Проверка статистических гипотез

Лирическое отступление для самых маленьких

Случайная величина

> 💡 Случайная величина - событие, связанное с появлением того или иного значения в ходе проведения эксперимента.

СВ бывает ***дискретной и непрерывной***.

Есть кубик, у которого 6 граней: 1, 2, 3, 4, 5, 6.

Дискретная СВ принимает только счетные значения, то есть значения кубика 1, 2, 3…

Все возможные значения дискретной СВ можно записать в виде таблицы, которая называется `рядом распределения`.

x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6

Первый ряд - возможные значения СВ, второй ряд - вероятность того, что СВ примет данное значение.

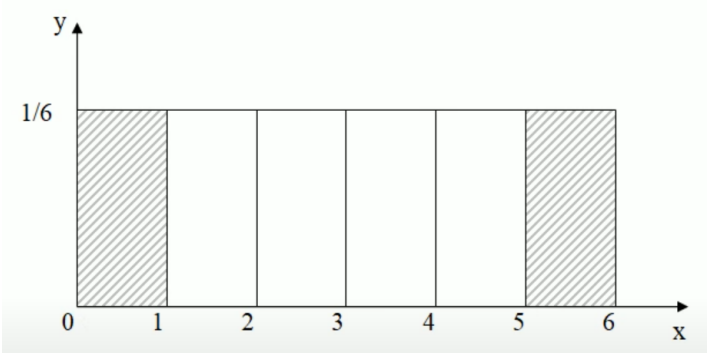
Непрерывная СВ принимает диапазон значений.

Например при стрельбе по мишени можно вычислять расстояние от точки попадания до центра мишени при помощи формулы Евклидова расстояния, тогда СВ X будет принимать любые значения в некотором диапазоне.

$$X = l = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

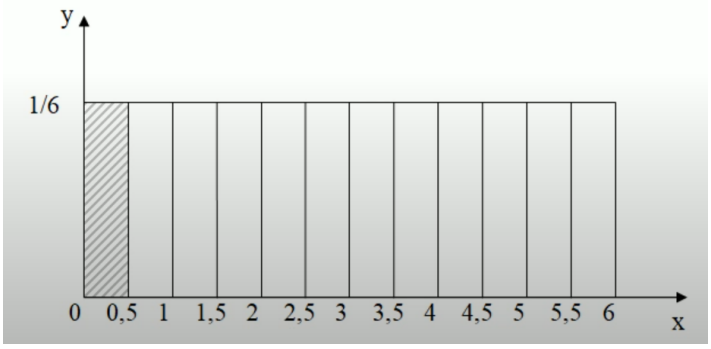
Плотность распределения вероятностей

Полное вероятностное описание кубика можно описать при помощи X и вероятностей выпадения значений кубика. Вероятности можно представить в виде графика плотности распределения вероятности:



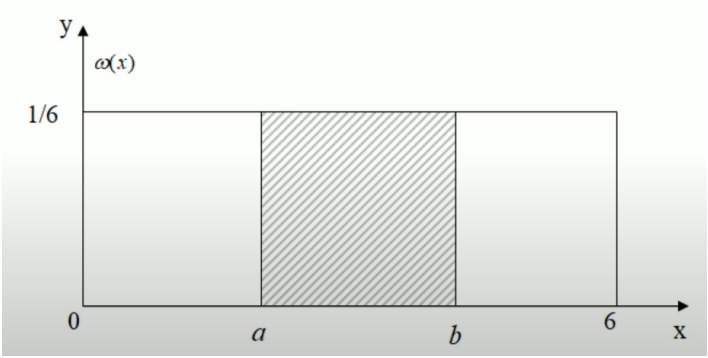
В данном случае вероятность выпадения 1 равна площади под графиком от 0 до 1 - 1/6. Соответственно, по оси Y вероятность, по оси X возможные значения СВ. Представить график можно по разному (с разными значениями). Например, возьмем кубик с 12 гранями, где значения кубика будут с шагом 0.5, т.е. 0.5, 1, 1.5, 2…

Тогда график плотности распределения можно показать так:



Граней 12, значит вероятность выпадения 0.5 равняется 1/12. Это видно на графике: 1/6 * 0.5 = 1/12. Если бы значения кубика были 1, 2… до 12, то удобнее было бы представить график в другом виде: вероятность по оси Y была бы 1/12, а СВ от 0 до 12. **Данные примеры представляют графики вероятности распределения для дискретной СВ.**

Увеличим число граней кубика до бесконечности, то есть кубик станет шаром. Если пронумеровать его грани от 0 до 6, то получим г**рафик плотности распределения непрерывной СВ**



Данный график показывает, что СВ может принять любое значение на отрезке от 0 до 6

Вероятность того, что СВ примет значение от a до b - это площадь, которая находится по следующей формуле:

$$P(a < X < b) = \int_a^b \omega(x)\,dx$$

Если мы расширим наш диапазон от - бесконечности до + бесконечности, то вероятность выпадения будет составлять 1, т.к. СВ точно примет какое-то значение:

$$P(-\infty < X < +\infty) = 1 = \int_{-\infty}^{+\infty} \omega(x)\,dx$$

`Это ключевое свойство плотности распределения вероятности СВ: площадь под графиком всегда равна 1.`

Если же сузить наш диапазон до конкретного числа a, то вероятность будет равна 0. Так происходит из-за того, что точки бесконечно малы (или что дробь бесконечна, т.к. у вещественного числа бесконечное число цифр после запятой, и мы не сможем точно сказать, что СВ примет значение a).

> Однако это чисто математически, в реальности наши измерения будут иметь конечную точность, а значит и вероятность не будет равна 0, т.к. они будут дискретны.

Второе свойство - плотность вероятности не должна быть меньше 0:

$$\omega(x) \geq 0,\ \forall x$$

Функция распределения СВ

Вернемся к диапазону от a до b.

Мы знаем, что вероятность того, что СВ окажется в диапазоне, вычисляется по следующей формуле:

$$P(a < X < b) = \int_a^b \omega(x)\,dx$$

Эту же формулу можно представить через первообразные:

$$P(a<X<b)=\int_a^b \omega(x)dx=F(b)-F(a)$$

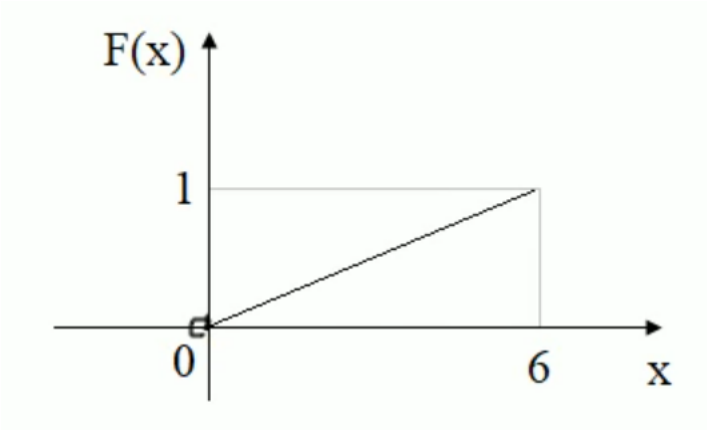
Если плотность распределения будет описывать так:

$$\omega(x)=\begin{cases}\frac{1}{6}, & x\in[0;6]\\ 0, & \text{иначе}\end{cases}$$

то ее первообразная так:

$$F(x)=\frac{1}{6}x, \quad x\in[0;6]$$

тогда график первообразной будет выглядеть следующим образом:



Из этого можно сделать вывод, что график первообразной это не что иное, как вероятность того, что СВ не примет значение больше x:

$$F(x)=P(X<x)$$

F(x) - функция распределения случайной величины.

Например, найдем вероятность того, что СВ будет в диапазоне от 3 до 4 при F(x) = 1/6 * x, то:

$$F(x)=\frac{1}{6}x,\quad x\in[0;6]$$

$$P(3<X<4)=\int_3^4\omega(x)\,dx=F(4)-F(3)=\frac{1}{6}\cdot 4-\frac{1}{6}\cdot 3=\frac{1}{6}$$

Функция распределения - первообразная для функции плотности вероятности СВ:

$$\omega(x)=\frac{dF(x)}{dx}$$

Ее свойства:

1. Функция распределения от - бесконечности равна 0, т.к. нет такого значения, которое было бы меньше - бесконечности
2. Отсюда второе свойство - для + бесконечности 1, т.к. все числа меньше + бесконечности
3. Функция распределения всюду больше или равна 0

$$F(x)=P(X<x)$$

$$F(-\infty)=0$$

$$F(+\infty)=1$$

$$F(x)\geq 0,\quad \forall x$$

Задача

Дана случайная величина X, заданная функцией распределениях F(x). Найти:

1. Плотность вероятности
2. Математическое ожидание и дисперсию X
3. Построить графики функции распределения и функции плотности распределения

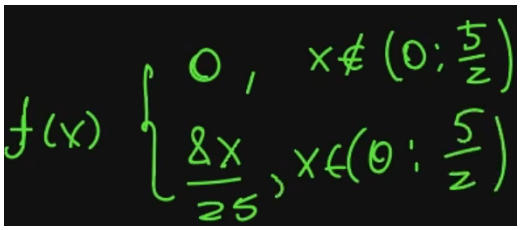
$$F(x)=\begin{cases}0, & \text{при } x\leq 0\\ 4x^2/25, & \text{при } 0<x\leq 5/2\\ 1, & \text{при } x>5/2\end{cases}.$$

1) Функция плотности вероятности - производная от функции распределения:

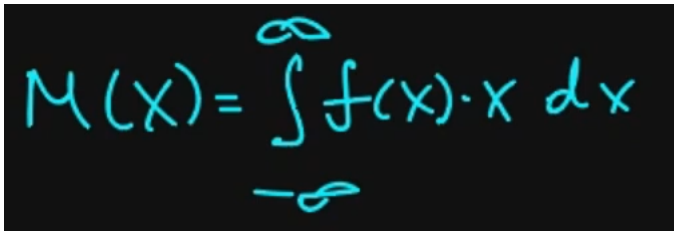
f(x) = F'(x).

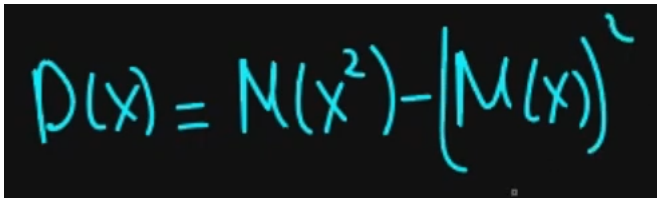
Производная от константы 0, значит f(x) = 0, тогда x не принадлежит интервалу (0, 5/2).

Производная от 4x^2 / 25 = 8x / 25, тогда:

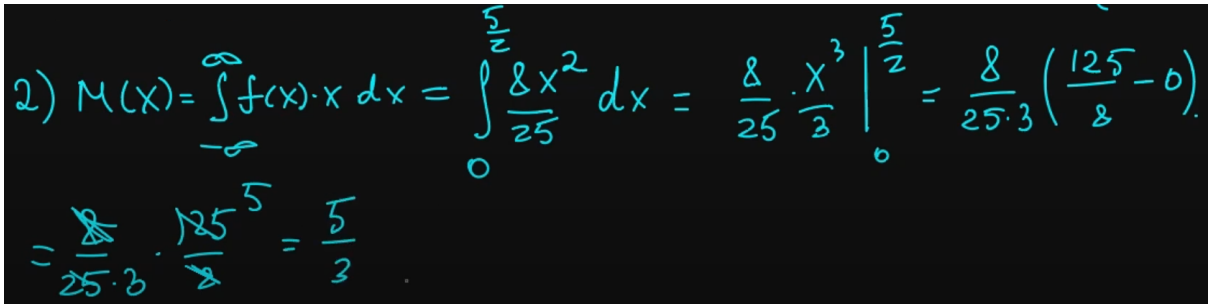


2) Формулы для мат ожидания и дисперсии:



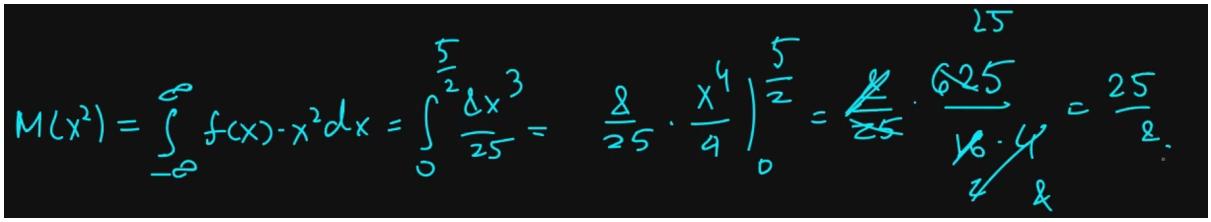


Найдем мат. ожидание. Интеграл можно сузить до 0 и 5/2:

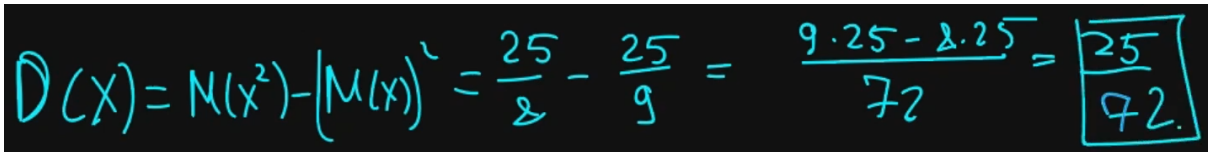


Дисперсия:

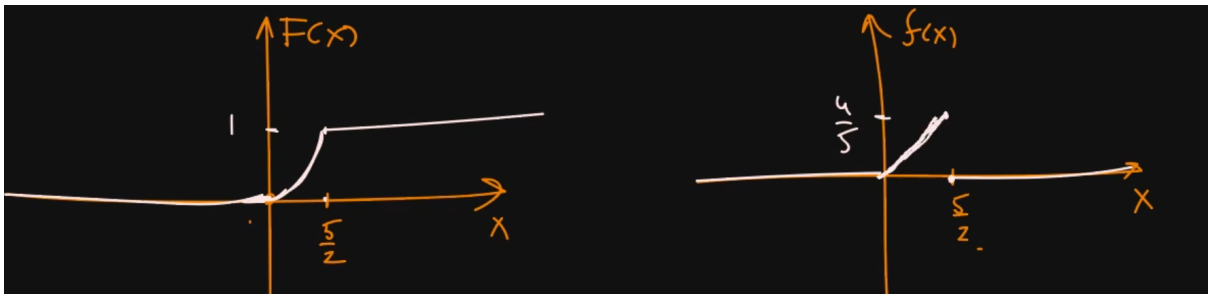
Найдем мат ожидание от квадрата x:



Подставляем в формулу дисперсии и получаем



3) строим графики



Проверка статистических гипотез

Статистическая гипотеза - утверждение о свойствах распределения вероятностей случайной величины (или случайног наблюдениях.

Проверка гипотез может происходить в следующих случаях.

Гипотеза согласия

Обозначим $F_X(t)$ функцию распределения случайной величины X.
Пусть $F_0(t)$ - некоторая заданная функция распределения.
Гипотеза : функции распределения совпадают, то есть $F_X(t) = F_0(t)$

То есть гипотеза согласуется. Например гипотеза о нормальности распределения.

$$F_0(t) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{t} \exp\left(-\frac{(s-a_X)^2}{2\sigma_X^2}\right) ds$$

Почему гипотеза о нормальном распределении важна?

1. Нормальное распределение часто встречается (особенно благодаря ЦПТ).
2. Если оно встретилось, то можно экономить на числе наблюдений (и времени, деньгах и т.д., например, опросим меньше покупателей)

Второй пример - гипотеза об экспоненциальности распределения

$$F_0(t) = \begin{cases} 1 - e^{-\lambda t}, t > 0 \\ 0 \quad , t < 0 \end{cases}$$

Экспоненциальное распределение характеризует время ожидания. Например, время до поломки устройства, время обслуживания покупателя, время до какого-то события (например, в страховой часто используется).

Опять таки, если данные распределены экспоненциально, то это существенно экономит время и деньги.

Гипотеза однородности

Гипотеза проверят, совпадают ли распределения. При этом саму функцию знать необязательно.

Обозначим $F_X(t)$ функцию распределения случайной величины X.
Обозначим $F_Y(t)$ функцию распределения случайной величины Y
Гипотеза : функции распределения совпадают

$$F_X(t) = F_Y(t)$$

Используется в случаях, когда хотим сравнить распределения до и после какого-то события. Например, рекламной акции.

Гипотеза независимости

Гипотеза: СВ X и Y независимы. То есть влияет ли одна величина на другую.

Однако не так интересно, как X влияет на Y, как то, как именно влияет. Это задача регрессионного анализа. Иногда зависимость бывает неочевидной: длина волос и рост людей - зависимые величины.

Очень часто зависимость между величинами объясняется скрытыми параметры, или т.н. `латентные переменные` (или фактор), который объясняет зависимость. Ищутся эти переменные при помощи факторного анализа.

Часто бывает, что в каких-то срезах есть влияние, а в других нет. Например, наличие балкона влияет на цену квартиры в низком сегменте, а в премиум классе нет.

Гипотезы о параметре распределения

Часто не так интересно распределение СВ, как ее какая-то характеристика. Например, мат. ожидание (вероятностная модель для среднего) или медиана. Вероятность - мат. модель частоты чего-либо.

Гипотеза: мат. ожидания для СВ X и Y одинаковы:

$$EX = EY$$

То есть средние равны.

Аналогично для медиан, которые используются чаще:

$$Med(X) = med(Y)$$

Условия применения статистических тестов

1. Вопрос должен касаться какой-либо характеристики массового явления. Например, продажа танкеров - не массовое явления, а продажи пива - массовые.
2. Характеристика меняется случайным образом от наблюдения к наблюдению
3. Вопрос простой и четко сформулирован

Технологии проверки статистических гипотез

Тезисно:

1. Гипотезы проверяются парами, то есть есть нулевая и альтернативная гипотеза.
2. У гипотез нет вероятностей того, верна она или нет.
3. Правильно говорить “основная гипотеза отвергнута” или ”гипотеза не отвергнута”, а не принята, т.к. обычно проверяют лишь достаточное условие:

Гипотеза: число делится на 6. У нас есть инструмент, который проверяет делимость на 2. Если число делится не делится на 2, то оно и не делится на 6. А если на 2 делится, то это не значит, что делится на 6. ***Поэтому гипотезы***

только отвергаются или не отвергаются

4. Часто бывает, когда не хватает данных для того, чтобы проявился изучаемый эффект. Например в медицине. В таком случае, отвергнуть или не отвергнуть гипотезу недостаточно.
5. Ошибка первого рода - принимается альтернативная гипотеза, которая на самом деле неверна. Она тяжелее по последствиям. Типичные значения альфа: 0.05, 0.01, 0.005, т.е. ошибаться будем с частотой 20, 100 и 200 случаев. При этом альфа зависит от индустрии: в эконометрике 0.05, в ядерной индустрии 0.000000000…1
6. Ошибка второго рода - отвергается альтернативная гипотеза, хотя она верна. Контролировать ее сложнее: нужны состоятельные критерии, где с увеличением числа наблюдений ошибка второго рода будет уменьшаться.
7. Если увеличивается альфа, то ошибка второго рода возрастает. Профессионалы часто увеличивают ошибку первого рода, чтобы сделать обе ошибки сопоставимыми. Но игнорировать ошибка второго рода нельзя!

Алгоритм проверки гипотез

1. Есть n наблюдений
2. Заранее задан уровень значимости альфа

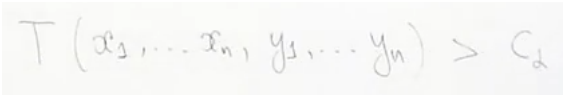
Как проверяли статистические гипотезы раньше

Идея проверки гипотез состоит в следующем. пример:

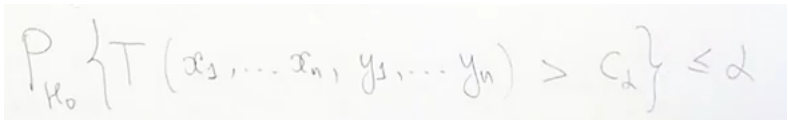
H0: X и Y независимы

H1: X и Y зависимы

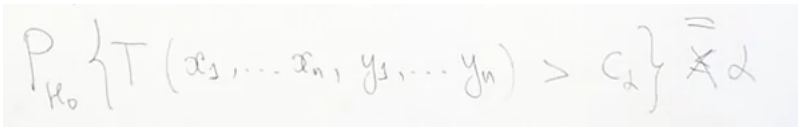
Идея состоит в том, чтобы придумать некоторую функцию T от X и Y, чтобы вычислить, согласуется ли гипотеза - статистика критерия. Чем она больше, тем больше склоняемся к альтернативной гипотезе. C альфа - критическое значение.



Нужно найти вероятность отвергнуть H0 при условии, что она верна:

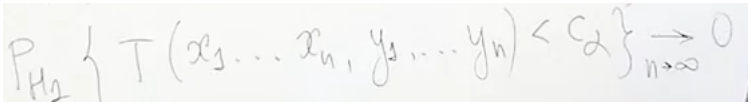


Из неравенства делаем уравнение. По факту нужно найти C *альфа* - критическое значение.

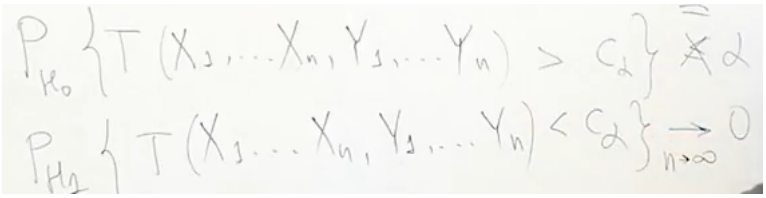


Откуда равно? По аналогии: если разрешен определенный процент брака, то процент брака таким и будет. Если сначала брака нет, то он наберется в дальнейшем.

Также нужно использовать только состоятельные критерии, то есть с ростом числа наблюдений вероятность ошибки второго рода будет стремиться к 0:



Есть неточность: вероятность превысить T *эксп* (которое какое-то число) для определённых чисел (в данном случае наблюдений) либо 0, либо 1. Мф же рассматриваем, как часто значение превысит порог, поэтому вместо наблюдений в формуле должны быть **случайные величины** Xi и Yi:

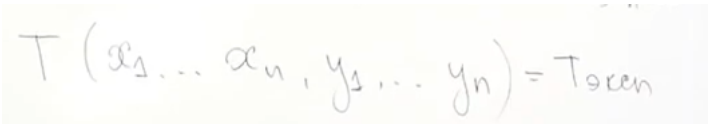


Как сейчас проверяются гипотезы: С *альфа* не ищется, ищется сразу p-value.

3. Задан статистический критерий
4. Найдено **p-value**

Что такое **p-value?**

Есть функция от наблюдений. Результатом этой функции будет значение Т *эксп*. Это именно конкретное вычисленное число, для конкретных наблюдений



Так вот вероятность превысить Т эксп и будет p-value:

Допустим бросаем кубик. Вероятность выпадения грани 1/6. Так вот при числе наблюдений, которое будет стремиться к бесконечности, вероятность выпадения какой-либо грани будет 1/6, т.е. доля будет 1/6 из всех наблюдений.

Как читать формулу выше: как часто статистика критерия (функция) Т будет превышать Т*эксп*, которое наблюдалось у нас (предпоследняя формула) при условии, что верна H0.

5. Проверяются условия, при которых критерий будет работать корректно
5. Если p-value меньше альфа, то H0 отвергается.

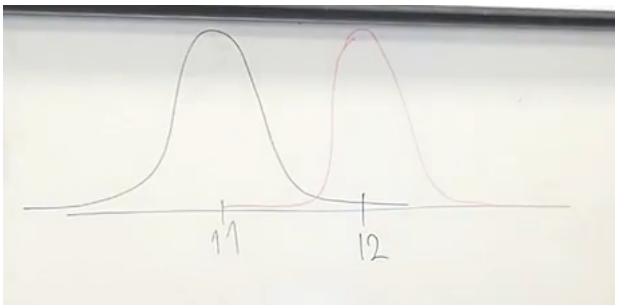
Пример

Есть нормальное распределение с D=1, n наблюдений. H0: математическое ожидание = 11. H1: математическое ожидание = 12.

Придумаем функцию, которая будет измерять, насколько данные согласуются с H0? Это среднее арифметическое. Мат ожидание - математическая модель для среднего арифметического.

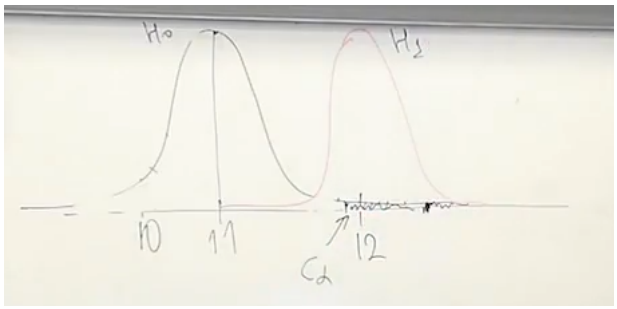
> Лемма: среднее арифметическое n независимых одинаково распределенных СВ с общим нормальным распределением N(a, b) имеет нормальное распределение N(a, b/n).

Нарисуем график плотности при H0 и H1:



Плотность распределения помогает посчитать вероятности - интеграл от плотности (площадь). Площадь фигуры от 11 и левее 1/2 (по определению, что это половина).

Зададим уровень значимости альфа = 0.05. Т в задаче - среднее арифметическое. Вероятность ошибки первого рода 0.05, на графике она будет на черном графике от плюс бесконечности до того момента, пока площадь не окажется равной 0.05

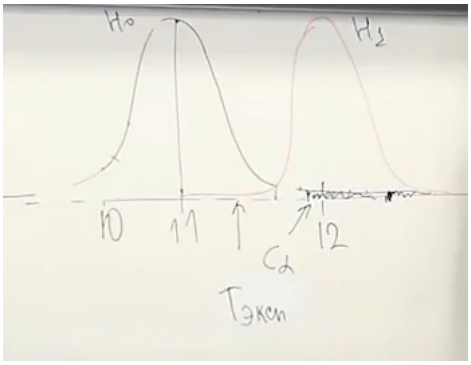


При этом С альфа может лежать правее 12 - красный график игнорируется.

На пересечении черного графика с красным вероятность ошибки первого рода будет равна вероятности ошибки второго рода (факт для понимания).

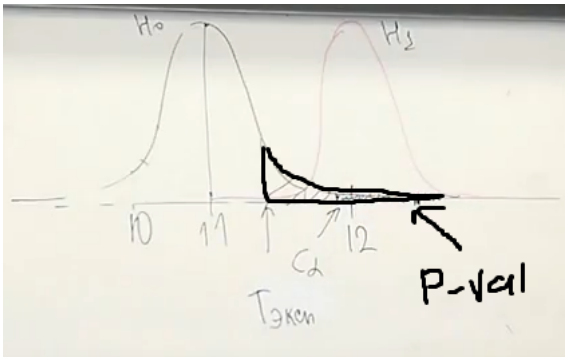
Далее нужно найти на графике p-value. Для этого надо найти Т*эксп*

Допустим, Т*эксп* будет в этой точке:



Тогда по старой схеме Н0 не отвергается, так как Т*эксп* меньше С *альфа*

При том, что p-value это вероятность того, что Т > Т*эксп,* то на графике это площадь участка графика, которая правее (больше) Т*эксп:*



В данном случае площадь p-value больше, чем вероятность ошибки первого рода, а значит Н0 не отвергается

Промежуточные итоги:

По старой схеме проверки гипотез:

Если Т*эксп* > С *альфа*, то Н0 отвергается,

если Тэксп < С *альфа*, то Н0 не отвергается. Здесь сравниваются именно точки (значения) на графике.

По новой схеме сравниваются площади (вероятности):

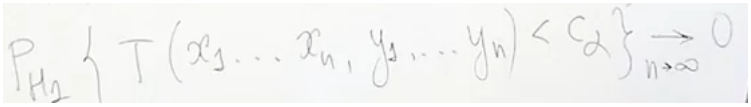
Если p-value > альфа, то Н0 не отвергается

если p-value < альфа, то Н0 отвергается

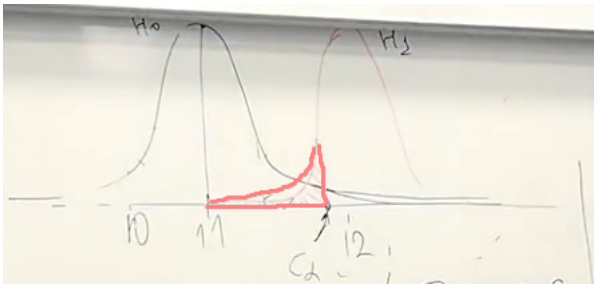
Преимущество новой схемы в том, что мы можем менять альфа и не пересчитывать результаты, а просто сравнить с p-value.

Вероятность ошибки второго рода

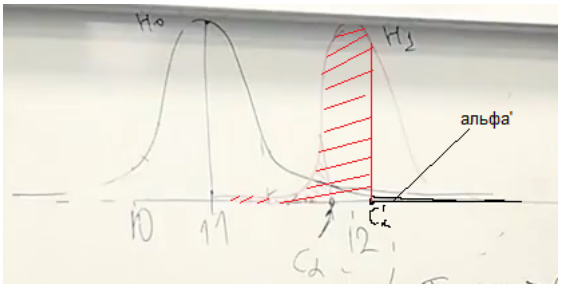
По старой схеме:



В данном случае будем смотреть на красную плотность. Это вероятность оказаться меньше (левее), чем С *альфа*. Вероятность - площадь от С *альфа* до минус бесконечности:



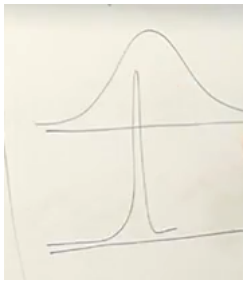
Если альфа мала (альфа'), то вероятность ошибки второго рода будет больше:



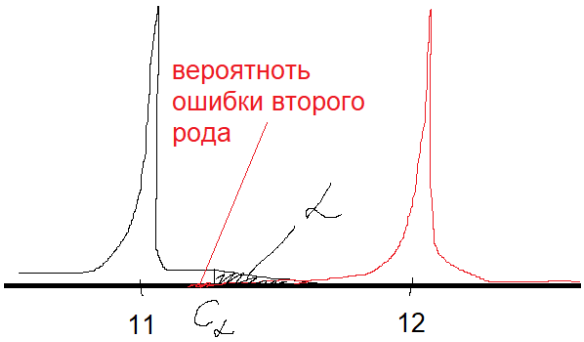
Состоятельность теста и увеличение числа наблюдений

Тест считается состоятельным, если при увеличении выборки вероятность ошибки второго рода стремится к 0.

С ростом числа наблюдений дисперсия будет уменьшаться. Почему? Потому что график плотности будет сжиматься (из леммы, где b/n). В результате, при D=1 график плотности будет классическим, а при увеличении числа наблюдений сужается:

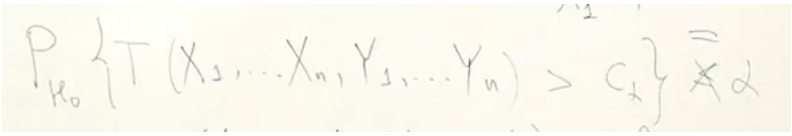


Получается, что увеличении выборки С *альфа* по оси уйдет левее (у альфы же фиксированная вероятность, площадь останется той же), а значит вероятность ошибки второго рода будет уменьшаться и стремиться к 0:



Интерпретация статистики критерия

Нулевая гипотеза всегда проще, ведь необходимо уметь решать уравнение:



Уравнение должно быть достаточно простым. Например, нормальное распределение или ненормальное? В H0 мы проверим только то, что распределение нормальное. В другом случае ненормальных распределений очень много, а значит проверить гипотезу сложнее.

Критерий Шапиро-Уилка и гипотезы о нормальности

Критерий проверяет распределение на нормальность.

> 💡 При этом если число наблюдений небольшое (меньше 2000) то применятся он, а если наблюдений больше, то лучше использовать критерий **Колмогорова-Смирнова.**

Стоит отметить, что в учебниках работа многих критериев описан для данных, распределение у которых нормальное. Однако есть допущения, когда критерии применяются и **при несущественных отклонениях от нормального распределения.**

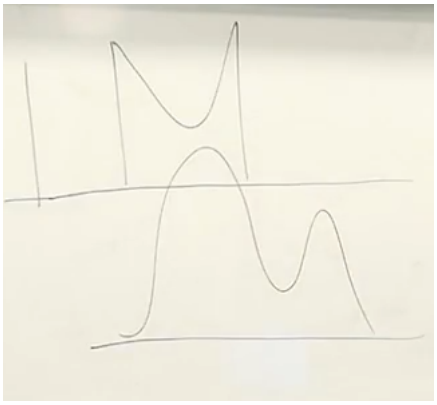
Замечание: если данных много (миллион и выше), то гипотеза о нормальном распределении будет отвергаться всегда. Так происходит потому, что критерии требовательны к точному следованию нормальному распределению, и чем больше наблюдений, тем более требовательны критерии. Выходы из положения:

1. Можно взять подвыборку из этих данных и посмотреть, нормальное ли распределение или нет.
2. Уменьшить уровень значимости. Однако подобрать коэффициент исходя из размера выборки сложно.
3. Посмотреть на глаз, есть ли существенные отклонения от нормальности.

Существенные отклонения от нормальности

При следующих признаках не рекомендуется использовать критерии, которые будут рассмотрены далее:

1. Наличие выбросов в данных (хотя бы нескольких). Строгое соблюдение
2. Явная асимметрия гистограммы. Более снисходительно
3. Очень сильное отклонение формы гистограммы от колоколообразной формы:



Равномерное распределение - пограничная ситуация, когда можно применять критерии.

При этом важно отметить, что если наблюдений меньше 30, то отношение самое либеральное, если от 30 до 150, то более строгое, если больше 150 - строгое отношение к колоколообразной форме.

Выходы из ситуации

1. Если есть выбросы, то можно осторожно удалить. И то если эти наблюдения ничем не отличаются от остальных, то способ достаточно спорный
2. Если есть асимметрия, то можно преобразовать данные - логарфимирование и преобразование Бокса-Кокса
3. Если форма распределения не колоколообразная, бимодальная, то можно разделить выборку на подвыборки (возможно были использованы разные источники данных, кластеры). Также отклонение от колоколообразной формы может быть от неудачного выбора числа столбцов гистограммы.

Сравнение типичных значений

Можно сравнивать как средние, так и медианы, так и усеченное среднее (есть и много других способов).

При этом, если распределение ненормальное, то лучше сравнивать медианы - критерий Манна-Уитни, если среднее, то т-тест.

Парные и независимые выборки

Существуют случаи, когда значения в выборках образуют пары.

Пример: проверить то, как люди справляются с нагрузками до и после тренировки. Ключевой показатель - среднее. В данном случае легче сравнить разность между выборками - если среднее разности равно 0, то средние в двух выборках тоже равны (в критерии Стьюдента параметр `paired` **= True**)

В случае же независимых выборок каждое наблюдение - отдельный объект. Принадлежность этих объектов будет определяться по значением дополнительной группирующей переменной.

Критерий Манна-Уитни

Работа критерия состоит в следующем. Наблюдения из двух выборок смешиваются в одну выборку. Выборка упорядочивается и дается ранг (номер наблюдения). После этого наблюдения заменяются рангами. А далее сравниваем первоначальные выборки (но уже с рангами вместо первоначальных значений) и смотрим, как часто ранг из одной выборки выше, чем в другой.

Критерий Манна-Уитни в действительности не сравнивает медианы, однако практически во всех случаях этим пренебрегают, и это не всегда будет ошибкой.

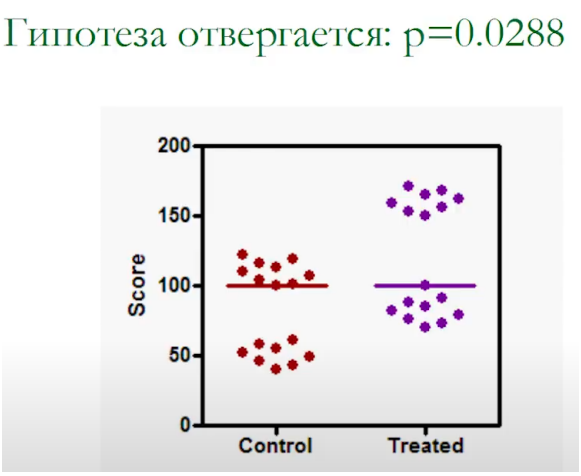
Критерий проверяет не медианы, а насколько часто X больше Y. То есть гипотеза выглядит так:

- Имеются две выборки наблюдений случайных величин X и Y.
- Гипотеза: $P\{X>Y\}=P\{X<Y\}$.
- Альтернативная гипотеза: $P\{X>Y\} \neq P\{X<Y\}$.

То есть проверяются вероятности, а именно их равенство о том, как часто значения в одной выборке больше, чем в другой. При этом критерий сравнивает не сами значения, а их ранги.

Что это за случаи, когда тест Манна-Уитни при проверке на равенство медиан будет давать ложные результаты?

1. Медианы двух распределений равны, но тест покажет, что это не так.



Почему так? Потому что фиолетовые в намного чаще будет превосходить красные по рангу.

2. В одной из выборке большой разброс.

Резюмируя:

> 💡 Если при сравнении медиан распределения одинаковы по форме, то можно применять критерий Манна-Уитни при проверке на равенство медиан.

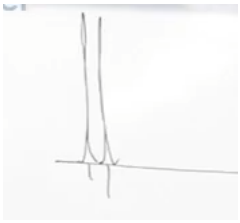
При этом критерий также может сравнивать средние арифметические. Однако он будет работать при нормальном распределении и немного проигрывать в мощности критерию Стьюдента.

Критерий Стьюдента

Помимо того, что критерий может использоваться и при ненормальном распределении, многие его используют и для вообще всех видов распределений, но при больших выборках.

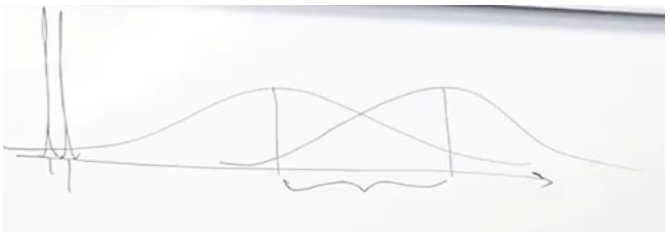
Как говорилось ранее, сравнивать средние арифметические. При этом важно понимать, что у выборок может быть разная дисперсия, что существенно влияет на конечный результат применения критерия:

может быть такая ситуация, когда разница между средними выборок мала, но различия статистически значимы, а может быть наоборот - разница большая, но т-тест говорит о том, что различий нет. Так может быть из-за дисперсий



Есть 2 выборки, средние которых отличаются не сильно, на каждое наблюдение одной выборки больше наблюдений другом - есть стат. значимая разница.

Другая ситуация:



Разница между средними намного больше, но различий нет - наложение гистограмм большое.

Поэтому дисперсии играют роль при сравнении средних. Поэтому есть много вариантов критерия Стьюдента - с разными дисперсиями и нет. **Если не уделить этому внимания, то возрастает ошибка второго рода.**

Проверка на гомогенность дисперсий

Есть множество тестов на проверку равенства дисперсий. Один из них долгое время был популярным - **F-test.** Однако сейчас он не рекомендуется к применению, т.к. слишком чувствителен к отклонениям от нормальности.

Также есть тест Бартлетта, который является лучшим выбором при нормальных распределениях. Однако в других случаях его тоже лучше не использовать.

Также есть тест **Левена** и **Флигнера-Киллена**. Последний робастен и рекомендуется при проверке равенства дисперсий.

В последнее время популярность приобрёл еще и тест Brown-Forsythe, но в питоне его нет.

Гипотеза независимости

H0: СВ X и Y независимы

H1: СВ X и Y зависимы

При этом понять, что от чего зависит, критерий не скажет. Также почти всегда интересует не сам факт зависимости, а функция зависимости.

Также корреляция важна в нейронных сетях и случайном лесе. Если в лесу много одинаковых деревьев (корреляция высокая), то голосующие деревья будут дублировать друг друга, что плохо. Тоже самое и в нейронных сетях: если есть дублирующиеся нейроны, то сетка слишком сложная.

Существует несколько коэффициентов корреляции:

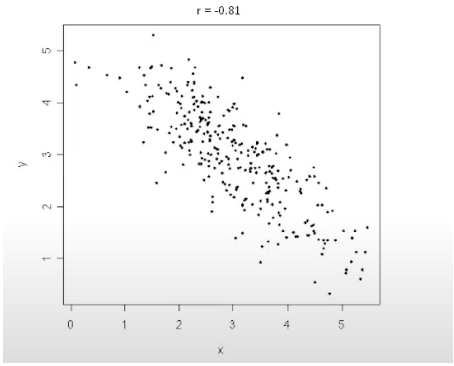
1. Пирсона - при нормальном распределении
2. Спирмена - во всех остальных случаях
3. Кендалла - во многом дублирует коэффициент Спирмена (при этом коэффициент получается ниже, поэтому используется редко).

Формула для коэффициента корреляции:

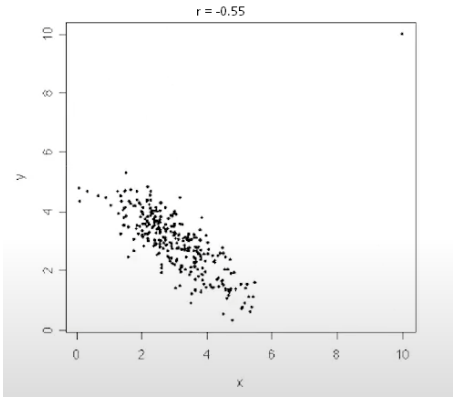
$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Ошибки при применении коэффициента корреляции

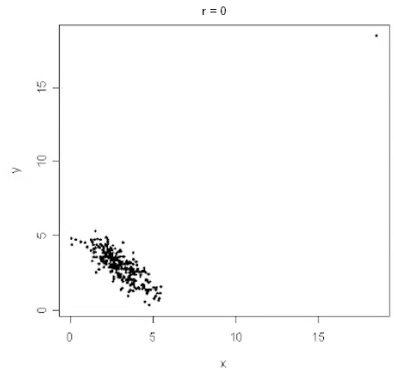
1. Зависимость должна быть линейной - только в таком случае r не будет равен 0 (если зависимость есть)
2. Выбросы существенно могут повлиять на r:



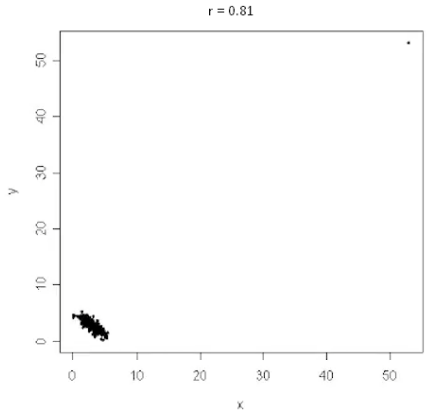
Добавим выброс:



Сделаем выброс сильнее:



Сделаем еще сильнее - коэффициент корреляции поменял знак:



3. Ложная корреляция - события, которые никак не связаны между собой, но сильно коррелируют. Возникает, когда есть монотонные временные ряды - по модулю r будет высок.