The Central Limit Theorem and its Implications for Statistical Inference

Jeremy Albright

Posted on Mar 1, 2019

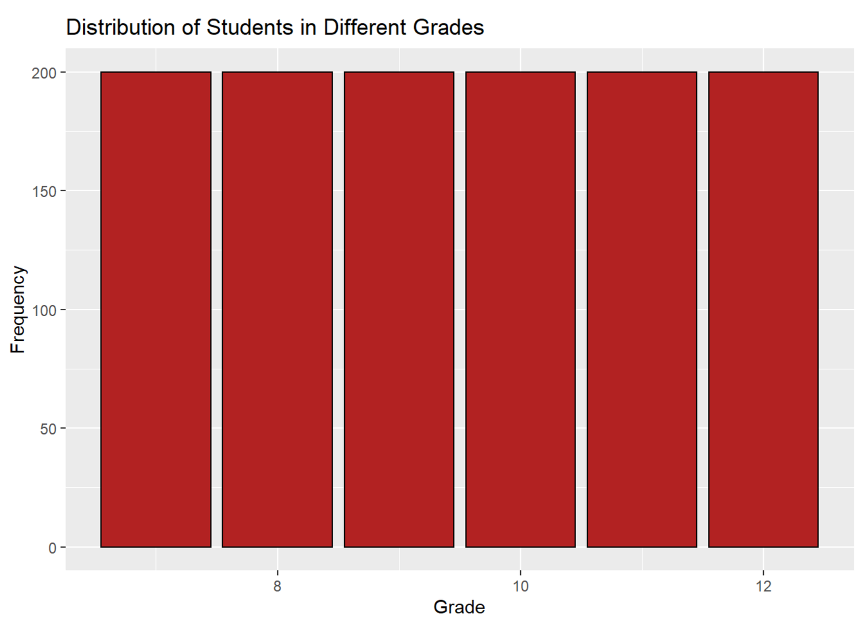
#central limit theorem #logic

Logic

The central limit theorem is perhaps the most fundamental result in all of statistics. It allows us to understand the behavior of estimates across repeated sampling and thereby conclude if a result from a given sample can be declared to be "statistically significant," that is, different from some null hypothesized value. This brief tutorial explains what the central theorem tells us and why the result is important for statistical inference.

The central limit theorem tells us exactly what the shape of the distribution of means will be when we draw repeated samples from a given population. Specifically, as the sample sizes get larger, the distribution of means calculated from repeated sampling will approach normality. What makes the central limit theorem so remarkable is that this result holds no matter what shape the original population distribution may have been.

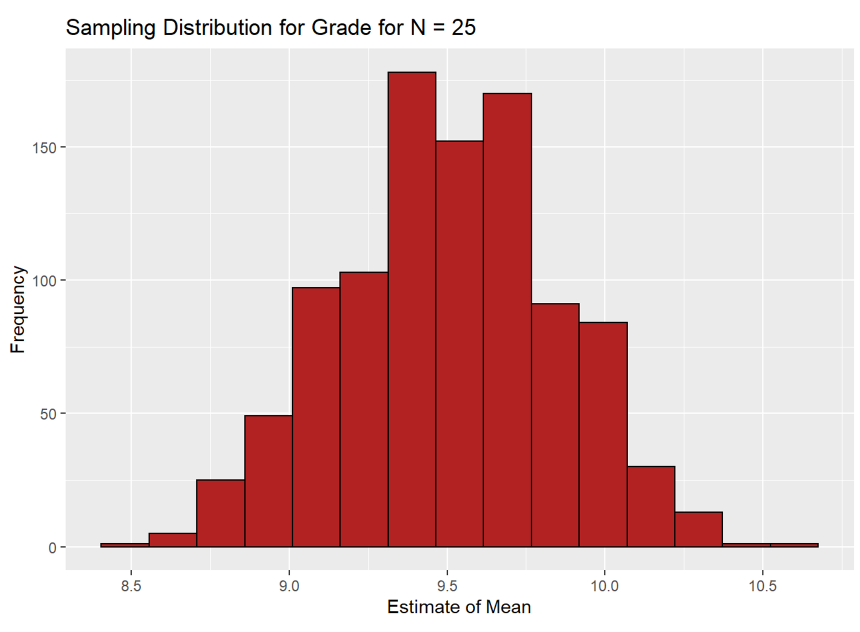
As an example, say that we find a school that has 1200 students, with exactly 200 students each in grades 7 through 12. The population distribution is, as the following figure shows, definitely not normal.



Say that we take a sample of 25 students and calculate the mean grade level for the sample, which we find to be 9.2. We then take another sample and find that its mean is 9.32. By the nature of random sampling, we will get a slightly different result each time we take a new sample. For example, the following table shows the mean we get from 10 separate samples, each of size n = 25, drawn from the population.

Sample(n=25)	*Average Grade*
1	10.12
2	9.20
3	9.68
4	9.52
5	9.16
6	9.20
7	9.88
8	9.12
9	9.36
10	8.84

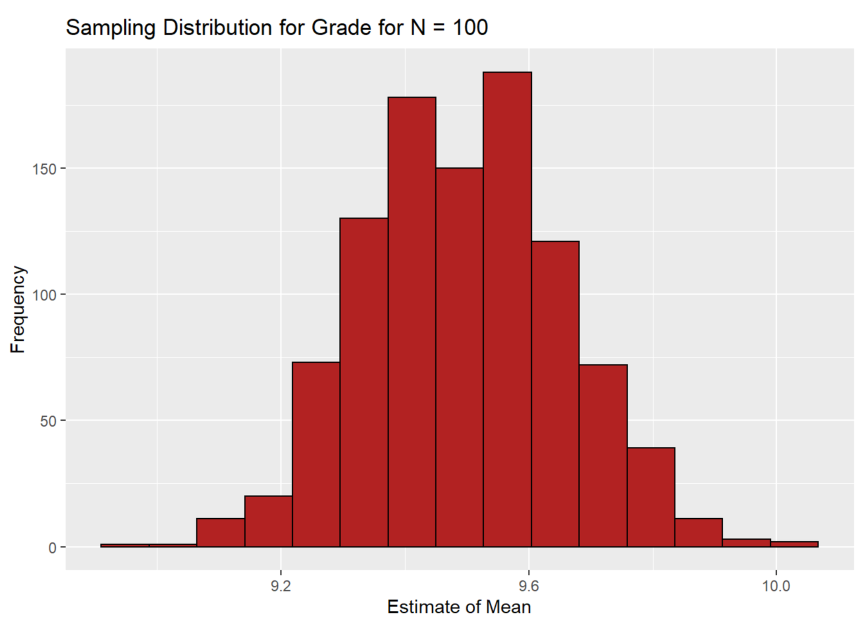
If we keep taking samples and calculating the mean each time, these means will begin to form their own distribution. We call this distribution of means the sampling distribution because it represents the distribution of estimates across repeated samples. In the case of this example, a histogram of sample means across 1,000 samples would look like the following.



Notice that the shape of the distribution looks something like a normal distribution, despite the fact that the original distribution was uniform! Regardless of the shape of the original population's distribution, the distribution of sample means calculated from that population will tend towards normal as n increases.

The central limit theorem therefore tells us that the shape of the sampling distribution of means will be normal, but what about the mean and variance of this distribution? It is easy to show (if you know the algebra of expectations and covariances) that the mean of this sampling distribution will be the population mean, and that the variance will be equal to the population variance divided by n. If we take the square root of the variance, we get the standard deviation of the sampling distribution, which we call the standard error. This information together tells us that the mean of the sample means will be equal to the population means, and the variance will get smaller when 1) the population variance gets smaller, or 2) the sample sizes get larger.

The second of these results has an easy intuition. As our samples get larger, we have more information about the population, and hence we should expect less sample-to-sample variation. Compare the following distribution of means from 1,000 samples to the previous histogram:



This is why our statistical inferences get better as we gather more data. The difference between our sample estimates and the true population value will get smaller as our sample sizes get larger, so we have more certainty in our estimates.

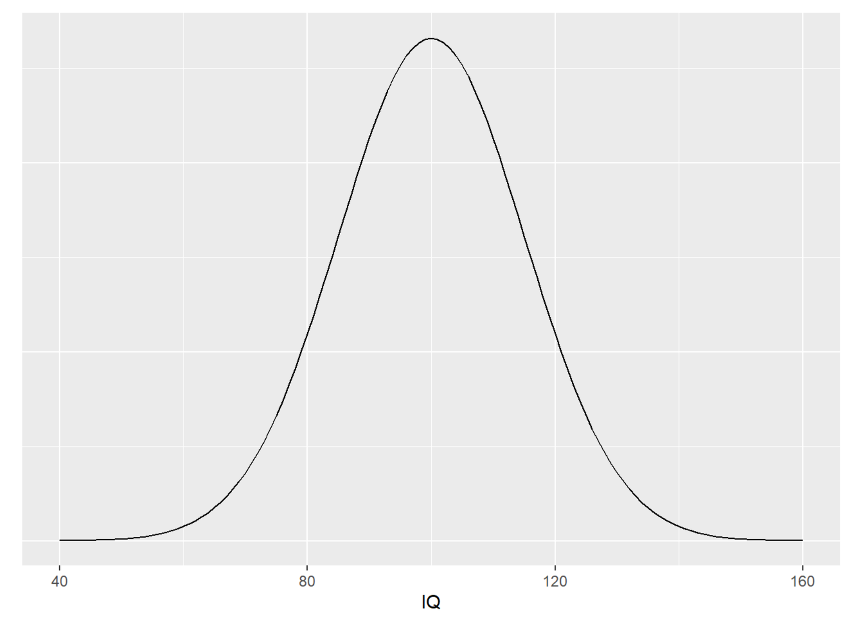
Why the Central Limit Theorem is Important

If we know the population mean and standard deviation, we know the following will be true:

"The distribution of means across repeated samples will be normal with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of n.

Since we know exactly what the distribution of means will look like for a given population, we can take the mean from a single sample and compare it to the sampling distribution to assess the likelihood that our sample comes from the same population. In other words, we can test the hypothesis that our sample represents a population distinct from the known population.

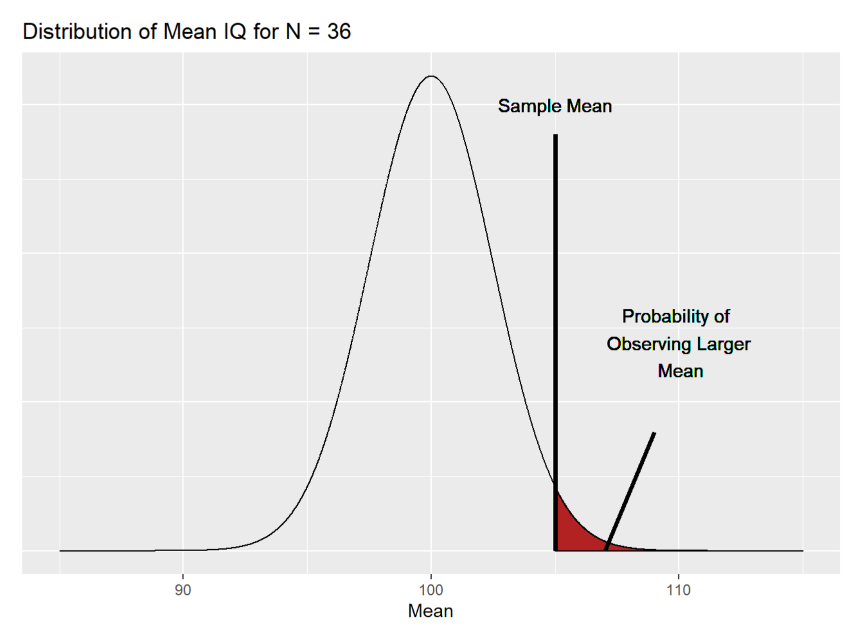
Here is an example. The population distribution of IQ in the general public is known to have a mean of 100 with a standard deviation of 15.



We take a sample of 36 students who have received a novel form of education and wish to determine if these individuals are systematically smarter than the rest of the population. To do so, we calculate the mean for our sample and consider how likely we would be to observe this value if the students were actually not any different (the null hypothesis).

The sample mean IQ we observe is 105. We know that, even if our students were not any different from the general public, we may still observe a 105 simply due to random sampling. Is this value sufficiently rare under repeated sampling that we can say our sample is different?

Given the central limit theorem, we know that the distribution of means will be normal with a mean of 100 and a standard deviation of $\frac{15}{\sqrt{36}} = 2.5$. We can compare our own mean to this distribution as follows:

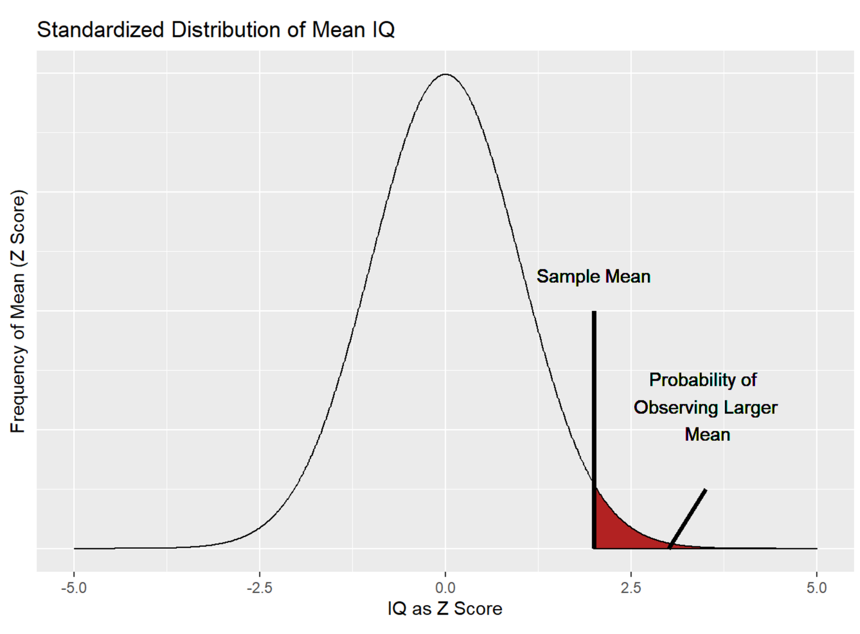


If the probability of observing our sample mean or something larger is sufficiently small (say, less than .05), then we can reject the assertion that our sample is just like the general public. This probability will be equal to the area under the normal curve above our observed sample value, indicated by the green shading in the figure.

To simplify the process of finding the area in the tail of the distribution, we typically convert our mean to a z-score as follows:

$$z = \frac{M - \mu}{\sigma/\sqrt{n}}$$

Here M is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size. This conversion rescales the distribution of means in the previous figure to have a mean of zero and a standard deviation of 1.



By making this conversion, we can rely on existing tables that have already found the area under the curve to the right or left of different z-scores. We call these probabilities *p-values*. If a p-value tells us that the probability of observing our sample mean is sufficiently small given the null hypothesis, we reject the null hypothesis. Consulting a table of z-scores tells us that the area to the right of 2 equal .023. In other words, the probability of observing a sample IQ of 105 is .023.

Still have questions? Contact us!

* * *

Jeremy Albright

Jeremy created Methods to make life a little simpler for those of us who aren't data nerds. Simplifying data into understandable insights is his passion. He remains dedicated to the education and knowledge of his clients every step of the way.

‹ Newer Older ›

* * *