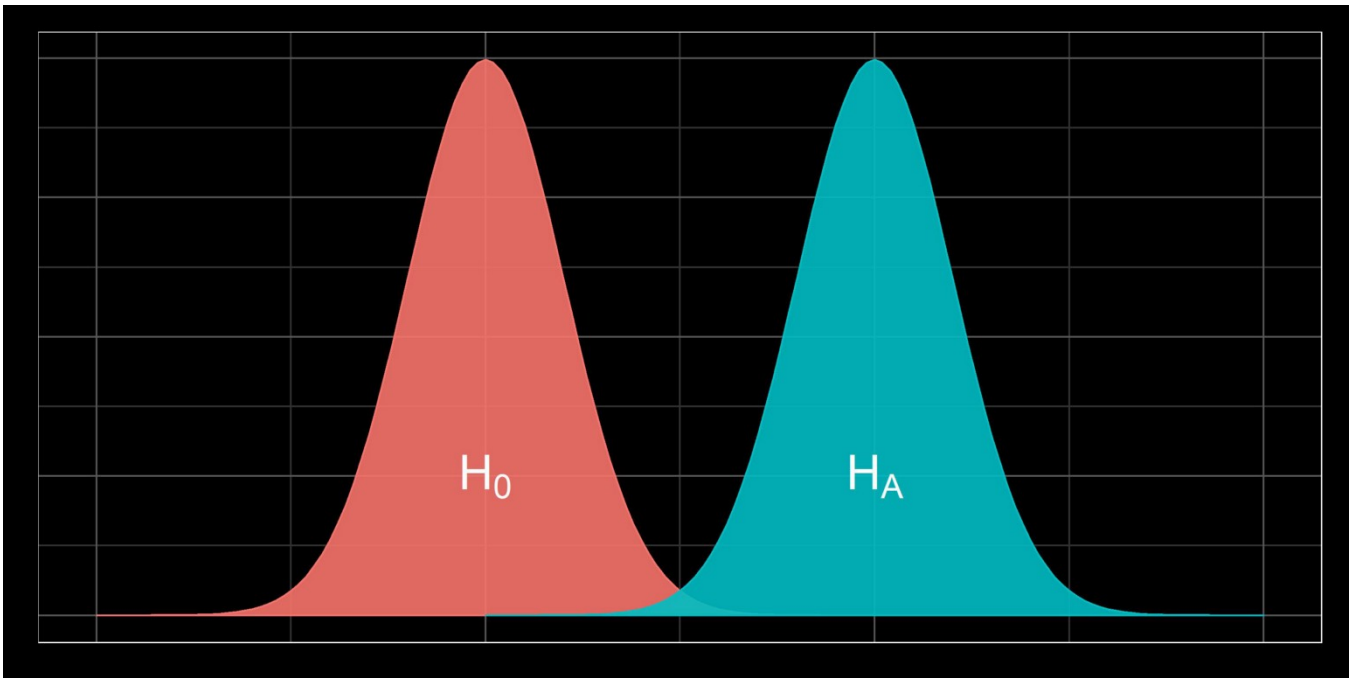
A/B Test Statistics Made Easy

Part 1: Continuous Metrics

Rezwan Hoppe-Islam Apr 10 · 12 min read ★



Your A/B test has just finished and it looks like the Test has performed better than the Control. But is this result just due to random noise? Or is the result large enough to be statistically significant? This guide will show you how to answer these questions using inferential statistics.

Statistics can often be an intimidating subject. Many textbooks, even those aimed at beginners, launch too quickly into seemingly complex formulae like this one: $\sigma d =$

sqrt(σ1^2 / n1 + σ2^2 / n2). Or they launch too quickly into abstract concepts like the "test statistic". Formulae and abstractions have an important place: they allow you to perform statistical analyses quickly and efficiently. But they aren't the best tools for *engaging* with the audience *or teaching* statistical concepts.

In this article we'll use simulations to bring key statistical concepts to life, rather than relying on formulae. And we'll deconstruct abstractions to illuminate all the hidden components. Altogether, we'll demystify this fascinating subject area and make it more accessible and more enjoyable to readers.

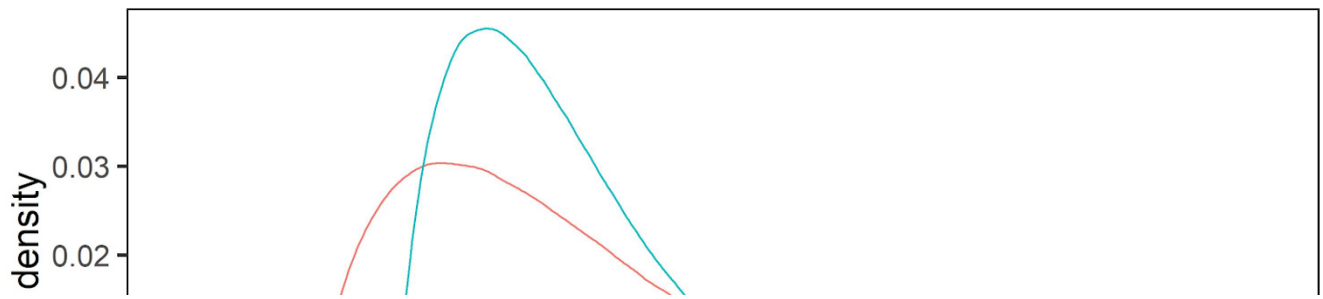
The only prerequisite knowledge is familiarity with the mean, the standard deviation, and density plots. This article (Part 1) focuses on the analysis of continuous metrics (think age, height, income, etc.). In a future article (Part 2) we'll examine proportion metrics (think conversion rate, % of people who smoke, fraction of customers that are satisfied, etc.).

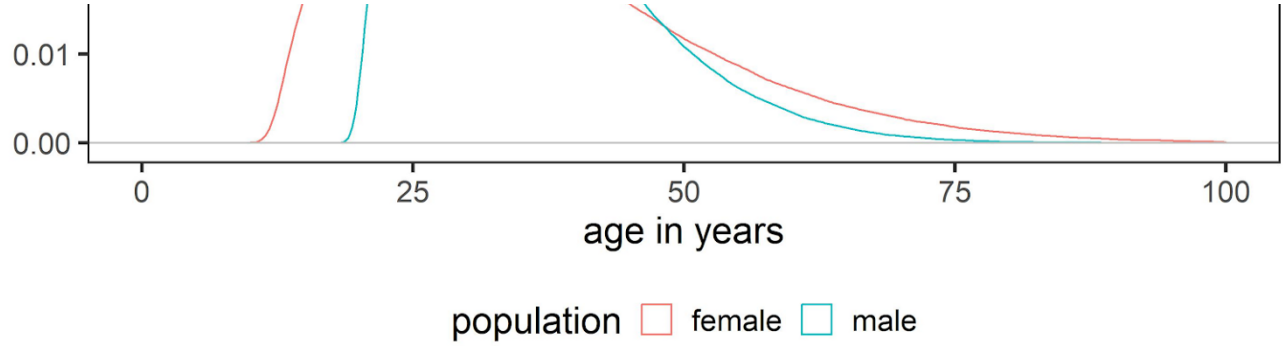
1. Creating a sampling distribution through simulation

Imagine a town with a population of 1,000,000 residents. Half the residents are male and the other half are female. The mean age of each group is 35 years old. The standard deviation of female ages is slightly higher than that of male ages. Here's a summary of the population parameters:

Group	Number of people	Mean age	Standard deviation
Male	500,000	35	10
Female	500,000	35	15

And if we were to create a density plot of each group's ages, this is what it might look like:





Now imagine that we wanted to explore the difference in mean age between the two populations. We know that it should be 0 since both populations have identical mean ages. Let's run the following procedure:

1. We take a random sample of 200 males and calculate their mean age
2. We take a random sample of 200 females and calculate their mean age
3. We calculate the difference in mean ages

After running through this procedure once we get the following results:

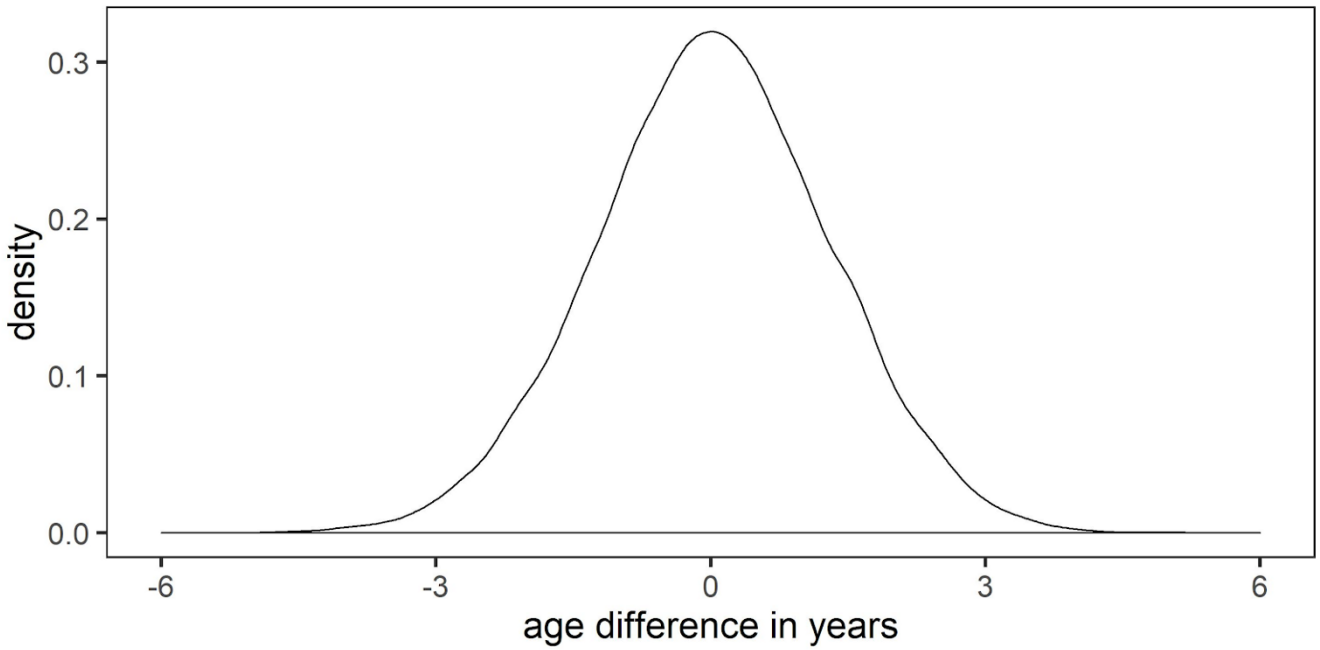
Sample	Mean age in years (across 200 males)	Mean age in years (across 200 females)	Age difference in years
1	35.4	35.8	-0.4

Now let's run a simulation that repeats the above procedure 10,000 times:

Sample	Mean age in years (across 200 males)	Mean age in years (across 200 females)	Age difference in years
1	35.4	35.8	-0.4
2	36.1	35.3	0.8
...	...	...	...
10,000	34.3	33.6	0.7

Even though we *know* that both populations have a mean age of 35, our estimates of the mean vary from sample to sample. Some estimates are higher than 35 and others are lower. This natural variation between samples is referred to as random ***sampling error***.

Sampling error also causes the estimated age difference (the last column) to vary between samples. Some estimates are higher than 0, others are lower. Let's examine the distribution of these values:



This plot is called a ***sampling distribution***. Specifically, it is a sampling distribution of the difference between means. The majority of difference values are close to zero which is as expected since both populations have identical means. There are also some samples which yield large differences, again due to random sampling error. These values are found in the tails of the distribution which tells us that they occur less frequently.

Recap: Using simulation we have created a sampling distribution. This sampling distribution illustrates the natural variation between samples that arises due to random sampling error.

2. A faster way to generate the sampling distribution

Simulations help us to understand what a sampling distribution is and how it is created. Once we understand these concepts, we can dispense with simulation altogether and compute the sampling distribution directly. This approach is faster and more accurate.

To compute the sampling distribution we just need to know three properties:

1. The **shape**. A sampling distribution is always symmetrical (given enough data points), regardless of the shape of the underlying population distribution. Specifically, the shape is referred to as ***normal*** (for sufficiently large samples).

2. The **mean**. The mean of the sampling distribution is the difference of the means of the two underlying distributions. In our case this will be zero as both the underlying populations have the same mean age.

3. The **standard deviation**. This is also called the ***standard error*** when referring to a sampling distribution, since a sampling distribution characterizes the sampling error. It turns out that we can calculate it as follows:

$$\text{Standard error} = \sqrt{\frac{(standard\ deviation\ of\ population\ A)^2}{sample\ size\ from\ population\ A} + \frac{(standard\ deviation\ of\ population\ B)^2}{sample\ size\ from\ population\ B}}$$

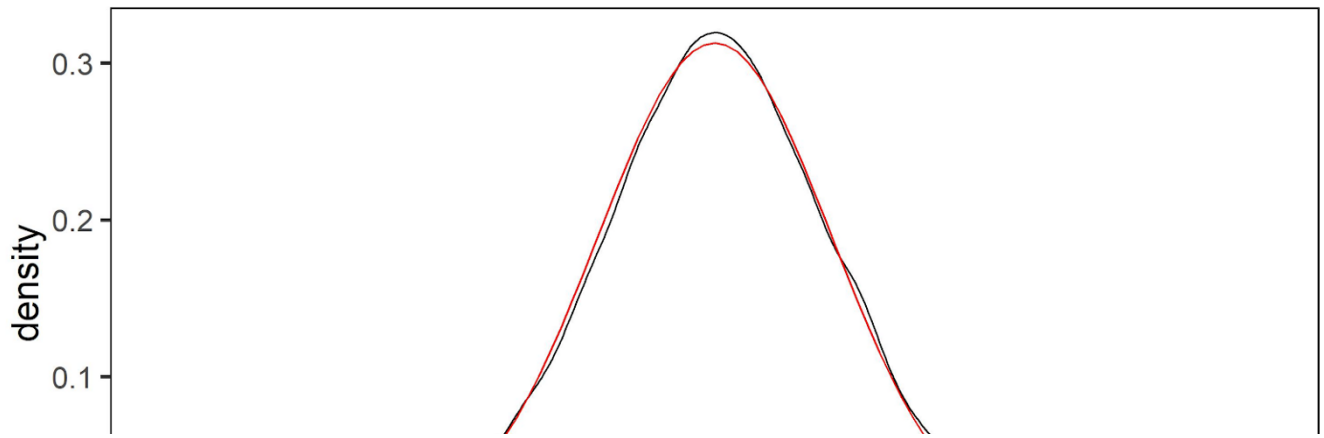
$$= \sqrt{\frac{10^2}{200} + \frac{15^2}{200}}$$

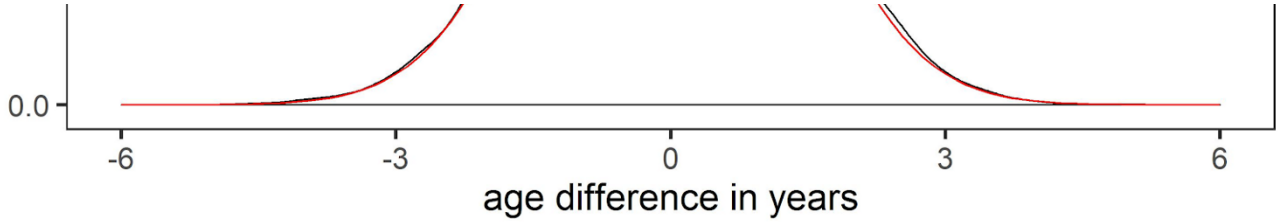
$$= 1.27\ years$$

Now that we know these three properties, we can quickly compute the sampling distribution using most programming languages. For example, in *R* we would simply write:

```
rnorm(n = 10000, mean = 0, sd = 1.27)
```

This single line of code would generate a normal distribution with 10,000 values, a mean of 0, and a standard deviation of 1.27. Let's plot the distribution of these computed values (red) over the previous distribution that we created through simulation (black):





Notice how the two distributions are very similar to one another other. The computed distribution is *exact*. The simulated distribution is an imperfect *estimate*. Hence the slight discrepancy between the two.

Recap: We no longer have to run many simulations to *estimate* the sampling distribution. It turns out that there exists a simple formula to calculate the standard deviation of our sampling distribution. And with this, we can now quickly compute the *precise* sampling distribution.

3. A first glance at some A/B test results

Now imagine you have to interpret the results of an A/B test. For example, you designed a brand new homepage for your website that you think will subsequently lead to a higher revenue per visitor. You set up an A/B test which ran for a few weeks where half of your visitors saw the new homepage (the Test group) and the other half saw the existing homepage (the Control group). At the end of the test, we see the following results:

Treatment group	Sample size	Sample mean	Sample standard deviation
Test	1,000	$39.8	$75.3
Control	1,000	$32.1	$72.7
Difference (Test - Control)		***$7.7***	

On first glance it appears that the Test actually performed better than the Control by $7.7 per visitor. But can we trust that this result is representative of the true difference between population means? After all, in the previous sections we observed that even when two populations have identical means we will still generate a range of estimates of the difference due to sampling error.

So, assuming that the Test and Control populations have identical means, what is the likelihood that we'd see a result as high as $7.7, due to sampling error alone? This assumption — that there is no real difference between Test and Control population means — is called the ***null hypothesis*** (***H0***). By convention, we always start with H0 and then proceed to examine the strength of the evidence for or against it.

Recap: Even though it first appears that the Test has performed better than the Control, we need to understand how likely we are to see a result of $7.7, given the null hypothesis.

4. Computing the sampling distribution for the null hypothesis

In order to understand the probability of seeing a result of $7.7 given the null hypothesis, we first need to understand the distribution of possible values that can arise assuming that the Test and Control populations have identical means.

In other words, we need to generate the sampling distribution of results assuming that the null hypothesis is true. We can compute this using the same steps as before — we just need to know three properties:

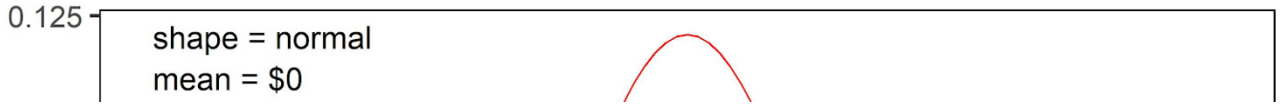
1. Shape = normal (as demonstrated)
2. Mean = zero (as stated by the null hypothesis)
3. Standard error (computed)

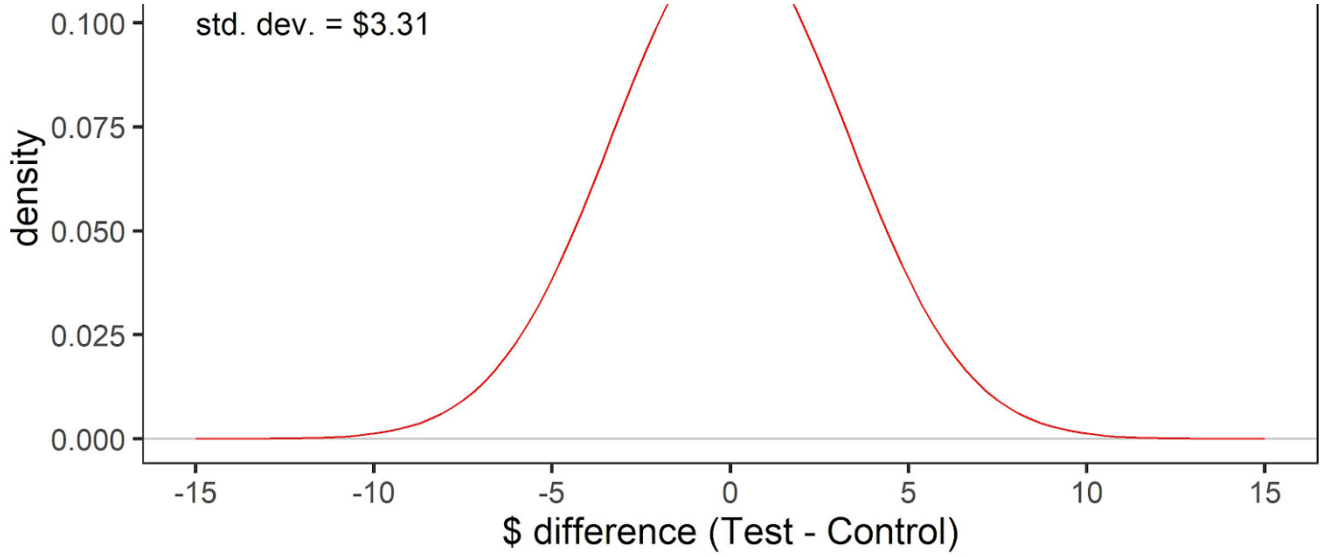
$$= \sqrt{\frac{(standard\ deviation\ of\ population\ A)^2}{sample\ size\ from\ population\ A} + \frac{(standard\ deviation\ of\ population\ B)^2}{sample\ size\ from\ population\ B}}$$

$$= \sqrt{\frac{75.3^2}{1000} + \frac{72.7^2}{1000}}$$

$$= \$3.31$$

Notice that we used the sample standard deviations as a best guess for the population standard deviations. Here's what our sampling distribution looks like:



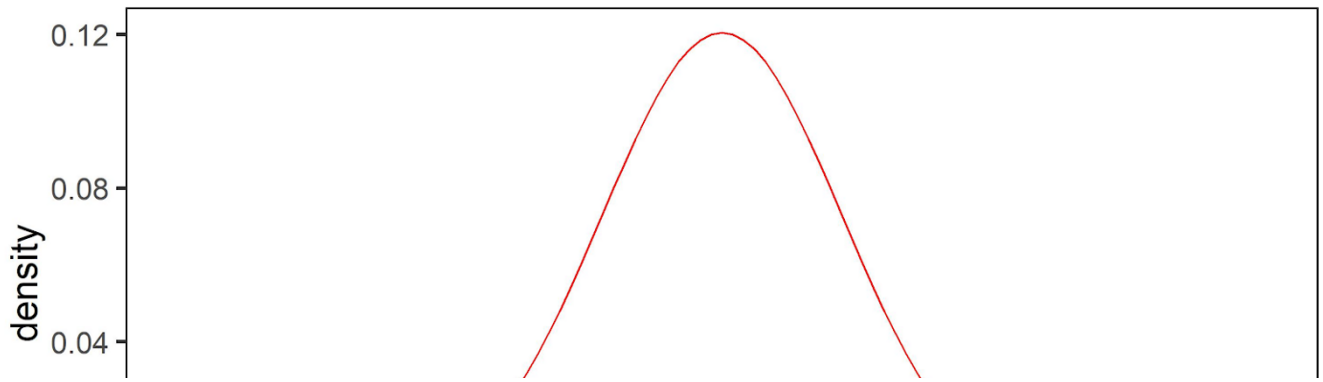


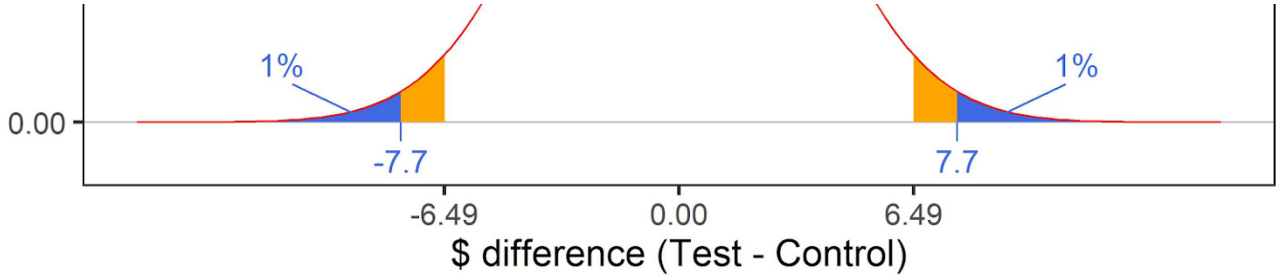
Recap: How likely are we to see a result of $7.7, given the null hypothesis? To answer this we first need to understand the sampling distribution under the null hypothesis. We can compute this assuming that a) the shape of the distribution is normal, b) the Test and Control populations have identical means, and c) our sample standard deviations are representative of the population standard deviations.

5. Interpreting the A/B test results

Given the sampling distribution under the null hypothesis, how likely are we to see a difference between Test and Control means of $7.7? Remember that if the null hypothesis is true then the distribution is symmetrical around a mean of zero. That means we have an equal chance of obtaining a result of $7.7 as we do of obtaining -$7.7. So the question becomes: what is the probability of obtaining a result *as extreme or more as* $7.7, given the null hypothesis?

Let's re-examine our sampling distribution:





Here we have shaded in blue the areas of the curve bound by values that are *as extreme or more as* $7.7. The sum of both areas is 2% and this value is known as the ***p-value***. It tells us that if the null hypothesis were true then we would expect to see a result as extreme or more as $7.7 in only 2 out of 100 A/B tests.

By contrast, the areas shaded in orange (you have to imagine that the orange areas continue behind the blue areas) represent 5% in total of the area under the curve. The value of 5% is an arbitrary choice and it is called the ***significance level*** or **α**. When $p < \alpha$, we reject the null hypothesis.

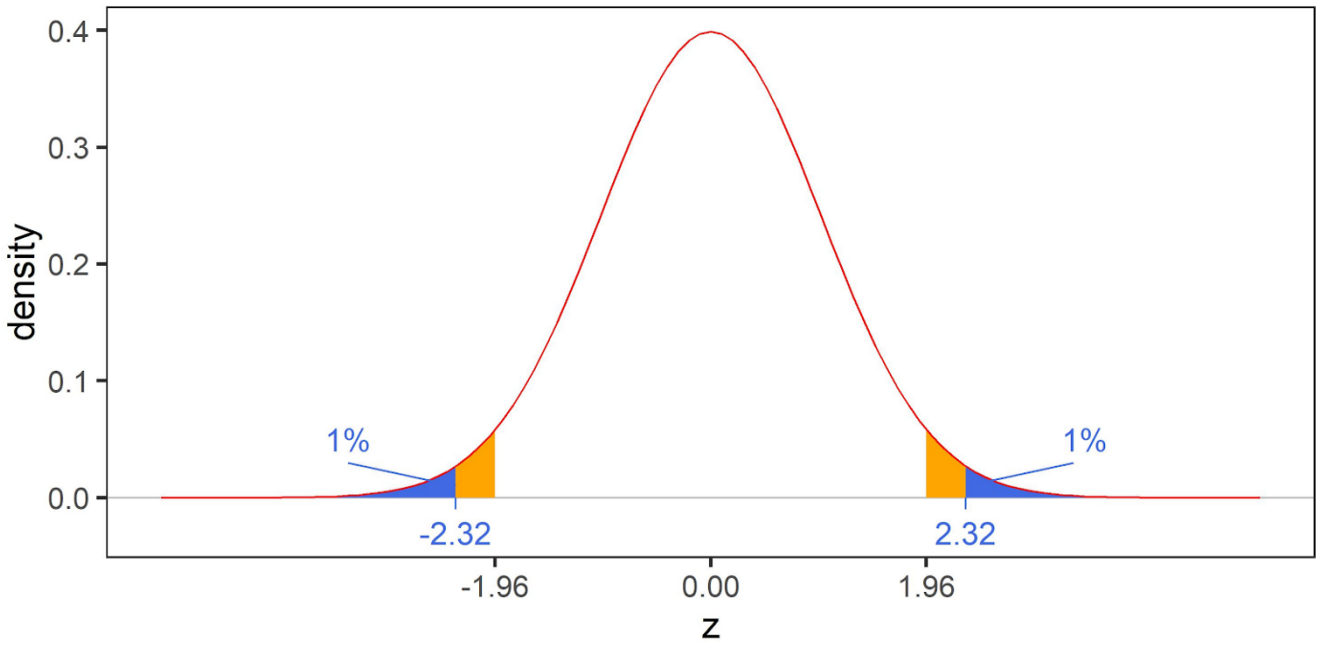
In other words, we are saying that if we see an A/B test result that's only likely to occur in less than 5 out of 100 A/B tests (given the null hypothesis) then we'll reject the null hypothesis (even when it is true). That means that we run the risk of *incorrectly* rejecting the null hypothesis 5% of the time. So **α** is also referred to as the Type 1 error rate.

Recap: The chance of seeing a result as extreme or more as $7.7 is improbable under the null hypothesis. Specifically, our p-value of 0.02 means that we're only likely to see a result as extreme or more as $7.7 in 2 out of 100 experiments. This ratio is less than our significance level (α) of 0.05 and, therefore, we reject the null hypothesis. This may or may not be the right conclusion and in the long run our Type 1 error rate = 0.05 (= α).

6. A faster way to calculate p-values

In practice, statisticians do not create a bespoke sampling distribution based on every set of A/B test results. They employ a single, ***standard normal distribution*** for sufficiently large sample sizes. This distribution has a mean of 0 and a standard deviation of 1. Any normal distribution with a mean of 0 can be transformed into a standard normal distribution by dividing by the standard error.

So if we divide all values in the previous distribution by the standard error ($3.31), we get the following:



The transformed values are referred to as **z values** and they are unit-less.

With a standard normal distribution, a significance level of 5% *always* corresponds to z values that are as extreme or more as 1.96 (approximately). We refer to 1.96 as the ***critical z value*** at 5% significance. Our p-value does not change, it is still 0.02. And here, it is the area under the curve that corresponds to z values as extreme or more as 2.32 ($7.7 / $3.31). This value is called the ***z score*** or ***test statistic***.

Recap: By convention, statisticians employ a standard normal distribution which has a mean of 0 and a standard deviation of 1. This eliminates the need to compute a bespoke sampling distribution for every A/B test. We only need to calculate the z score (observed difference / standard error). The p-value is then the area under the curve bound by the z scores.

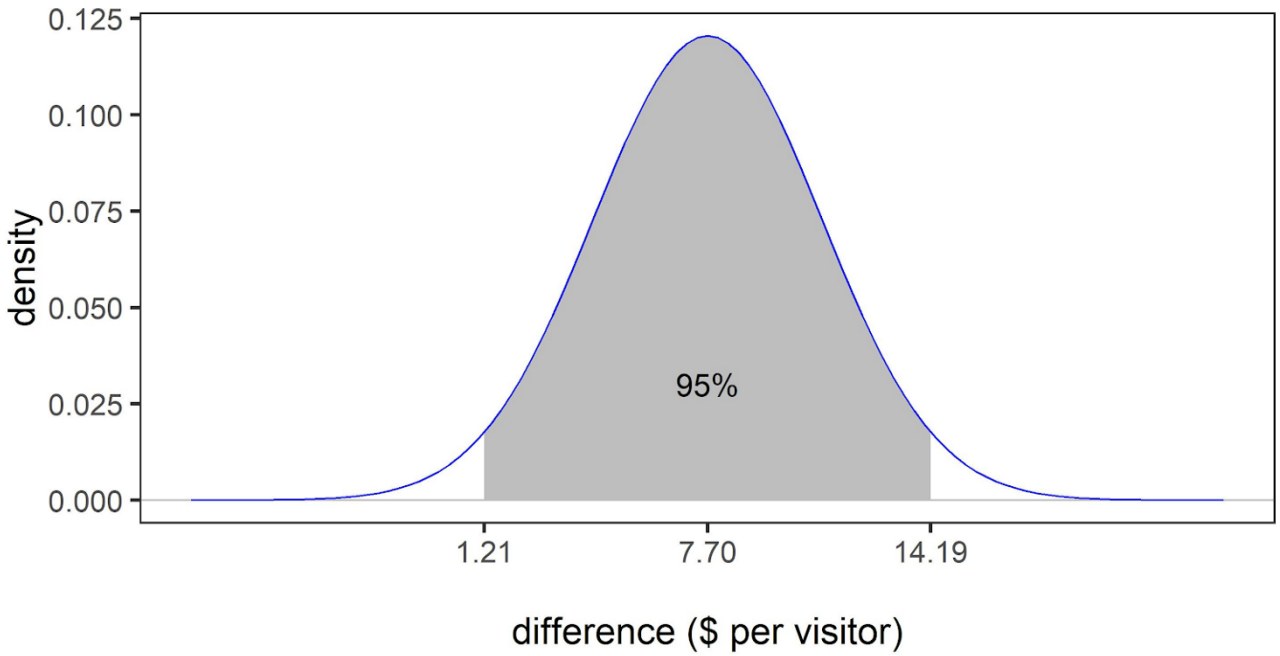
7. The confidence interval

To recap, we took our A/B test result of $7.7 per visitor and decided that this value is unlikely to occur if, in reality, both Test and Control populations have identical means. More formally, our p-value of 0.02 is less than our significance level of 0.05 and so we reject the null hypothesis.

We are rejecting the (null) hypothesis that the difference between population means = 0. Therefore, we automatically accept the ***alternative hypothesis (HA)*** which, here, states that the difference between means ≠ 0.

So if the difference between means *does not* equal 0, then what *does it* equal? One answer would be $7.7. A *better* answer would identify a *range* of possible values, given that we now recognize the impact of sampling error.

So what is the range of possible values given the alternative hypothesis? Again, we need only refer to the sampling distribution. The sampling distribution for the alternative hypothesis is identical to that for the null hypothesis with one exception: its mean is equal to the observed result ($7.7):



The area shaded in grey represents the middle 95% of the chart which corresponds to a ***confidence interval*** of (1.21, 14.19) or 7.7 ± 6.49. Every confidence interval has an associated ***confidence level*** which is predetermined by the significance level (**α**) and is equal to 1 — **α** = 1–0.05 = 95%.

Another interpretation of the confidence interval and confidence level is this: if we were to repeat the entire A/B test then we'd likely get Test and Control samples with slightly different sample standard deviations due to sampling error. And so we'd generate a slightly different sampling distribution, and thus a slightly different confidence interval. Now if we were to repeat the entire A/B test many times then, in the long run, 95% of our confidence intervals would include the *true* difference between population means. Note that this is *not the same* as saying that there's a 95%

probability that our single confidence interval includes the true difference between populations.

Recap: A confidence interval (at an associated confidence level) shows the range of possible values for our A/B test result. Here, we report a confidence interval of 7.7 ± 6.49 at a confidence level of 95%.

8. Power

Another key concept to understand regarding hypothesis testing is statistical power. Power is analogous to significance:

- Significance (α) = the probability of rejecting H0 when H0 is true.
- Power (1-β) = the probability of accepting HA when HA is true.

Since we've already accepted HA the concept of power is mostly redundant here. We'll cover it in a future article on sample size calculations.

Recap: The power of a test is the probability of accepting HA when HA is true. It is mostly relevant before the test is run, when you're calculating the required sample size of the test.

9. Statistical analyses in R

By this point you should feel fairly comfortable with the overall framework of a statistical analysis. Now let's see how we can reduce all the above-mentioned steps into just a few lines of code. In *R* we would write:

```
# Inputs
se.1 <- 75.3
se.2 <- 72.7
n1 <- 1000
n2 <- 1000
mean.delta <- 7.7

# Outputs
```

```r
print('standard error')
se <- sqrt((se.1/sqrt(n1))^2 + (se.2/sqrt(n2))^2)
se

print('p-value')
pnorm(q = -mean.delta/se, mean = 0, sd = 1) * 2

print('critical value at alpha = 95%')
crit <- qnorm(p=0.975, mean = 0, sd = 1) * se
crit

print('confidence interval, lower bound')
mean.delta-crit

print('confidence interval, upper bound')
mean.delta+crit
```

Continuous Metrics.r hosted with ❤️ by GitHub view raw

This produces:

```
[1] "standard error"
3.30989123688377

[1] "p-value"
0.0199993307409128

[1] "critical value at alpha = 95%"
6.48726761703693

[1] "confidence interval, lower bound"
1.21273238296307

[1] "confidence interval, upper bound"
14.1872676170369
```

Recap: Once you're comfortable with all the steps involved in a statistical analysis, you can perform the calculations very quickly using a computer program such as *R*.

10. Summary

In this article, we explored many of the key concepts and layers involved in a statistical analysis. And we reviewed some shortcuts to arrive at the answers more quickly once these concepts are understood. Hopefully, this has left you with a greater sense of comfort around A/B test statistics. In the next article, we'll use what we've learned here to analyze a different A/B test where the success metric is a proportion.

Statistics A B Testing Hypothesis Testing