A/B Testing: The Most Powerful Way to Turn Clicks into Customers
By Dan Siroker, Pete Koomen and Cara Harshman

APPENDIX

2

Metrics and the Statistics behind A/B Testing

By Bryan Gumm

As online marketers and product managers, we can choose to optimize our users' experience along several different metrics. For example, a product manager of a subscription service might be interested in optimizing retention rate (percent), and an online marketer of an e-commerce site might focus on optimizing average order value ($). While each of these is obviously valid, the statistics behind A/B testing are slightly different for each. Before delving into the nuances of each, we'll introduce a few core concepts.

Confidence Intervals

Suppose we know that 51.4 percent of the population of the City of San Francisco has a bachelor's degree or higher. If we were to choose 1,000 city residents at random, we'd expect that exactly 514 of those people would have a bachelor's degree or higher. In reality, of course, this rarely happens. Why not? First, depending on your sample size, it may not be mathematically possible to arrive at exactly 51.4 percent (try this example with a sample size of 100 instead of 1,000). Second (and more important), by using a small sample to represent a large population, we are introducing some error.

In reality, it's usually difficult or impossible to measure the *exact* value of a statistic for an entire population; hence the obvious value of sampling. It seems, then, that we need a way to quantify the reliability of our sample data. We do this using estimates.

When we talk about statistics from a sample, we tend to provide two types of estimates: point estimates (single numbers)

and interval estimates (two numbers). If we were to poll 1,000 city residents chosen randomly, and found that 509 respondents had earned a bachelor's degree or higher, our point estimate would be 50.9 percent (509/1,000). The interval estimate is slightly more complex and depends partly on how certain we need to be in our estimate. The latter, often called the desired confidence level, varies by application, but for most A/B testing and other business analytics in general, 95 percent confidence is the standard. In the next section, we'll dive more into confidence levels.

For the time being, let's assume a 95 percent desired confidence level. The formula for the 95 percent interval estimate is given by:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{E.1}$$

where:

$\hat{p}$ = Our point estimate (50.9% or 0.509)

1.96 = A normal curve Z-value estimate corresponding to 95% significance

n = Our sample size (1,000)

For our example, the interval estimate would be:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.509 \pm 1.96\sqrt{\frac{0.509(1-0.509)}{1,000}}$$

$$0.509 \pm 0.031$$

We interpret that by saying we are 95 percent confident that the rate of San Francisco residents having a bachelor's degree or better is between 47.8 and 54.0 percent.

If we are instead interested in the average age of a San Francisco resident, the approach is the same and the formula very similar.

$$\bar{x} \pm 1.96\sqrt{\frac{s^2}{n}} \tag{E.2}$$

where:

$\bar{x}$ = The average age from our sample
n = Our sample size
s^2 = The variance in age from our sample

The sample variance calculation is given by:

$$s^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{E.3}$$

This formula says:

x_i	Take the age value for each resident.
$x_i - \bar{x}$	Subtract from that value the average age of the sample.
$(x_i - \bar{x})^2$	Square the result.
$\sum_{i=1}^{n} (x_i - \bar{x})^2$	Sum all of those values together.
$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$	Divide by the sample size less 1.

If your data is in Excel, you can compute the sample mean and variance by using the average( ) and var( ) formulas, respectively.

The curious reader might ask why the confidence interval formula used when measuring an average is different from the formula used when measuring a percentage. In fact, they're exactly the same! To see this for yourself, try simplifying the variance term (s^2) when every data point (x_i) is known to be either 1 or 0, which is the case when measuring a percentage. (Hint: When x_i is either 1 or 0, note that $x_i^2 = x_i$.)

For both the average age and the percentage of degree holders measurements, there are three conditions that result in a large confidence interval:

1. High variance in our data (inconsistent data).
2. A small sample (not enough data points).
3. A high desired confidence (greater than 95 percent).

In other words, to reduce the size of the interval, we'll need to take a larger sample, select a metric with less variability, or accept lower confidence in our results.

A percentage measurement has a unique property that its variance depends only on the percentage value itself:

$$\boxed{s_{\hat{p}}^2 = \hat{p}(1 - \hat{p})} \tag{E.4}$$

If we plot this function (Figure A-2.1), we can easily see that the variance of a percentage measurement is maximized when the point estimate is 0.5.

Therefore, it is possible to reduce variance (and reduce overall required sample size) if a metric whose expected value is closer to 0 or 1 is selected.

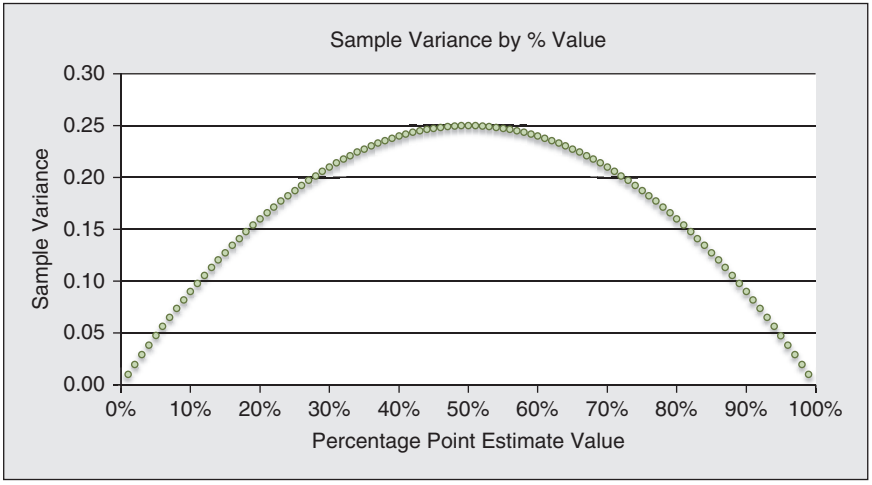

FIGURE A-2.1 Sample variance by percentage point estimate value.

Confidence Levels

As mentioned earlier, the standard confidence level for most business applications is 95 percent. Other than its pervasiveness, there is no reason statistical studies need to be limited to 95 percent. In other applications, such as biomedical research, manufacturing, or statistical computing, it is not uncommon to use other confidence levels (80 percent, 90 percent, and 99 percent are other common ones).

So what does a confidence level actually mean, anyway? Recall that earlier we said a 95 percent confidence level indicates we are 95 percent confident the true population value lies between 47.8 percent and 54.0 percent. In that example, we happened to know the true value was 51.4 percent, so our interval was correct. What a 95 percent confidence level actually tells us is that if we repeated the survey many, many times and computed a confidence interval for each, 95 percent of those confidence intervals would

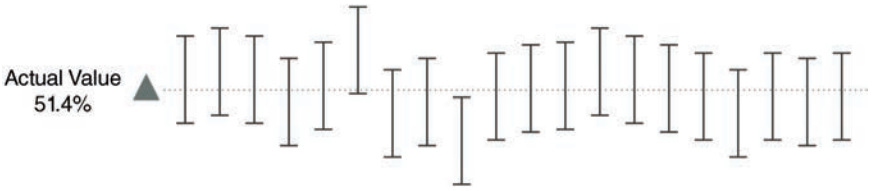

FIGURE A-2.2 Confidence intervals per sample.

contain the true population value. In the case depicted in Figure A-2.2, one of the 20 samples did not contain the true value of 51.4 percent.

In practice, we don't take multiple samples; we take one or two. For that reason, it becomes important to select the right metric and ensure that the sample size is sufficiently large to detect differences in the key performance indicator (KPI) of interest.

Recall from formula E.1 we said the value 1.96 corresponded to a 95 percent confidence interval. You can obtain this value for any confidence interval you want using the Excel functions found in Table A-2.1.

TABLE A-2.1 Excel Functions for Confidence Levels

Confidence Level	Excel Formula	Approximate Value
80%	=NORMSINV(0.90)	1.28
90%	=NORMSINV(0.95)	1.65
95%	=NORMSINV(0.975)	1.96
99%	=NORMSINV(0.995)	2.58

To determine what number to pass the NORMSINV function, first let your desired confidence level be represented by $(1 - \alpha)$. Then if b is the number passed to Excel, we find b by the following function:

$$b = 1 - \frac{\alpha}{2} \tag{E.5}$$

For example, for the 95 percent confidence level, compute:

$$(1-\alpha) = 0.95$$
$$\alpha = 0.05$$
$$b = 1 - \frac{0.05}{2} = 0.975$$

With the confidence interval foundation laid, we can now better explain how A/B test statistics work.

A/B Testing Metrics Framework

The preceding examples all dealt with a single sample. In A/B testing, however, we are interested in comparing multiple samples. For purposes of simplicity, here we will focus on two-sample comparisons (a simple A/B paradigm).

Suppose we wanted to compare conversion for two variants: current versus test. Using the principles applied earlier, we could derive point and interval estimates for each of the two versions. If we plot those side by side (Figure A-2.3), we then get a sense of how different the two experiences are in terms of that metric.

In the first case, we can visually see that versions A and B are not very different from one another. Case 2 shows a little more distinction between the two versions, and Case 3 shows the most differentiation between the two versions.

While visual inspection is always helpful, we need some way to empirically say whether A and B are "different enough." That is, we need a concrete formula to tell us when the observed difference between two observed conversion rates in an A/B test is large enough that it can reasonably be attributed to the difference in experience between A and B, and not just to a sampling error.

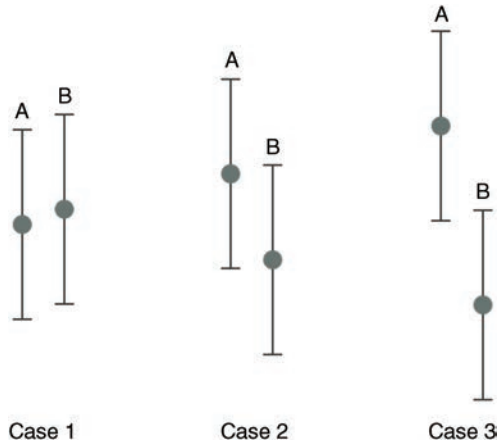

FIGURE A-2.3 Conversion comparisons for two variants.

There is an entire field of statistics devoted to defining "different enough" in multisample comparisons. That field is experimental design, a field in which there exist dozens of great books and an abundance of academic research (starting in the early 1900s). In most cases, we use the data we have to produce a test statistic. If that test statistic gets far enough away from 0, we conclude that the observed differences are likely not due to random chance. We will explore two such statistics.

The Z Statistic for Two Proportions

If our metric is a percentage or a proportion, the test statistic is given by:

$$Z = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{p(1-p)\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}} \tag{E.6}$$

From an earlier section, the denominator should look familiar to you. In its most basic form, it can be thought of as a measure of the overall variation in the metric we are examining (called the standard error). In fact, most basic test statistics follow this same general form:

$$Test\ Statistic = \frac{Observed\ Differences}{Standard\ Error} \tag{E.7}$$

For two proportions, the standard error is obtained by combining all values (irrespective of original group), computing the combined proportion, and then normalizing by the two sample sizes. If X_A is the number of people who converted in the first sample and X_B is that for the second, the combined conversion rate is given by:

$$p = \frac{X_B + X_A}{n_B + n_A} \tag{E.8}$$

For a computational example, suppose we have:

$$\begin{array}{llll} n_B = 49{,}981 & n_A = 50{,}332 \\ X_B = 42{,}551 & X_A = 42{,}480 \\ \hat{p}_B = 0.851 & \hat{p}_A = 0.844 \end{array}$$

We first compute p.

$$\begin{aligned} p &= \frac{42{,}511 + 42{,}480}{49{,}981 + 50{,}332} \\ p &= \frac{85{,}031}{100{,}313} \\ p &= 0.8477 \end{aligned}$$

We then compute our test statistic.

$$Z = \frac{0.851 - 0.844}{\sqrt{0.8477(1 - 0.8477)\left(\frac{1}{50{,}332} + \frac{1}{49{,}981}\right)}}$$

$$Z = 3.09$$

Note that if you were following along on your own, you may have ended up with a Z-value of 3.24. The difference is due to rounding.

Based on this test statistic, we can compute the probability that our observed difference (−0.007) is due to random chance. This value, called the p-value, is the area underneath the standard normal (Z) distribution before or after a certain point. If the value of Z is negative, we compute the value *up to* that point, and multiply it by 2. If the value of Z is positive, we compute the value *after* that point, and then multiply it by 2. The multiplication by 2 is the application of what is called a two-tailed test, which will be clarified later.

In Excel, it is simple to produce the two-tailed p-value using this formula: NORMSDIST(-ABS(Z))∗2. For our value of 3.09, this results in a p-value of 0.0021. This means that there is only a 0.2 percent chance that the difference we observed is due to random chance. If not due to random chance, it must be due to the differential experience. Put another way, there is a 99.8 percent chance that version B increases conversion above version A. This is the number many statistical packages, including Optimizely, publish as the "chance to beat original" metric (though Optimizely uses a one-tailed test).

The difference between a one- and two-tailed p-value is that in the latter, we are curious to know whether the two metrics

differ at all, whereas in the former, we care only if one metric is greater than the other. The latter is often the case in A/B testing. Institutional learning aside, having a current version (A) that outperforms a new version (B) isn't very actionable, but if we find a new version (B) that outperforms the existing version (A), we will likely take action and make the new version become the default. For that reason, many online marketers and product managers use one-tailed tests.

If conducting a one-tailed test, it is important to order the numerator of the Z statistic such that the new version's metric is first, as is the case in formula E.6. This becomes important in the computation of the "chance to beat original" estimate. For a properly formed one-tailed test, the p-value can be obtained in Excel using the formula 1-NORMSDIST(Z). The "chance to beat original" is simply NORMSDIST(Z).

Why is the order of the metrics in the numerator important? If we applied formula E.9 instead, our statistic would be −3.09. If that value is plugged in to the Excel functions above, the p-value and "chance to beat original" would be reversed.

$$Z = \frac{\hat{p}_A + \hat{p}_B}{\sqrt{p(1-p)\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}} \tag{E.9}$$

The Z test for proportions is acceptable to use as long as the sample size is sufficiently large and random sampling is applied. Common test statistics books call for a sample size of at least 30 to 50 in each group being compared, but in the A/B testing space, we often greatly exceed this.

The *t* Statistic for Two Averages

If our metric of interest is an average instead of a proportion, the basic concepts are the same, though there are more assumptions, legwork, and input to do these tests properly.

The basic t statistic is given by

$$t = \frac{\bar{x}_B - \bar{x}_A}{s_p\sqrt{\frac{1}{n_B} + \frac{1}{n_A}}} \tag{E.10}$$

where:

$\bar{x}_B$ = The sample average for experience B

n_B = The sample size for experience B

s_p = The combined sample standard deviation for both experiences

The combined sample standard deviation is given by:

$$s_p = \sqrt{\frac{(n_B - 1)s_B^2 + (n_A - 1)s_A^2}{n_B + n_A - 2}} \tag{E.11}$$

The sample variance formulas were given earlier.

Finally, because we are interested in the difference between two averages but must compute two standard deviate ons in order to arrive at a decision, we must account for that secondary estimation by incorporating a concept known as degrees of freedom. For this t test, the degrees of freedom are given by:

$$df = n_B + n_A - 2 \tag{E.12}$$

The "2" in formula E.12 represents the number of other parameters we must estimate.

In Excel, the formula to compute the *p*-value is TDIST(ABS (t),df,tails) where tails is 1 or 2 depending on the type of *p*-value we want. As was the case with the *p*-value in the last section, the "chance to beat original" is 1 minus the *p*-value. Again, if we are employing a one-tailed test, the order in which the terms appear in the numerator of the *t* statistic is important.

There are far more assumptions involved in using this test than there were for the *Z* test for proportions.

Random sampling is used to determine who gets experience A versus B.

The two population variances are equal.

The sample sizes are sufficiently large (>50 or so for each experience).

The underlying populations are approximately normally distributed (i.e., they follow a bell curve).

When statisticians talk about the assumptions of a test, they are referring to the original assumptions under which the test statistic was developed. It is often the case that those assumptions are made, the test statistic is developed, and then those assumptions are challenged. In the latter stage, the statistician will purposely violate the assumptions and analytically determine what happens to the overall measure. If the statistician finds that the overall test statistic is still valid, then the statistic is said to be robust to that assumption (i.e., it can be relaxed).

There is a lot of published research on what to do when assumptions aren't met. If you are concerned with this, further reading can be found in the Wikipedia article "Student's *t*-test" with some referencing links on that site.