03/25/2023

Confidence Intervals
One-Sample Hypothesis Tests

Last time we left off with sampling distributions.

A sampling distribution is the distribution of a statistic taken from repeated samples of the same size. For example, I take repeated samples of American adult women and measure their heights and collect the mean of 5 people. Create a distribution of those means. The central limit theorem tells us how these sample means behave: they center around the population mean, and the standard deviation of the sampling distribution is reduced by the square root of the sample size (for means: the standard error is $\frac{\sigma}{\sqrt{n}}$, and for proportions, the standard error is $\sqrt{\frac{p(1-p)}{n}}$). And the distribution will be more normal as the sample size increases, regardless of the shape of the original distribution.

We don't actually do repeated samples, but we use the properties of the sampling distribution to estimate when we take one sample about how good that estimate is from the true value of the population.

The margin of error will be calculated by using the standard error of the sampling distribution and a scaling factor that will measure the level of confidence (essentially, the probability that our mean is inside the given range). A typical confidence level is 95%.

Confidence interval is a range of values in which we think, based on our sample, that the true mean (or proportion) of the population likely falls. The mean of our sample is the center of the interval, and we add/subtract the margin of error from the center.

Take a sample, calculate the mean and the standard deviation. (some problems will give you information about the population, but when working with most data, we will calculate it from the sample.) If it's a proportion problem, the sample proportion will be calculated.

To calculate the standard error.
If it is a means problem (we have both the mean and the standard deviation), then $SE = \frac{\sigma}{\sqrt{n}}$ or $SE = \frac{s}{\sqrt{n}}$.
If it is a proportion problem, we calculate: $SE = \sqrt{\frac{p(1-p)}{n}}$.

Then we calculate the scaling factor that measures our confidence level. If the sample size is either smaller than 40 or there is no population standard deviation, then we use the t-distribution to calculate the multiplier. If we have both a large sample size and a population standard deviation, then a means problem can use the normal distribution instead of the t-distribution. If the problem is a proportion problem, we use the normal distribution.

From the scaling factor, we can calculate the margin of error.

$$ME = t^* SE$$
$$ME = z^* SE$$

Confidence interval is the last step.
Lower bound is mean minus the margin of error, and the upper bound is the mean plus the margin of error.

$$(mean - ME, mean + ME)$$
$$(proportion - ME, proportion + ME)$$

Why isn't a point estimate appropriate? Why do we need an interval? The point estimate doesn't give us any information of its level of accuracy. The interval gives me that information on accuracy. Wide intervals are poor estimates, and narrow intervals are better estimates.

Confidence intervals will get narrower if the sample size goes up (narrower is better). Will get wider if you want more confidence.

Suppose that we have a sample of 50 students who took the SAT and got a mean score of 510. The population standard deviation is 100. Construct a confidence interval for the true population mean.

Can use the normal distribution for the scaling factor because BOTH the sample size is bigger than 40, and the population standard deviation is known. For a 95% confidence interval, the z-score is 1.96.

$$SE = \frac{100}{\sqrt{50}} \approx 14.1421 \ldots$$

$$ME = 1.96 \times 14.1421 \ldots = 27.71858 \ldots$$

Confidence interval: $(510 - 27.7, 510 + 27.7) = (482.3, 537.7)$

This suggests that the population mean of 500 is inside the confidence interval. The sample is pretty average and does not indicate the sample has a mean which is above average (the increase is probably just noisiness in the data, not a sign that the students are smarter than most students).

If you want to construct a confidence but have not yet collected the data, you can select the margin of error as a target and then calculate the required sample to obtain that margin of error.

For means:

$$n = \left(\frac{z^* s}{ME}\right)^2$$

Round up to the nearest whole number.

For proportions:

$$n = p(1-p)\left(\frac{z^*}{ME}\right)$$

If you don't know what the proportion is, guess $p = 0.5$ because that will give you the largest margin of error. Round up to the next whole number.

Hypothesis Testing
Rooted in the same math of sampling distributions as confidence intervals, but instead of calculating a range of possible values the population mean could be, we are going to make an assumption about the population mean, and then calculate a probability that the data we have fits that assumption.

If that probability is relatively "large" then the result is likely due to chance (that it's different from the assumption) and we conclude that the assumption is not necessarily incorrect. If the probability is small enough, then we conclude that it's unlikely to be based on the assumptions we started with.

In the latter case, we are saying there is good evidence that the assumptions are not correct. But, we can't say the assumption is correct in the other case, only that we don't have the evidence to say otherwise.

The default assumptions are called the null hypothesis. And the thing we are trying to prove is the alternative hypothesis. The null hypothesis always includes an equal sign. The alternative will include inequalities: $<$, $>$ or $\neq$.

Next time we will talk about notation, Type I and Type II errors. And, one and two-sample testing.