Урок 5
Основы математической статистики

На этом уроке

- Поговорим о важности статистики для A/B тестов
- Пройдемся по базовым понятиям статистики
- Посмотрим, что такое Центральная предельная теорема и для чего она применяется в A/B тестировании
- Разберем как оценивать по выборке в каких границах лежат реальные значения ваших метрик
- Рассмотрим как минимизировать выкатку в продакшн неработающих решений и отклонение работающих

Оглавление

Важность статистики в A/B тестах

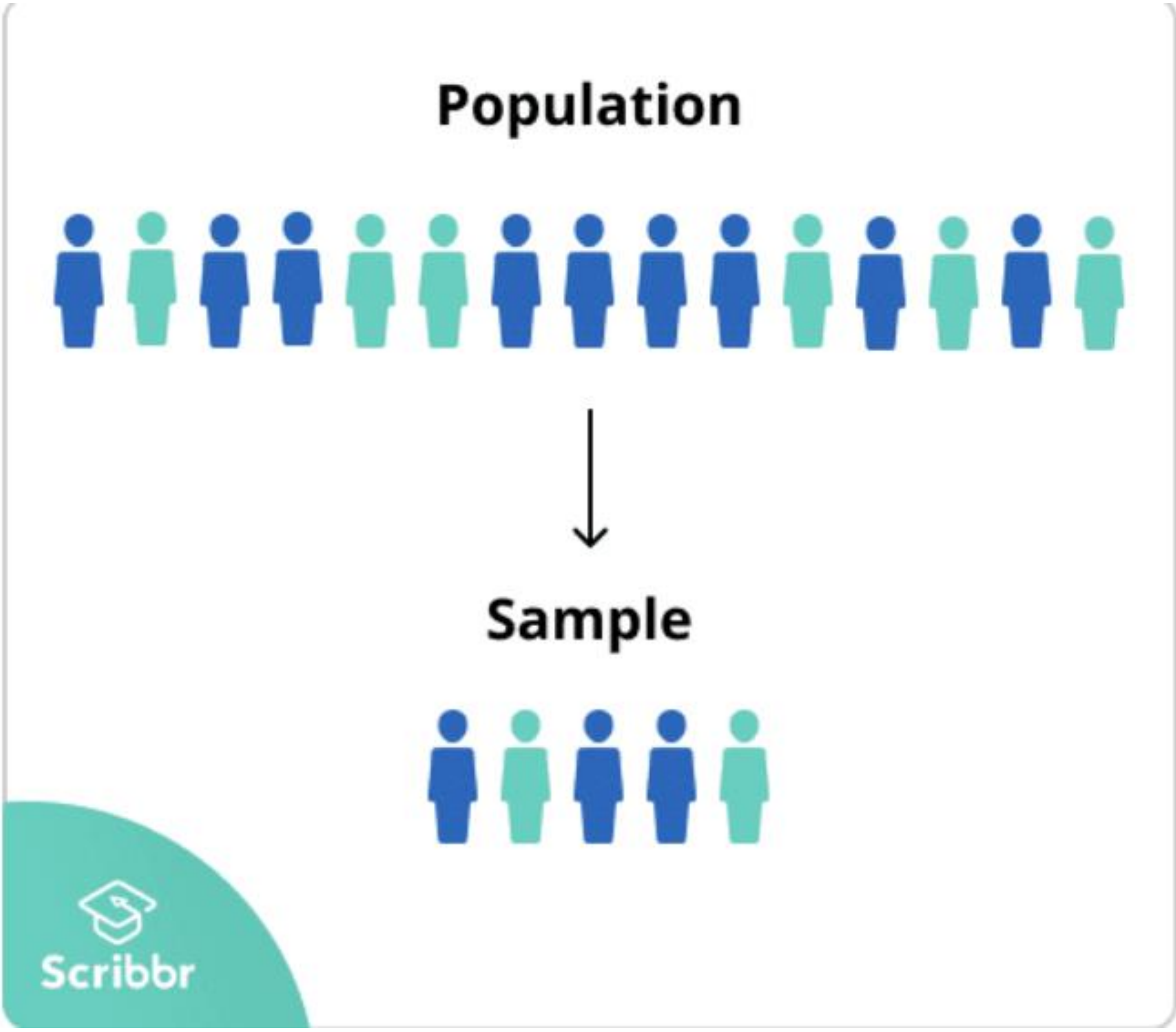
Как вы думаете возможно ли построить хороший дом без фундамента? Тут очевидно напрашивается единственно верный ответ.

Аналогично, и в A/B тестах нельзя обойтись без базовых знаний по математической статистике, если вы хотите проводить в своем продукте качественные и полезные для бизнеса эксперименты .

Зная основы математической статистики **вы сможете:**

- Оценивать по выборке в каких границах лежат реальные значения ваших метрик
- Сравнивать между собой выборки по различным параметрам
- Понимать, где вы получили случайную разницу в метриках, а где нет
- Рассчитывать длительность теста
- Минимизировать выкатку в продакшн неработающих решений
- Экономить команде время и деньги :)

Базовые понятия статистики



Начнем с базовых понятий:

Генеральная совокупность - совокупность всех объектов или наблюдений, относительно которых исследователь намерен делать выводы при решении конкретной задачи. В ее состав включаются все объекты, которые подлежат изучению.

Выборка - часть генеральной совокупности с помощью определённой процедуры выбранных из генеральной совокупности для участия в исследовании.

Могут возникнуть возникнуть вопросы: зачем же нужна выборка если есть генеральная совокупность? Почему сразу все не считать по ней?

В реальной жизни, у исследователей к сожалению, почти всегда нет доступа ко всем данным генеральной совокупности , это может быть долго, дорого , а иногда и физически невозможно.

Чтобы переносить выводы с выборки на генеральную совокупность, выборка должна быть **репрезентативной,** отражать пропорции и особенности генеральной совокупности.

На репрезентативность сильно влияет метод отбора выборки(их существует довольно много). На практике часто используется [рандомизация.]

Пример: Допустим в нашем исследовании - мы хотим знать средний рост мужчин в Москве живущих на данный момент .

Генеральная совокупность в данном случае - это все мужчины живущие мужчины в Москве сейчас (будет исчисляться миллионами).

Выборка - например, случайно отобранные 10 к мужчин из разных районов.

По этим мужчинам мы замерим рост и сможем оценить его истинное значение с определенной точностью, поговорим дальше как это сделать:)

Описываем выборку и ГС:

В большинстве случаев как в бизнесе так и в научных исследованиях вы будете считать метрики и делать выводы по выборке, а не по всей генеральной совокупности. Вам нужно уметь описывать и то и другое, в этом вам помогут базовые понятия:

Желающие могут освежить знания ознакомившись со статьей или посмотреть видео.

Случайная величина (ξ) – это математическое понятие, служащее для представления случайных явлений, когда для них может быть определена их вероятность, то есть мера возможности наступления.

По сути это переменная со значениями. Для каждого значения есть своя вероятность исхода

Математическое ожидание (*μ, M*) — это среднее арифметическое значение случайной величины.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Дисперсия (*S2*) – рассчитанное расстояние, на которое значения случайной величины находятся вокруг его математического ожидания

$$\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Стандартное отклонение (*SD*)— это квадратный корень от дисперсии

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

Пример: Мы хотим узнать среднюю цену блюда в ресторане 1 и то насколько сильно в нем колеблются цены, для этого можем подсчитать базовые описательные статистики.

Блюдо	Цена руб
1	500
2	450
3	400
4	470

Мат ожидание	Дисперсия	Стандартное отклонение
455	1 767	42

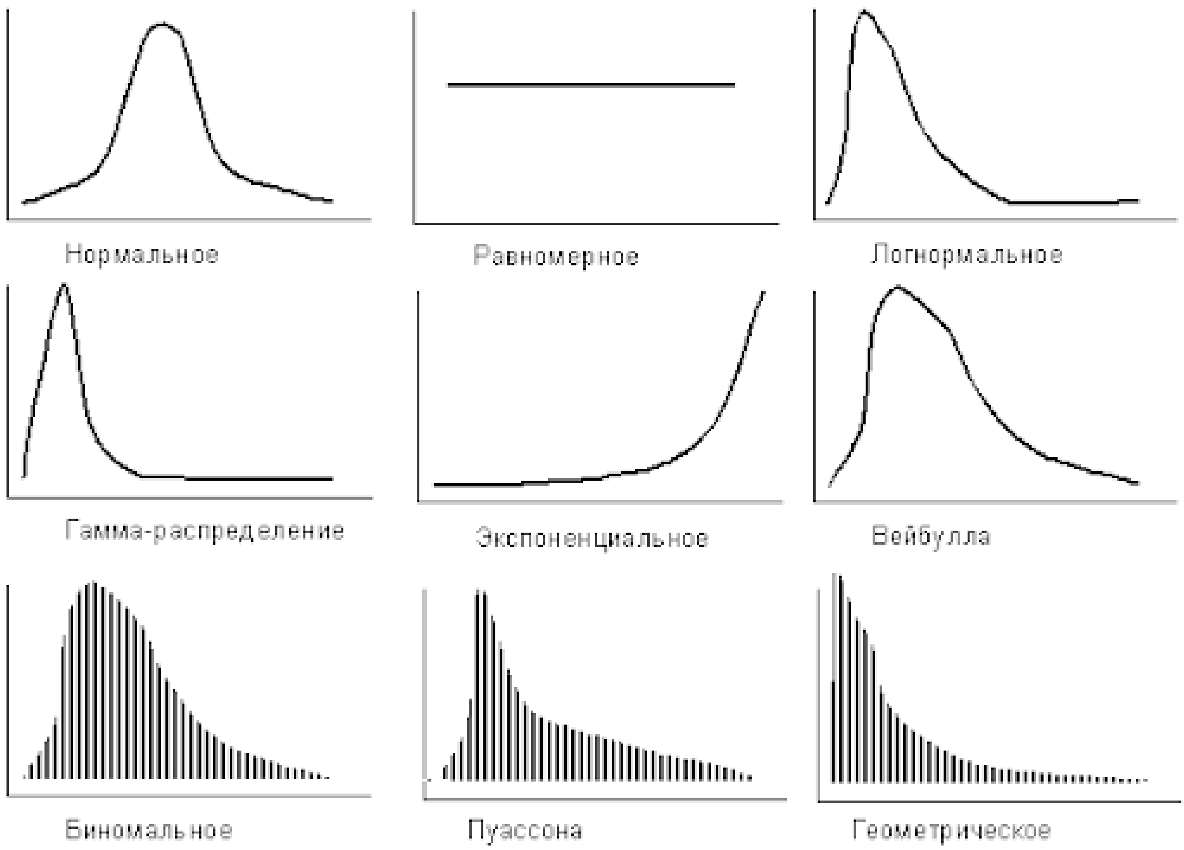
Посчитаем дисперсию:((500-455)^2 + (450-455)^2 + (400-450)^2 + (470-450)^2) / (4-1) = 1767

Законы распределения

Для описания случайных величин используются законы распределения с параметрами.

Распределение вероятностей — это закон, описывающий область значений случайной величины и соответствующие вероятности появления этих значений.

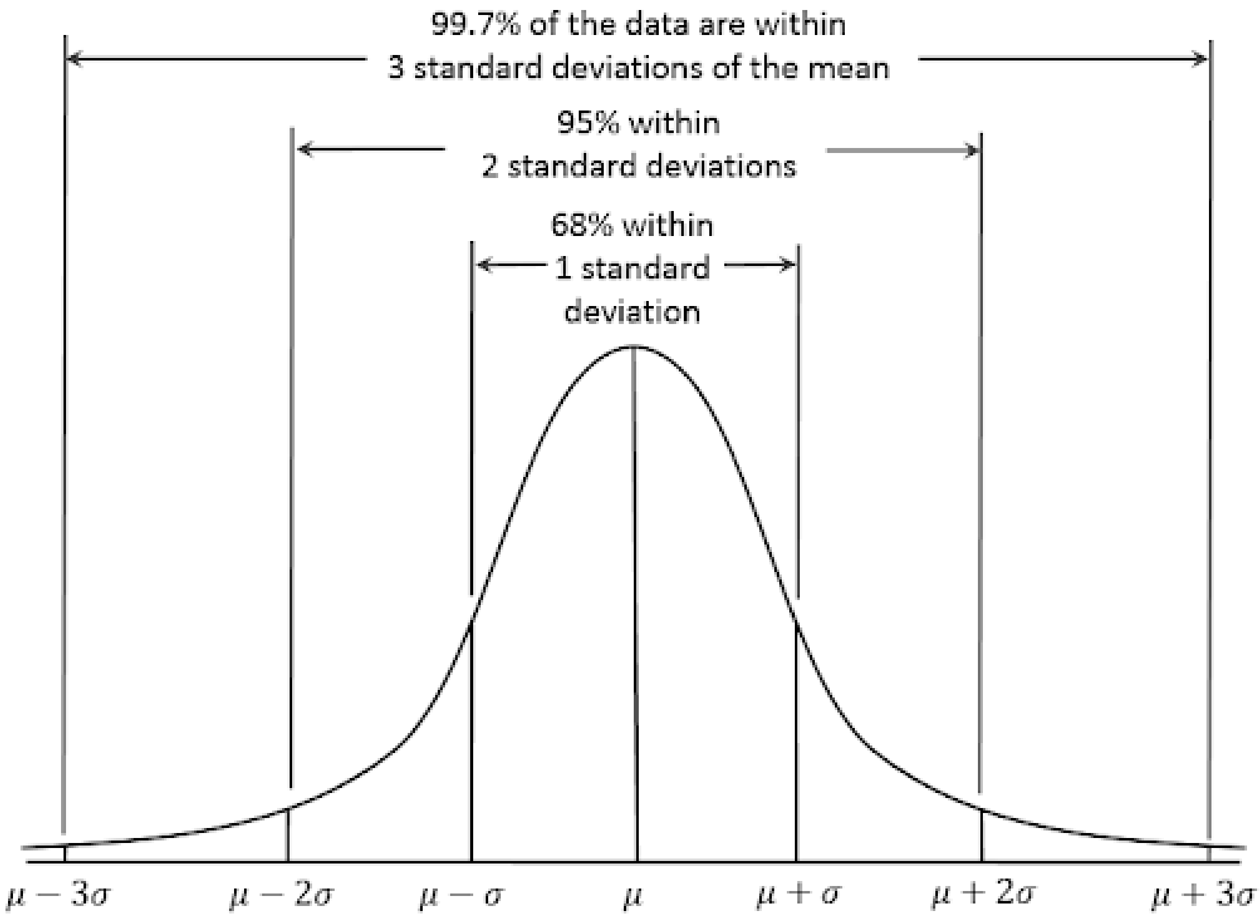
На рисунке ниже визуализированы различные законы распределения:



Нормальное распределение:

Нормальное распределение самый частый вид, который используется в статистике. Закон распределения задается по формуле:

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



Где, $\boldsymbol{\mu}$ - математическое ожидание и $\boldsymbol{\sigma}$ - стандартным отклонением. Эти параметры моделируют, как будет выглядеть распределение. Большая часть тестов при проверке гипотез требует в качестве условий корректности работы- нормальность распределения или принадлежность к какому-либо закону распределения.

Центральная предельная теорема

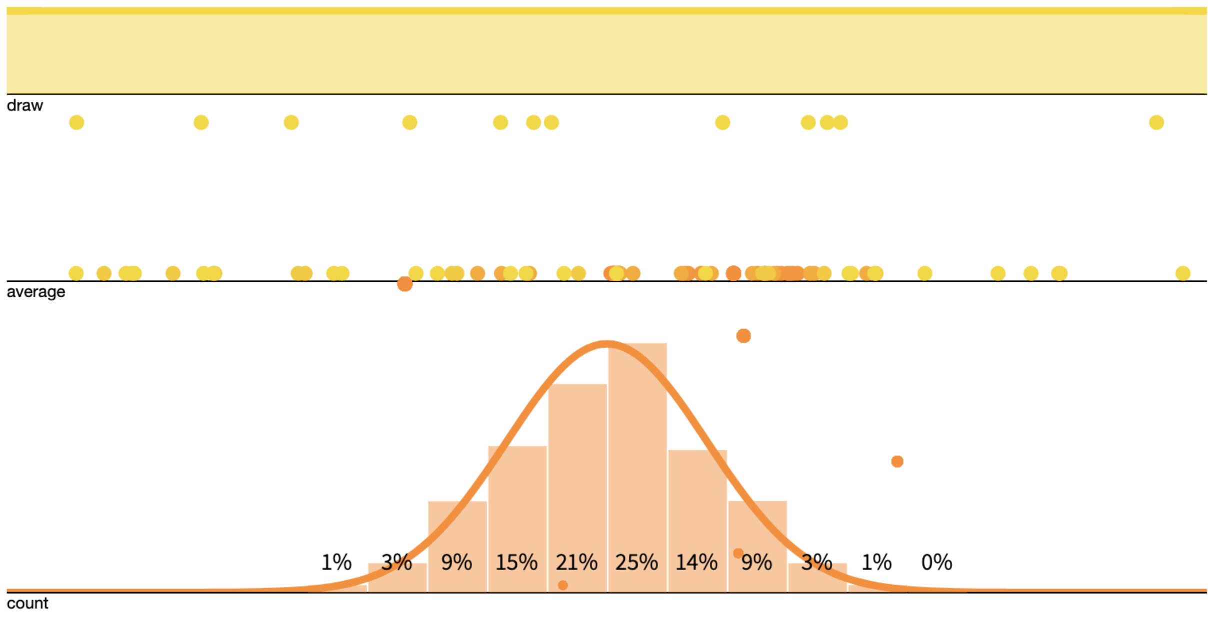
Как можно судить о свойствах генеральной совокупности по выборке?

ЦПТ говорит, что если мы возьмем достаточно большую выборку из независимых, одинаково распределенных случайных величин(i.i.d из генеральной совокупности , то среднее значение будет нормально распределено с ***μ(мат ожидание) и SD(стандартное отклонение) .***

Причем мы можем посчитать оценить параметры *μ и SD по нашей выборке.*

Визуализация ЦПТ

Берем и многократно извлекаем выборки определенного размера и считаем по ним среднее. Распределение средних при соблюдении предпосылок выше будет нормально распределенным и мы сможем оценивать истинное значение в ГС.



(промежуточный скрин из визуализатора показывает что мы постепенно будем приближаться к нормальному распределению при многократном извлечении)

Что нам позволяет делать ЦПТ ?

ЦПТ дает возможность строить доверительные интервалы и проверять статистические гипотезы. Поговорим об этом ниже. При желании вы можете прочитать [хороший разбор ЦПТ тут.]

Доверительные интервалы

Допустим у вас e-commerce магазин и вы хотите посчитать какое истинное значение конверсии(по всей ГС) из добавления товаров в корзину в покупку для вашего продукта .

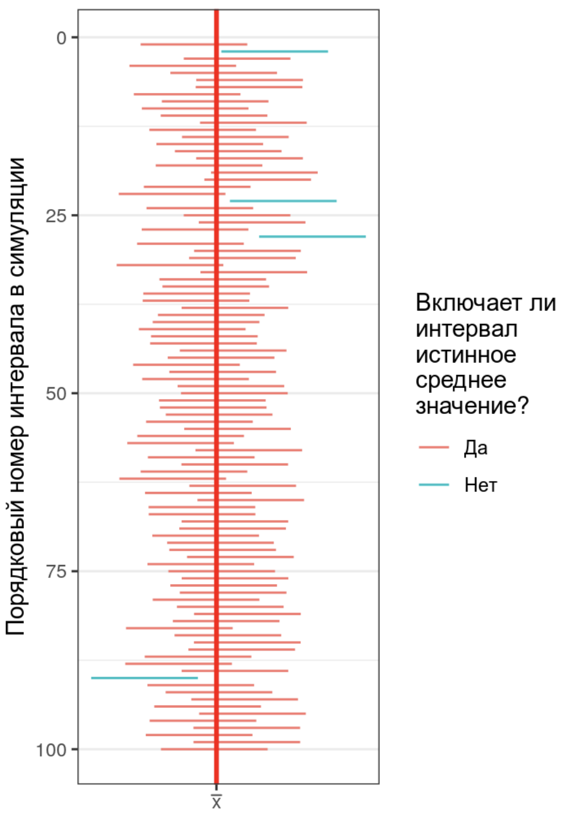
К сожалению, истинное значение конверсии вы не знаете - но зато у вас есть доступ к данным ваших систем аналитики и вы сможете посчитать выборочное значение и по нему оценить в каком диапазоне может лежать истинное значение конверсии. В этом нам поможет доверительный интервал, который мы можем применять благодаря ЦПТ.

Доверительный интервал для долей

Что это такое ? Это способ оценки метрики, используя который, мы получим диапазон значений [x,y] , внутри которого будет лежать истинное значение метрики ГС в 95% случаев.

(Если провести очень большое количество независимых экспериментов с аналогичным построением доверительного интервала, то в 95% экспериментов доверительный интервал будет содержать оцениваемый параметр ген совокупности.

В оставшихся 5% экспериментов доверительный интервал не будет содержать параметр ген совокупности.)



(Мы можем регулировать точность и выбирать 99% , 90% и другие диапазоны.)

Для метрик долей*(ретеншн, конверсия и тп) используется формула

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

*Типы метрик которые бывают в экспериментах:

1)Доли - (ретеншн, конверсии) [0,1,0,0,0,1]

2)Непрерывные - таймспент в сек / деньги

3)Отношения - (клики на сессию , поездки на юзера))

Вы можете рассчитать доверительный интервал на калькуляторе:

[Ссылка на калькулятор]

Пример расчета доверительного интервала для конверсии:

Допустим по выборке 1000 юзеров мы посчитали конверсию из добавления товаров в корзину в покупку для нашего e-commerce магазина , конверсия равнялась 10%.

Мы хотим с уровнем доверия 95% оценить истинное значение конверсии. Для этого используем формулу:

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

0.10 +- 1.96 * (0.10 * (1-0.10) /1000) ^ 1/2

Доверительный интервал будет следующий:

[8.14% , 11.86%]

Давайте попробуем проинтерпретировать полученный результат:

В нашем e-commerce магазина истинная конверсия лежит в диапазоне [8.14% , 11.86%] с 95% уровнем доверия .

Если бы мы много раз оценивали по выборкам конверсию - ее истинное значение лежало бы в 95% случаев в границах доверительного интервала в каждом из расчетов. При желании чуть глубже погрузиться вы можете посмотреть видео 1. + видео 2

Как вы могли заметить по формуле доверительного интервала: Чем больше выборка, тем точнее оценка.

Ошибки первого и второго рода

Рассмотрим базовые понятия:

Нулевая гипотеза – принимаемое предположение о том, что не существует связи между наблюдениями в двух (или более) событиях (выборках, феноменах, совокупностях). Гипотезу отвергают, если данные показывают разницу между выборками.

True Positive = говорим истина, когда по факту истина (факт)**True Negative** = говорим не истина, когда по факту тоже не истина (факт)

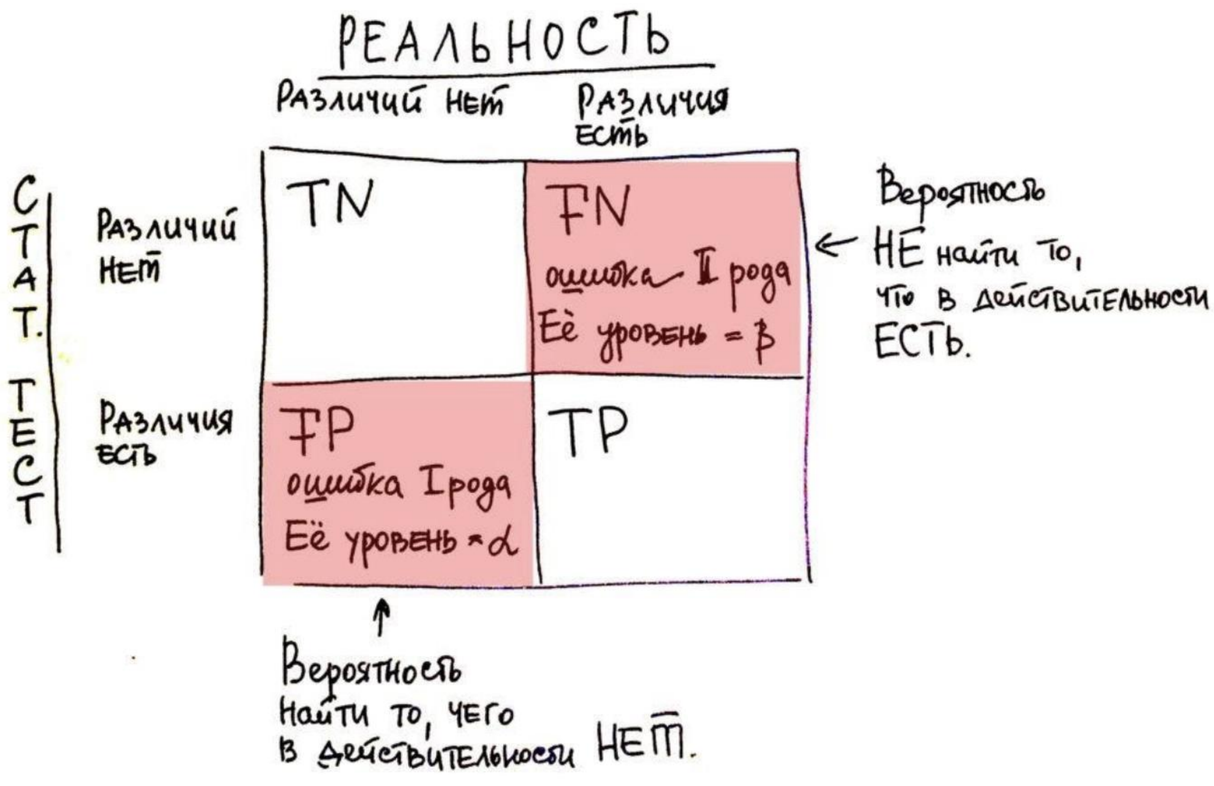
False Positive (ошибка I рода) = говорим истина, когда по факту не истина. Отклонение верной нулевой гипотезы. Риск совершить такую ошибку равен выбранному уровню статистической значимости (например, α=0.05) (ошиблись)

False Negative (ошибка II рода) = говорим не истина, когда по факту истина (ошиблись). Принятие неверной нулевой гипотезы. Вероятность ошибочно сохранить неверную нулевую гипотезу обозначают буквой β (например, 0.2).

Статистический критерий(тест) — это математическое правило, которое помогает нам понять, отвергнута или нет наша гипотеза. Мы поговорим о них подробно на следующем занятии.

Мощность критерия/теста - это вероятность, что тест правильно засечёт эффект там, где он и правда есть. (т.е. 1-β)

Допустим мы сформулировали нулевую гипотезу для эксперимента(поговорим о них подробнее на следующих занятиях). У нас могут быть следующие исходы:



Рассмотрим пример :

Однажды воскресным вечером через пару дней после веселых посиделок в баре вы замечаете что у вас пропало обоняние и появилась легкая усталость . Вы пошли и сдали тест на COVID и получили положительный результат. Но не спешите грустить и лучше по возможности сдайте его повторно:)

В реальности могут быть следующие исходы :

Вы реально больны и тест показал что вы больны **True positive**

Вы реально не больны и тест показал что вы больны **False positive**

Вы реально не больны и тест показал что вы не больны **True negative**

Вы реально больны и тест показал что вы не больны **False negative**

У всех тестов разное качество что может влиять на False Positive и False negative. Поэтому лучше перепроверяться

еще один наглядный пример:

Мы в продуктах не хотим принимать ложноотрицательные(отклонять крутые гипотезы) и ложноположительные(раскатывать плохие гипотезы) решения и хотим минимизировать ошибку первого и второго рода .Выбираем alpha и ищем самый мощный критерий среди корректных.

Мощность и точность зависят от размера выборки - чем большую мощность и точность мы хотим - тем больше выборка нам нужна - поговорим об этом в следующей главе. Вам сейчас важно понять суть различных ошибок - поскольку в дальнейшем они будут использоваться регулярно.

Концепция проверки статистических гипотез

В этой главе поговорим о том, как устроен процесс проверки статистических гипотез и какие нюансы существуют .

Начнем с определения статистической гипотезы.

Статистическая гипотеза - это выдвигаемое предположение о свойствах и характеристиках генеральной совокупности(ей).

После выдвижения гипотезы собираются реальные данные выборок(выборки) и к ним применяются статистические методы(критерии) например t test или z test, о которых детальнее мы поговорим позже.

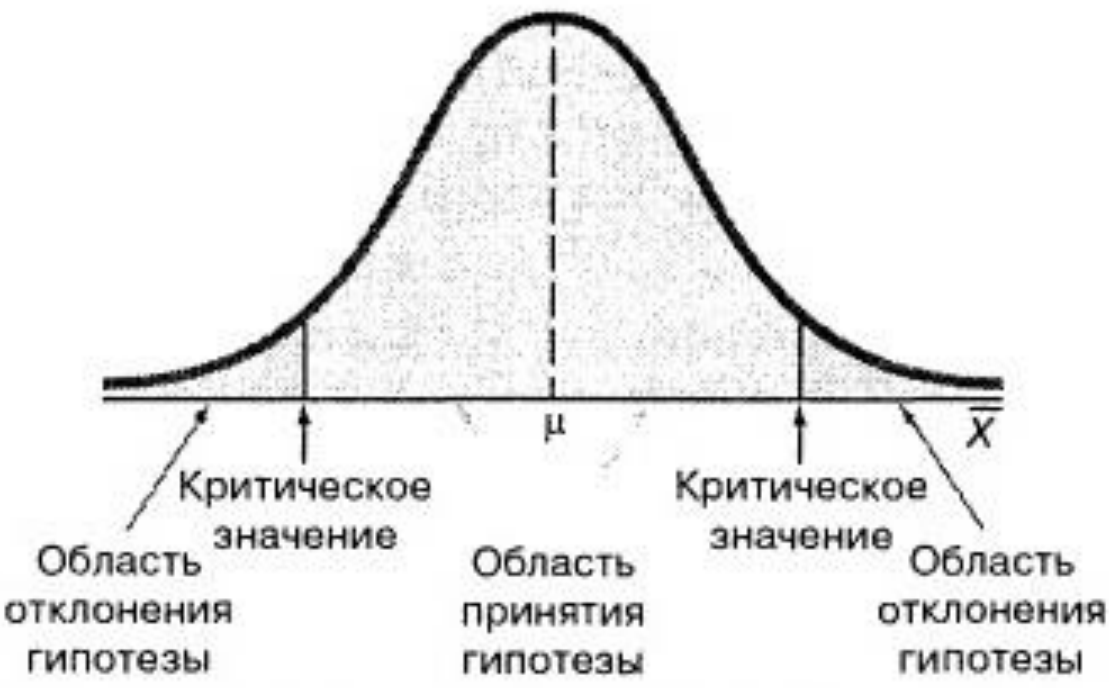
После применения критериев мы либо отвергаем, либо не отвергаем гипотезу. Эта гипотеза называется нулевой и мы дадим ей подробное определение ниже.

Примеры гипотез:
Среднее время звонка для всех наших клиентов равно 3 минуты
Уровень брака в текущей партии продукта равен 1%
Retention 5 дня в контрольной группе не отличается от Retention 5 дня в тестовой группе.
Важно! : Корректно говорить не отвергаем нулевую гипотезу, а не принимаем .

Проверка гипотез в статистике строится по методу доказательства от противного. Сначала выдвигают нулевую гипотезу и предполагают что она верна. Ее придерживаются пока наблюдаемые данные на докажут обратное.
Нулевая гипотеза – принимаемое предположение о том, что не существует связи между наблюдениями в двух (или более) событиях (выборках, феноменах, совокупностях). Гипотезу отвергают, если данные показывают разницу между выборками.
Нулевой гипотезе сопутствует альтернативная гипотеза - принимаемая в случае отвержения нулевой гипотезы.
Если предположение противоречит наблюдаемым данным, то гипотезу отклоняют, как ложную в пользу альтернативной, если не противоречит, то говорят что оснований отвергнуть нулевую гипотезу нет.
Сила противоречия зависит от того, того насколько далеко выборочная статистика(например для выб.среднего) отклоняется от гипотезы.

Благодаря ЦПТ, которая рассматривалась на предыдущем занятии знаем что выборочные средние при соблюдении определенных предпосылок имеют нормальное распределение.
Чтобы погрузиться тему чуть глубже /освежить знания можно почитать:

статья 1 + статья 2 + видео 1 или видео 2



Можно рассматривать площадь под кривой в определенном диапазоне как вероятность получить тот или иной диапазон значений.

Поизучать по теме статья про z score № 1 и статья № 2

На графике выше можно заметить что существуют область принятия и область отклонения гипотезы.

Областью принятия гипотезы называют совокупность значений статистики критерия, при которых гипотезу H0(нулевую) принимают.

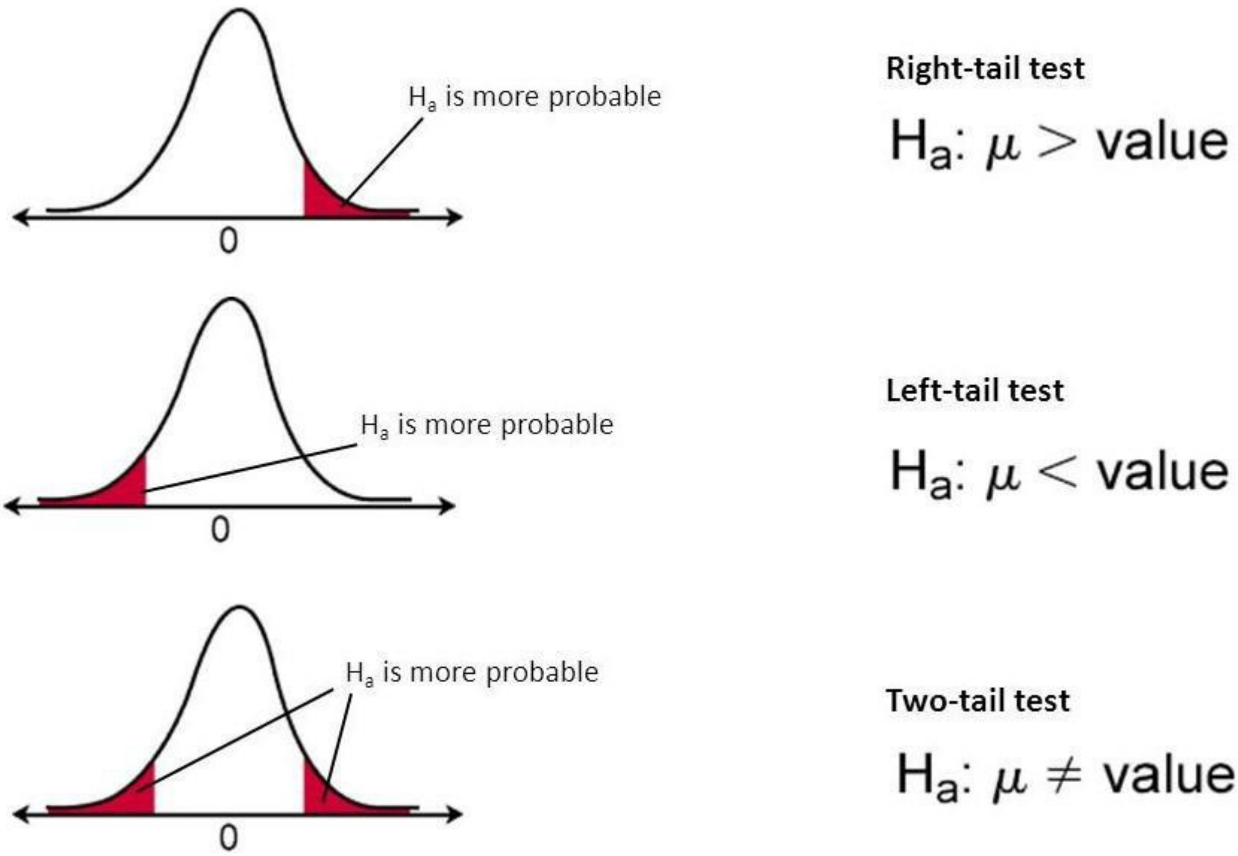
Областью отклонения гипотезы(критической областью) - называется область значений статистики критерия, при которых отвергается H0(нулевую).

Критические значения - это граница критической области, они зависят от выбранного alpha. Вид критической области определяется видом альтернативной.

Мы смотрим куда у нас попадает выборочное значение. Если рассчитанное значение окажется в зоне близкой к центру(в области принятия гипотезы), то оно не противоречит гипотезе(могли быть просто небольшие колебания в силу случайности).

Но если оно окажется далеко от центра за пределами критических значений(слева или справа) - это значит что вероятность получить такое или еще более высокое отклонение крайне мало и гипотеза скорее всего ложная.

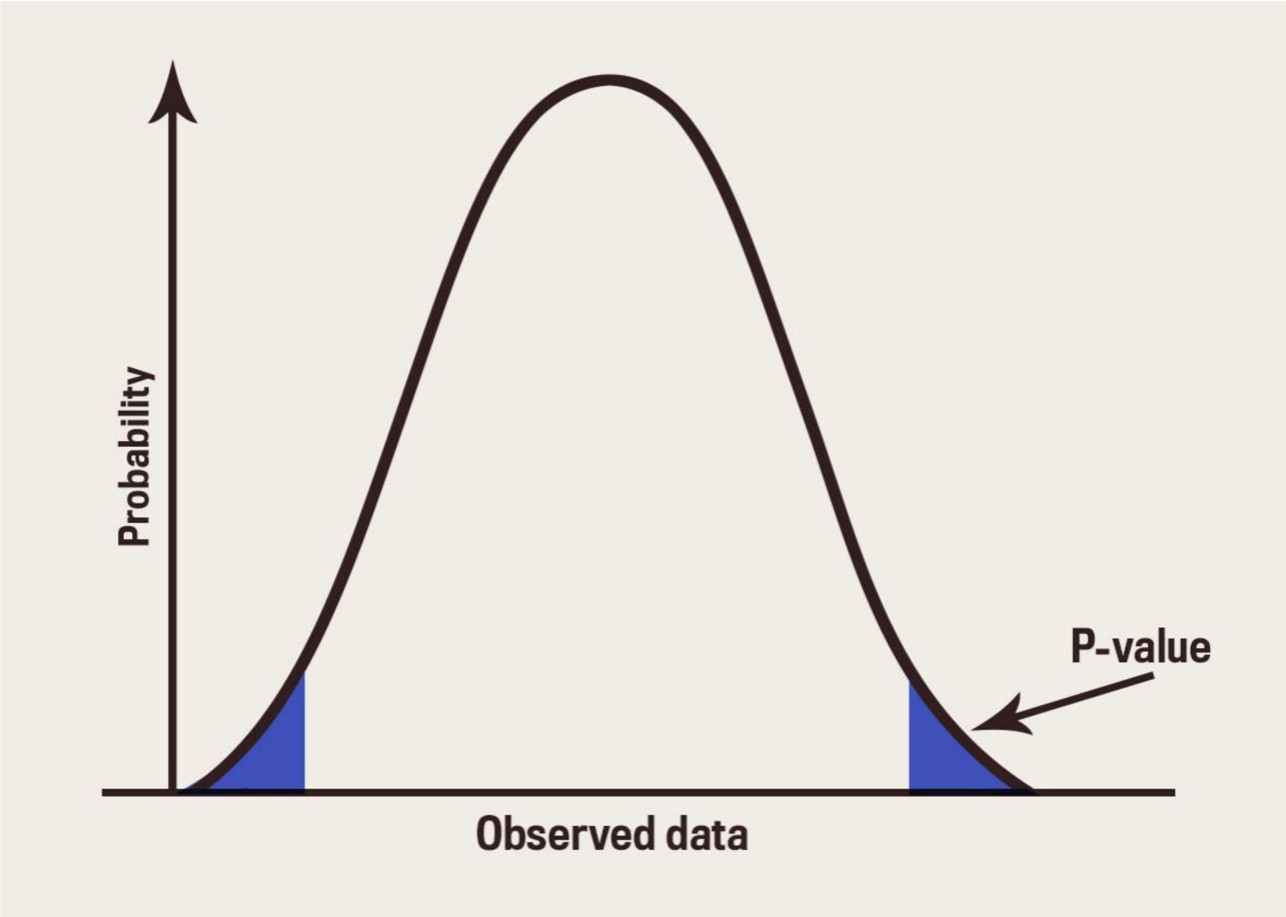
Гипотезы могут быть **односторонними и двухсторонними** (см.график ниже). На практике почти всегда используют двухсторонние тк мы почти никогда не знаем заранее перед запуском теста в какую сторону отклонится метрика в тестовой группе. А после получения данных менять двухсторонний на односторонний критерий некорректно - таким образом мы подгоняем теорию под наши данные.



Для углубленного изучения при желании можно почитать лекции ВШЭ и статью.

В современном мире используется понятие:

p-value – вероятность при справедливости нулевой гипотезы получить наблюдаемое или еще большее отклонение оценки статистики(например среднего) по сравнению с предполагаемым в гипотезе.



Многие некорректно интерпретируют p-value. Основная проблема - для многих людей это не интуитивная сущность и возникают проблемы с интерпретацией. Я предлагаю вам по желанию ознакомиться с парой интересных статей о нем, чтобы не совершать ошибок при интерпретации данного понятия .

Статья на русском на хабре и на английском.

Если p-value < alpha (false poistive), то нулевая гипотеза отвергается в пользу альтернативной!.

Пример: проверяем гипотезу о равенстве конверсии в покупку магазина e-commerce В одной группе CR1 =1 % в другой CR2 = 1,01%.

Выдвигаем нулевую гипотезу что между CR1 и CR2 нет разницы, это двухсторонняя гипотеза, alpha(false positive) возьмем равную 5%.

Допустим , применив критерий мы вычислили p-value =0,7% .

Сравнив p-value и alpha : 0,7 > 0,05 - значит нулевая гипотеза не отвергается.(исторически сложился бенчмарк 5%, но также допустимы и другие отсечки например 1%,10% и другие в зависимости от объема выборки в продукте и готовности получать false positive ошибки.)

К p-value мы еще вернемся в следующих главах. В соседних главах рассмотрим статистические критерии , алгоритм принятия решений и идею статистического вывода по данным.

Статистические критерии

Чаще всего в A/B тестах проверяют гипотезу о равенстве двух показателей(средних) - используя для этого статистические критерии.

Статистический критерий — математическое правило, в соответствии с которым отвергается либо не отвергается та или иная статистическая гипотеза с заданным уровнем значимости.

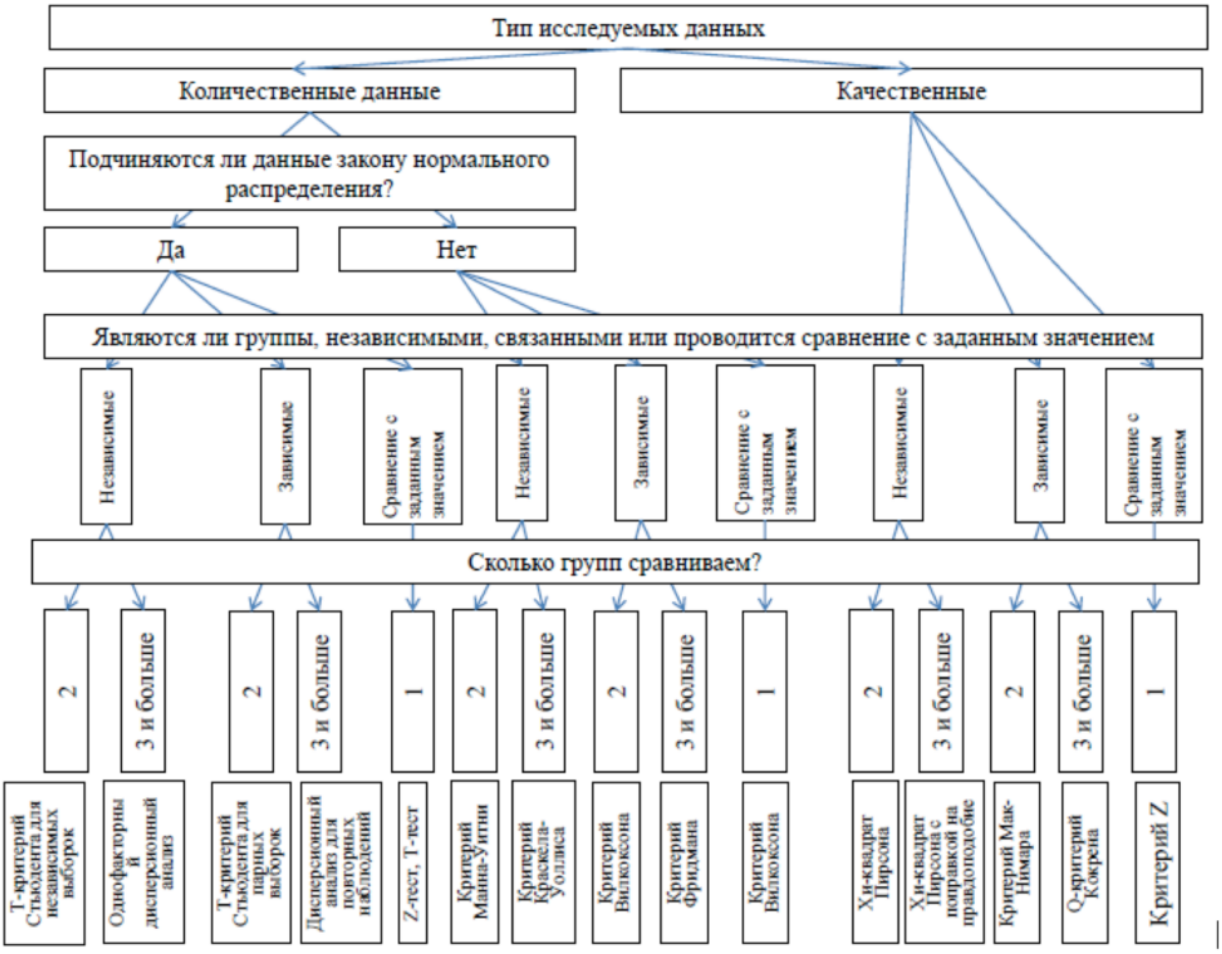
Для того, чтобы проверить гипотезу о равенстве показателей, применяется два типа критериев оценки: *параметрические* и *непараметрические*.

Если просто описывать основную суть их различия - то можно сказать что:

Параметрическими называются критерии, в которых мы можем сделать предположение о распределении , относящееся к какой-то выборке. В большинстве случаев в качестве распределения используется нормальное.(t тест)

Непараметрические не используют предположения о распределении, а оперируют *рангами* и *частотами.(*U-критерий Манна — Уитни*)*

У каждого критерия есть предпосылки - и можно использовать таблицу ниже для выбора подходящего критерия.



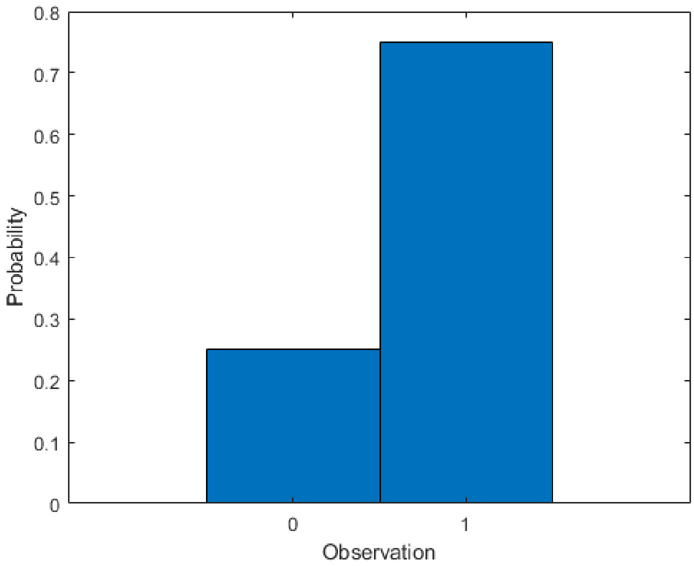
Необязательно знать все критерии - на практике чаще всего используется [лишь их часть.]

Часто используются [t-test Уэлча] для сравнения средних , z - критерий , для сравнения метрик долей(конверсии/ретеншн) , критерий Хиквадрат тоже подойдет для долей + Критерий Манна Уитни для сравнения ненормально распределенных величин.

Рассмотрим z - критерий для метрик долей как с помощью него проверять гипотезы для конверсий /ретеншн итд используя p-value и доверительный интервал.

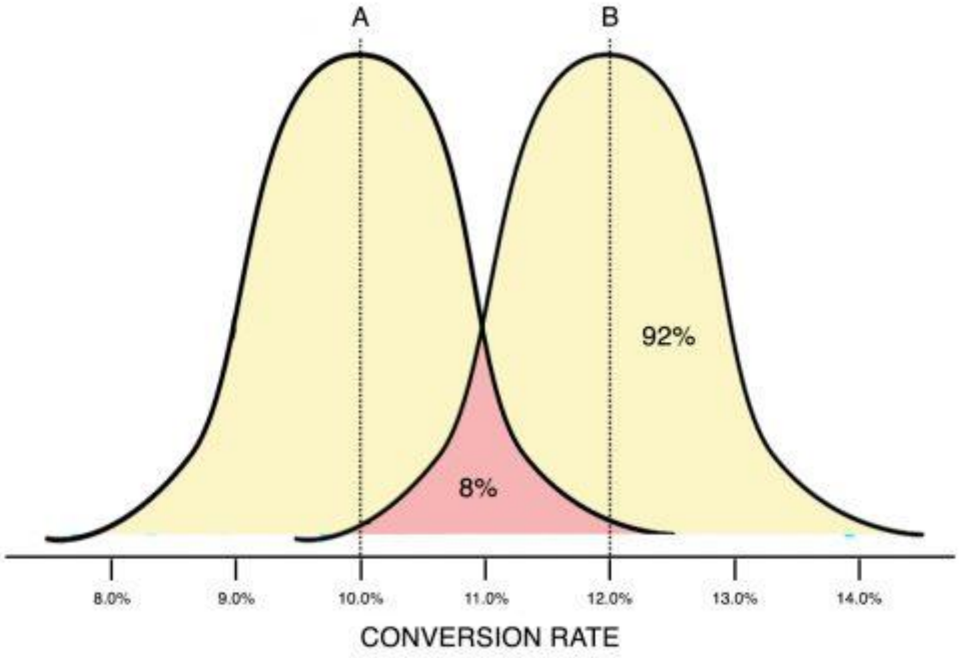
Критерии долей — это критерии, которые работают с распределениями Бернулли.

Они принимают на вход выборки из нулей и единиц и проверяют гипотезы о вероятности появления единицы в выборке



Гипотезы можно проверять с помощью доверительного интервала или p-value.

Допустим мы запустили эксперимент направленный на улучшение конверсии на каком-нибудь участке продукта , получили результаты конверсий в обеих группах и хотим понять эта разница получена не случайно, а действительно существует закономерность.



У каждой из метрик в двух выборок есть сигнал(разница средних) и шум (дисперсия). Мы хотим понять несмотря на наличие шума а есть ли действительная разница между средними ? В этом нам помогают статистические критерии.
В случае с критериями долей мы можем использовать следующий принцип:
Пусть p̂ 1 и p̂ 2 посчитанные в А/Б-тесте конверсии, в контроле p̂ 1 , а в тесте p̂ 2.
n1 и n2 — количество элементов в каждой
1) Используя доверительный интервал: выберем уровень значимости alpha = 5%
Тогда 95% доверительный интервал для их разницы будет считаться следующим образом:

$$\hat{p}_2 - \hat{p}_1 \pm 1.96 \times \sqrt{\hat{p}_1 \times (1 - \hat{p}_1) \div n_1 + \hat{p}_2 \times (1 - \hat{p}_2) \div n_2}$$

Мы получим диапазон значений [x;y]

Если 0 входит в интервал значит на основе собранных данных статистически значимой разницы между группами не обнаружено.

Если не входит, то между группами есть стат.значимая разница.

2) Альтернатива - можем также по нашим данным посчитать p-value:

Выберем уровень значимости alpha = 5%

Посчитаем p-value на полученных в эксперименте данных.

Помним про нулевую гипотезу:(Разницы между средними в двух группах нет.)

Если p-value > 5% значит на основе собранных данных статистически значимой разницы между группами не обнаружено.

Если p-value < 5% , то между группами есть стат.значимая разница

Рассмотрим пример:

Допустим мы запустили эксперимент направленный на улучшение конверсии на одном из кусков сайта:

Aa Группа	Посетители	Кликнувшие посетите...	Конверсия %
Тестовая группа	2400	178	7,42%
Контрольная группа	2400	199	8,29%

Хотим понять есть ли стат.значимая разница между группами с достоверностью 95%:

Пусть p̂ 1 и p̂ 2 посчитанные в А/Б-тесте конверсии, в контроле p̂ 1 = 8,29%

, а в тесте p̂ 2. = 7,41%.

n1 и n2 — количество посетителей в каждой из них.

Посчитаем доверительный интервал по формуле:

$$\hat{p}_2 - \hat{p}_1 \pm 1.96 \times \sqrt{\hat{p}_1 \times (1 - \hat{p}_1) \div n_1 + \hat{p}_2 \times (1 - \hat{p}_2) \div n_2}$$

0.0742-0.0829 ± 1.96 * (0.0742 × (1 − 0.0742) ÷ 2400 + 0.0829 × (1 − 0.0829) ÷ 2400) ^1/2 = [-2.392% ; 0.652%]

0 входит в интервал - значит на основе собранных данных статистически значимой разницы между группами не обнаружено.

Также можно было бы посчитать p-value воспользовавшись например калькуляторами:

[Калькулятор 1]

[Калькулятор 2]

p value = 0.1299 > 0.05 - на основе собранных данных статистически значимой разницы между группами не обнаружено.

Похожий принцип принятия решений действует и для других критериев.

Алгоритм проверки гипотез

Давайте рассмотрим алгоритм проверки гипотез, который можно использовать при проверке гипотез.

1. Выбираем метрику и формулируем нулевую и альтернативную гипотезы
2. Выбираем параметр alpha (например 5%) равный вероятности допустить ошибку первого рода
3. Выбираем критерий, подходящий под наши условия
4. Считаем p - value и(или) доверительный интервал и делаем вывод:

Если p-value < alpha - разница между группами стат.значима . Либо если доверительный интервал для разницы не включает 0.

5. Даем (рекомендации лицам принимающим решения /принимаем решение) выкатывать или не выкатывать новое изменение - если ключевая метрика в группе с новой фичей стат значимо лучше чем в контроле а дополнительные метрики не упали- можно раскатывать , помня про ошибку первого рода).

   Критерии принятия решений определяются заранее на этапе планирования эксперимента.

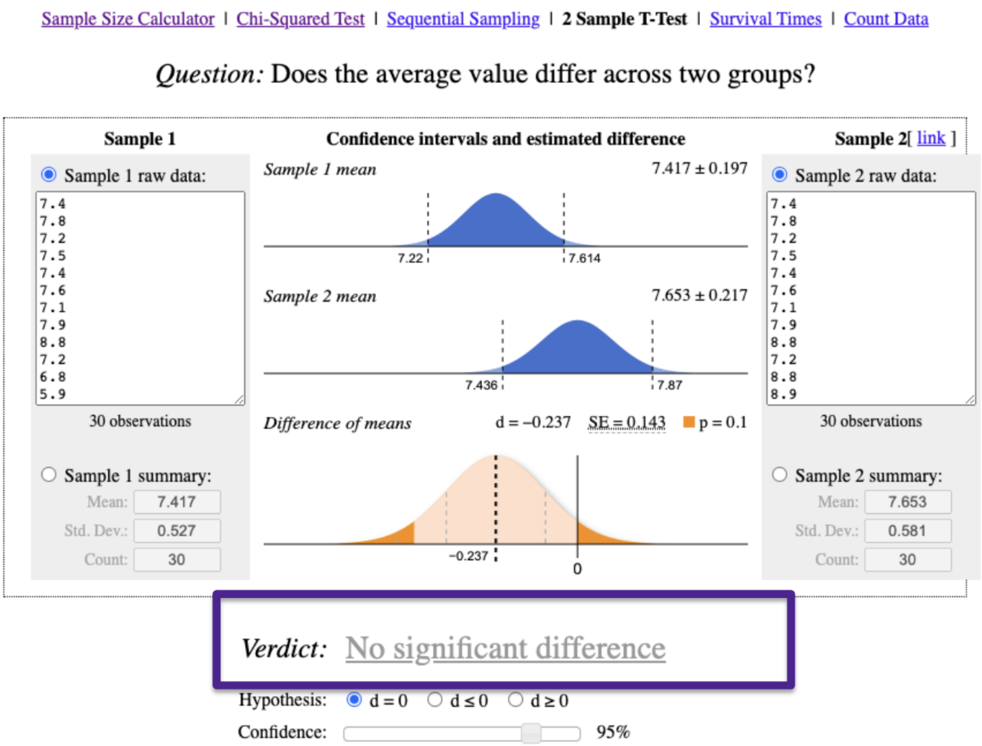
Калькуляторы, которые вы можете использовать:

1. https://www.evanmiller.org/ab-testing/t-test.html

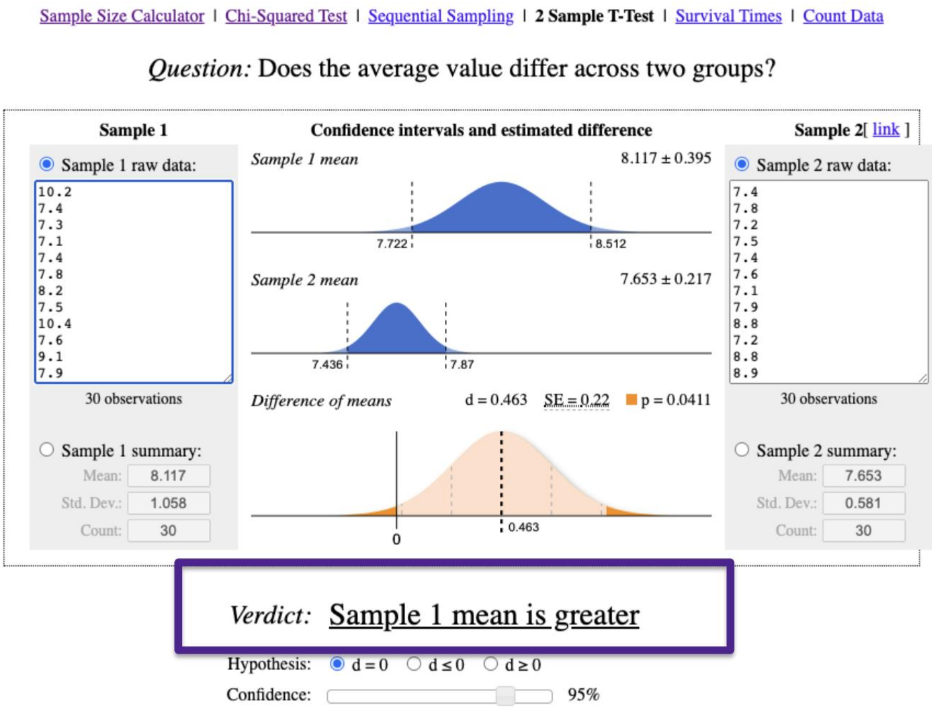
2. https://abtestguide.com/calc/

3. https://www.surveymonkey.ru/mp/ab-testing-significance-calculator/

Пример №1



Пример №2


Sample Size Calculator | Chi-Squared Test | Sequential Sampling | 2 Sample T-Test | Survival Times | Count Data
Question: Does the average value differ across two groups?
Sample 1
Confidence intervals and estimated difference
Sample 2[link]
Sample 1 raw data:
10.2
7.4
7.3
7.1
7.4
7.8
8.2
7.5
10.4
7.6
9.1
7.9
30 observations
Sample 1 summary:
Mean: 8.117
Std. Dev.: 1.058
Count: 30
Sample 1 mean
8.117 ± 0.395
7.722
8.512
Sample 2 mean
7.653 ± 0.217
7.436
7.87
Difference of means
d = 0.463
SE = 0.22
p = 0.0411
0
0.463
Sample 2 raw data:
7.4
7.8
7.2
7.5
7.4
7.6
7.1
7.9
8.8
7.2
8.8
8.9
30 observations
Sample 2 summary:
Mean: 7.653
Std. Dev.: 0.581
Count: 30
Verdict: Sample 1 mean is greater
Hypothesis: d = 0 d ≤ 0 d ≥ 0
Confidence: 95%

Дополнительные материалы

1) Видео про описательную статистику
2) Базовые концепты статистики
3) Хорошее описание ЦПТ
4) Про доверительные интервалы от Демешева
5) Про доверительные интервалы от Кирилла Мильчакова
6) Про MDE и ошибки 1 и 2 типа - вводная статья
7) Про MDE посложнее
8) Книга Голая статистика
9) Книга статистика и котики
10) Книга Statistics in Nutshell

Используемые источники

1) Видео про описательную статистику
2) Базовые концепты статистики
3) Хорошее описание ЦПТ
4) Про доверительные интервалы от Демешева
5) Про доверительные интервалы от Кирилла Мильчакова
6) Про MDE и ошибки 1 и 2 типа - вводная статья
7) Про MDE посложнее
8) Книга Голая статистика
9) Книга статистика и котики
10) Книга Statistics in Nutshell