A/B Testing and Beyond

Designed Experiments for Data Scientists

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Week 3

Tuesday September 18th, 2018

Outline

- Recap
- Experiments with Two Conditions
  - Comparing Means
    - The two-sample *t*-test
    - Example
    - Sample size calculations
  - Comparing Proportions
    - The Z-test for proportions
    - Example
    - Sample Size Calculations

RECAP

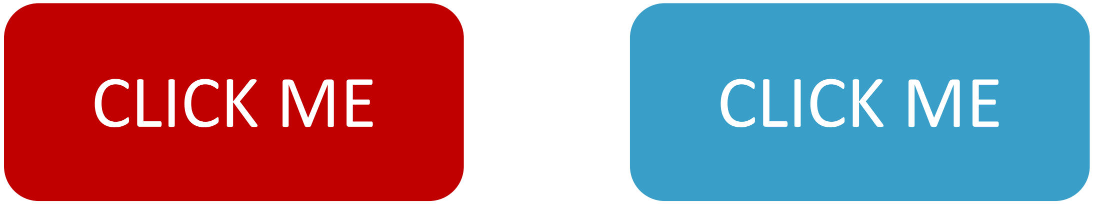
Recap

- Random variables – discrete, continuous
- Probability functions – PMF, PDF
- Probability calculations
- Expectation, variance and moments
- Special distributions – binomial, normal, t, χ^2
- Statistical Inference – population, sample
- Point/ interval estimation
- Hypothesis testing – null, alternative, Type I/II error, test statistic, null distribution, p-value, significance level, power

EXPERIMENTS WITH TWO CONDITIONS

Preliminary

- We now consider the design and analysis of an experiment consisting of two experimental conditions i.e., an A/B Test
- The typical goal is to decide which condition is optimal with respect to some metric of interest
- Canonical A/B test:



- Given two options, which one is best?

Preliminary

Designing an A/B test:

- Choose the metric θ that will answer your question of interest
- Choose the response variable (y) that will be used to calculate θ
- Choose a design factor and two levels to experiment with
- Choose n_1 and n_2 the number of units to assign to each condition

Preliminary

Data Collection:

- Randomly assign n_1 units to the first condition and randomly assign n_2 units to the other condition
- Measure the response (y) on each unit and summarize the measurements with the metric of interest θ in both conditions

Goal:

- Identify the optimal condition

Preliminary

Return to the canonical A/B test

- Suppose θ_1 represents the **probability** that the red button is clicked and θ_2 represents the probability that the blue button is clicked
- We estimate these probabilities with $\hat{\theta}_1$ and $\hat{\theta}_2$ which are the observed **proportions** of units that that clicked the buttons in each condition
- Suppose $\hat{\theta}_1 = 0.12$ and $\hat{\theta}_2 = 0.03$
- Does this mean $\theta_1 > \theta_2$?

Preliminary

To decide, we must formally test a statistical hypothesis

$$H_0\!: \theta_1 = \theta_2 \text{ vs. } H_A\!: \theta_1 \neq \theta_2$$

$$H_0\!: \theta_1 \leq \theta_2 \text{ vs. } H_A\!: \theta_1 > \theta_2$$

$$H_0\!: \theta_1 \geq \theta_2 \text{ vs. } H_A\!: \theta_1 < \theta_2$$

Which statement is appropriate depends on the question the test is designed to answer

Preliminary

We will now discuss the design and analysis of experiments that are meant to test hypotheses like these.

In particular, we will

- Discuss how to choose the number of units to assign to each condition

- Describe different analysis techniques that are appropriate for different metrics of interest, and different types of response variables

Comparing Means

- Here we assume the response variable of interest is measured on a continuous scale

- But this methodology is also commonly applied when the response variable is discrete with a large support set

- We assume that the n_j response measurements in condition $j = 1,2$ follow a normal distribution:

$$Y_{ij} \sim N(\mu_j, \sigma^2)$$

for $i = 1,2, \dots, n_j$

Comparing Means

What does this mean?

- Y_{ij} represents the observation for the i^{th} unit in the j^{th} condition

- We believe the two samples of observations could reasonably have been drawn from a normal distribution

- These normal distributions have the same variance σ^2 but potentially different means μ_1 and μ_2

Comparing Means

To formally decide whether $\mu_1 = \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$, we test one or more of the following:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$

$$H_0: \mu_1 \leq \mu_2 \text{ vs. } H_A: \mu_1 > \mu_2$$

$$H_0: \mu_1 \geq \mu_2 \text{ vs. } H_A: \mu_1 < \mu_2$$

Comparing Means

Every hypothesis test is composed of the following components:

- There is a test statistic that we calculate from the data
- This test statistic is compared to the null distribution
- The extremity of the observed test statistic is quantified with a p-value
- Conclusions are drawn based on the size of the p-value in relation to the significance level of the test

Comparing Means

The Two-Sample *t*-Test

Based on the distributional assumptions we have made about the data, it is true that

$$\bar{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \sim N\left(\mu_j, \frac{\sigma^2}{n_j}\right)$$

And hence that

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

Comparing Means

The Two-Sample *t*-Test

And thus

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Which looks like it could be a test statistic, but the problem here is that we don't know σ.

What if we replace σ in this expression with its sample estimate $\hat{\sigma}$?

Comparing Means

The Two-Sample *t*-Test

Recall that we have assumed σ^2 is the same in each condition

Thus we estimate σ^2 with a pooled estimate (i.e., a weighted average of the sample variances from the two conditions):

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_1^2 and S_2^2 are the sample variances in the two conditions

Comparing Means

The Two-Sample *t*-Test

Substituting $\hat{\sigma}$ for σ in the previous expression yields

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

which follows a *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom

This is the test statistic for this test, and the null distribution is $t_{(n_1+n_2-2)}$

Comparing Means

The Two-Sample *t*-Test

To actually test a hypothesis we must calculate an observed value of T based on the data that have been collected through the experiment:

$$\{y_{11}, y_{21}, \dots, y_{n_1 1}\} \text{ and } \{y_{12}, y_{22}, \dots, y_{n_2 2}\}$$

Which are summarized with

$$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij},\ s_j = \sqrt{\frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j\right)^2}$$

for $j = 1{,}2$

Comparing Means

The Two-Sample *t*-Test

The observed value of the test statistic is calculated as

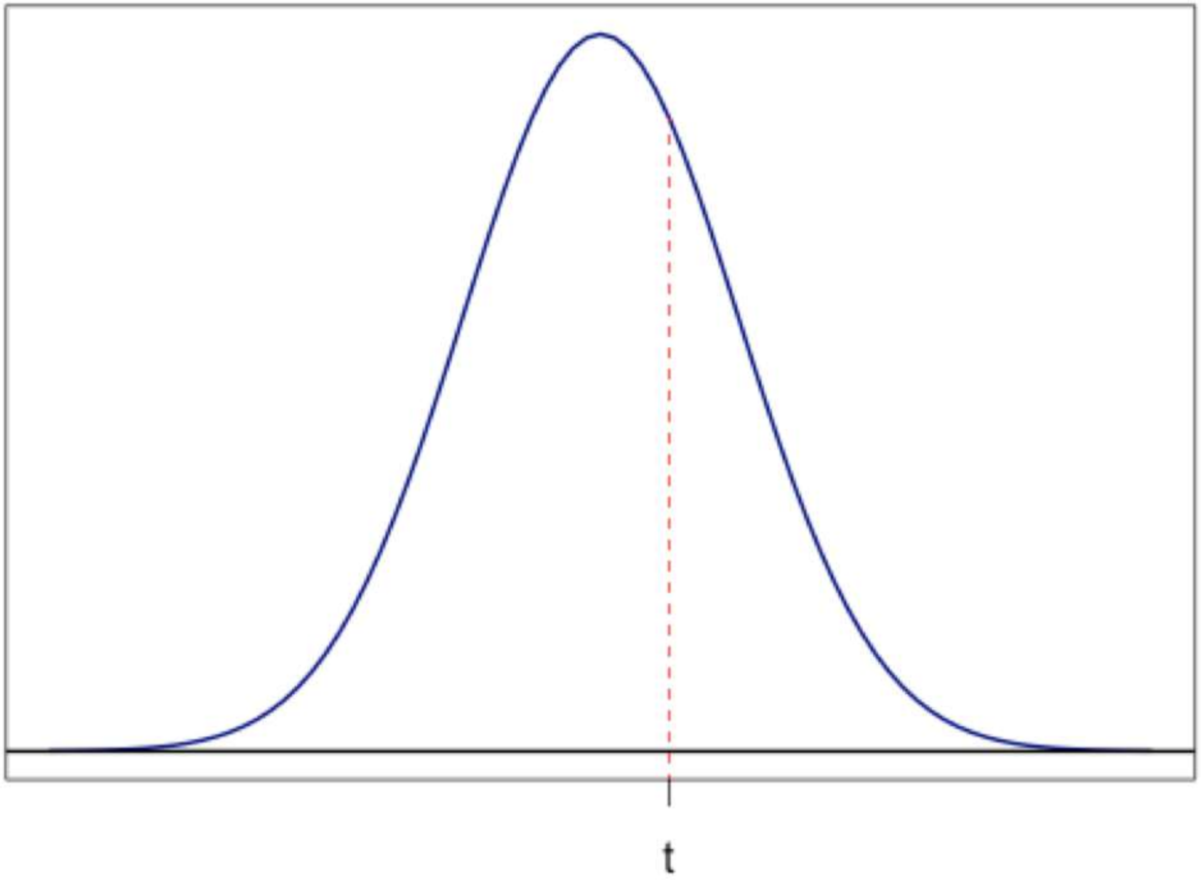
$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Next we evaluate whether t seems extreme in the context of the $t_{(n_1+n_2-2)}$ distribution

We quantify the extremity of t with the p-value

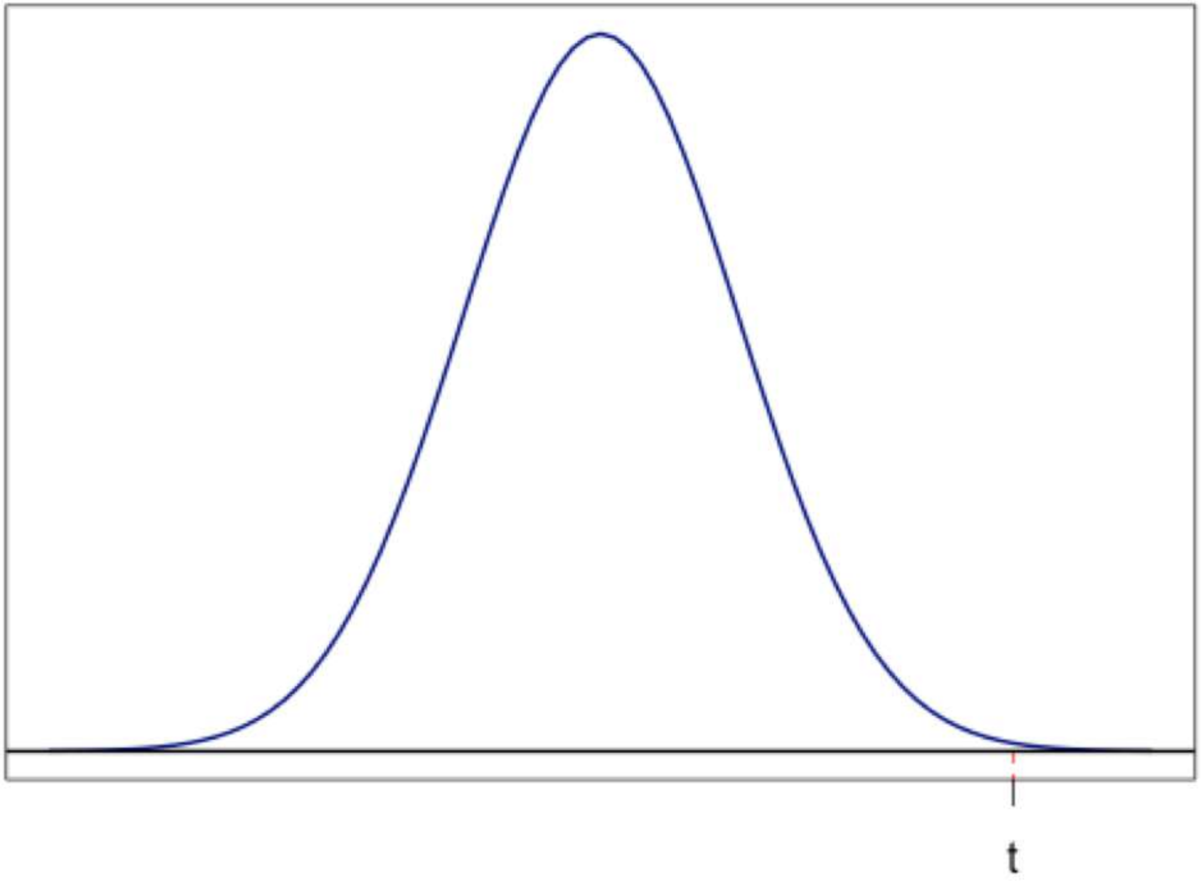
Comparing Means

The Two-Sample *t*-Test



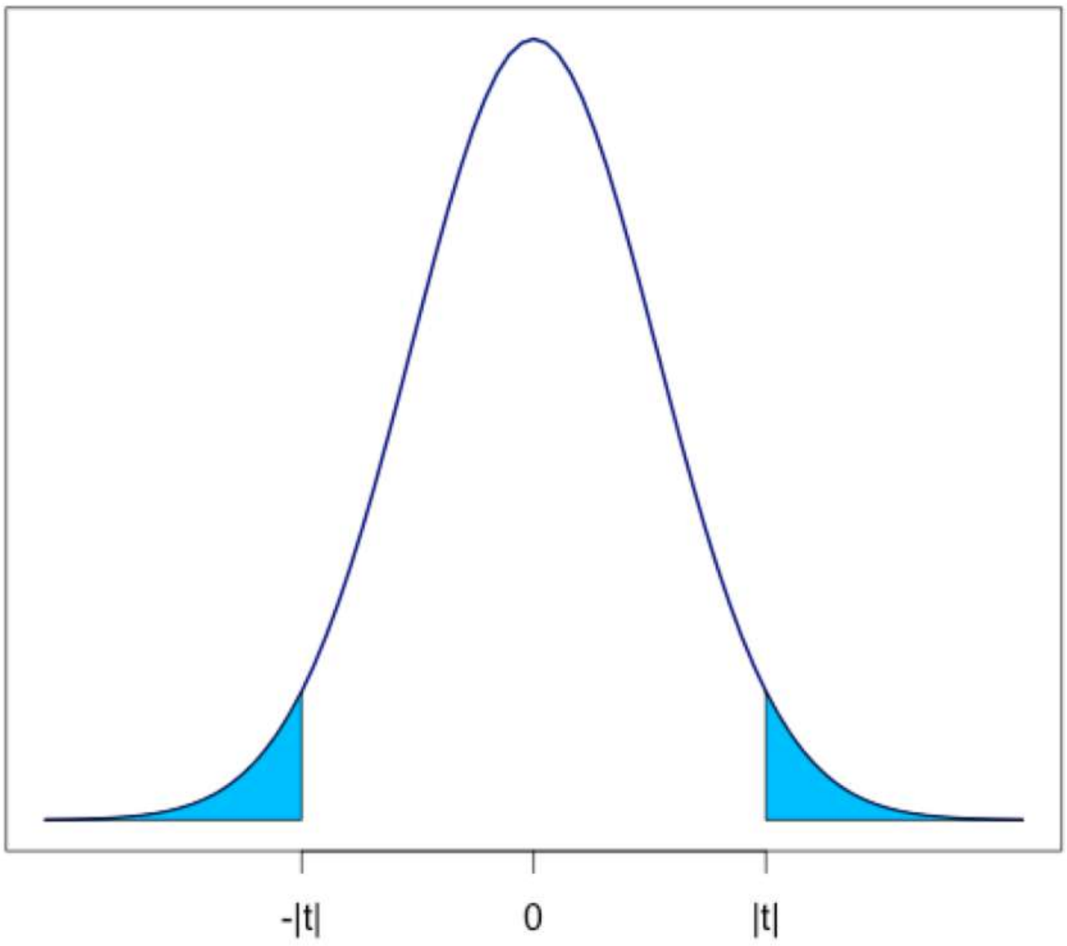
Hypothesis Testing

The Two-Sample *t*-Test



Comparing Means

$H_0: \mu_1 = \mu_2$ versus $H_A: \mu_1 \neq \mu_2$

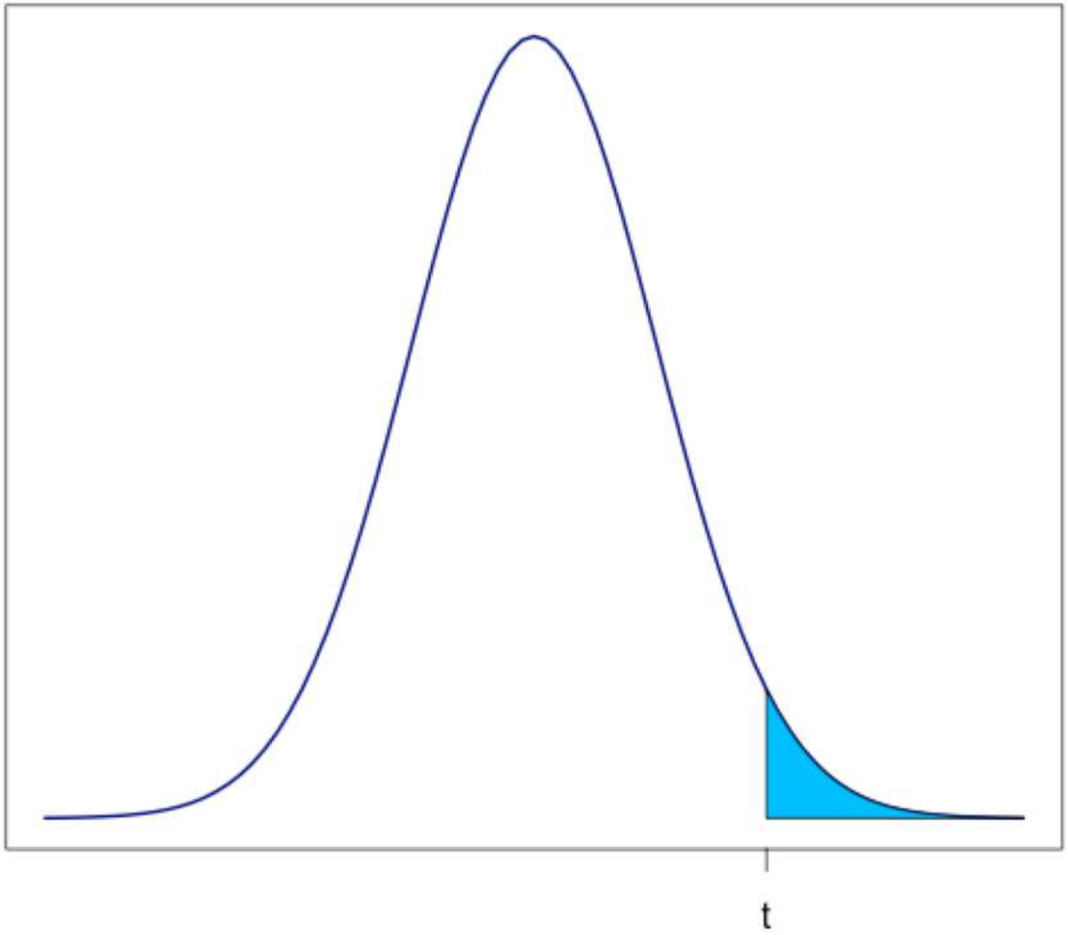


p-value = 2P(T > |t|)

Comparing Means

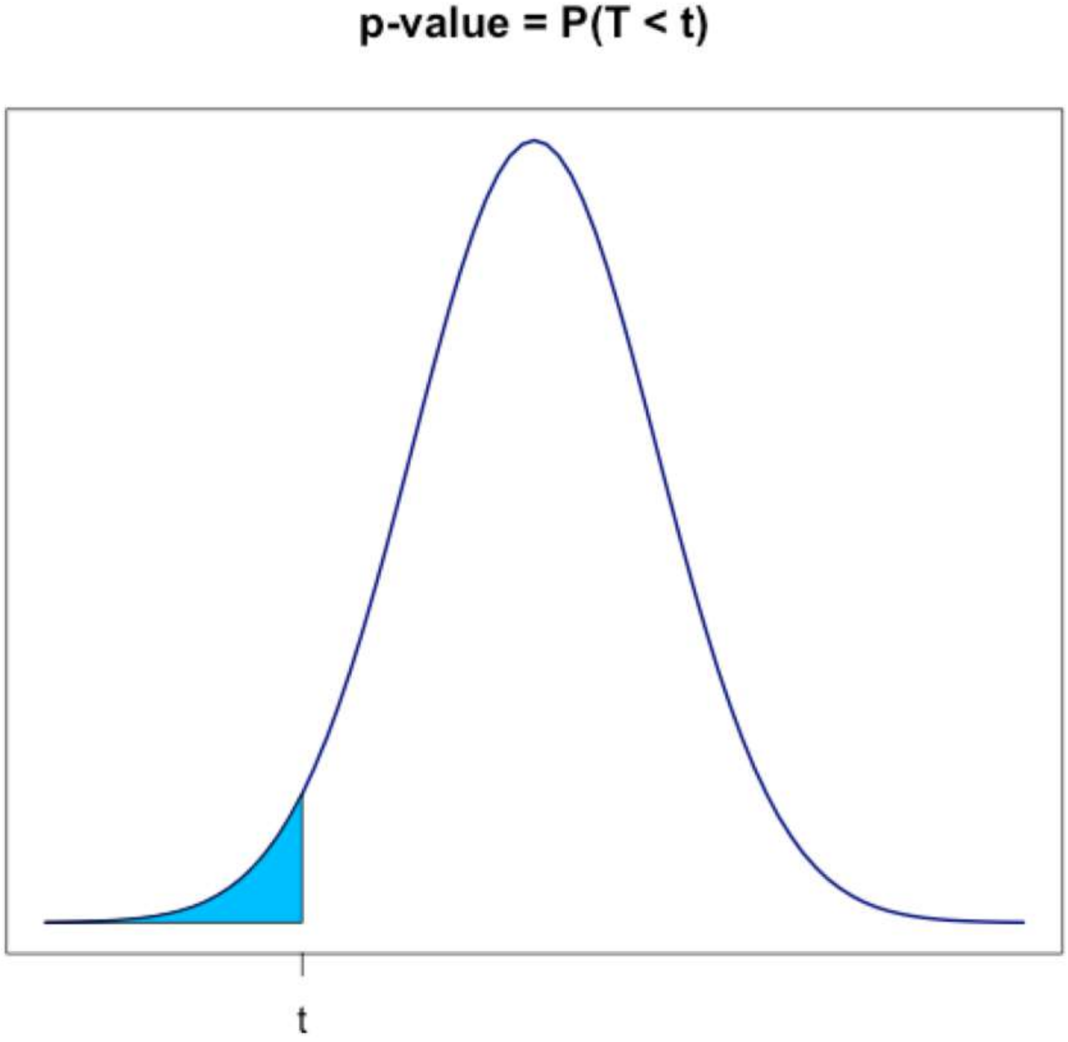
$H_0: \mu_1 \leq \mu_2$ versus $H_A: \mu_1 > \mu_2$

p-value = P(T > t)



Comparing Means

$H_0: \mu_1 \geq \mu_2$ versus $H_A: \mu_1 < \mu_2$



Comparing Means

Recall: how “extreme” t must be, and hence how small the p-value must be, to reject H_0 is determined by the significance level of the test, which we denote by α.

In particular, if

- p-value $\leq \alpha$ we reject H_0 in favor of H_A
- p-value $> \alpha$ we do not reject H_0

Note that $\alpha = 0.01$ and 0.05 are common choices.

Comparing Means

Example: Instagram Ad Frequency

- You are a data scientist at Instagram, and you are interested in understanding how user engagement is influenced by ad frequency

- Currently users see one ad every 8 posts in their social feed (i.e., a ratio of 7 non-ads : 1 ad)

- Management is pushing for one ad every 5 posts (i.e., a ratio of 4 non-ads : 1 ad)

- You worry this will hurt user engagement and so you decide to run an experiment

Comparing Means

Example: Instagram Ad Frequency

- You choose your response variable: $y =$ session length (the length of time, in minutes, that a user engages with the app)

- You define two experimental conditions:
  - Condition 1 – the current ad regime (7:1)
  - Condition 2 – the proposed ad regime (4:1)

- Interest lies in comparing μ_1 and μ_2 – the average session length in the two conditions

Comparing Means

Example: Instagram Ad Frequency

- You hypothesize that condition 1 will have a significantly longer average session time than will condition 2

- Formally

$$H_0: \mu_1 \leq \mu_2 \text{ versus } H_A: \mu_1 > \mu_2$$

- This null hypothesis assumes what your manager believes – so we expect to collect data that provides evidence against this statement, allowing us to reject it in favor of H_A

Comparing Means

Example: Instagram Ad Frequency

- In order to test this hypothesis you randomize $n_1 = 500$ users to the 7:1 condition and $n_2 = 500$ users to the 4:1 condition
- The data you collect is summarized as follows

$$\hat{\mu}_1 = \bar{y}_1 = 4.9162 \text{ and } s_1 = 0.9634$$

$$\hat{\mu}_2 = \bar{y}_2 = 3.0518 \text{ and } s_2 = 0.9950$$

$$\hat{\sigma} = \sqrt{\frac{499 \cdot 0.9634^2 + 499 \cdot 0.9950^2}{998}} = 0.9793$$

Comparing Means

Example: Instagram Ad Frequency

- The observed test statistic is calculated to be

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{4.9162 - 3.0518}{0.9793\sqrt{\frac{2}{500}}} = 30.1013$$

- And the p-value is calculated to be

$$P(T \geq 30.1013) = 1.84 \times 10^{-142} \cong 0$$

where $T \sim t_{(998)}$

- This probability can be calculated in R and Python (scipy.stats) using the commands

```
        1-pt(30.1013, df = 998)
   1-t.cdf(30.1013, df = 998)
```

Comparing Means

Example: Instagram Ad Frequency

- Conclusion: for any reasonable significance level α, this p-value will be smaller than it and so we choose to

  Reject H_0

- What this means: increased ad frequency significantly reduces the amount of time users engage with the app.

- Specifically, a 2 minute reduction in average session time can be expected when you move from a 4:1 ad frequency to a 7:1 frequency

Comparing Means

Example: Instagram Ad Frequency

- In fact, this whole hypothesis test is trivially carried out in R using the `t.test()` function and in Python (scipy.stats) using the `ttest_ind()`

- Let’s take a look.

Comparing Means

Power Analysis & Sample Size Calculations

Once the response variable and conditions have been chosen, the most important question when designing an A/B Test is:

"How many units do I need in each condition?"

The answer to this question depends on the frequency with which you are comfortable making the following types of errors:

- Type I Error: Reject H_0 when it is in fact true
- Type II Error: Do not reject H_0 when it is in fact false

Comparing Means

Power Analysis & Sample Size Calculations

Clearly we would like to reduce the likelihood of either type of error happening.

We define

- $\alpha = P(\text{Type I Error})$
  $= P(\text{Reject } H_0 | H_0 \text{ is true})$
- $\beta = P(\text{Type II Error})$
  $= P(\text{Do not reject } H_0 | H_0 \text{ is false})$

which reflect the chances that a Type I or Type II error will occur.

Comparing Means

Power Analysis & Sample Size Calculations

We call α the significance level of the test and $1 - \beta$ the power of the test.

Thus a test with a **small significance level** and **large power** is desirable.

Common choices are $\alpha = 0.05$ and $\beta = 0.2$ although one's own risk tolerance should dictate these choices

The goal of a power analysis is to determine the sample size necessary to fix α and β at particular values

Comparing Means

Power Analysis & Sample Size Calculations

We begin our first derivation assuming:

- $H_0: \mu_1 = \mu_2$ vs. $H_A: \mu_1 \neq \mu_2$
- σ is known or that we have a reasonable guess
- $n_1 = kn_2$

With these assumptions we use

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

as our test statistic

Comparing Means

Power Analysis & Sample Size Calculations

Why are the test statistic and null distribution important here?

We require a clear notion of what it means to reject the null hypothesis – in terms of the test statistic

We know that we reject H_0 if p-value $\leq \alpha$ and we do not reject otherwise

Can this criteria be stated in terms of the observed test statistic t?

Comparing Means

Power Analysis & Sample Size Calculations

The answer is yes and we use something called rejection regions to do this

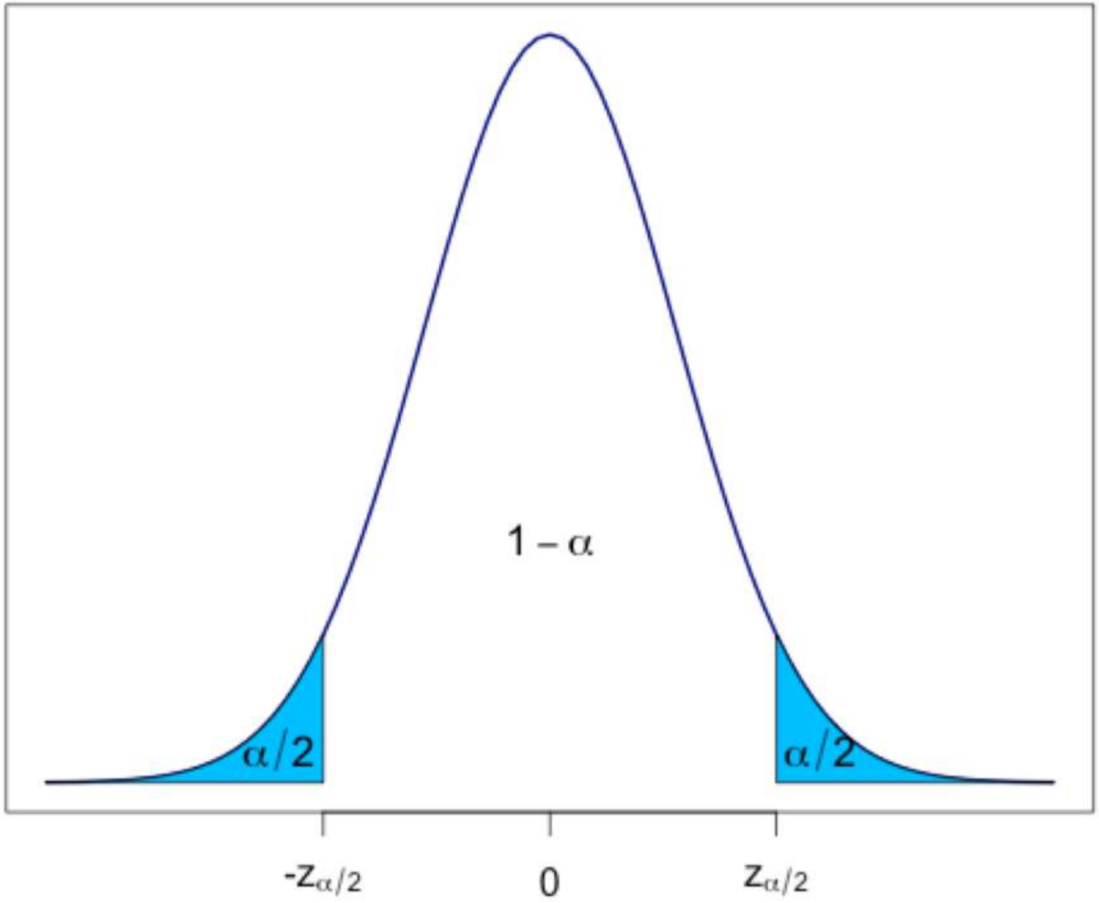
A rejection region for a given hypothesis test is the set of values of t that would lead to a rejection of the null hypothesis

Let’s make this clear by visualizing it

Comparing Means

Power Analysis & Sample Size Calculations

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$

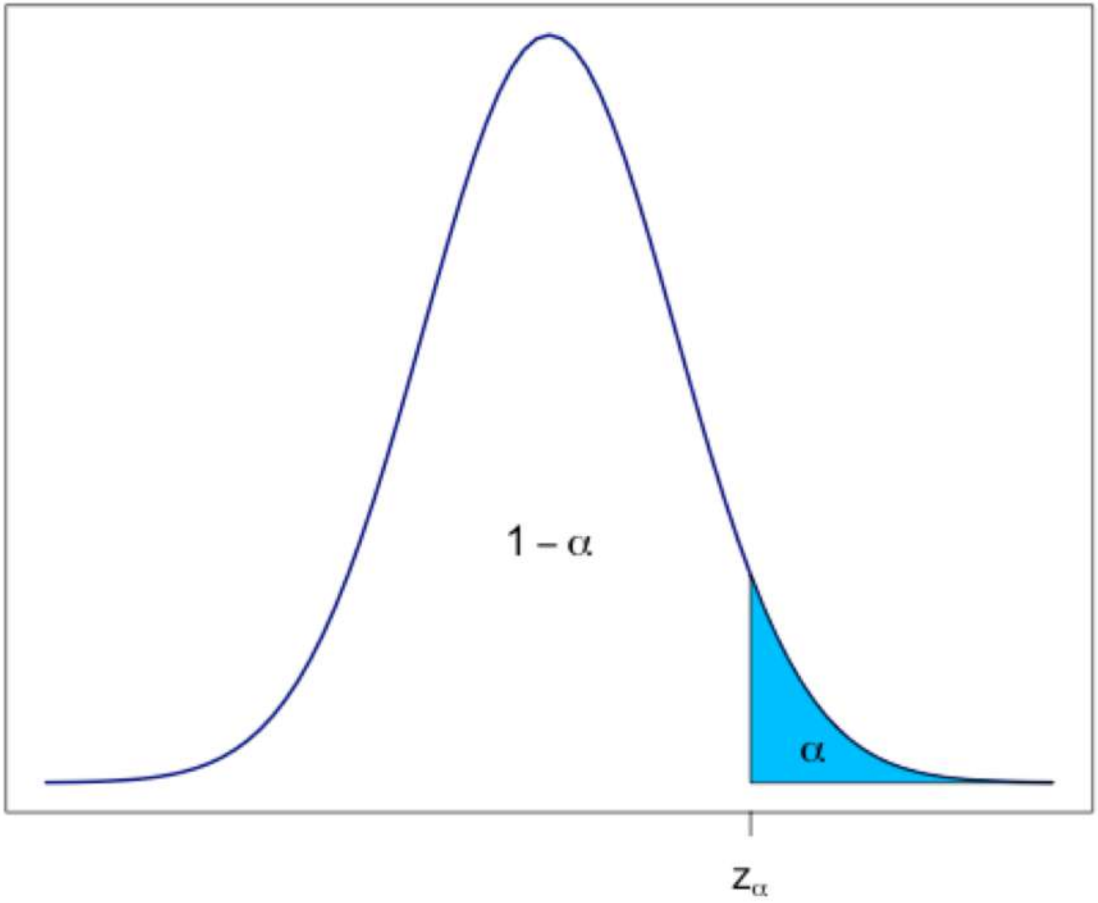


$$R = \{t | t \geq z_{\alpha/2} \text{ or } t \leq -z_{\alpha/2}\}$$

Comparing Means

Power Analysis & Sample Size Calculations

$$H_0\!: \mu_1 \leq \mu_2 \text{ vs. } H_A\!: \mu_1 > \mu_2$$

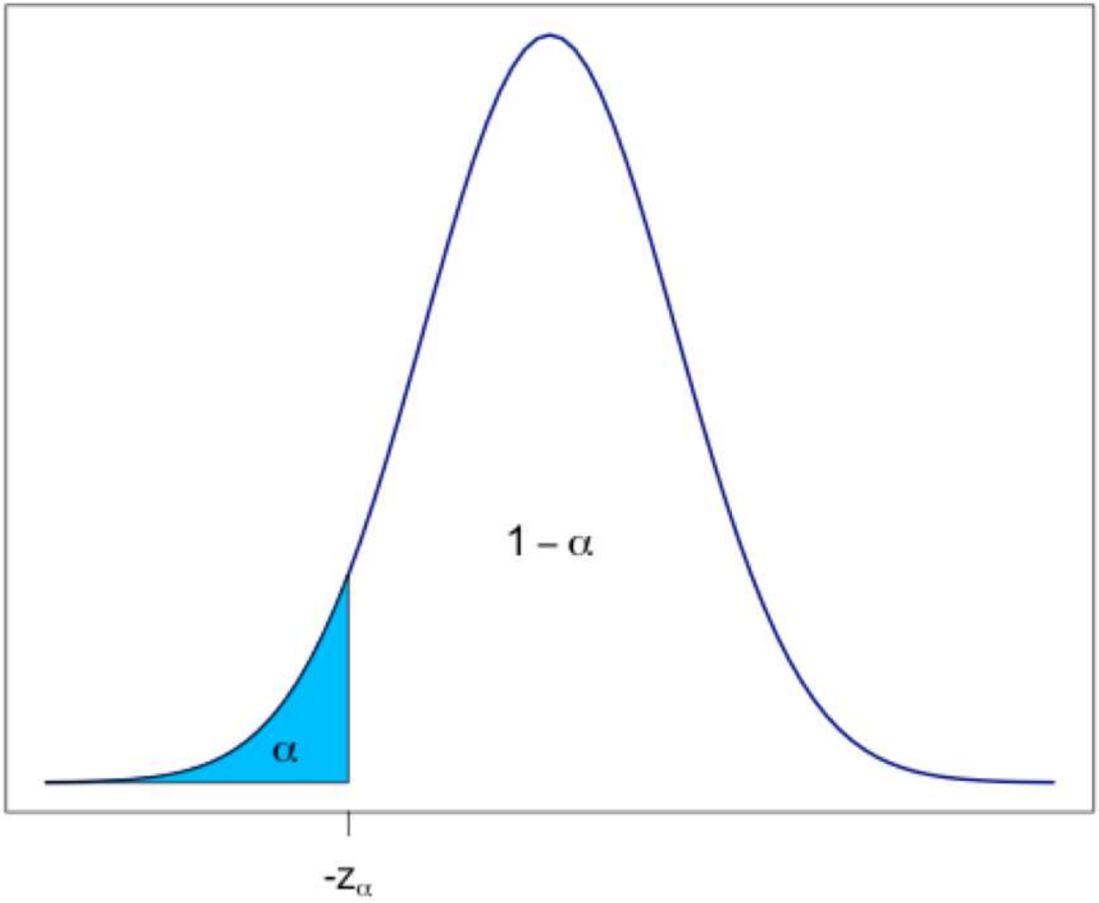


$$R = \{t | t \geq z_\alpha\}$$

Comparing Means

Power Analysis & Sample Size Calculations

$$H_0: \mu_1 \geq \mu_2 \text{ vs. } H_A: \mu_1 < \mu_2$$



$$R = \{t | t \leq -z_\alpha\}$$

Comparing Means

Power Analysis & Sample Size Calculations

$$
\begin{aligned}
1-\beta &= P(\text{Reject } H_0 | H_0 \text{ is false}) \\
&= P(T \in R | H_0 \text{ is false}) \\
&= P\left(T \geq z_{\alpha/2} \text{ or } T \leq -z_{\alpha/2} | H_0 \text{ is false}\right)
\end{aligned}
$$

$$
= P\left(T \geq z_{\alpha/2}\ | H_0 \text{ is false}\right) + P\left(T \leq -z_{\alpha/2} | H_0 \text{ is false}\right)
$$

$$
= P\left(\frac{(\bar{Y}_1 - \bar{Y}_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq z_{\alpha/2} | H_0 \text{ is false}\right) + P\left(\frac{(\bar{Y}_1 - \bar{Y}_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -z_{\alpha/2} | H_0 \text{ is false}\right)
$$

Comparing Means

Power Analysis & Sample Size Calculations

Assuming H_0 is true, then $\mu_1 - \mu_2 = 0$ and

$$\frac{(\bar{Y}_1 - \bar{Y}_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

But here H_0 is false which means $\mu_1 - \mu_2 = \delta$ for some $\delta \neq 0$, and so

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

We need to take this into account in our derivation

Comparing Means

Power Analysis & Sample Size Calculations

$$1-\beta = P\left(\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) +$$

$$P\left(\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$$

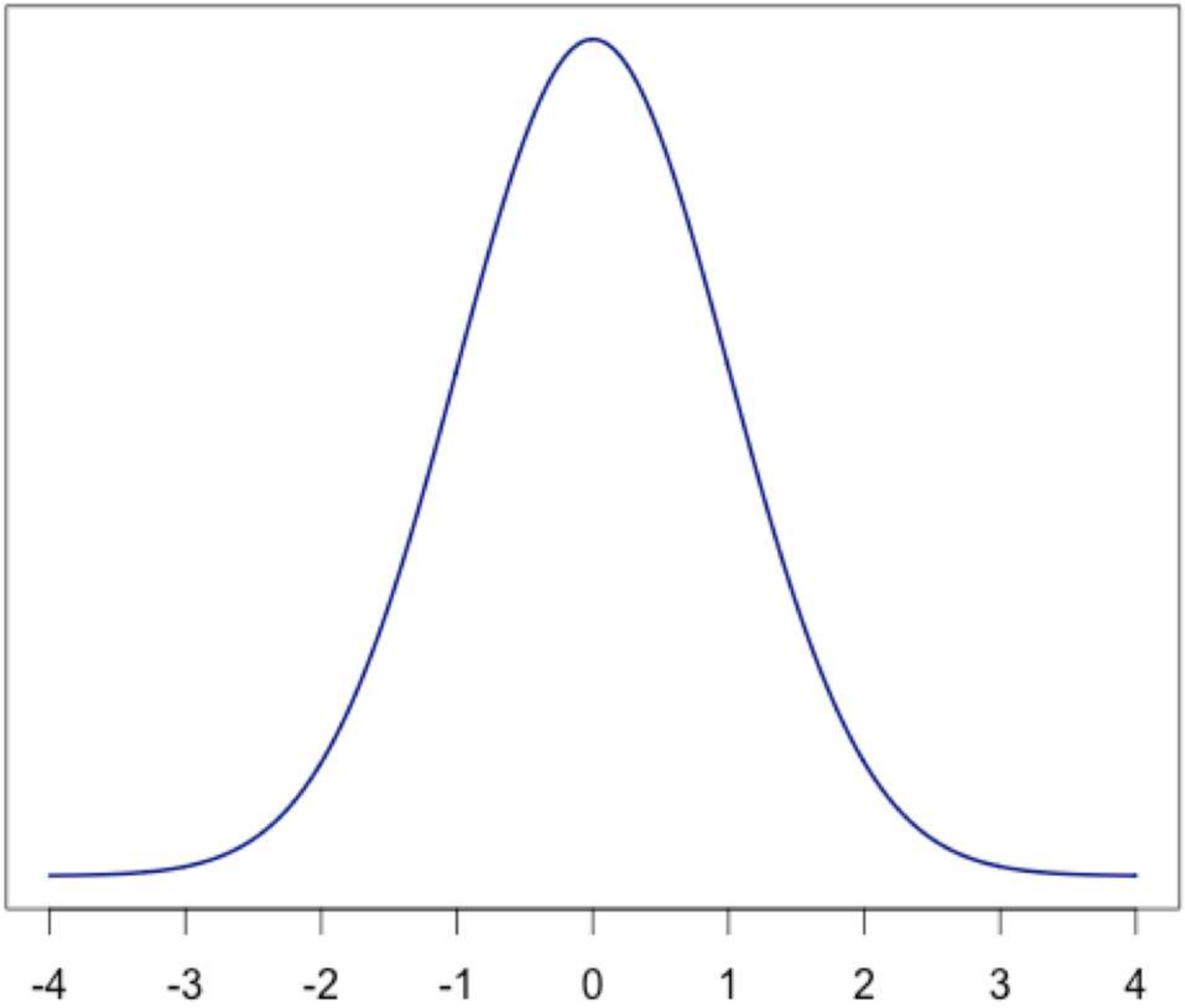
$$= P\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) +$$

$$P\left(Z \leq -z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$$

where $Z \sim N(0,1)$

Comparing Means

Power Analysis & Sample Size Calculations



Comparing Means

Power Analysis & Sample Size Calculations

Without loss of generality, assume $\delta > 0$:

$$1 - \beta = P\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)$$

But we know that $P(Z \geq z_{1-\beta}) = 1 - \beta$ and so

$$z_{1-\beta} = z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Substituting $n_1 = kn_2$ and solving for n_2 yields:

Comparing Means

Power Analysis & Sample Size Calculations

$$n_2 = \frac{\left(\frac{1}{k} + 1\right)\left(z_{\alpha/2} - z_{1-\beta}\right)^2 \sigma^2}{\delta^2}$$

Thus, to ensure the Type I and Type II error rates are fixed at α and β

- We use this formula to find out how many units we need in condition 2
- We use $n_1 = kn_2$ to find out how many units we need in condition 1

Comparing Means

Power Analysis & Sample Size Calculations

If we want equal sample sizes ($n_1 = n_2 = n$) we take $k = 1$ in the previous formula

Thus, to ensure the Type I and Type II error rates are fixed at α and β we require n units in each condition where

$$n = \frac{2\left(z_{\alpha/2} - z_{1-\beta}\right)^2 \sigma^2}{\delta^2}$$

Comparing Means

Power Analysis & Sample Size Calculations

But how do we choose δ?

- We define δ to be the effect size of the test
- The effect size of a hypothesis refers to the minimal difference between conditions that would be practically relevant and that we would like to detect as being statistically significant
- It is the answer to the question

What is the minimal difference between μ_1 and μ_2 that is practically important?

Comparing Means

Power Analysis & Sample Size Calculations

Note that sometimes δ is defined on a standardized scale:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

In which case effect size is communicated as 'numbers of standard deviations'

This approach is advantageous because it means that we do not need to plug in a value of σ when doing sample size calculations

Comparing Means

Power Analysis & Sample Size Calculations

To see this, notice that when δ is defined on a standardized scale, the sample size formulae simplify to

$$n_2 = \frac{\left(\frac{1}{k} + 1\right)\left(z_{\alpha/2} - z_{1-\beta}\right)^2}{\delta^2}$$

and

$$n = \frac{2\left(z_{\alpha/2} - z_{1-\beta}\right)^2}{\delta^2}$$

Comparing Means

Power Analysis & Sample Size Calculations

In the context of these formulae, it becomes clear that there is an interdependent relationship between

- Sample size (n_1, n_2, n)
- Significance level (α)
- Power $(1 - \beta)$
- Effect size (δ)

Let's play with a sample size calculator for an interactive demonstration of these dependencies

Comparing Proportions

Very often the response variable in an A/B test is binary, indicating whether an experimental unit did, or did not, perform some action of interest

$$Y_{ij} = \begin{cases} 1 \text{ if unit } i \text{ in condition } j \text{ does action} \\ 0 \text{ if unit } i \text{ in condition } j \text{ doesn't do action} \end{cases}$$

For $i = 1,2, \ldots, n_j, j = 1,2$

Comparing Proportions

Examples of “actions of interest” include

- Opening an email
- Clicking a button
- Watching an ad
- Leaving a webpage with no interaction

Interest lies in determining the **optimal condition**:

- the one for which the likelihood that a unit performs the action is highest/lowest

Comparing Proportions

To make this decision formally (i.e., with a hypothesis test) we must make an assumption about the distribution of the response

Because the Y_{ij}'s are binary it is common to assume that they follow a Bernoulli distribution:

$$Y_{ij} \sim BIN(1, \pi_j)$$

where π_j represents the probability that a unit in condition j performs the action of interest.

Comparing Proportions

The goal of the experiment, then, is to decide whether $\pi_1 = \pi_2, \pi_1 > \pi_2$ or $\pi_1 < \pi_2$

We do this formally by testing hypotheses of the form

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_A: \pi_1 \neq \pi_2$$

$$H_0: \pi_1 \leq \pi_2 \text{ vs. } H_A: \pi_1 > \pi_2$$

$$H_0: \pi_1 \geq \pi_2 \text{ vs. } H_A: \pi_1 < \pi_2$$

Comparing Proportions

Z-Test for Proportions

Due to the Central Limit Theorem we can say that for large n_j the following distributional result will be approximately correct

$$\bar{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \overset{\cdot}{\sim} N\left(\pi_j, \frac{\pi_j(1-\pi_j)}{n_j}\right)$$

And hence that

$$\bar{Y}_1 - \bar{Y}_2 \overset{\cdot}{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

Comparing Proportions

Z-Test for Proportions

And thus

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} N(0,1)$$

Which looks like it could be a test statistic, but we don’t know the values of π_1 and π_2 individually, even though we assume $\pi_1 - \pi_2 = 0$.

We need to do something with the denominator…

Comparing Proportions

Z-Test for Proportions

Let's replace π_1 and π_2 in the denominator with estimates.

Since we assume $\pi_1 = \pi_2$ under H_0, it makes sense to use the following pooled estimate

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2}$$

Comparing Proportions

Z-Test for Proportions

Substituting $\hat{\pi}_1 = \bar{Y}_1$ and $\hat{\pi}_2 = \bar{Y}_2$ and $\hat{\pi}$ into the previous expression yielding

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \dot{\sim} N(0, 1)$$

which still approximately follows a standard normal distribution

This is the test statistic for this test, and the null distribution is $N(0,1)$

Comparing Proportions

Z-Test for Proportions

To actually test a hypothesis we must calculate an observed value of T based on the data that have been collected through the experiment:

$$\{y_{11}, y_{21}, \dots, y_{n_1 1}\} \text{ and } \{y_{12}, y_{22}, \dots, y_{n_2 2}\}$$

Which are summarized with

$$\hat{\pi}_j = \bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

for $j = 1,2$

Comparing Proportions

Z-Test for Proportions

The observed value of the test statistic is calculated as

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

We evaluate the extremity of t in the context of the $N(0,1)$ distribution using a p-value

Comparing Proportions

Z-Test for Proportions

The p-value is calculated in exactly the same manner as before:

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_A: \pi_1 \neq \pi_2$$

- p-value = $2P(T \geq |t|\ )$

$$H_0: \pi_1 \leq \pi_2 \text{ vs. } H_A: \pi_1 > \pi_2$$

- p-value = $P(T \geq t)$

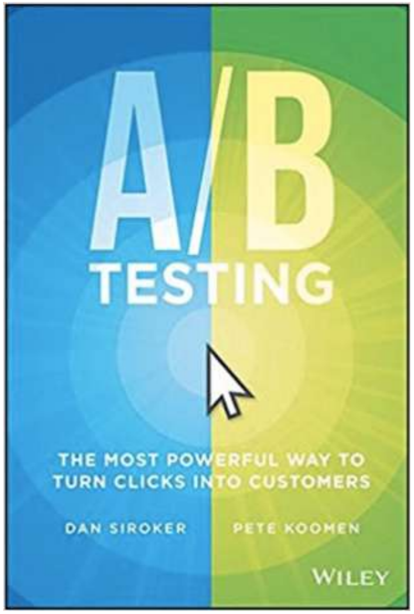
$$H_0: \pi_1 \geq \pi_2 \text{ vs. } H_A: \pi_1 < \pi_2$$

- p-value = $P(T \leq t)$

But here $T \sim N(0,1)$

Comparing Proportions

Example: Optimizing Optimizely



Siroker and Koomen describe an A/B test they ran in the midst of updating Optimizely's website

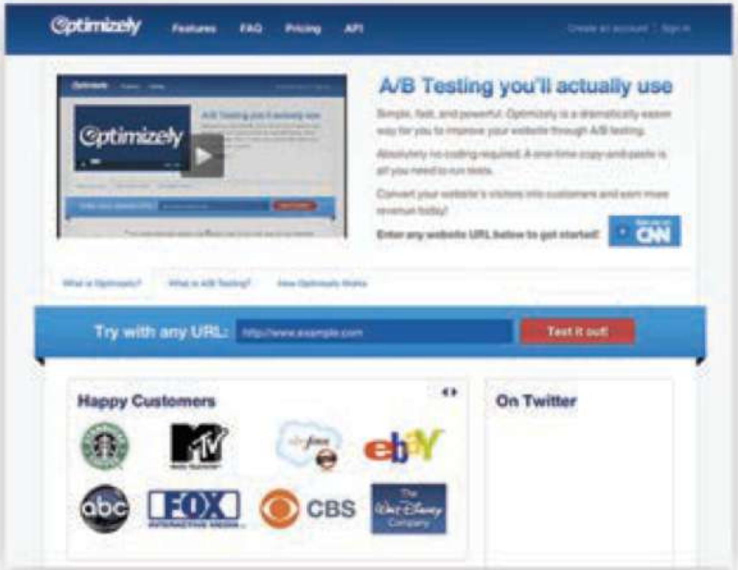
They wanted to see whether redesigned webpages influenced conversion and engagement

In particular they were interested in determining whether the redesigned homepage lead to a significant increase in the number of new accounts created

Comparing Proportions

Example: Optimizing Optimizely

ORIGINAL



A/B Testing you'll actually use

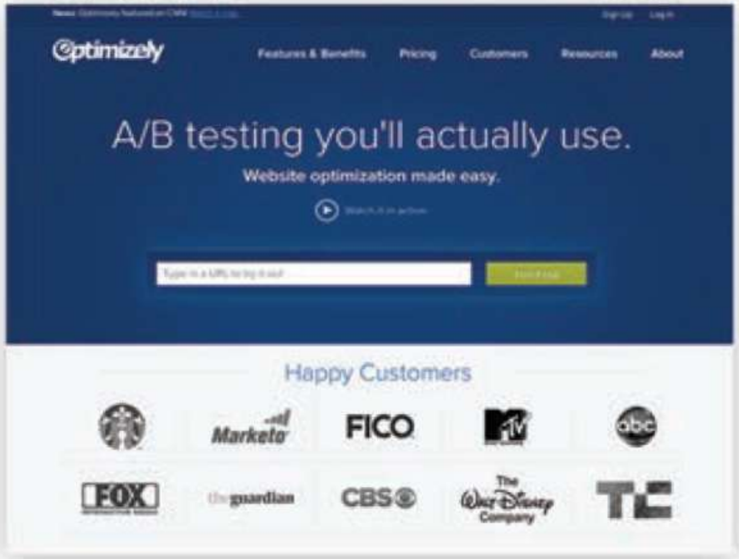
Enter any website URL below to get started!

Try with any URL: http://www.example.com

Happy Customers

On Twitter

REDESIGN



A/B testing you'll actually use.

Website optimization made easy.

Happy Customers

Comparing Proportions

Example: Optimizing Optimizely

- Their response variable was defined to be $y =$ user did or not create a new account

- They defined the two experimental conditions as
  - Condition 1 – current homepage (‘control’)
  - Condition 2 – redesigned homepage (‘redesign’)

- Interest lies in comparing π_1 and π_2 – the probabilities that a user creates a new account in the two conditions

Comparing Proportions

Example: Optimizing Optimizely

- They hypothesized that condition 2 will have a significantly larger likelihood of account creation than will condition 2

- Formally

$$H_0: \pi_1 \geq \pi_2 \text{ versus } H_A: \pi_1 < \pi_2$$

- This null hypothesis assumes that the redesigned webpage is not better than the original – so we expect to collect data that provides evidence against this statement, allowing us to reject it in favor of H_A

Comparing Proportions

Example: Optimizing Optimizely

- In order to test this hypothesis they randomized $n_1 = 8872$ users to the 'control' condition and $n_2 = 8642$ users to the 'redesign' condition
- They find that 280 users in the 'control' condition created new accounts and 399 users in the 'redesign' condition created new accounts
- This sample data is summarized as

$\hat{\pi}_1 = 280/8872 = 0.0316$

$\hat{\pi}_2 = 399/8642 = 0.0462$

$\hat{\pi} = (280+399)/(8872+8642)$
$= 0.0388$

- Users in the 'redesign' condition are 46% more likely to create accounts than 'control' users

Comparing Proportions

Example: Optimizing Optimizely

- The observed test statistic is calculated to be

$$t = \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$= \frac{0.0316 - 0.0462}{\sqrt{(0.0388)(0.9612)\left(\frac{1}{8872}+\frac{1}{8642}\right)}}$$

$$= -5.002$$

- And the p-value is calculated to be

$$P(T \leq -5.002) = 2.84 \times 10^{-7} \cong 0$$

where $T \sim N(0,1)$

Comparing Proportions

Example: Optimizing Optimizely

- Conclusion: for any reasonable significance level α, this p-value will be smaller than it and so we choose to

Reject H_0

- What this means: the redesigned homepage has a significantly larger likelihood of user sign-up than does the original homepage

- Specifically, a 46% increase in sign-ups can be expected with the redesigned homepage relative to the original.

Comparing Proportions

Example: Optimizing Optimizely

- Note that the p-value can be calculated in R using the command

```
pnorm(-5.002)
```

- And in Python (scipy.stats) the p-value can be calculated as

```
norm.cdf(-5.002)
```

Comparing Proportions

Power Analysis & Sample Size Calculations

We begin our second derivation assuming:

- $H_0: \pi_1 = \pi_2$ vs. $H_A: \pi_1 \neq \pi_2$
- $n_1 = kn_2$
- An effect size of $\delta = \pi_1 - \pi_2$ (and if we have a good guess of π_1, it means we also know π_2)

With these assumptions we use

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} N(0,1)$$

as our test statistic

Comparing Proportions

Power Analysis & Sample Size Calculations

Fortunately, because the null distribution is the same one used in the previous derivation, the rejection regions here are the same as before:

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_A: \pi_1 \neq \pi_2$$

- $R = \left\{t | t \geq z_{\alpha/2} \text{ or } t \leq -z_{\alpha/2}\right\}$

$$H_0: \pi_1 \leq \pi_2 \text{ vs. } H_A: \pi_1 > \pi_2$$

- $R = \{t | t \geq z_\alpha\}$

$$H_0: \pi_1 \geq \pi_2 \text{ vs. } H_A: \pi_1 < \pi_2$$

- $R = \{t | t \leq -z_\alpha\}$

Comparing Proportions

Power Analysis & Sample Size Calculations

$$1-\beta = P(\text{Reject } H_0 | H_0 \text{ is false})$$

$$= P(T \in R | H_0 \text{ is false})$$

$$= P\left(T \geq z_{\alpha/2} \text{ or } T \leq -z_{\alpha/2} | H_0 \text{ is false}\right)$$

$$= P\left(T \geq z_{\alpha/2} \; | H_0 \text{ is false}\right) + P\left(T \leq -z_{\alpha/2} | H_0 \text{ is false}\right)$$

$$= P\left(\frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \geq z_{\alpha/2} | H_0 \text{ is false}\right) + P\left(\frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \leq -z_{\alpha/2} | H_0 \text{ is false}\right)$$

Comparing Proportions

Power Analysis & Sample Size Calculations

Assuming H_0 is true, then $\pi_1 - \pi_2 = 0$ and

$$\frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} N(0,1)$$

But here H_0 is false which means $\pi_1 - \pi_2 = \delta$ for some $\delta \neq 0$, and so

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} N(0,1)$$

We need to take this into account in our derivation

Comparing Proportions

Power Analysis & Sample Size Calculations

$$1-\beta = P\left(\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \geq z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right) +$$

$$P\left(\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \leq -z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right)$$

$$= P\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right) +$$

$$P\left(Z \leq -z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right)$$

where $Z \sim N(0,1)$

Comparing Proportions

Power Analysis & Sample Size Calculations

Without loss of generality, assume $\delta > 0$:

$$1 - \beta = P\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}\right)$$

But we know that $P(Z \geq z_{1-\beta}) = 1 - \beta$ and so

$$z_{1-\beta} = z_{\alpha/2} - \delta \Big/ \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

Substituting $n_1 = kn_2$ and solving for n_2 yields:

Comparing Proportions

Power Analysis & Sample Size Calculations

$$n_2 = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2 \left[\frac{\pi_1(1-\pi_1)}{k} + \pi_2(1-\pi_2)\right]}{\delta^2}$$

Thus, to ensure the Type I and Type II error rates are fixed at α and β

- We use this formula to find out how many units we need in condition 2
- We use $n_1 = kn_2$ to find out how many units we need in condition 1

Comparing Proportions

Power Analysis & Sample Size Calculations

If we want equal sample sizes ($n_1 = n_2 = n$) we take $k = 1$ in the previous formula

Thus, to ensure the Type I and Type II error rates are fixed at α and β we require n units in each condition where

$$n = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2 [\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\delta^2}$$

Comparing Proportions

Power Analysis & Sample Size Calculations

Again we see that there is an interdependent relationship between

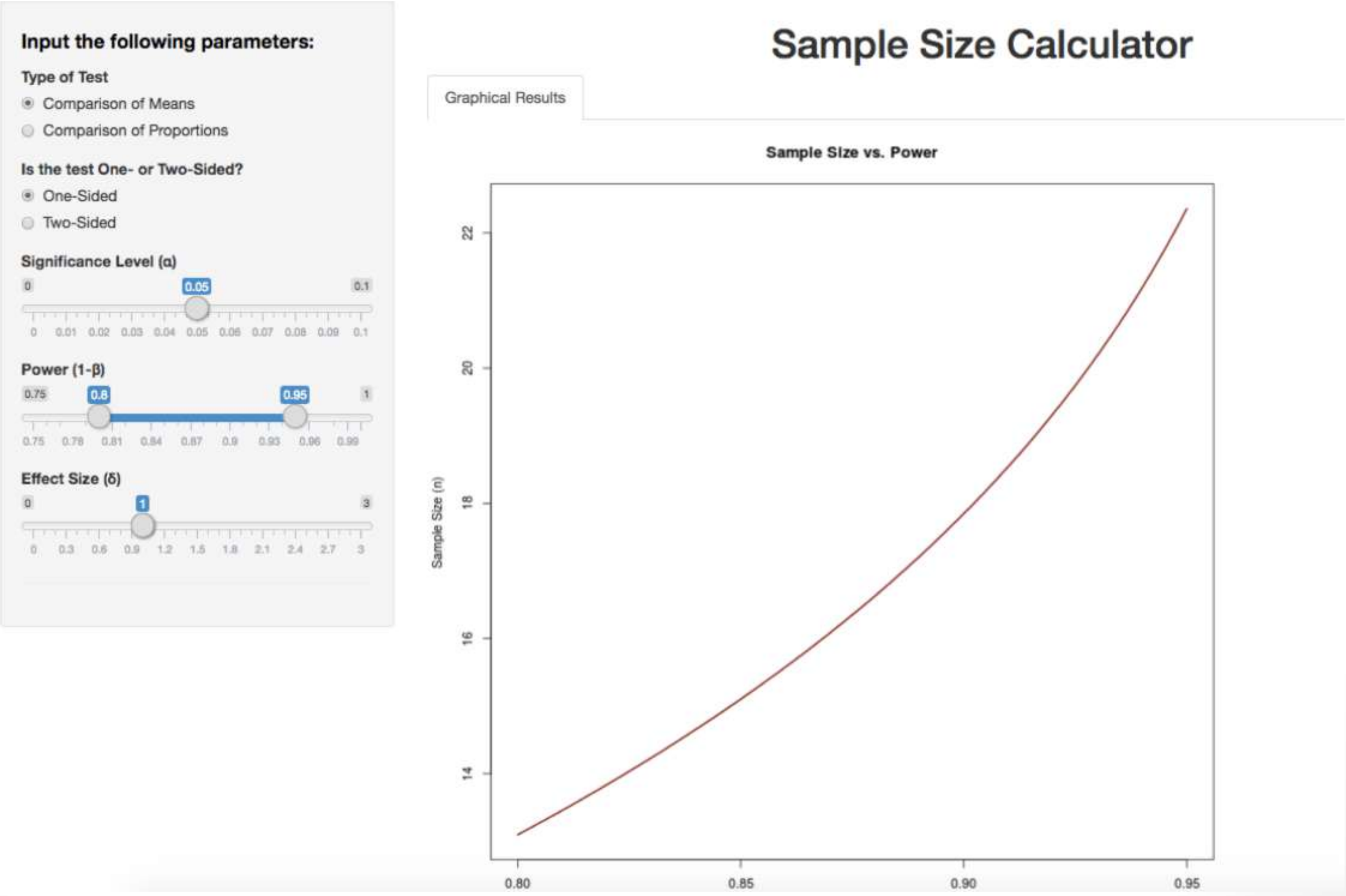
- Sample size (n_1, n_2, n)
- Significance level (α)
- Power $(1 - \beta)$
- Effect size (δ)

This is true of **every** hypothesis test

We can use the same sample size calculator to explore these interdependencies in this setting as well.

Take Home Task

Play with this app:



Input the following parameters:

Type of Test

● Comparison of Means

○ Comparison of Proportions

Is the test One- or Two-Sided?

● One-Sided

○ Two-Sided

Significance Level (α)

0 — 0.05 — 0.1

Power (1-β)

0.75 — 0.8 — 0.95 — 1

Effect Size (δ)

0 — 1 — 3

Sample Size Calculator

Graphical Results

Sample Size vs. Power

See you next week!