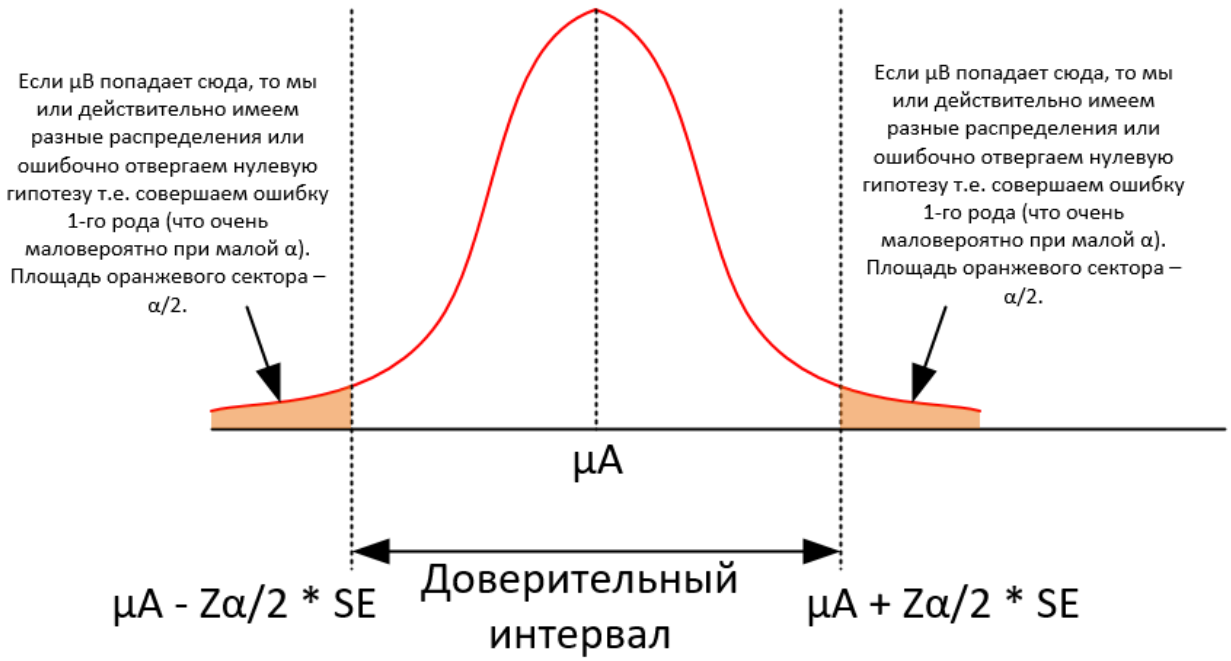
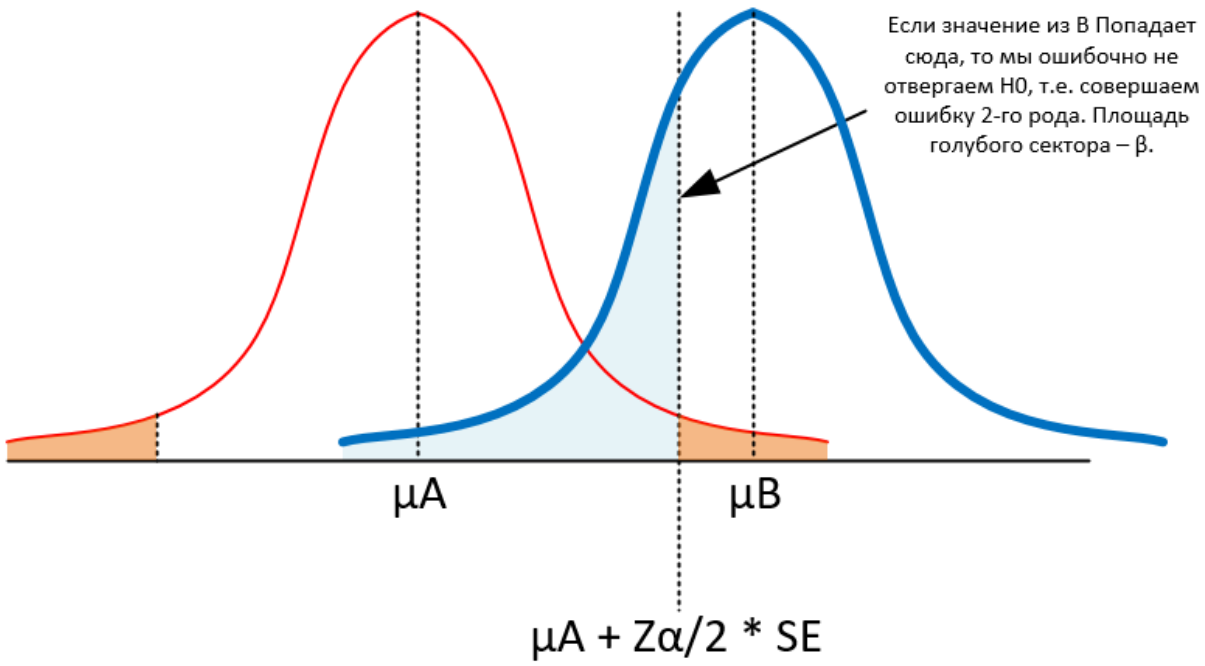
Расчет длительности AB-тестов

Сначала вспомним, что такое ошибки первого и второго рода. Пусть у нас есть распределение выборочного среднего для группы A. Согласно ЦПТ, оно будет распределено нормально со средним μA и некоторым стандартным отклонением SE (standard error - стандартное отклонение выборочного среднего):



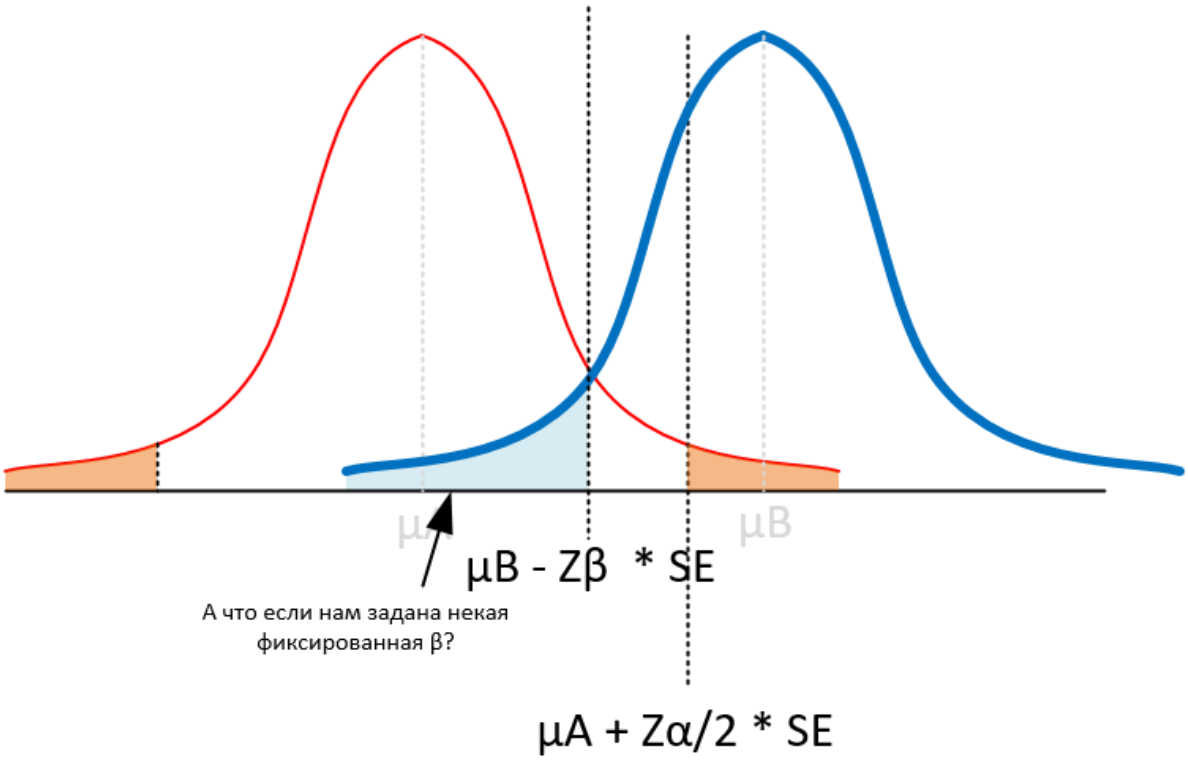
Z это Z-распределение – некоторая табличная функция, которая позволяет нам «реконструировать» нормальное распределение. Ее значения при умножении на SE дают нам границы доверительного интервала. Это интервал, в который значение выборочного среднего попадает с вероятностью 1- α (чаще всего 95% или 99%).

Добавим группу B:



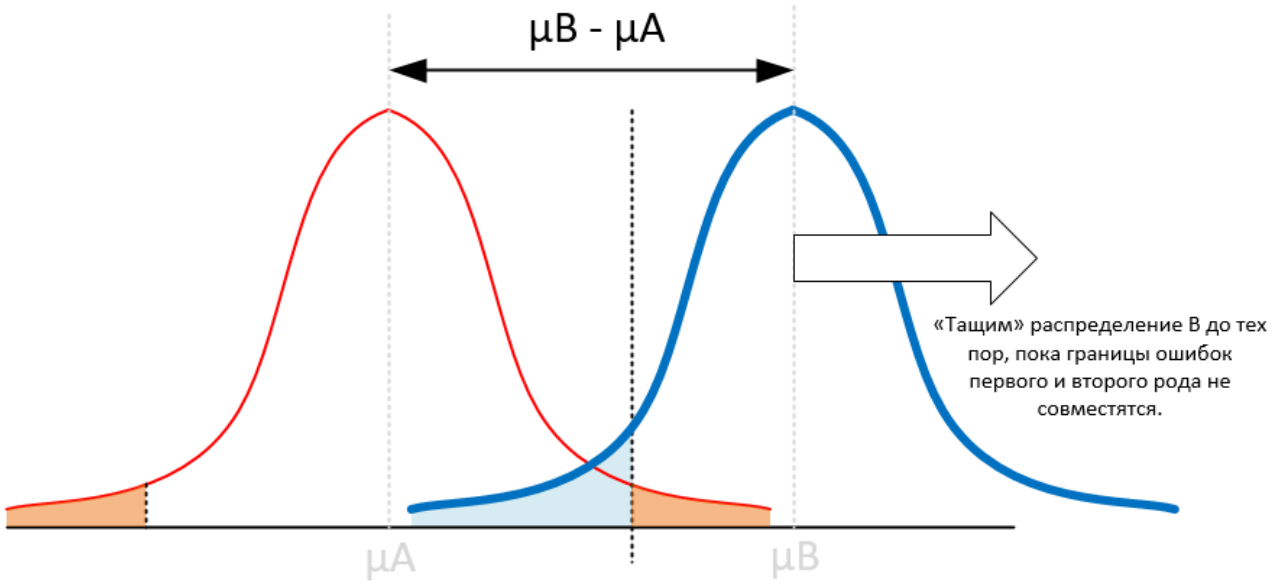
В этом случае параметр β как бы зависит от α.

Теперь посмотрим ситуацию, когда нам жестко задан β. Допустим, заказчик требует, чтобы ошибка 2-го рода не превышала 0.2 (т.е. мощность 0.8):



Границы секторов ошибок не совпадают (чего не может быть). Это нужно как-то решать.

Самое простое решение - «перетащить» распределение B вправо, чтоб границы секторов ошибок совпали:



$$\mu A + Z\alpha/2 * SE = \mu B - Z\beta * SE$$

Тогда, у нас получается уравнение:

$$\mu_A + z_{\alpha/2}SE = \mu_B - z_\beta SE$$

В этом случае полагаем, что SE равны т.к. природа поведения групп A и B одна и та же, меняется только среднее, а не дисперсия. Немножко переформатируем:

$$\mu_B - \mu_A = (z_{\alpha/2} + z_\beta)SE$$

Левая часть это засекаемый эффект – разница между выборочными средними в группах. Назовем его d:

$$d = (z_{\alpha/2} + z_\beta)SE$$

Существует теорема, доказывающая, что (считаем размеры и дисперсии в группах идентичными):

$$SE = \sqrt{\frac{{\sigma_A}^2}{n_A} + \frac{{\sigma_B}^2}{n_B}} = \sqrt{\frac{2\sigma^2}{n}} = \sigma\frac{\sqrt{2}}{\sqrt{n}}$$

Доказательство:

http://onlinestatbook.com/2/sampling_distributions/samplingdist_diff_means.html

https://web.stanford.edu/~kcobb/hrp259/lecture11.ppt

Здесь сигма, это оценка стандартного отклонения параметра, для которого мы оцениваем разницу средних. Т.е. если вы оцениваете разницу средних для индивидуального чека, сигма

– это дисперсия индивидуального чека, а не дисперсия выборочного среднего. n – объем выборки (количество наблюдений, количество пользователей в тесте и т.д.).

Итак, у нас есть выражение:

$$d = (z_{\alpha/2} + z_{\beta})\sigma\frac{\sqrt{2}}{\sqrt{n}}$$

Если мы решим его относительно n:

$$n = 2\left(\frac{(z_{\alpha/2} + z_{\beta})\,\sigma}{d}\right)^2$$

В результате, у нас есть формула, которая позволяет нам найти размер выборки, зная заданные уровни ошибок первого и второго рода (они управляются соответствующими значениями z из таблиц), имея заданный уровень различия между средними в группах (d) и имея выборочную дисперсию для изучаемого параметра (ее м обычно получаем из исторических наблюдений). Вот тут можно посмотреть, как эти параметры влияют друг на друга и на размер выборки:

https://rpsychologist.com/d3/NHST/

Функцию расчета довольно просто написать в Питоне:

```
def getSampleSize(mean, var, relativePracticalSignificance, alpha, power):

    z = norm.ppf(power) + norm.ppf(1 - alpha / 2)
    absolutePracticalSignificance = mean * relativePracticalSignificance

    sigma = np.sqrt(2 * var)

    return math.ceil((sigma * z / absolutePracticalSignificance) ** 2)
```

Здесь:

- mean и var – среднее и дисперсия интересующей нас величины, полученные из исторических наблюдений;
- relativePracticalSignificance – относительная разница между средним в группе A и B в процентах. Например, ожидаем, что среднее в группе B будет отличаться от A не менее чем на 5%, тогда relativePracticalSignificance = 0.05;
- alpha, power – уровень значимости и мощность.

Теперь нам нужно понять, какие распределения мы можем тестировать т.к. для разных видов распределений среднее и дисперсия получаются по-разному:

Измеряемый параметр	Конверсии (в покупку, в регистрацию, в подписку, в прохождение туториала и т.д.)	Среднее (средний чек, средняя длинна сессии и т.д.)
Распределение	Бернулли (где p - вероятность конверсии)	Нормальное

Среднее (mean)	p	Выборочное среднее
Дисперсия (var)	p (1-p)	Выборочная дисперсия

Итого, получается вот такой алгоритм определения длительности теста:

1. Задаем нужные уровень значимости и мощность (alpha, power);
2. Задаем минимальный обнаруживаемый эффект, он же относительная практическая значимость;
3. По нужному параметру берем выборку исторических данных, находим среднее и дисперсию (в соотв. с тем, какое распределение исследуем);
4. Считаем потребное количество наблюдений n;
5. Умножаем n на количество групп в тесте;
6. Из исторических данных находим, сколько пользователей мы можем лить в тест в единицу времени, находим длительность теста:

   Длительность = (n * число_групп) / юзеры_в_единицу_времени

Для проверки:

- Калькулятор длительности для конверсий: http://www.evanmiller.org/ab-testing/sample-size.html
- Калькулятор длительности для средних: https://select-statistics.co.uk/calculators/sample-size-calculator-two-means/