Confidence Intervals

Owen G. Ward

2021-05-24

Confidence Intervals for Estimation

As we saw, we can get an estimate of a parameter and of the uncertainty of that estimate.

A further step is to quantify the range this parameter can take explicitly, such as saying there is a large probability it is in some interval.

Suppose we have data about union support, $X_1, \dots, X_n$. A confidence interval for p is some numbers a, b, based on $X_1, \dots, X_n$, such that

$$P(a \leq p \leq b) = 1 - \alpha,$$

for α close to 0, typically $\alpha = 0.1, 0.05, 0.01$.

We will see confidence intervals for Normal, Binomial and Poisson data.

The normal is widely used because of the central limit theorem.

Normal Mean

Suppose we have data $X_1, \dots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$.

We know σ but not μ. We want a confidence interval for μ. To construct one we need an estimate for μ.

We know

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1).$$

Suppose we want a $1 - \alpha$ confidence interval for μ. We first find values a, b for the standard normal such that

$$P\left(a \leq Z \leq b\right) = 1 - \alpha.$$

These numbers a, b are going to depend on the data we have observed.

We saw that a normal is symmetric, and want equal probabilty on each end, so we can write this using

$$P\left(-z \leq Z \leq z\right) = 1 - \alpha.$$

How do we get z?

For a given $z_{\alpha/2}$ we have

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha.$$

Now we need to solve this for μ, because that's what we want the confidence interval of.

$$P\left(-z_{\alpha/2} \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le z_{\alpha/2}\right) = 1-\alpha.$$

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

We then rearrange this to the original form to give

$$P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

So our $1-\alpha$ confidence interval for a normal mean is given by

$$\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

For a different α we just choose a different $z_{\alpha/2}$.

As n get's larger this confidence interval shrinks. It has length $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ where $\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{X}$.

If we have larger σ the interval gets wider.

If we have larger α then the interval gets shorter.

Common cut offs

- If $\alpha = 0.05$ then $z_{\alpha/2} = 1.96$.
- If $\alpha = 0.1$ then $z_{\alpha/2} = 1.64$.
- If $\alpha = 0.01$ then $z_{\alpha/2} = 2.58$.

Example

Suppose the time it takes runners to run a 10 Mile race is normally distributed with known SD of 16 minutes. 100 runners ran a race with average finishing time being 95.61 minutes. Construct a 95% confidence interval for the average time it takes all runners to run the race.

To get a 95% confidence interval we have $\alpha = 0.05$ so $z_{\alpha/2} = 1.96$.

We know $\sigma = 16$ and $n = 100$ so our interval will be

$$\left(95.61 - 1.96\frac{16}{10}, 95.61 + 1.96\frac{16}{10},\right) = (92.474, 98.746).$$

Say that we are 95% confident the true average run time falls in this interval.

Confidence Interval for Binomial Data

We know that if we have samples $X_1, X_2, \dots, X_n \sim Binomial(1, p)$, (so each a single coin trial), that

$$\mathbb{E}(\bar{X}) = p,\ Var(\bar{X}) = \frac{p(1-p)}{n}.$$

By the central limit theorem $\bar{X}$ is approximately normal with mean p and variance $\frac{p(1-p)}{n}$.

So

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1),$$

and we can use the same ideas as the previous example.

For a $1 - \alpha$ confidence interval we can use the same cut-offs so that

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha,$$

and we want to isolate the p in the numerator.

When we do this we get

$$P\left(\bar{X} - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2} \leq p \leq \bar{X} + \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}\right).$$

So our confidence interval is

$$\left(\bar{X} - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}, \bar{X} + \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}\right).$$

What is the problem with this interval?

Our confidence interval contains p, which we're trying to estimate. But we have a method to estimate p. We can just use $\bar{X}$ instead of p in our confidence interval.

This gives our confidence interval as

$$\left(\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} z_{\alpha/2}, \bar{X} + \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} z_{\alpha/2}\right).$$

This is now an approximate confidence interval for two reasons:

- We used the normal approximation for the distribution of $\bar{X}$.
- We had to plug in our estimate of p into the confidence interval above also.

This means our confidence interval is only approximate here. This makes sense anyway as p can take on any value, $\bar{X}$ can only get so close.

As we get a larger n this interval will become more correct.

In general a confidence interval for a parameter is given by an estimate of that parameter plus and minus a z score times the standard error of the estimate.

Example - Union Vote

For the previous union support example, we have 225 out of 500 supported the union. Construct a 99% confidence interval for p the true proportion.

We have $\alpha = 0.01$ so $z_{\alpha/2} = 2.58$ and $n = 500$.

Plugging these into our formula gives

$$\left(0.45 - 2.58\sqrt{\frac{0.45(0.55)}{500}}, 0.45 + 2.58\sqrt{\frac{0.45(0.55)}{500}},\right)$$

$$= (0.393, .507)$$

Say we are 99% confident the true value of p lies in this interval.

Interpreting Confidence Intervals

Confidence intervals capture a range of values that are likely to contain the true unknown value. But it is not guaranteed!

The cutoffs are based on the sample mean of the data which is random.

We had a definition for probability in terms of the long run average of repeated experiments. The same is true here.

A 95% confidence interval means that if we got many many 95% confidence intervals, we would expect 95% of them to contain the true value.

One Sided Intervals

Sometimes we are only interested in the range of values a parameter can take on in one direction. Say you only care about the upper limits of support in some poll.

For a normal mean with known σ, we want a value b such that

$$P(\mu \leq b) = 1 - \alpha.$$

Call this a $1 - \alpha$ upper confidence interval.

Procedure very similar, just need to adjust z correctly.

For a two sided interval, we wanted cutoffs on both sides which gave total area α.

No we only want an upper cutoff, and we want the area beyond that cut off to be α.

So instead of using $z_{\alpha/2}$ we use z_α, getting a confidence interval of the form

$$\left(-\infty, \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right).$$

For a two sided 95% confidence interval we had $z_{\alpha_2} = 1.96$. For a one sided 95% interval we use $z_\alpha = 1.68$.

Similarly for a lower confidence interval we want a such that

$$P(a \leq \mu) = 1 - \alpha,$$

which for a normal mean is of the form

$$\left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$$

Example

For the same runner example, construct a 95% upper confidence interval for the mean finishing time.

We have $\bar{X} = 95.61, n = 100, \sigma = 16$ and $\alpha = 0.05$. Here $z_{ }

So our upper interval is of the form

$$\left(-\infty, 95.61 + 1.64\frac{16}{\sqrt{100}} = (-\infty, 98.234).\right)$$

We our 95% confident the mean time is less than 98.234. Notice this is different to the limit of the two sided interval for the same α.

One sided Binomial

We can do the same thing with Binomial data. We just need to be sensible with the bounds (as we were in the previous example).

To get a one sided 99% upper confidence interval for the union support, our interval would be

$$\left(0, 0.45 + 2.32\sqrt{\frac{0.45(0.55)}{500}}\right) = (0, 0.501).$$

We used the CLT to get our confidence intervals for Binomial data. A homework problem will work through how this also works for confidence intervals for the rate of a Poisson distribution.

Recap

Confidence intervals are confusing at first but there is a standard procedure to follow.

- Figure out if a one or two sided interval.
- Figure out the confidence level (α) and the corresponding cut off for the normal distribution.
- Get the bounds of the confidence interval using the estimate, the standard error of the estimate and the cut off.

Further Points

We have only seen two specific forms of confidence intervals but the above procedure is pretty general.

We can extend it to cases like the difference in Binomial proportions, or the difference in means. The standard deviation of the estimate will change.

Similarly, if we don't know σ we can use something other than a z-score to account for this additional uncertainty.

Confidence intervals are closely related to the next topic, hypothesis testing.