AB testing has many different application including optimizing the website or mobile app as well as determining the probability threshold for a second-time conversion rate for example. We can measure after building churn prediction what is the best probability threshold to encourage the user to contact with the business. AB testing cannot tell you if you miss something. It compares old to new design / method / style.

In this project due to data availability I will concentrate on introducing new feature on the website. You will exposure new and existing user to 2 different versions old, called control and the new method, called test group and will measure response from the user. In the current version we have gate on level 30 and we want to test if this gate would be good to put on level 40 or keep in the old place.

Important is to check if the number of users in control and each of the tests groups is similar. In our case it is but if the number of user in each of the cases is different it is better to convert into probabilities. For example CTR = number of clicks / number of page views or similarly we could use number of retention / number of users in the group. Since the CTR may count users more than once depending on how many times they visit the website we may want to use click through probability, rather than CTR, defined by unique visitors who clicked / unique visitors to page.

1. Formulating hypothesis.

Hypothesis is made of two words: hypo – under / less than and thesis – refer to place / generally held view. It is effective way to use this fact when writing hypothesis. I will refer to this way of writing the hypothesis:

**Writing the “If” section of your Hypothesis**1. Start your sentence with the word “If”  
2. Write down one of the variables  
3. Connect statement with one of the following:  
\* is related to  
\* is affected by  
\* causes  
**Write down other variables**  
**Writing the “then” section of your Hypothesis  
1.** Make a comment on the relationship between those two variables.

**Primary hypothesis** - main hypothesis of our interest

H0: If sum\_gamerounds is related to version then by placing the gate at level 40 we will change the average sum\_gamerounds.

HA: If sum\_gamerounds is not related to version then by placing the gate at level 40 we will not change the average sum\_gamerounds.

**Secondary hypothesis** – this can be related to the variable from primary hypothesis. For example with the increase of sum game rounds that player played can decrease the retention date if we will take on account time that user needs to play for completing each round.

H0: If retention rate is affected by version of game where gate is placed on different stage, then by placing the gate at level 40 we will change the retention rate.

HA: If retention rate is not affected by the version of game then placing the gate at level 40 will not change the retention rate.

**Statistics behind AB testing.**

The distribution of AB test can be seen as Binomial – success or failure of the experiment. Either retention rate increased or not, either number of games played increased or not.

Mean = µ = and SEmean = We can use binomial distribution if: there are two outcomes success or failure, Ho or HA, the events are independent and the events follow an identical distribution <probability of success needs to be identical for all of them>

First we calculate statistics in the control group as that’s what we know. In our example we can see that probability of retention day 1 in the control group is 0.448 while retention day 7 is 0.190 . 99% confidence interval means that if we would theoretically repeat the experiment over and over again we would expect our interval around the sample mean to cover true value of population 99% of time. Use normality assumption for calculating confidence interval for retention rate at control group I need to use the rule of thumb that if N\*p > 5 I can use normal distribution In our case is much more than that. This means that if you run experiment again with similar number of players the number of retention in day 1 would be between 19,763 and 20,305 players and in day 7 between 8,288 and 8,716 players.

Important is to check if the probability is due to chance only or because of the treatment.

H0: P(due to chance) pcont = pexp => pcont-pexp = 0

HA: pexp – pcont

We chose α = 0.01

We need to first calculate pooled probability of the click.

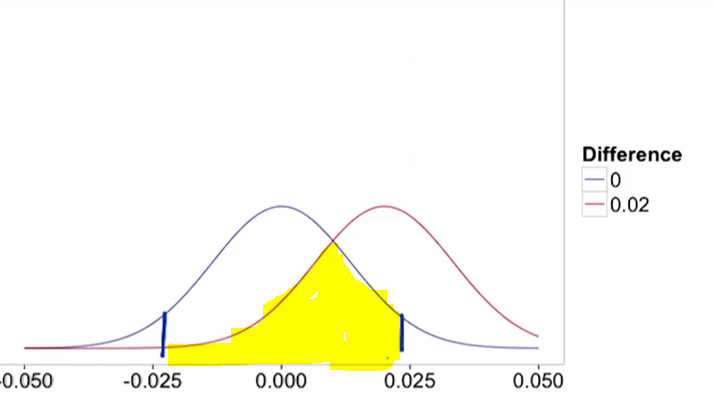
We would need then to decide what change is important for us. What essential in addition to statistically significant. You need to take in consideration other things before implementing the change such as business investment and time efficiency.

We need to pick up practical significant boundary. Statistical significant is about repeatability. You want to get this quarantee that yes the measure is repeatable but also you want to be sure that if you see difference between business stand point so it’s practically significant it’s also statistically significant. To pick up practical significance let’s first pick up boundaries. Let say that from business perspective x% of change would be practically significant. Before running the experiment you need to decide how big the sample size you want it to be. How big the control and test group should be.

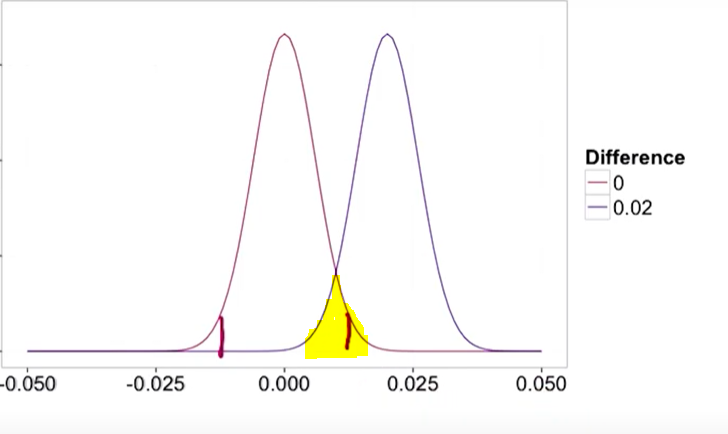
No true difference between two groups. You reject the null and conclude that there was true difference

Alpha = P(falsely conluding there is a difference) = P(reject null | null true) = 0.05

If you would increase sample your standard error would decrease so the distribution around the mean would be more narrower. Consider that there is true difference of the mean. The difference is equal to practical significant of 0.02 You fail to reject the null hypothesis and you conclude there is not statistically significant between two cases. Beta = P(fail to reject | null false) is pretty high (yellow shaded area)

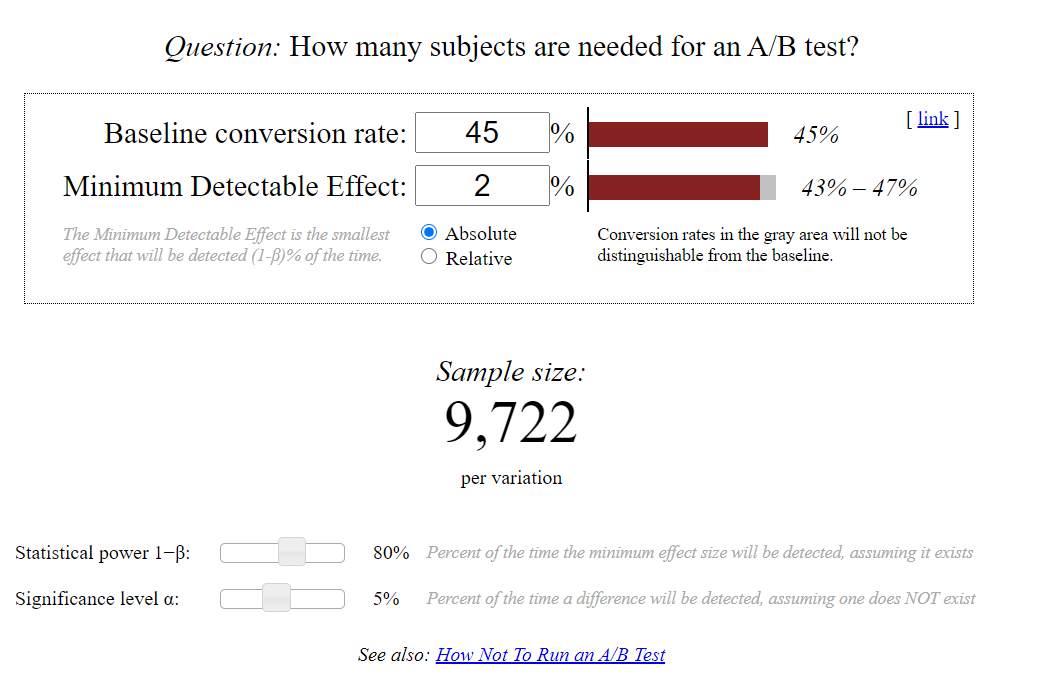


Then you would fail to reject the experiment that actually did have the difference you care about. Risk of small sample is that alpha is low and beta is high. Beta depends on how big your effect really was As your true change grows larger and larger then your beta would go down. You consider beta at your practical significant boundary as you don’t care about any smaller changes and any larger changes will have lower beta that means lower chance of error. With larger sample both distributions got tighter alpha doesn’t change but you are less more likely to reject the null and commit beta error



Baseline conversion rate is the estimated retention probability before making any change. In our case it’s 45%

Minimum detectable effect is the practical significance that we talked about



With increase of retention probability but still less than 50% you would need to increase the sample size. The standard error depends on the probability it’s proportional to so I will need to increase the sample size to decrease the standard error

If you decide to increase your practical significance level you can decrease the sample size. If you increase your practical significance level you look for larger change to detect. Larger changes are easier to detect so you won’t need so big sample size

If you would consider increase your confidence level you would need to increase the sample size. You’re saying that you want to be more certain that the change occurred before you reject the null. If you want to keep your sensitivity the same you would need to increase the sample size.

In our case we would conclude that this is highly probable that the change won't be bigger than 2% so we would not implement the change.

This case is called neutral. You are confident that the result is not different than 0 since the confidence interval contains 0 and you are confident that there is no practically significant change. 