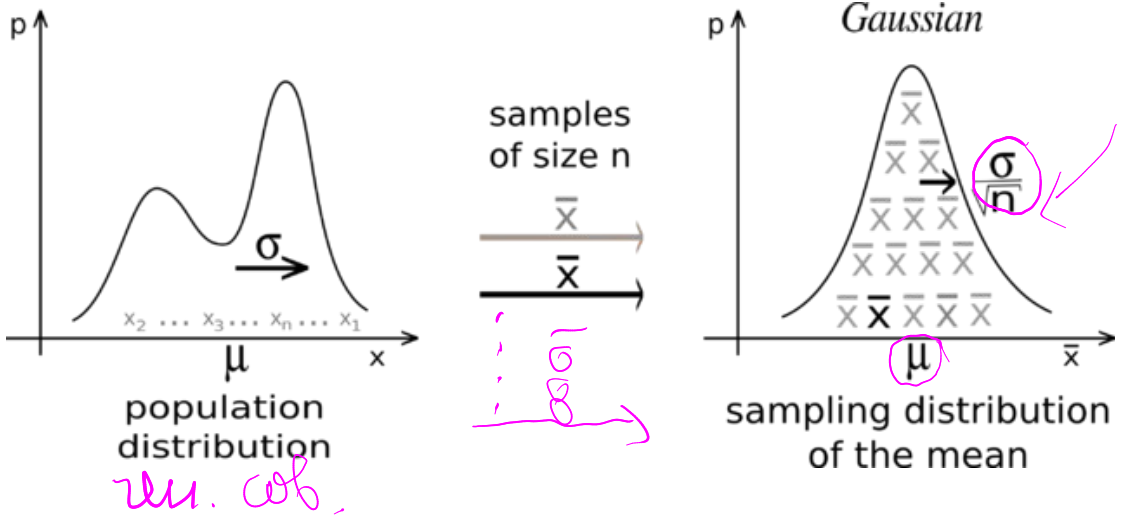
Пример: Среднее значение равняется 150, а стандартное отклонение равно 8. Какой процент наблюдений превосходит значение, равное 154?

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:

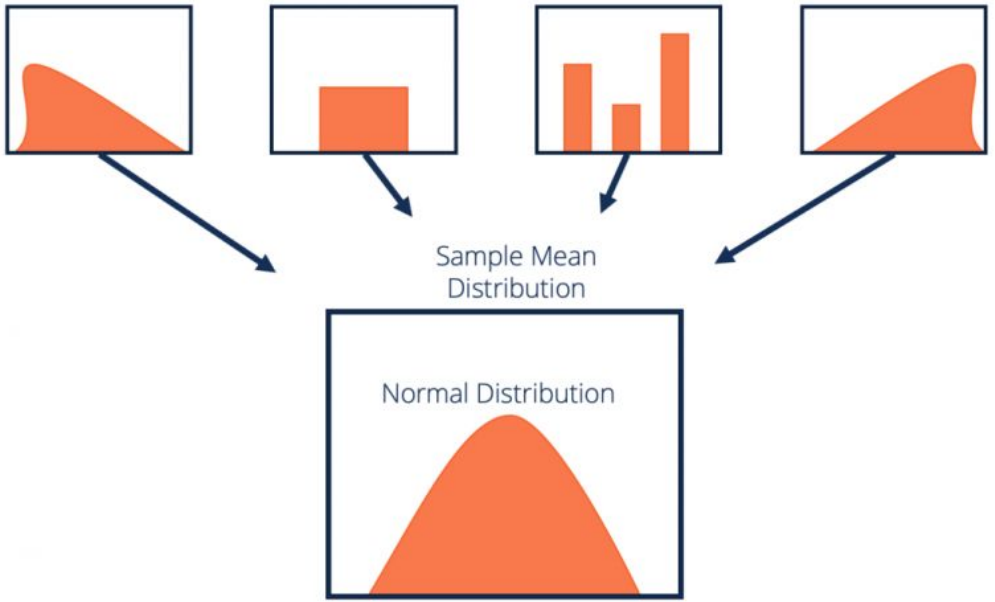
$$\frac{154 - 150}{8} = \frac{4}{8} = 0.5$$

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

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.

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

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

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

Стандартная ошибка среднего

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

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

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

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

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

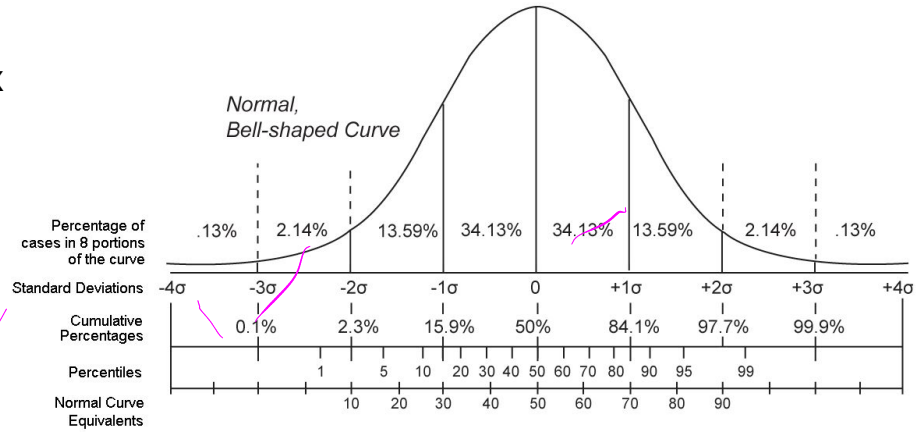
Доверительный интервал при известной дисперсии

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%:

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



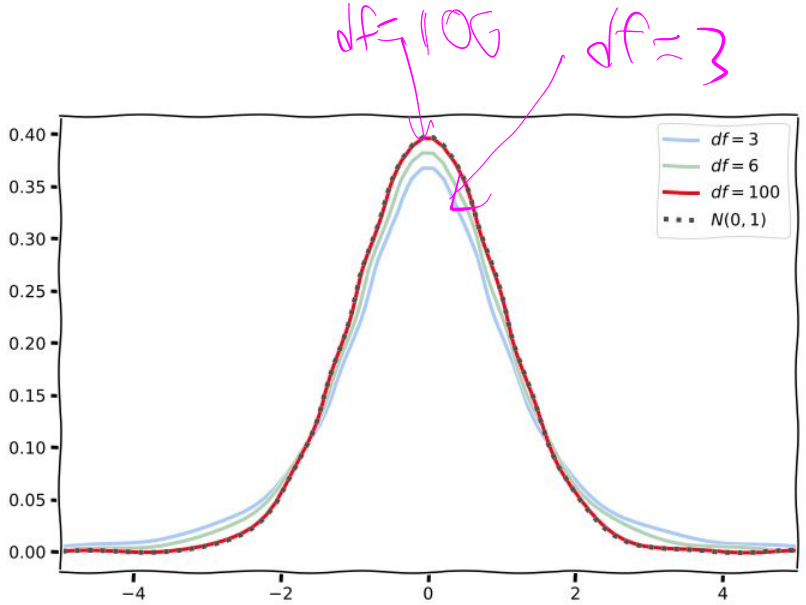
$$se = \frac{\sigma}{\sqrt{n}}$$

Тренируемся

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



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

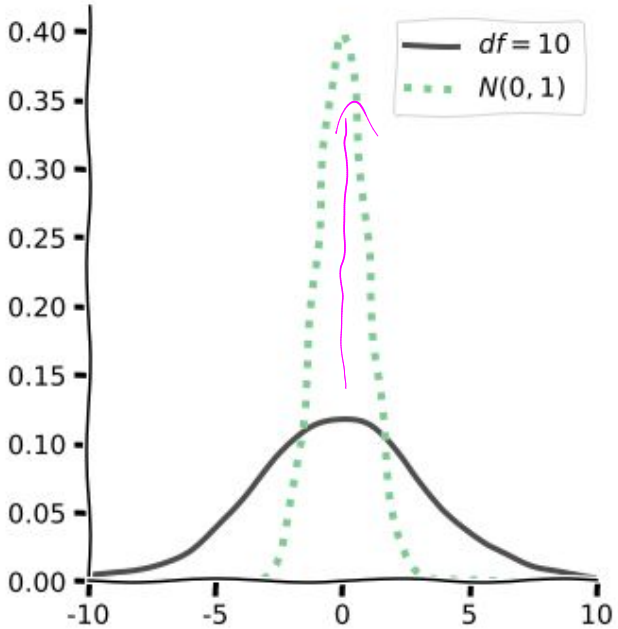
$$\overline{x} \pm t_{0.99} \cdot se$$

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

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

...

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