

Основы анализа и визуализации данных для медиа

Лекция 7
23 сентября 2019

Рогович Татьяна
rogovich@gmail.com

Тестирование
гипотез на разницу
средних.

Структура лекции

- Немного поговорим о том, как сравнивать результаты разных выборок (в том числе и визуально), и делать вывод о том, есть ли разница, например, между их средними.

Тестирование гипотез на разницу средних выборок

Зачем это нужно?

Мы с вами научились находить меры центральной тенденции и меры разброса и говорить. Но как понять, что если у двух выборок, например, средний доход отличается на какую-то сумму, что это действительно статистически значимая разница, а не случайность?

Rule of thumb (очень осторожно и без уверенности)

Если **между средними** двух выборок расстояние больше двух среднеквадратичных отклонений, как для одной, так и для другой выборки, то скорее всего различие статистически значимо.

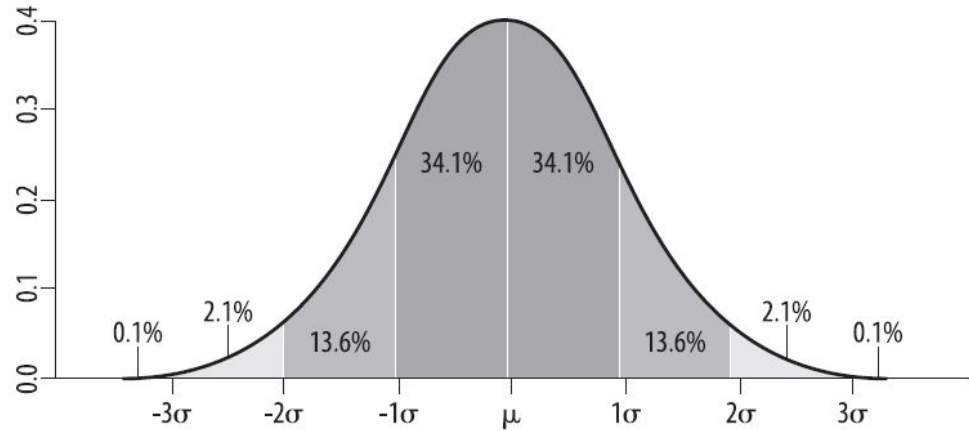
Для интерквартильного размаха и медианы: среднее одной выборки является выбросом по отношению к другой. И наоборот.

А теперь,
как это посчитать

Z-нормальное распределение

Это такое распределение, которое выражается в общих терминах среднего и среднеквадратичного отклонения.

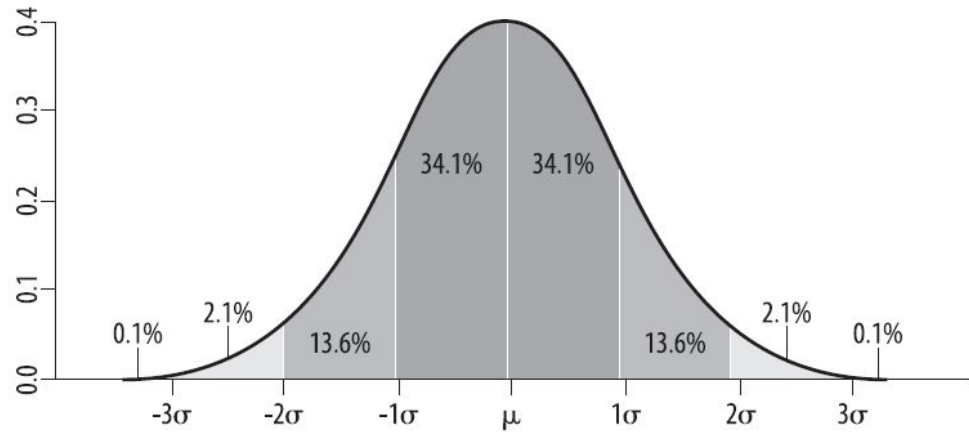
- около 68% данных находятся в интервале $\pm$ одно стандартное отклонение от среднего;
- около 95% данных находятся в интервале $\pm$ два стандартных отклонения от среднего;
- около 99% данных находятся в интервале $\pm$ три стандартных отклонения от среднего.



Z-нормальное распределение

Соответственно, если мы среднее второй выборки окажется по отношению к среднему первой в каком-нибудь мало-вероятном интервале (например, дальше чем 2 или 3 стандартных отклонения), то мы можем сделать вывод, что разница между ними статистически значима.

Чтобы это выяснить, мы считаем z-статистику, если знаем параметры генеральной совокупности, и t-статистику, если знаем только значения выборок.

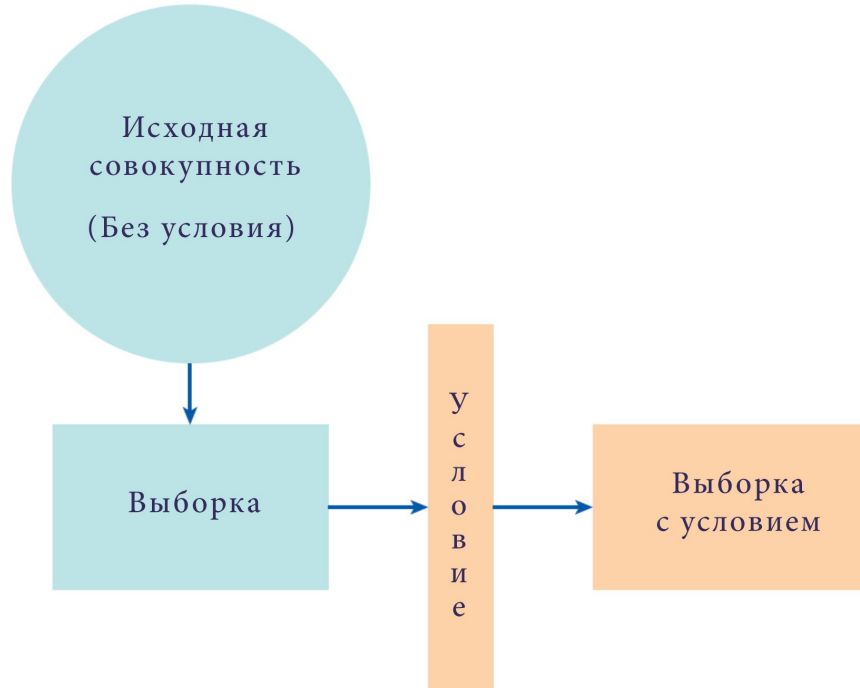


Тестирование гипотезы

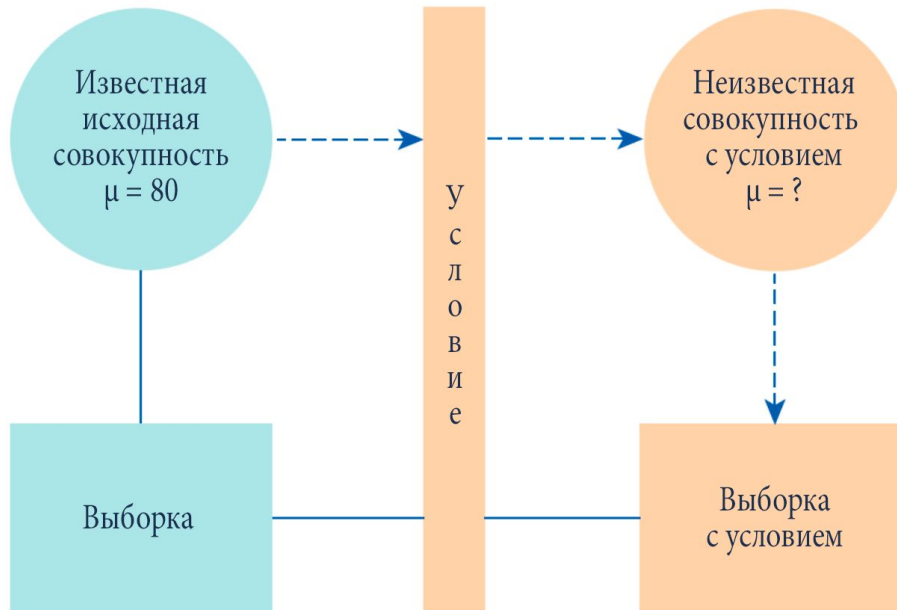
Статистический метод, когда с помощью выборки мы проверяем наши предположения о генеральной совокупности.

Например, что курение влияет на вероятность наличия сердечно-сосудистых заболеваний.

Процесс экстраполяции выводов



Процесс экстраполяции выводов



Тестирование гипотезы: этапы

1. Сформулируем нулевую и альтернативную гипотезы.

Нулевая гипотеза говорит о том, что в генеральной совокупности не будет изменений/нет взаимосвязи/нет разницы.

Альтернативная гипотеза утверждает, что эта разница существует.

Гипотеза о влиянии подготовки к экзамену на оценку

Студенты, которые готовились к экзамену

Студент	Оценка
Студент 1	91
Студент 2	86
Студент 3	90
Студент 4	93
Студент 5	84

Студенты, которые НЕ готовились к экзамену

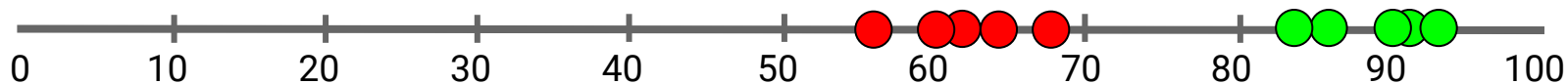
Студент	Оценка
Студент 1	56
Студент 2	62
Студент 3	64
Студент 4	60
Студент 5	68

Гипотеза о влиянии подготовки к экзамену на оценку

Нулевая гипотеза - подготовка студентов не влияет на их оценку на экзамене

Альтернативная гипотеза - подготовка к экзамену положительно влияет на оценку

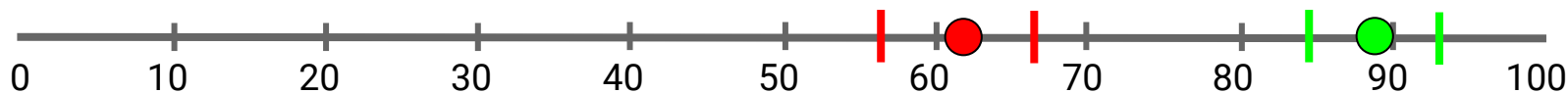
Гипотеза о влиянии подготовки к экзамену на оценку



Первая группа: зеленые

Вторая группа: красные

Гипотеза о влиянии подготовки к экзамену на оценку



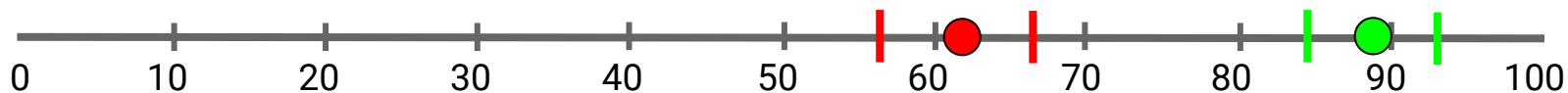
Среднее первой группы:
88.8

Стандартное отклонение:
3.70

Среднее второй группы: **62**

Стандартное отклонение:
4.47

Гипотеза о влиянии подготовки к экзамену на оценку



«На глаз» средние двух групп находятся на расстоянии больше двух стандартных отклонений друг от друга. Можем предположить, что подготовка значимо положительно влияет на оценку.

А теперь протестируем это формально через через t-статистику

Хороший пошаговый разбор по ссылке:

<https://ru.wikihow.com/%D0%BE%D1%86%D0%B5%D0%BD%D0%B8%D1%82%D1%8C-%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D1%83%D1%8E-%D0%B7%D0%BD%D0%B0%D1%87%D0%B8%D0%BC%D0%BE%D1%81%D1%82%D1%8C>

Вычислим дисперсию между двумя группами.

Дисперсия вычисляется по следующей формуле:

$$s_d = \sqrt{((s_1/N_1) + (s_2/N_2))}$$

- s_d — дисперсия между двумя группами.
- s_1 — стандартное отклонение в группе 1, N_1 — объем выборки в группе 1.
- s_2 — стандартное отклонение в группе 2, N_2 — объем выборки в группе 2.

Тест гипотезы через t-статистику

Находим дисперсию:

- $s_d = \sqrt{((s_1)^2/N_1) + ((s_2)^2/N_2))}$
- $s_d = \sqrt{((3.70)^2/5) + ((4.47)^2/5))} = \sqrt{(13.69/5) + (19.98/5)} =$
 $= \sqrt{(2.74 + 3.99)} = \sqrt{6.728} = 2.59$

Тест гипотезы через t -статистику

t -оценка дает возможность перевести данные в форму, которая позволяет сравнить их с другими данными.

t -оценки позволяют проверить t -критерий и найти, насколько вероятности двух групп отличаются друг от друга

По сути t -распределение - это стандартизированное распределение средних выборок, которое позволяет нам сравнить значения средних двух разных групп.

Тест гипотезы через t-статистику

Для вычисления t-оценки используется следующая формула:

$$t = (\mu_1 - \mu_2)/s_d$$

- μ_1 — среднее значение для первой группы.
- μ_2 — среднее значение для второй группы.
- s_d — дисперсия между двумя выборками.

Тест гипотезы через t-статистику

В качестве μ_1 используйте большее среднее значение, чтобы t-оценка не получилась отрицательной.

Находим t-оценку: $t = (\mu_1 - \mu_2)/s_d = (88,8 - 62)/2.59 = 10.34$

Тест гипотезы через t-статистику

Проверять статистическую значимость нашего результата, мы будем по специальной таблице. А ей, кроме параметра t , нужен второй параметр - степень свободы выборки, который зависит от размера (чем выборка больше, тем он будет больше и результат будет более статистически значимым).

Определим степень свободы выборки. В случае t -оценки степень свободы определяется по объему выборки. Сложим объемы двух выборок и вычитаем из суммы 2. В нашем примере степень свободы равна 8, поскольку первая и вторая выборки содержат по пять значений: $(5 + 5) - 2 = 8$

Тест гипотезы через t -статистику

Оценим значимость по таблице значений критерия Стьюдента (t -критерия). Отыщем строку, в которой содержится найденная степень свободы, и определим соответствующее t -оценке P -значение.

Таблица значений критерия Стьюдента

- При степени свободы 8 и t -оценке 10.34 P -значение для одностороннего критерия попадает в промежуток выше 0.0005. Вернее самое большое значение t в этом ряду 5.041, а наше значение выше. Это значит, что наш результат статистически значимый и вероятность, что мы ошиблись меньше сотых долей процента.
- Ввиду этого мы отказываемся от нулевой гипотезы и принимаем альтернативную гипотезу, которая гласит: те ученики, которые готовились к экзаменам, получают более высокие оценки.

Для тех, кто хочет
узнать про это все
побольше
(на контрольной это не будет)

Распределение средних значений выборок

Теорема центрального предела:

Для любой генеральной совокупности со средним значением μ и стандартным отклонением σ , распределение средних значений для выборок размера n будет иметь среднее μ и стандартное отклонение $(\sigma/\sqrt{n})$ и будет приближаться к нормальному распределению, как n приближается к бесконечности.

Пример: совокупность из 4 наблюдений

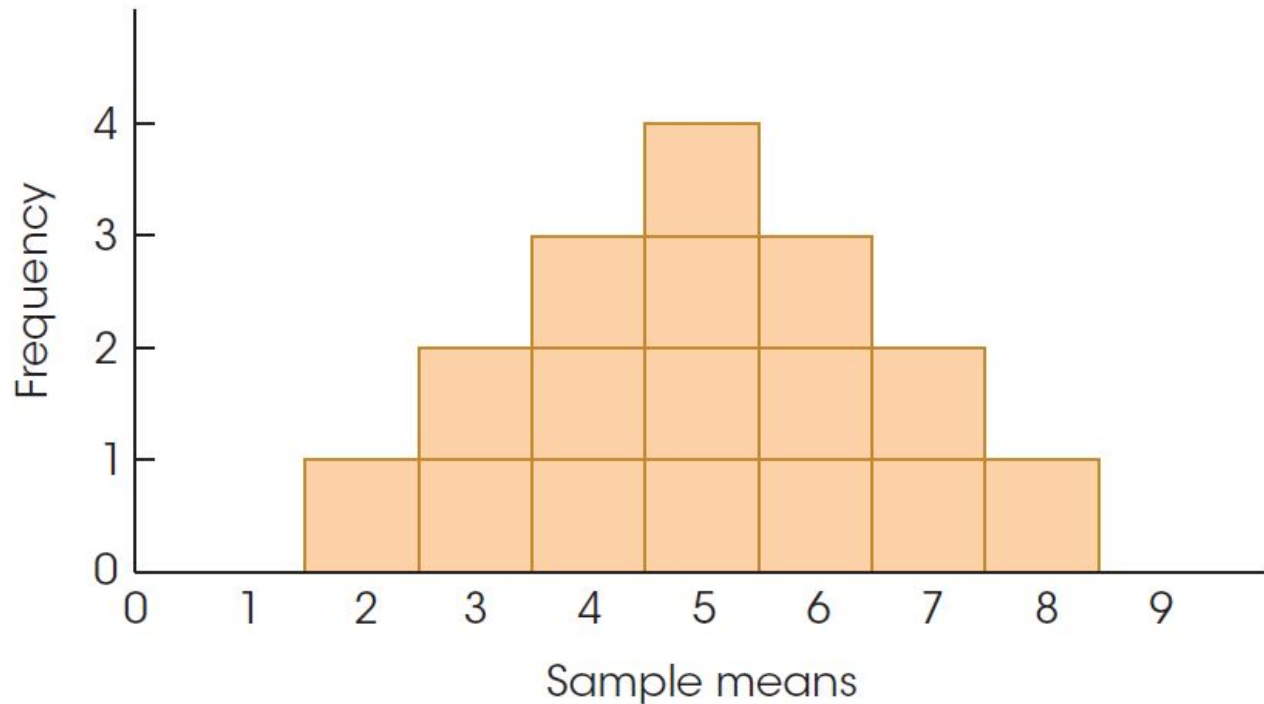


Пример: составим все возможные выборки размера 2

Sample	Scores		Sample Mean (<i>M</i>)
	First	Second	
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

Все
возможные
выборки
 $n = 2$

Пример: распределение значений наших 16 выборок



Когда это работает?

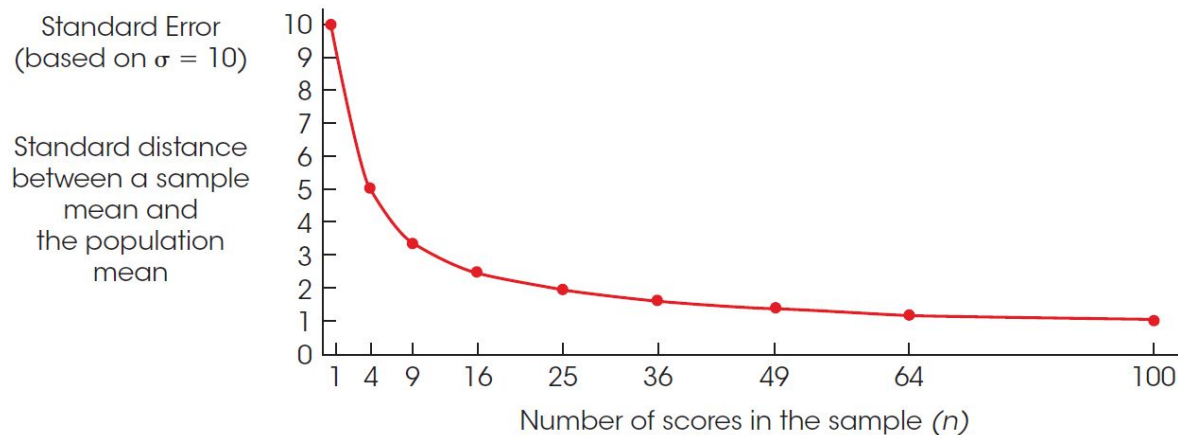
Распределение средних значений выборок будет нормальным

1. Если у генеральной совокупности нормальное распределение.
2. Если количество наблюдений в выборке 30 или больше (независимо от изначальной формы распределения генеральной совокупности).

Среднее такой выборки равно среднему генеральной совокупности - ожидаемое среднее значение.

Закон больших чисел

Чем больше выборка, тем ближе ее среднее будет к среднему генеральной совокупности. А значит, если среднее значение, какой-то из выборок отличается от среднего нашей первой выборки, то мы предполагаем, что она пришла из другой генеральной совокупности с другим средним значением.



Тестирование гипотезы: этапы

1. Сформулируем нулевую и альтернативную гипотезы.

Нулевая гипотеза говорит о том, что в генеральной совокупности не будет изменений/нет взаимосвязи/нет разницы.

Альтернативная гипотеза утверждает, что эта разница существует.

Тестирование гипотезы: этапы

2.

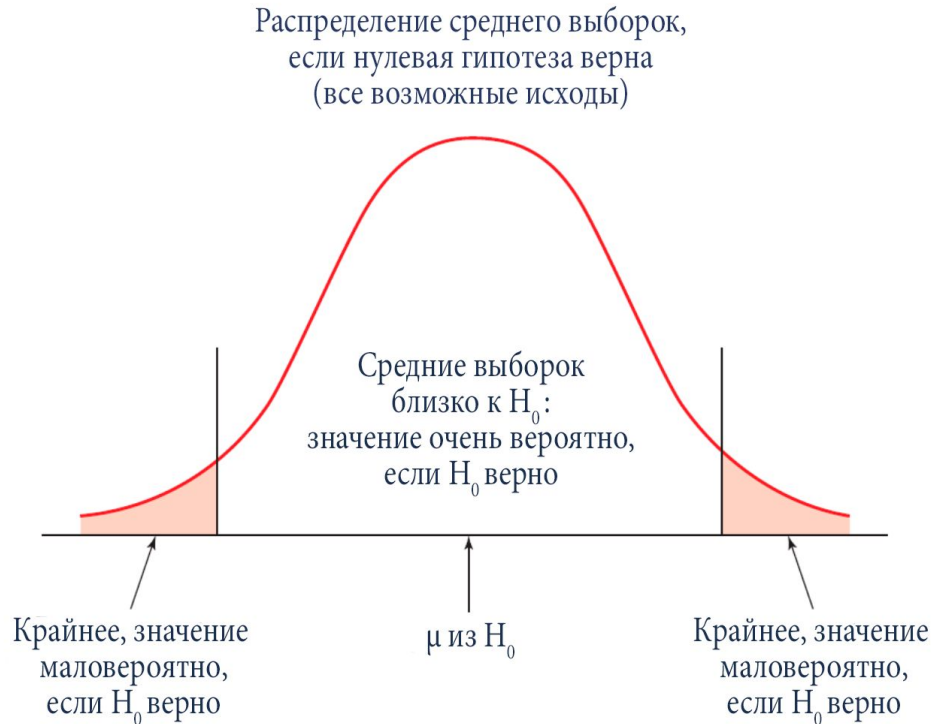
Объявляем критерий проверки гипотезы

Мы делим распределение средних выборок на две части - те, которые имеют вероятность быть выбранными, если нулевая гипотеза правильная, и те, которые имеют очень маленькую вероятность быть выбранными, если нулевая гипотеза корректная.

Уровень значимости (α -уровень) - значение вероятности, который “отсекает” те значения, которые маловероятны для нулевой гипотезы.

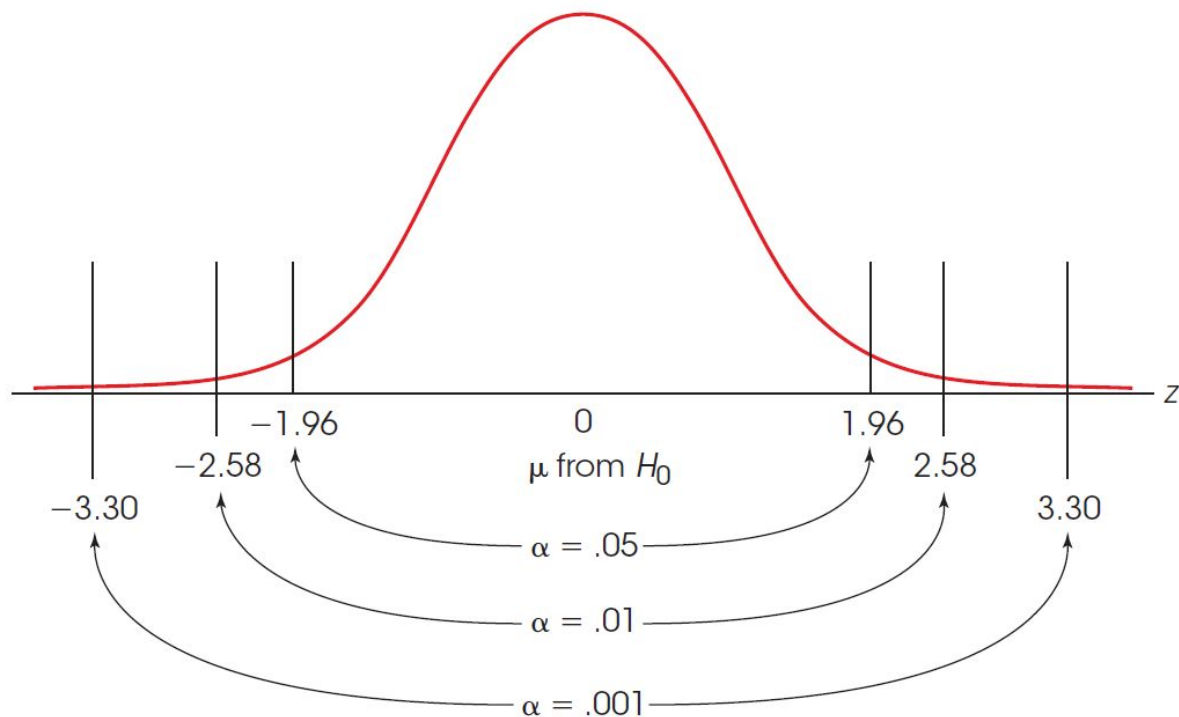
Традиционно: 0.05, 0.01, 0.001

Тестирование гипотез



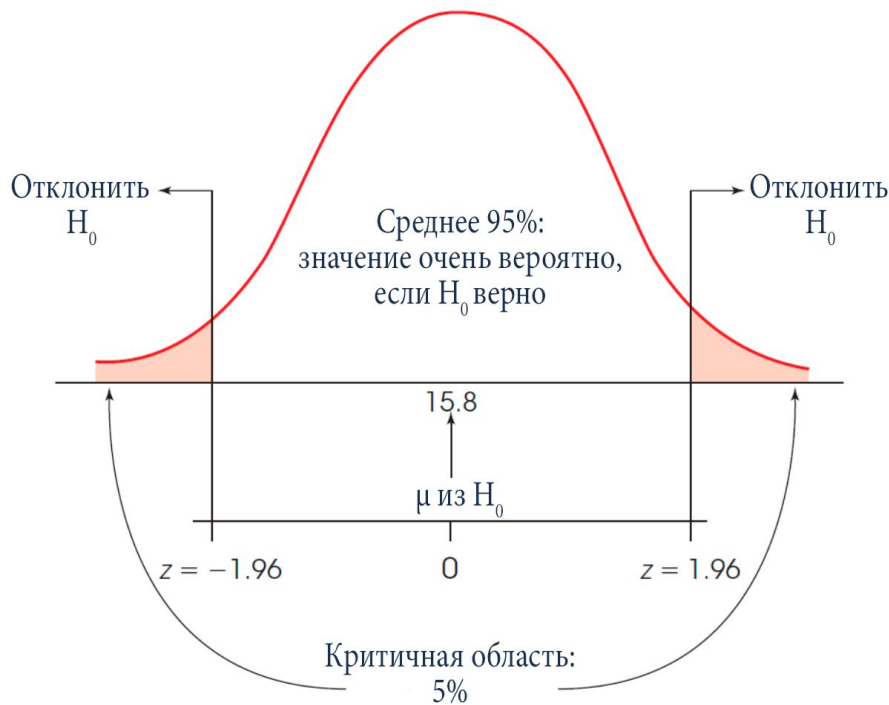
Если мы хотим отвергнуть нулевую гипотезу, то мы хотим, чтобы среднее второй выборки оказалось таким значением, которое было бы маловероятно, если бы считали, что разницы быть не должно (нулевая гипотеза).

Выбор уровня значимости



Уровень значимости (α -уровень) говорит нам в каком регионе должно находиться среднее второй выборки, чтобы мы считали, что между ними есть статистическая разница. Чем альфа меньше, тем значимость больше. Здесь пример для z-статистики.

Проверка гипотезы z-статистикой.



Находим z-значение для среднего выборки ("генеральной совокупности" после вмешательства), где в знаменателе стандартное отклонение средних значений выборок от генеральной совокупности. Т-статистика работает очень похоже, только учитывает тот факт, что мы не знаем среднего генеральной совокупности.

$$z = \frac{M - \mu}{\sigma_M} \quad \sigma_M = \frac{\sigma}{\sqrt{n}}$$

Ошибки при тестировании гипотез

Гипотеза H_0	Принимается	Отвергается
Верна	Правильное решение	Ошибка 1-го рода
Неверна	Ошибка 2-го рода	Правильное решение

Уровень значимости (α -уровень) - вероятность совершить ошибку первого рода.

Где узнать про
это подробнее?

<https://www.coursera.org/lecture/znakomstvo-r-bazovaya-statistika/tyestirovaniye-ghipotiez-4akxk>

<https://www.amazon.com/Essentials-Statistics-Behavioral-Frederick-Gravetter/dp/1133956572>