ST 516: Foundations of Data Analytics

Confidence Intervals

Confidence intervals

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate.

Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* for the parameter.

A plausible range of values for the population parameter is called a **confidence interval**.

Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

An approximate 95% confidence interval

Our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate.

The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

When the sampling distribution is approximately Normal, 95% of the time the estimate will be within 1.96 standard errors of the parameter.

If the interval spreads out 1.96 standard errors from the point estimate, we can be roughly 95% **confident** that we have captured the true parameter:

Example

A 95% confidence interval for an estimate with a Normal sampling distribution is:

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{1}$$

The sample mean of days active per week from `yrbss.samp` is 3.75. The standard error, as estimated using the sample standard deviation, is $SE = \frac{2.6}{\sqrt{100}} = 0.26$ days.

Calculate an approximate 95% confidence interval for the average days active per week for all YRBSS students.

Solution

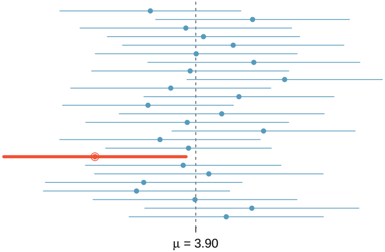
We will estimate the population mean with the sample mean. Since we have a large sample, we know from the CLT, the sampling distribution of the sample mean is approximately Normal, and can apply the formula from the previous slide:

$$\begin{array}{cccccc} \overline{x} & \pm & 1.96 \times SE_{\overline{x}} & & \\ 3.75 & \pm & 1.96 \times 0.26 & \rightarrow & (3.25, 4.25) \end{array}$$

Based on these data, we are about 95% confident that the average days active per week for all YRBSS students was larger than 3.25 but less than 4.25days.

Interpreting confidence intervals

But what does “95% confident” mean? Suppose we took many samples and built a confidence interval from each sample. Then about 95% of those intervals would contain the actual mean, μ.



CIs for the population mean number of active days constucted from 25 different samples. (The true population values is 3.9days).

Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

> *We are XX% confident that the population parameter is between...*

Incorrect language might try to specify a probability that the parameter is in the reported interval.

Another important consideration of confidence intervals is that they *only try to capture the population parameter*. A confidence interval says nothing about the confidence of capturing individual observations, a proportion of the observations, or about capturing point estimates.

Pieces of a confidence interval

The general form of confidence intervals for many (but not all) parameters is:

$$\text{point estimate} \pm \text{critical value} \times SE$$

The critical value depends on the sampling distribution and the level of confidence. For an estimate with a Normal sampling distribution, and 95% level of confidence, the value was 1.96 because, 95% of the time, a Normal random variable lies within ± 1.96 standard deviations of its mean.

Pieces of a confidence interval

We follow Ott & Longnecker's notation for this value. When the sampling distribution is Normal, and we require a $(1-\alpha)100\%$ confidence interval,

$$\text{critical value}: z_{\alpha/2} = \text{the } 1-\alpha/2 \text{ quantile of a standard Normal}$$

For example, for a 95% CI, $\alpha = 0.05$, $z_{0.025} = 1.96$

Put in picture

Confidence levels

If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?

If we want to be more certain we will capture the fish, we might use a wider net. Likewise, we use a wider confidence interval if we want to be more certain that we capture the parameter.

Increasing the confidence level of an interval (for example, from 95% to 99%), involves increasing the critical value in the interval calculation.

Finding the critical value for Normal based CIs in R

For an estimate with a Normal sampling distribution, the critical value in a 95% confidence interval, is the 0.975 quantile of a Normal distribution

```
qnorm(0.975) # the usual 95% CI critical value
```

```
## [1] 1.959964
```

Why 97.5?

We want the value, z, such that the probability of falling between $-z$ and z is 0.95.

Or, since the Normal is symmetric, that is the same as saying, find z such that the probability of falling above z is 0.025.

Or, find z such that the probability of falling below z is 0.975, the definition of the 0.975 quantile.

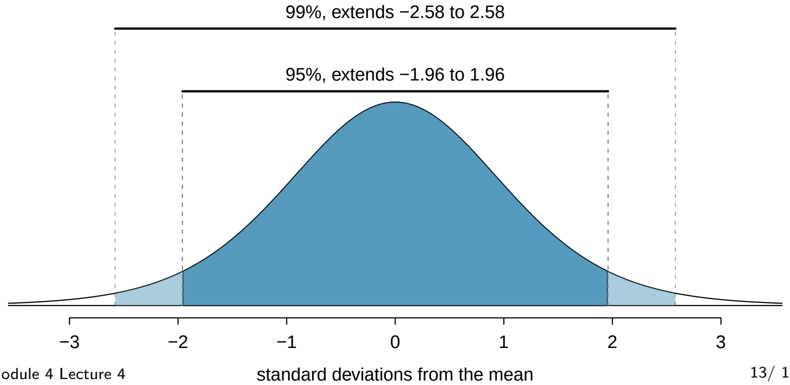
Finding the critical value for Normal based CIs in R

What quantile would we need for a 99% confidence interval?

0.995

```r
qnorm(0.995) # 99% CI critical value
```

```
## [1] 2.575829
```



Other population parameters

What about generating point estimates and confidence intervals for other population parameters, such as the population median or population standard deviation?

Once again we might estimate parameters based on sample statistics. For example, the population standard deviation of *active* using the sample standard deviation, 2.56 days.

Finding the standard error of the estimate and building confidence intervals for the parameter requires us to know the sampling distribution of the estimate.