Lecture 13: Confidence Intervals

Chapter 4.2

September 29, 2014

Goals for Today

- Introduce confidence intervals
- Interpretation
- Give an informal description of the central limit theorem

Recap

- Point estimates are based on a sample $x_1, \ldots, x_n$ and are used to estimate population parameters.
- The sampling distribution characterizes the (random) behavior of point estimates.
- The standard deviation of a sampling distribution is the standard error: it quantifies the uncertainty/variability of point estimates.

Intuition of a Confidence Interval

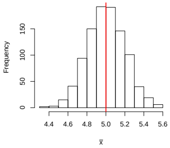
Our Goal: we want estimate a population parameter (e.g. μ).
Analogy from book: imagine the population parameter is a fish in a murky river. We want to capture this fish.

Using just the point estimate:

Using a confidence interval:

Intuition of a Confidence Interval

Recall that we had the following sampling distribution of 1000 instances of $\overline{x}$ where each $\overline{x}$ is based on $n = 100$. Keep in mind this was an unrealistic situation where we know $\mu = 5$ & $\sigma = 2$.



We observe the sampling distribution

- is centered at μ
- has $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$

Intuition of a Confidence Interval

A plausible range of values for the population parameter is called a confidence interval (CI). Since

- the SE is the standard deviation associated with $\overline{x}$ i.e. the SD of the sampling distribution
- roughly 95% of the time $\overline{x}$ will be within 2 SE of the parameter μ

If the interval spreads out 2 SE from $\overline{x}$, we can be roughly "95% confident" that we have captured the true parameter μ.

Intuition of a Confidence Interval

So a rough confidence interval is:

$$\begin{aligned}\bar{x} \pm 2SE &= [\bar{x} - 2SE,\ \bar{x} + 2SE] \\ \bar{x} \pm 2\frac{\sigma}{\sqrt{n}} &= \left[\bar{x} - 2\frac{\sigma}{\sqrt{n}},\ \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right]\end{aligned}$$

But since we will typically not know σ, assuming the conditions hold we throw in s in place of σ

$$\bar{x} \pm 2\frac{s}{\sqrt{n}} = \left[\bar{x} - 2\frac{s}{\sqrt{n}},\ \bar{x} + 2\frac{s}{\sqrt{n}}\right]$$

Confidence Intervals

A 95% confidence interval for the mean is more precise using $z^* = 1.96$, and not 2.

$$[\bar{x} - 1.96SE,\ \bar{x} + 1.96SE] = \left[\bar{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}\right]$$

A 99% confidence interval for the mean

$$[\bar{x} - 2.58SE,\ \bar{x} + 2.58SE] = \left[\bar{x} - 2.58 \times \frac{s}{\sqrt{n}},\ \bar{x} + 2.58 \times \frac{s}{\sqrt{n}}\right]$$

Conditions Needed

Important conditions to help ensure the sampling distribution of $\overline{x}$ is nearly normal and the estimate of $SE = \frac{s}{\sqrt{n}}$ is sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The distribution of sample observations is not strongly skewed.

Additionally, the larger the sample size, the more lenient we can be with the sample's skew.

Everyday Example: Political Polls

You look at a news article online and it says:

> *We polled the electorate and found that 45% of voters plan to vote for candidate X. The margin of error for this poll is ±3.4 percentage points 19 times out of 20.*

What does this mean?

- "19 times out of 20" indicates 95%
- The margin of error of ±3.4% indicates that 95% C.I. is:

$$45 \pm 3.4\% = [41.6, 48.4]$$

Crucial: How to Interpret a Confidence Interval

The confidence interval has nothing to say about any particular calculated interval; it only pertains to the method used to construct the interval:

- Wrong, yet common, interpretation: There is a 95% chance that the C.I. captures the true population mean μ
- Correct, interpretation: If we were to repeat this sampling procedure 100 times, we expect 95 (i.e. 95%) of calculated C.I.'s to capture the true μ

Illustration: How to Interpret a Confidence Interval

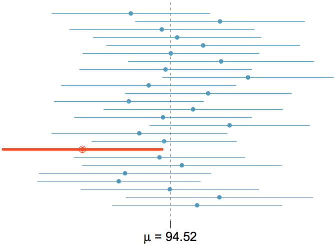
At the outset of Chapter 4, there is an example data set of finish times (in minutes) from the 2012 Cherry Blossom 10 mile run, which had 16,924 participants. In this case, we can compute the true population mean $\mu = 94.52$.

Say we take 25 (random) samples of size $n = 100$ (less than 10% of 16,924), and each time we compute:

- $\overline{x}$
- s
- and hence the 95% CI: $\left[\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right]$

We then observe

How to Interpret a Confidence Interval



How to Interpret a Confidence Interval

In this case, of the 25 confidence intervals we generated based on 25 samples of size $n = 100$, one of them (in red) did not capture the true population mean μ.

As stated earlier, if we were to repeat this whole procedure (collect sample of size $n = 100$ and compute $\overline{x}$, s and the 95% CI), we expected 95 of them to capture the true population mean μ.

Back to Political Polls

You look at a news article online and it says:

> *We polled the electorate and found that 45% of voters plan to vote for candidate X. The margin of error for this poll is ± 3.4 percentage points 19 times out of 20.*

What does this mean?

- "19 times out of 20" indicates 95%
- The margin of error of $\pm 3.4\%$ indicates that 95% C.I. is:

$$45 \pm 3.4\% = [41.6, 48.4]$$

Intrepretation: the interpretation is not that there is a 95% chance that $[41.6, 48.4]$ captures the true %'age of people who will vote for candidate X. Rather, that if we were to take 20 such polls, 19 of them would capture the true %'age.

Central Limit Theorem

Central Limit Theorem (informal description, more thorough treatment in Chapter 4.4):

If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean $\overline{x}$ is well approximated by a normal model.

Next Time

Hypothesis Testing: we can perform statistical tests on population parameters such as μ:

- Null and alternative hypotheses.
- Testing hypotheses using confidence intervals.
- Types of errors