Unit 2: Foundations for Inference

5. Confidence Intervals

10/24/2016

Quiz 3 - The Normal Distribution

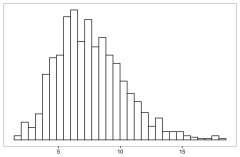
Recap from last time

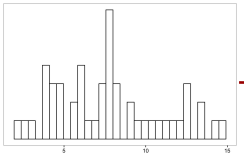
1. We can use Z-scores to compare points on two different normal distributions
2. We can use Quantile-Quantile Plots to check for Normality
3. We are really thinking about three distributions: the sample, the population, and the test statistic

Key Ideas

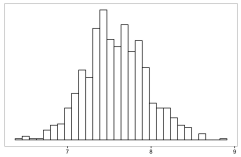
1. Statistical inference methods based on the CLT depend on the same conditions as the CLT
2. We can use confidence intervals to estimate population parameters
3. Critical values depend on the confidence interval

A reminder about the Central Limit Theoreom





Take the mean,
Repeat many times...



When I draw **independent samples** from the population, as sample size **approaches infinity,** the distribution of means approaches normality

Many statistical methods we use leverage this relationship
(t-test, linear regression, ANOVA, etc)

These methods inherit the conditions of the CLT

Always check these in context of the data and the research question!

1. **Independence**: Sampled observations must be independent.
   - This can be difficult to verify
2. **Sample size/skew**: Either the population distribution is normal or $n > 30$ and the population distribution is not extremely skewed (the more skewed , the higher n necessary for the CLT to apply).
   - This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample is representative

Confidence Intervals

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

Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

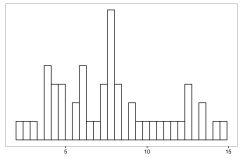
We can throw a spear where we saw a fish, but we'll probably miss. If we toss a net, we have a good chance of catching it.

If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Where does the confidence interval come from?

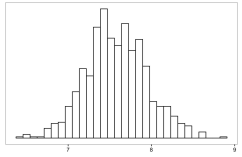
When I collect a sample, and take it's mean, what is my best guess for the mean of the population?

The sample mean



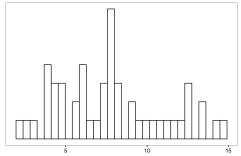
What does my confidence in the estimate depend on?

The sample standard deviation



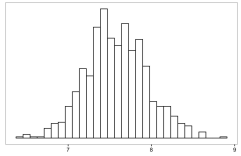
Defining a confidence Interval

CI: point estimate ± margin of error



If the parameter of interest is the population mean,

$$\mu = \bar{x} \pm Z^* \frac{s}{\sqrt{n}}$$

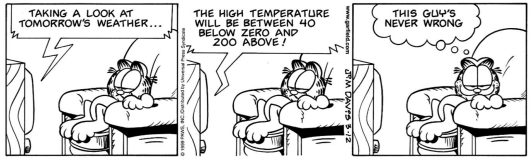


Picking a Z*

If we want to be more certain that we capture the population parameter,
i.e. increase our confidence level,
should we use a wider interval or a smaller interval?

A wider interval

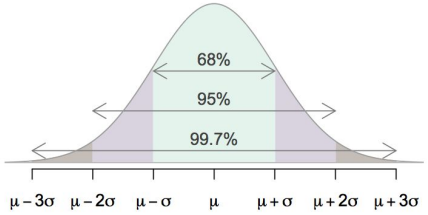
Can you see any drawbacks to having a wide interval?



Picking a Z*

The Z-score we pick for Z* works just like picking a critical p-value.

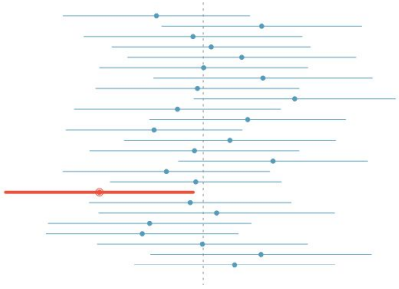
The larger the Z*, the wider the interval. E.g., if we set Z* to 1, we have a ~68% confidence interval. If we set Z* to 1.96, we have a 95% confidence interval.



What does a 95% confidence interval mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate ± 2 x SE*.

Then about 95% of those intervals would contain the true population mean (μ).



What a confidence interval does NOT mean

The confidence level of a confidence interval is the probability that the true population parameter is in the confidence interval.

Actually, The confidence level is equal to the proportion of random samples that result in confidence intervals that contain the true population parameter.

A narrower confidence interval is always better.

Actually, width is a function of both confidence level **and** the standard error.

A wider interval means less confidence.

Actually, you can make very precise statements with very little confidence.

Key Ideas

1. Statistical inference methods based on the CLT depend on the same conditions as the CLT
2. We can use confidence intervals to estimate population parameters
3. Critical values depend on the confidence interval

We're halfway done!

Unit 1: Data and where it comes from

1. Why inferences depend on how you collected your data
2. How to use graphs and descriptive statistics to describe your data
3. Why the right statistics depend on the data

Unit 2: Randomization and Sampling

1. The difference between samples and populations
2. How to do a null hypothesis test by comparing an empirical difference to a Null difference
3. Why Normal Distributions are everywhere and how we profit from that
4. What we are doing when use a sample to draw an inference about a population