**2.8 AB Testing Math**

1. 1. The form of the cut-off would be *‘Greater than’*.

Since the theoretical non-null value under is greater than the null value under , we will reject if the test statistic is greater than the cut-off value.

* 2. + According to the equation, remains constant when remains constant. Therefore, as increases, the cut-off value decreases.

As increases, the distribution curves for become thinner. While the area under the null hypothesis curve and right to the cut-off value () remains the same, the cut-off value would shift left.

* + - According to the equation, as increases, decreases. While remains constant, the cut-off value decreases.

As increases, the area under the null hypothesis curve and right to the cut-off value increases. While the distribution curves for remains the same, the cut-off value would shift left.

* + - According to the equation, as increases, decreases. While the cut-off value remains constant, decreases.

As increases, the area under the null hypothesis curve and right to the cut-off value increases. While the cut-off value remains the same, the distribution curves for would become wider, which implies a decrease in .

* 2. + According to the equation, remains constant when remains constant. Therefore, as increases, the cut-off value decreases.

As increases, the distribution curves for become thinner. While the area under the null hypothesis curve and right to the cut-off value () remains the same, the cut-off value would shift left.

* + - According to the equation, increases as increases. While and remain the same, the cut-off value increases.

As increases, the area under the alternative hypothesis curve and left to the cut-off value increases. While and the effect size remain constant, the shape and position of alternative hypothesis curve remains the same. Therefore, the cut-off value would shift right.

* + - According to the equation, increases as increases. While cut-off value and remain the same, increases when and decreases when . Often the case, and decreases as increases.

As increases, the area under the alternative hypothesis curve and left to the cut-off value increases. While the cut-off value and the effect size remain constant, the position of the alternative hypothesis curve and the distance between its center and cut-off value remain the same. Therefore, the alternative hypothesis curve becomes wider, which implies an increase in if the cut-off value is to the right of and a decrease in if the cut-off value is to the left of .

* 1. Subject: Email campaign test results on conversion

Hi Nick,

We analyzed the information on users converting from a set of two email campaigns and found that the group of users who received *Offer A* has a significantly higher rate of conversion than the group who received *offer B*. The conclusion was drawn from a two-tailed proportion test at significance level of 0.10.

Here is a summary of the data and the test statistics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | # success | # total | proportion | test statistic\* | p-value |
| Offer A | 19 | 126 | 15.08% | 1.9192 | 0.0275 |
| Offer B | 9 | 121 | 7.44% |

\* test statistic is computed as

Best,

XXX

* 1. Subject: Email campaign test results on revenue

Hi Nick,

In addition to the previous analysis on conversion rates, we further investigated the effects of the ad campaigns on revenue. Specifically, we performed two hypothesis tests to see if the revenue generated by the paying customers as well as by all users from each sample group is statistically different.

Both tests show that the revenue generated by the group of users who received *Offer A* doesn’t differ from the group who received *offer B*. The conclusion was drawn from the two-tailed test of means with sample variance at significance level of 0.10.

Here is a summary of the data and the test statistics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | # users | mean rev | std rev | test statistic\*\* | degrees of freedom | p-value |
| paying users | Offer A | 19 | 42.39 | 14.39 | 0.6122 | 8 | 0.2787 |
|  | Offer B | 9 | 48.30 | 27.22 |
| all users | Offer A | 126 | 6.39 | 16.18 | 1.4316 | 120 | 0.0774 |
|  | Offer B | 121 | 3.59 | 14.54 |

\*\* test statistic is computed as

Best,

XXX

1. 1. Assume that the amount of time each user from sample #1 and #2 spent on our web page respectively. We will test on the following hypothesis:

The test statistic is given by:

At significance level .05, we have:

The associated rejection region is defined as .

To achieve a power level of .8 with an effective size of .2, we have:

Consider the case when and solve for the above equation.

The rejection region:

* 1. Suppose the true probability of heads is , we would like to perform the following test:

Let variable denote the number of heads in experiments, then

Consider rejecting when , since this is a two-tailed test and , should also be rejected when . Therefore, the type I error is given by

While the least possible cut-off value is 1, the minimum is given by

If we set the sample size to , with a reasonable effect size the power of the test would be very low and the probability of making a Type II error is very high, which means that we are more likely to believe , even through is the actual case. Thus, this test does NOT make sense.

* 2. Suppose the true probability of heads is , we should perform the following test:

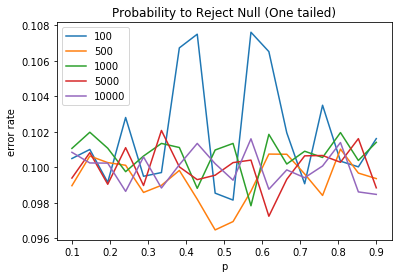
Let variable denote the number of heads in experiments, then

Consider rejecting when or when

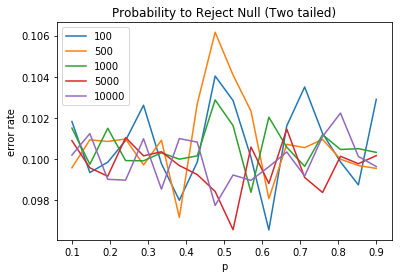
Since the null value is , . Therefore, the minimum can be achieved when setting and

If we set the sample size to , with a reasonable effect size the power of the test would be very low and the probability of making a Type II error is very high, which means that we are more likely to believe , even through is the actual case. Thus, this test does NOT make sense.

* 1. Solving with R library pwr () and we have
  2. From the simulation, we have observed several things:
     + More observations (higher ) will lead to lower variability of the error rate. This is what we have expected. With the increase of n, the error rate will less likely to be affected by random noise, thus we expect the error rate be more stable for all .
     + For all values of and , the error rate is close to .1, which is close to significance level alpha as we expected. This means that we will reject null hypothesis 10% of the time and result in a Type I error. Here, the error rate is by definition the significance level.
     + For extreme (close to .1 or .9), the error rate is close to the significance level. This behavior is expected as the variance of a binomial distribution is , which is higher when is .5, and lower when is extreme. With high variance, we expect to see more extreme error rates.



* 1. From the simulation, we observed similar effects with the two-tailed hypothesis test. Comparing with the one-tailed test, we noticed that two-tailed test presented less variation with the error rate. We believe that this is a trivial improvement from one-tailed hypothesis test that will be discussed later.



* 1. As two plots shown above, two-tailed hypothesis test gave us less extreme error rate than one-tailed hypothesis test, therefore, the two-tailed hypothesis test did a slightly better job in not making the Type I error that reject the null hypothesis.

However, this improvement is trivial regarding changing the result of the test. From the experiment above, we conclude that one-tailed and two-tailed hypothesis test gave us the same result, which is different from the idea delivered in the “How Optimizely (Almost) Got Me Fired”. The article believes that two-tailed hypothesis test is almost always better than one-tail hypothesis test, which is not what we’ve observed. Thus, we will believe the article less.