**How Big Should Your Sample Size Be?**

**A handy little formula that every data scientist should know**

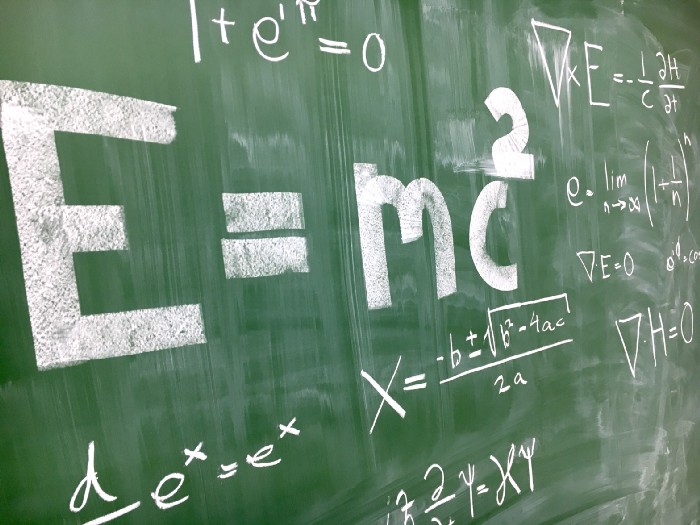


Photo by [Artturi Jalli](https://unsplash.com/@artturijalli?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

You are in a meeting with the CEO and the executive team.

The Product team is pitching a new product idea. It has lots of cool features and will work smoothly with the company’s current products. They think existing customers will love it. They think it is a great way to increase share of wallet from the customer base. They are asking the CEO for budget to build the product.

The Sales team isn’t so sure. They think there’s no burning reason for an existing customer to buy the new product —they think it is a vitamin, not a painkiller. They think most customers won’t buy. They think it will be a waste of money to build it.

They all look to the CEO but she is not ready to decide. She is very data-savvy. She wants to survey a *random* *sample* of existing customers and see what % of them want to buy the new product\*.

She turns to you. *Can you run the numbers and let me know by end of day how many customers we should survey? I want the margin of error to be plus-minus 3 percentage points.*

You reply at lightning speed: *About 1100 customers*.

She is dazzled by the speed of your response. *Wow, that was fast.*

You try to look modest :-).

She continues. *Surveying 1100 customers will be too expensive. Let’s say that I am OK with a margin of error of plus-minus 5 percentage points. How many customers then?*

Another instant response from you: *400 customers*.

*That’s better.* She turns to the Marketing team. *Let’s do a random survey of 400 customers and meet again when the results are in.*

The CEO leaves. The meeting ends.

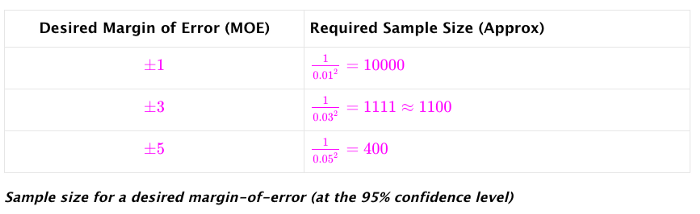
People gather around you. *How did you do that?*

Simple. This is the handy little formula ([source](https://en.wikipedia.org/wiki/Sample_size_determination#Estimation_of_a_proportion)).

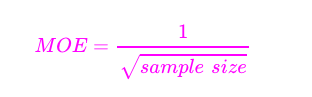
https://miro.medium.com/max/346/1*FYN3b6ZivI3RJBJcbswHNg.png

With this, you can use your phone calculator to rapidly find the approximate sample size for any margin-of-error.

***But some margin-of-error numbers — 1%, 3% and 5% — are so common in the business world that you should memorize the results***. After all, appearing to do lightning-fast mental math is more impressive than taking out your phone ;-).



And there’s one more thing. You can flip the formula around too.



Now you can handle the ‘reverse’ question: *I can only afford to survey 250 customers. What is the margin of error in that case?*

No problem. Pull out your phone’s calculator and calculate 1 / (square root of 250) which is 0.063 (i.e., an MOE of plus-minus 6.3).

Nice, huh?

That’s it. Go forth and impress! :-)

***Where does this formula come from?***

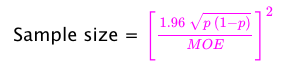
Let’s start with an age-old stats question.

*How many people should we randomly sample so that we can estimate a population proportion within a margin-of-error of plus-minus MOE with 95% confidence\*\*?*

In our example,

* the population = all existing customers
* population proportion = % of all existing customers who want to buy the new product
* MOE = recall that the CEO asked for 3% margin-of-error first and later 5%

The age-old question above has an age-old [answer](https://en.wikipedia.org/wiki/Sample_size_determination#Estimation_of_a_proportion):



*p* is the *sample proportion* — the % of people **surveyed** who say they will buy the product.

But we haven’t done the survey yet so we don’t know *p (*in fact, the whole point here is to figure out *p*, right?). So what will we plug into the formula for *p*?

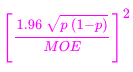
A little math comes to our rescue.

Turns out, for *p* between 0 and 1*,* this is always true:

https://miro.medium.com/max/186/1*UcFEI64hT8oXjIVwEtm27A.png

(you can plot it and confirm for yourself or just do a little calculus and prove it)

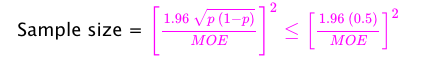
Since the sample size



is **at most**

https://miro.medium.com/max/90/1*Lcllfu-0Kvw5PsGwnbT-vg.png

the original formula can be approximated as:



Note that since there is no *p* in

https://miro.medium.com/max/90/1*Lcllfu-0Kvw5PsGwnbT-vg.png

we have solved the “I don’t know what to plug in for *p*” problem.

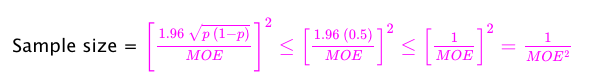
Next, notice that 1.96 (0.5) = 0.98 is just slightly less than 1.0. So this

https://miro.medium.com/max/90/1*Lcllfu-0Kvw5PsGwnbT-vg.png

can be approximated by



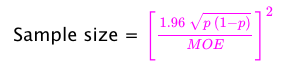
Putting it all together:



We are done!

*Important caveat*: Because of the simplifications we are making, this formula is ***conservative.*** It will recommend a more-than-needed sample size if the true value of *p* is far from 0.5.

If you can guess at the likely value of *p* from previous similar surveys, you can use that value of *p* in this formula instead.



\*Yes, she is aware that what people *tell* you in surveys may not be what they will actually *do.* Here’s a fun recent example ([source](https://medium.com/magnetic/52-things-i-learned-in-2022-db5fcd4aea6e)):

Researchers asked 100 people whether a reasonable person would unlock their phone and give it to an experimenter to search through. Most said no. Then the researchers asked 103 other people to unlock their phone and give it to them. 100 of them complied.

\*\*The confidence level doesn’t have to be 95% but it is used for two reasons.

* It is very common in the business world
* Conveniently for us, the associated Z-score of 1.96 nicely multiplies the 0.5 to give us a number close to 1.0, which leads to the super-simple formula for the sample size.

To use a different confidence level, you will need to replace 1.96 with other values in the formula above. [More](https://en.wikipedia.org/wiki/Sample_size_determination#Estimation_of_a_proportion).

If you liked this article, you may like these too (I realize I am sounding like a recommendation engine :-)).

[I have data. I need insights. Where do I start?](https://medium.com/p/7ddc935ab365)

[Create a Common-Sense Baseline First](https://medium.com/p/e66dbf8a8a47)

[How to Use Causal Inference In Day-to-Day Analytical Work](https://medium.com/p/b5efbdbf8ab0)

[Lessons from a Deep Learning Master](https://medium.com/p/1e38404dd2d5)

[From Prediction to Action — How to Learn Optimal Policies from Data](https://medium.com/p/1edbfdcb725d)