How do I estimate my sample size?

Chelsea Green
October 1, 2021

Resources for this slideshow

Blair, Edward and Johnny Blair. *Applied Survey Sampling*. Los Angeles: Sage (2015).

Check out Chapter 4 on estimating sample size, which includes information about the confidence interval and hypothesis testing methods.

How do I establish my sample size?

Math ≠

How do I establish my sample size?

Math is for everyone.

How do I establish my sample size?

What are you trying to explain?

1. Estimate a quantity of interest (confidence interval approach)
   - What is the % of people that believe X?
2. Test whether a difference exists between two groups' sample means (hypothesis test approach)
   - Great for experiments, comparing treatment v. control groups
3. Evaluating whether a relationship exists between two variables (regression)
   - If you're interested in evaluating whether one variable predicts your quantity of interest *more* than another, then you're using multiple linear regression!

How do I establish my sample size?

What is a **sample**?

A subset of the target population.

N= the number of elements in your target population

n= the number of elements in your sample

How do I establish my sample size?

What is a **sample mean** or **sample proportion**?

The key statistic you're interested identifying: the average or percentage of a specific quantity of interest in your sample.

- Example 1. 45% of individuals in your sample held favorable views toward China.
- Example 2. Large firms in your sample had an average of 15 employees taking 6 weeks or more of family medical leave.

How do I establish my sample size?

What is **sampling error?**

The standard deviation of the distribution of sample means (or proportions) across samples of a particular size.

...But what is the standard deviation again?

How do I establish my sample size?

If I were able to sample the Harvard population ten times, I might get ten different mean proportions of those that believe climate change is "highly concerning."

40% 44% 45% 47% 53% 55% 55% 56% 57% 70%

The standard deviation of these proportions is my sampling error.

σ = standard deviation (for our purposes, the sampling error)

How do I establish my sample size?

But Chelsea…

I'm only going to end up taking one sample, so how can I determine my sample error?

How do I establish my sample size?

Two methods for estimating sample error:

1. If you are interested in absolute numbers as sample means (eg. 15 employees, $20), either:
   a. Estimate the number based on what previous studies use.
   b. Do a pre-test of your survey to estimate sample error.
2. If you are interested in percentages/proportions **(π)** as sample means (eg. 47% of students, 50% of employees), then use **50% (.5 for the formula) when calculating your sample error**

How do I establish my sample size?

What is a **confidence interval?**

A estimated range of values which is likely, at some threshold that we set, to contain our population statistic (eg. population mean)

How do I establish my sample size?

Establish your confidence interval. Then use that information to determine what size your sample should be!

For example, if your sampling procedure is unbiased, then there is a 95% chance that any given sample mean **will fall within ± *some interval*** of the true population value.

You pick the most relevant interval for your subject and your audience.

How do I establish my sample size?

You might want to be able to say...

- There is a 95% chance that this sample % of students that consider climate change "highly concerning" **falls within ± *5%*** of the true % of Harvard students that believe climate change is "highly concerning."
- There is a 95% chance that this sample mean number of employees within large firms taking family medical leave **falls within ± *5 employees*** of the true number of employees in all large firms taking family medical leave.

$\mathbf{I_{95\%}}$= your chosen interval quantity

How do I establish my sample size?

Throughout this presentation, we will assume that we want to establish a **95% confidence interval.** (ie. There is a **95%** chance that our sample mean captures the true population sample mean.)

This is pretty standard in social science, at least for now.

How do I establish my sample size?

If our sample **less than 10% of the population** and your quantity of interest is a **number**, then we use this formula:

$$n = \left(\frac{1.96 * \sigma}{I_{95\%}} \right)^2$$

How do I establish my sample size?

If our sample **less than 10% of the population** and your quantity of interest is a **percentage,** then we use this formula:

$$n = \left(\frac{1.96 * \sqrt{\pi(1 - \pi)}}{I_{95\%}} \right)^2$$

How do I establish my sample size?

If our sample **more than 10% of the population** and your quantity of interest is a **number**, then we use this formula:

$$n = \frac{\sigma^2 * \left(\frac{N}{N-1}\right)}{\left(\left(\frac{I_{95\%}}{1.96}\right)^2 + \frac{\sigma^2}{N-1}\right)}$$

How do I establish my sample size?

If our sample **more than 10% of the population** and your quantity of interest is a **percentage**, then we use this formula:

$$n = \frac{\pi(1-\pi) * \left(\frac{N}{N-1}\right)}{\left(\left(\frac{I_{95\%}}{1.96}\right)^2 + \frac{\pi(1-\pi)}{N-1}\right)}$$

How do I establish my sample size?

So that's the confidence interval approach (for those of you interested in estimating a quantity of interest). Next, the hypothesis testing approach!

Use this approach when you're testing whether there is **a significant difference in sample means between groups.**

Eg. Harvard students that have taken environmental classes are more likely to believe that climate change is 'highly concerning' than those that have not.

How do I establish my sample size?

Here is the formula that Blair and Blair give us for estimating sample size for our two groups ($\mathbf{n}_1$ and $\mathbf{n}_2$), where $\mathbf{p}_1$ and $\mathbf{p}_2$ are the *sample mean percentages* for groups 1 and 2:

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}}$$

How do I establish my sample size?

We are solving for (**n**$_1$ and **n**$_2$).

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}}$$

How do I establish my sample size?

First, we should assume for sake of having a very conservative sample error that **p_1** and **p_2 in our denominator each equal .5. Put aside the numerator for now.**

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}}$$

How do I establish my sample size?

Let's clean this up a little bit.

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{.5\,(1-.5)}{n_1}\right) + \left(\frac{.5\,(1-.5)}{n_2}\right)}}$$

How do I establish my sample size?

Almost there!

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{.5\,(.5)}{n_1}\right) + \left(\frac{.5\,(.5)}{n_2}\right)}}$$

How do I establish my sample size?

Phew, that looks better! Now, let's assume that we are going to select samples for each group (**n_1** and **n_2**) that are the **same size**. That will allow us to solve for a new value that will make our lives easier: **n.**

$$t = \frac{\mathrm{p}_1 - \mathrm{p}_2}{\sqrt{\left(\frac{.25}{n_1}\right) + \left(\frac{.25}{n_2}\right)}}$$

How do I establish my sample size?

Perfect! Now we can add those two values in the denominator together:

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{.25}{n}\right) + \left(\frac{.25}{n}\right)}}$$

How do I establish my sample size?

That looks way more simple!

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{.5}{\mathbf{n}}\right)}}$$

How do I establish my sample size?

If we use the standard employed widely through social science for establishing whether a value is statistically significant, we can plug in the value of **t** (again assuming we are using that standard 95% confidence approach used across social science):

$$t = \frac{p_1 - p_2}{\sqrt{\left(\frac{.5}{\mathbf{n}}\right)}}$$

How do I establish my sample size?

Done! Now, we are going to get **n** by itself and out of that square root symbol:

$$1.96 = \frac{p_1 - p_2}{\sqrt{\left(\frac{.5}{\mathbf{n}}\right)}}$$

How do I establish my sample size?

$$1.96\sqrt{\left(\frac{.5}{n}\right)} = p_1 - p_2$$

How do I establish my sample size?

$$1.96^2\left(\frac{.5}{n}\right) = (p_1 - p_2)^2$$

How do I establish my sample size?

$$1.96^2 * .5 = (p_1 - p_2)^2 * n$$

How do I establish my sample size?

Wow, now that is **exactly** what we were looking for!

$$n = \frac{.5(1.96)^2}{(p_1 - p_2)^2}$$

How do I establish my sample size?

So finally, what do we do with **$\mathbf{p_1}$** and **$\mathbf{p_2}$?**

$$n = \frac{.5(1.96)^2}{(p_1 - p_2)^2}$$

How do I establish my sample size?

Remembering that $\mathbf{p_1}$ and $\mathbf{p_2}$ are the *sample mean percentages* for group 1 and group 2, **plug in the difference that you hope to meaningfully detect.**

$$n = \frac{.5(1.96)^2}{(p_1 - p_2)^2}$$

That difference could be 5% (.05), 10% (.1)...your choice.

Are you still with me?

How do I establish my sample size?

If you're evaluating whether a relationship exists between two variables (regression), or whether one variable predicts your quantity of interest *more* than another (multiple linear regression), you need to establish several components to estimate your sample size.

Check out this awesome regression sample size calculator for linear regression. This is a fantastic resource for those not yet well-versed in the tricky stats, but still want to get started with estimating sample size.

How do I establish my sample size?

Effect size: the size of the relationship between X and Y.

These range from 0 to 1. Most of you using regressions will use an F test to determine effect sizes.

When using the calculator, work with the preset small, medium, and large effect sizes. Select based on the effect that you anticipate. For those of you hoping to more conservatively estimate your sample sizes, select "small."

How do I establish my sample size?

Number of predictors: how many independent variables are you including in your analysis?

Eg. Estimating whether party identification, taking an environmental class, and income level influences belief that climate change is "highly concerning." I have three independent variables!

How do I establish my sample size?

What is **power**?

The probability of accurately accepting the null hypothesis if it's true.

Power: The standard is typically .8 (or .9 for those of you wanting to be more ambitious).

How do I establish my sample size?

What is Alpha (**α**)?

The probability of mistakenly rejecting the null hypothesis if it's true.

Alpha (α): .05 is widely used across studies in social science. (Hint: **α** is simply equal to 1 minus your confidence level! 1 minus .95, our preferred confidence level, equals .05.

In-class activity

What are you trying to explain?

Talk about your dependent variables and begin to consider which method discussed today you might utilize to establish your sample size. For those of you with less defined dependent variables, chat with your classmates about how you're thinking about making them measurable and concrete.