Confidence Intervals

Jeffrey Woo

School of Data Science, University of Virginia

1. Confidence Intervals
2. Finding Multipliers
3. t Distributions

From Sampling Distributions...

- We know that the sample mean, $\bar{x}$, describes our particular sample. However, if we select another random sample, the sample mean will probably be different.
- We do know that under some circumstances, the distribution of the sample means can be approximated by a normal distribution.
- We also know that with a larger sample size, the sample means will be closer to the population mean, on average.

Reality: we will not know the value of the population mean, μ. So we will apply the facts above to use the sample mean to estimate the population mean.

To Confidence Intervals

Goal(s) of confidence intervals:

- Provide an estimate for the unknown parameter of interest
- Provide a **range of plausible values** for the unknown parameter of interest
- Provide a measure of **uncertainty**

General Form of Confidence Intervals

Confidence intervals generally take the following form:

$$\text{Estimate} \pm \text{margin of error.} \tag{1}$$

The **margin of error** reflects how precise we believe our estimate is, and is calculated using the confidence level $C = 1 - \alpha$. $C = 0.95$ is considered the standard.

Confidence Levels and Margin of Error

- **Confidence Level**: If we obtain many random samples of the same sample size n, and construct a confidence interval with $C\%$ confidence level based on each sample, $C\%$ of samples will have a confidence interval that contains the population mean μ.
- **Margin of Error**: Suppose we obtain many random samples of the same sample size n, and construct a confidence interval with $C\%$ confidence level based on each sample. The difference between the sample mean and population mean in $C\%$ of samples will be no greater than the value of the margin of error.

Confidence Interval for Population Mean

The confidence interval for population mean is

$$\bar{x} \pm z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}. \tag{2}$$

- $z_{1-\alpha/2}$ denotes the value of the standard normal distribution that corresponds to the $(1 - \frac{\alpha}{2})$th percentile. In a confidence interval, this is also called a **multiplier**.
- Generally speaking, the margin of error can be viewed as multiplier $\times$ standard deviation of estimate.

Finding Multiplier in CI

Finding Multiplier using R

Type qnorm(percentile) in R to find $z_{percentile}$.

Finding Multiplier

- Find the z multiplier at 90% confidence
- Find the z multiplier at 98% confidence
- Find the z multiplier at 99% confidence

Question: Do you notice a trend in the z multiplier as confidence level increases? Does this make sense?

Confidence Interval for Population Mean

Look back at (2). Do you notice anything strange about this formula?

When σ is Unknown

Recall that the population variance is

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

and the sample variance is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}.$$

When σ is **unknown**, we use the sample standard deviation, s, to estimate σ.

Standard Error

- Previously, we computed the standard deviation of the sample mean, $sd(\bar{x})$, as $\frac{\sigma}{\sqrt{n}}$.
- When σ is unknown, we compute the **standard error** of the sample mean: $se(\bar{x}) = \frac{s}{\sqrt{n}}$.

When the standard deviation of a statistic is estimated from the data, the result is the **standard error of the statistic**.

The t Distribution

Scenario: a random sample of size n is drawn from $N(\mu, \sigma)$.

- When σ is known, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$, and so $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.
- When σ is unknown and estimated using s, the sampling distribution of $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ is approximated by a t **distribution with degrees of freedom** $n-1$.
- If we do not have a normal population, the approximation to the t distribution works well if we have a large enough sample size.

Degrees of Freedom

- t distributions are specified by their **degrees of freedom**.
- We specify t distributions using t_k, where k is the degrees of freedom.

t Distribution Vs Standard Normal

Both distributions are centered at 0, symmetric, bell-shaped. Their differences are:

- t_k has an associated degrees of freedom.
- t_k has slightly **larger spread**.

As the sample size increases, t_k approaches the standard normal.

Confidence Interval for Population Mean

We use s to estimate σ when it is unknown. The level C CI for a population mean becomes

$$\bar{x} \pm t_{1-\alpha/2,k} \times \frac{s}{\sqrt{n}} \tag{3}$$

where $t_{1-\alpha/2,k}$ is the value from the t_k curve with area C between $t_{\alpha/2,k}$ and $t_{1-\alpha/2,k}$. The degrees of freedom is $k = n - 1$.

Finding Multiplier

In R, type qt(percentile, df) to find $t_{percentile,df}$.

- Find the t multiplier at 90% confidence with 10 df
- Find the t multiplier at 92% confidence with 35 df
- Find the t multiplier at 98% confidence with 50 df

Worked Example: Banks' Loan-to-Deposit Ratio (LTDR)

Question: The sample mean LTDR for 110 randomly selected American banks is 76.7 and the sample standard deviation is 12.3. Compute a 95% CI for the population mean LTDR. Based on this CI, is it reasonable to say that the average LTDR is less than 80 for the population?