Minimum Sample Size & Sufficient Statistical Significance

How many samples are needed for a good model?

G.G.Ashbrook 2022-2023

How many samples are needed for a good model? Or how many samples does it take to model and represent a large population?

The Short Version

How many data points or 'samples' are needed for a good model?

139 samples **could** be ok.
385 samples **should** be ok.
500 samples **will** be ok.

The Longer Version

See: https://www.calculator.net/sample-size-calculator.html , and more calculator links below.

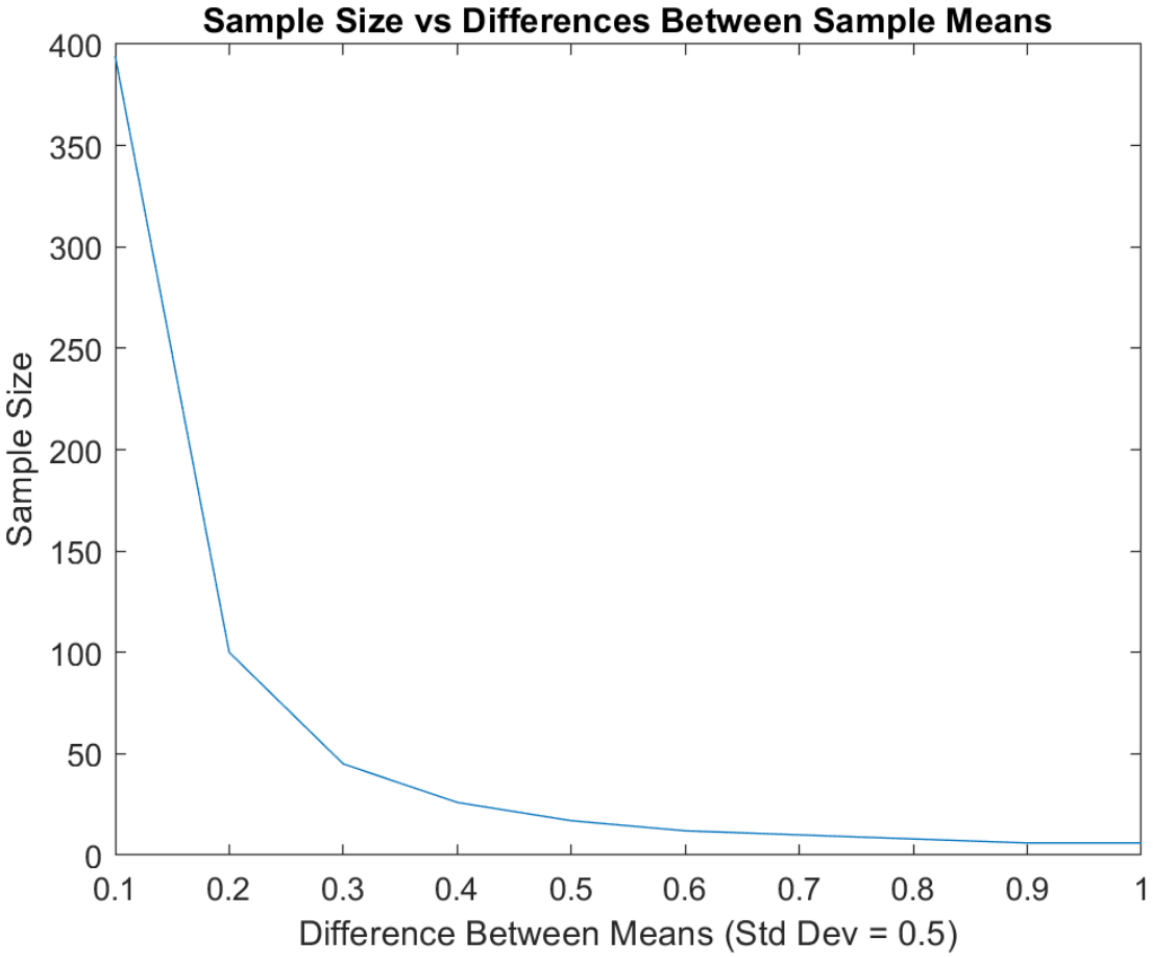
According to articles and the tool above, 385 data points is the number you need for the standard very-high 95% confidence level, and a 5% margin of error, for large data sets (or populations) including both OSHA data ~200,000 rows, and EPA data with millions of rows. But if you drop the confidence requirement to 90% and raise the margin of error a bit to 7%, the required number is only 139 data points.

Note: Machine Learning is **not** always the same as these statistical measures, but this may be a fair general guide. Once you have made a machine learning model you can run many tests on it to evaluate precision, accuracy, false-negative, false-positives, etc. See 'Confusion Matrix' for one set of tools for classification models (or you may need to pick the 'loss function' before training). Some model-methods work better for smaller data sets, other models are excellent but require huge datasets. Each data set is different too. If the pattern is clear, maybe you need just 20 samples, but if the data are very messy you might need more. You do not know for sure what a model will show until you actually make the model and test it out. This is sometimes called the 'no free lunch hypothesis' in data science, you can't say what the models will do before you actually do the modeling.

In general, the more data you train with then the better the model will be (especially neural Networks).

Note: You may also want to look into methods of removing outliers such as IQR-1.5 for standard regression type models, or methods such as K-fold cross-validation for eliminating ambiguous data in neural network training, again, especially when your dataset is smaller and can be negatively biased by unusual data points or samples.

As a visualization (from article linked below) see the chart below, though the details will get a bit into the weeds as to why 'differences between sample means' is used as the X axis. To simplify: If you are trying to show the difference between two things (e.g. that one business choice is better) the greater the variation in the data then the fewer samples you need to show a significant difference. This graph shows that even if the variations between samples is tiny, you are still able to show significant patterns as you get closer to 400 samples.

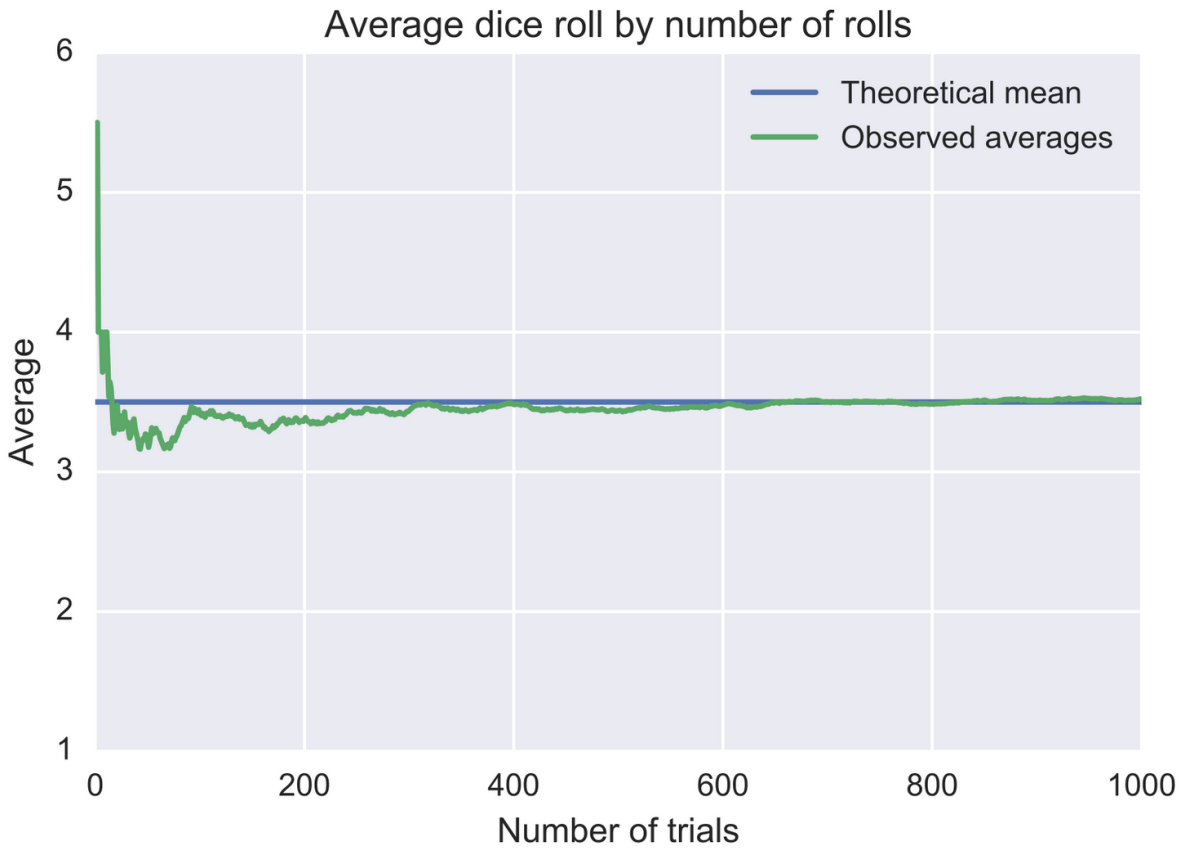


A lot of the discussion about sample size is when you have a tiny sample size and you want to know if it is good enough, so you can get into the weeds of analyzing how much you can do with your 20 or 50 or 60 samples, which is very case-by-case. As your sample size gets larger you can be more confident in patterns not being due to random noise in the samples, or other problems with AI-ML such as over-fitting your model (also see the Manifold Hypothesis).

For another visualization showing the same pattern around 400 samples, here is a figure from the Wikipedia article on The Law of Large Numbers. Again, this is a different specific case, but the overall theme is the same: once you get past 400 samples, the chance of random variation disrupting the overall results becomes very small.

The article says:

> *In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value and tends to become closer to the expected value as more trials are performed.[1]*
>
> *The LLN is important because it guarantees stable long-term results for the averages of some random events.*



Links

Standard Sample Size Calculator
https://www.calculator.net/sample-size-calculator.html

Simpler Size Calculator:
https://www.surveymonkey.com/mp/sample-size-calculator/

More Technical Calculators:
https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

https://www.gigacalculator.com/calculators/power-sample-size-calculator.php

Article:
https://www.designreview.byu.edu/collections/how-many-samples-do-i-need-determining-sample-size-for-statistically-significant-results

https://en.wikipedia.org/wiki/Law_of_large_numbers

More Articles:
https://blogs.nature.com/methagora/2013/08/giving_statistics_the_attention_it_deserves.html

https://en.wikipedia.org/wiki/Manifold

https://machinelearningmastery.com/k-fold-cross-validation/

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

https://en.wikipedia.org/wiki/Confusion_matrix