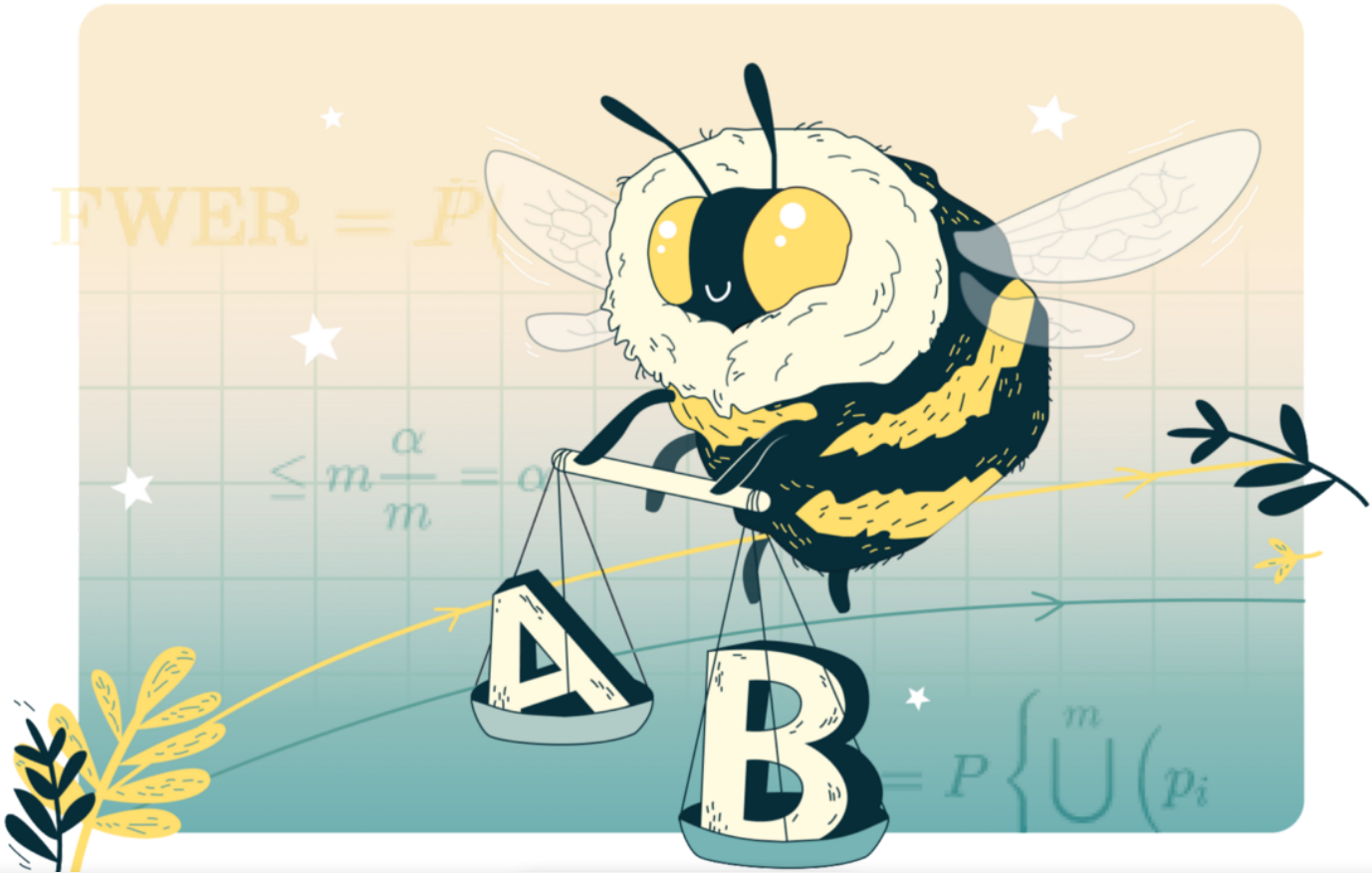
Published in Towards Data Science

Irene P Follow

Oct 16, 2020 · 7 min read · Listen

A/B testing statistics: true and estimated value of conversion rate

How to ensure your split test's result is statistically significant (in simple words)



— I can tell you.
— I can tell too. How do I check?

The A/B test is one of the most popular user experience research methodology for evaluating the effectiveness of UX or UI changes on your website or mobile app. So say you changed something, started the test — and it seems the modified version wins. But can we be sure of that? And what percent out of 100 we sure?

To answer let's look at this method A/B test not as a way to measure user engagement and satisfaction of the changes, but as a statistical experiment. So, what do you need to know to say for sure that we can go with the results obtained?

A/B test result doesn't show conversion rate

I mean it doesn't show the actual value of CR, but estimated. No matter how many users you have in the experiment, the sample size is still less than the total number of users. Therefore, when you determine the required sample size for the test, the next logical question is "how many users do we need to be sure the test result (estimated CR value) really describes the actual state of things (true CR value)?

In other words, how close is the **estimated CR (T)** to the **true value (Q)**. And to answer this question, we need confidence intervals.

What is the confidence interval?

So, we ran the A/B test and now have some estimated conversion (let's designate it with T). But in fact, the actual conversion rate may be slightly less, or slightly more, and we don't know an exact value, but can calculate an interval from [a;b] which contains it. This is called the confidence interval.

The larger the sample, the smaller the interval. That is why it's so important to determine the minimum number of users in each group of the experiment before




Here's what: if we run the test 100 times on 100 different samples, get 100 estimated CRs and calculate the confidence interval for each of them, then 95 times out of 100 the interval contains true CR value.

It's important to understand that the confidence level is a characteristic of the confidence interval calculation procedure, not the interval itself. And we can choose which confidence level is enough for us to trust the test result.

> So, a confidence interval is a way of CR estimating, which results in not just a single value, but an interval of values that can contain the actual CR value. And the confidence level shows how likely it's to be contained there.

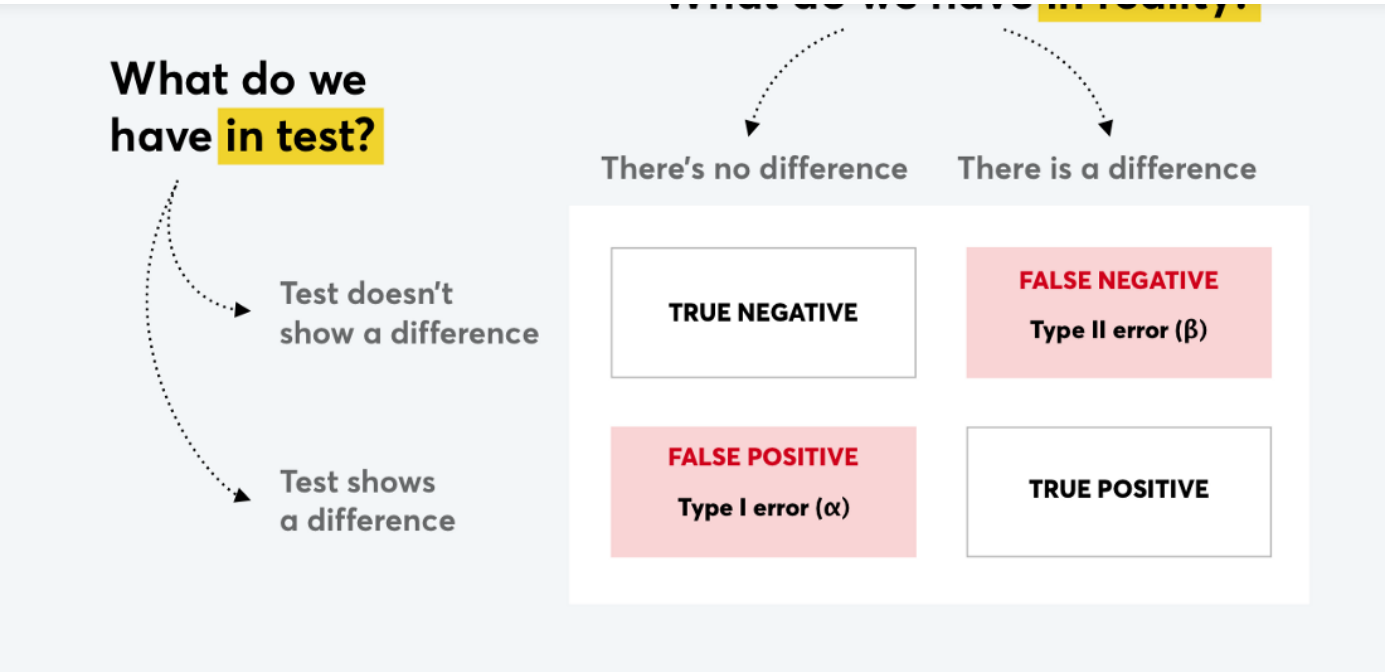
Where did the 95% come from?

I hope I have succeeded in conveying the meaning of the concept of confidence level, so now let's see why it is usually 95%, and not 75% or 80%?

And here I have to introduce two more terms from mathematical statistics (the last ones for today, I promise!) — significance level and statistical power.

But first, let's have a look at the pic below. This is a statistical test matrix (a/b test is a statistical test), and it shows 4 scenarios:



Statistic test matrix: 4 cases

Let's look closer.

Say, we had a red button on the site, and we decided that it should be repainted in blue. We hope that the blue button will be clicked more often, so we run an a/b test to check this hypothesis. What result can we get?

- **true-negative:** the test shows that both buttons have the same CR, and this is true — there's no difference for our users.
- **false-negative:** the test shows that both buttons have the same CR, but this isn't true: in fact, users click the blue (or red, no matter) button more often. The test is wrong.
- **true-positive:** the test shows that one button is better than the other, and it's true.
- **false-positive:** the test shows that users like one and hate another button, but in fact both buttons have the same CR. Misleading findings, the test is wrong.

So, in 2 cases the result is true, and in 2 other cases test distorts the real picture.

difference while the variants are the same.

What do such misleading findings mean for us? Well, apparently, we're happy, we believe the test, we spend time and money to roll out the changes and introduce it to all users, and — no effect, the CR is the same which is only to be expected considering there was no difference from the very start. Whether you are ready for this in 5% of cases or in 1% — it's up to you :) Usually, people agree on α=5%, so the confidence level, which is calculated as a 1-α, is 95%.

> Thus, the confidence level (1-α) is the probability of NOT getting a Type I error (α). It's the percentage of confidence in the result if the test shows that variants A and B do not differ.

What if I want 99% instead of 95%?

Feel free. But you need more people in the experiment then, or a greater difference between the winner and the runner-up conversion rates. This can affect the time of the experiment, so start from how important it is for you to have such high confidence in the result.

Statistical power

However, another erroneous scenario is possible — when variants A and B have a true difference in its conversion rates, but the test failed to detect it. It's called a false negative result or a Type II error (denoted by β).

What are we risking here? Missing a good idea and not implementing it. This is painful, but slightly less than with the Type I error, so here we can set probability of mistake β=20% or less. Then statistical power, which is calculated as a 1-β, is 80% or more.

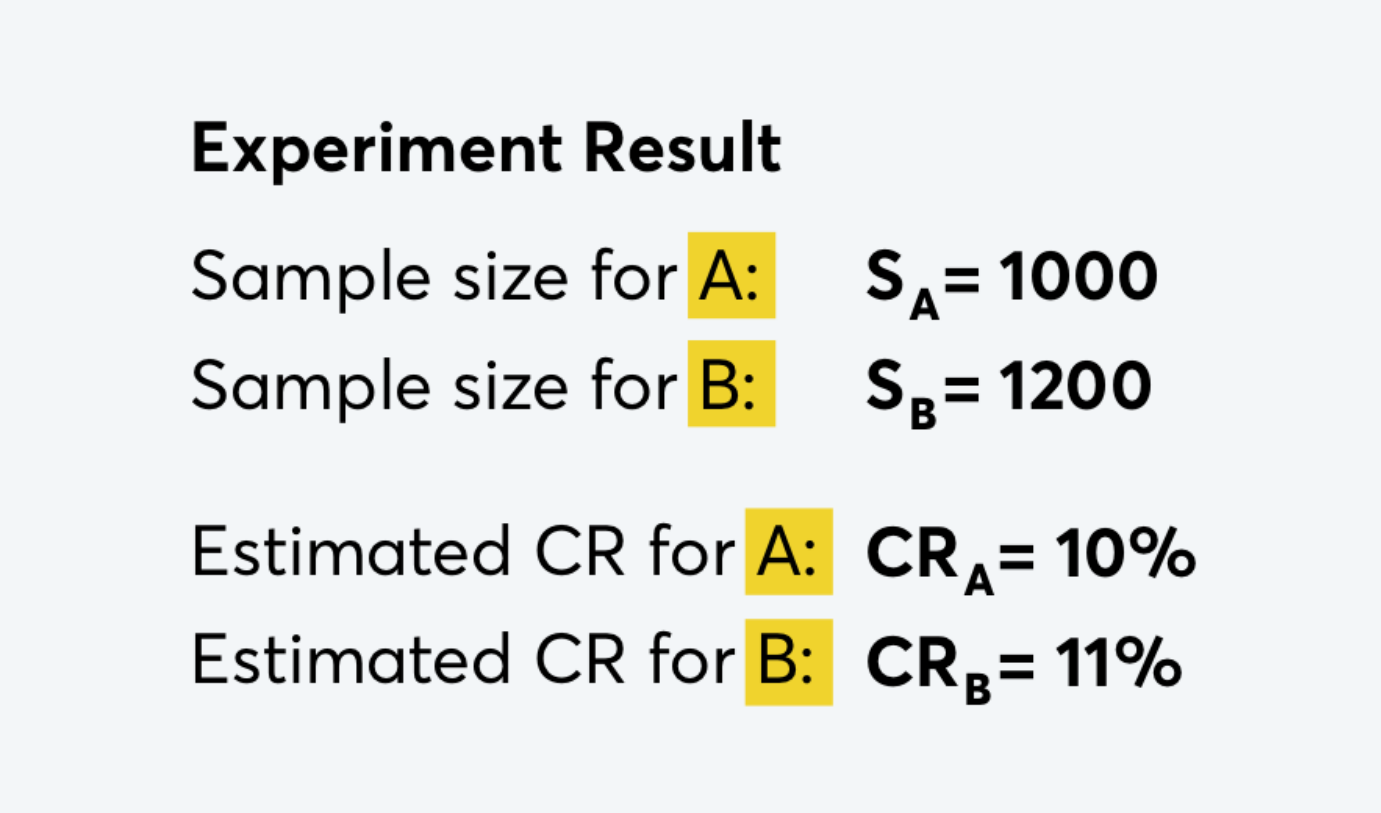



> Type II error (β), i.e. the percentage of confidence in the result if the test shows that there is a difference between variants A and B.

Back to confidence interval

Ok, now we're ready to calculate confidence interval, and here is how to do it:

1. Collect the data:



Experiment results: sample sizes and conversion rates

2. Calculate the difference between variants B and A conversion rates (it can be negative number, if the A variant is winner):

$$\text{CR uplift: } CR_U = CR_B - CR_A = 1\%$$

CR difference

3. Calculate the confidence interval for the difference:

Confidence interval for the difference between CR_A and CR_B:

$$[\,x\,;\,y\,]$$

If the interval doesn't contain 0 (zero), test result is valid.

$$x = CR_U - 1{,}96 \sqrt{\frac{CR_A * (1 - CR_A)}{S_A} + \frac{CR_B * (1 - CR_B)}{S_B}}$$

Z-score for confidence level = 95%

$$y = CR_U + 1{,}96 \sqrt{\frac{CR_A * (1 - CR_A)}{S_A} + \frac{CR_B * (1 - CR_B)}{S_B}}$$

Confidence interval for CR difference. If it doesn't contain zero — it's ok. If contains — you need more traffic or greater difference between conversion rates of your variants.

What conclusions can be drawn?

- as a result of A/B test, we get not a true values, but estimated values of CR for each

- after calculating the confidence interval for the difference in conversions between options A and B, we could conclude whether there is a true difference or not
- we can't be 100% sure in this conclusion because of two types of errors: Type I (misleading findings) and Type II (failing to detect a true difference in CR)
- however, though there's no 100% confidence, it can be 95% or even 99% if we decide so when choosing confidence level (usually still 95%, but can be greater i.e. for A/B/C test)
- if you want more confidence, you need more traffic or greater difference between the winner and the runner-up conversion rates

In closing

1. All of the above works only if we consider our sample distribution to be normal (it can be Gaussian distribution, Bernoulli, Pearson, etc.), which means that we have evenly distributed probabilities, i.e. there is a main mass (by some parameter), and there are minor deviations in both directions.
2. If you run multivariate test or A/B/n test (compare n different variants), you should use corrections (e.g. **Bonferroni correction** to keep confidence level at 95%. However, it might require you to run the test for an unreasonably long period of time.
3. A/B test may show wrong results not only due to statistical errors, but to technical errors also. Therefore, before you run the A/B test, don't forget to check the test settings by running the A/A test: let both traffic groups, A and B, be assigned to the same experience. If you make a mistake in your tool's settings, you'll see the difference in conversions rates (which can't actually be).