Statistics and Quantitative Analysis U4320

Segment 6: Confidence Intervals
Prof. Sharyn O’Halloran
URL: http://www.columbia.edu/itc/sipa/U4320y-003/

Review

Population and Sample Estimates:

- The mean defines central tendency distribution.
- The variance defines dispersion of the distribution.

Review: Sampling

- When we sample from a population, our sample should be representative of the underlying population.
  - That is, our sample should be unbiased.
  - A simple random sample is selected in such a way that each member of the population has the same chance of being included in the sample.
- Sampling variability is the variance of sample estimates around population parameters
  - This is inherent in the sampling process.

Review: Sampling (cont.)

- Two sources of sampling variability:
  - Sampling error occurs by chance
    - It is simply the difference between the value of a sample statistic and the value of the corresponding population parameter.
    - Sampling Error = $\bar{x} - \mu$
  - Non-sample Errors
    - Errors that occur in the collection, recording, and tabulation of that data.
      - Using non-random samples in polling
      - Over-sampling one class or group
      - Under-sampling other class or groups
**Review: Sampling (cont.)**

Example:
- Consider a population of five employees' salaries:

<table>
<thead>
<tr>
<th>Individual</th>
<th>Salary in Thousand $</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>43</td>
</tr>
</tbody>
</table>

Mean = 30.8

- **Sampling error** = $x - \mu = 31.67 - 30.80 = $.87 thousand
- **Non-Sampling error** = $x - \mu = 32.33 - 30.80 = $1.53 thousand

= $1.53 - $.87 = $.66 thousand

**Review: Central Limit Theorem (cont.)**

If a **simple random sample** is taken from any population with mean $\mu$ and standard deviation $\sigma$,

- **As $n$ increases**, the sampling distribution of $\bar{X}$ tends toward the true population mean $\mu$.

- **Implications**
  - From the **Central Limit Theorem** we are able to show that even if the population is not normally distributed, but the sample size is large,
  - the sampling distribution of $\bar{X}$ can be approximated by a normal distribution.
  - This allows us to use the standard normal tables to make inferences about the population from our sample estimates.
Review: Inference

To make inferences about the population from a given sample, though, we make one correction:

- Instead of dividing by the standard deviation $\sigma$, we divide by the standard error of the sampling process:
  \[ SE = \frac{\sigma}{\sqrt{n}} \]

- We can then standardize by converting observed values to z-values:
  \[ Z = \frac{X - \mu}{SE} \]

- And then use the standard normal table to find the probability of events.

Confidence Interval

Motivation

We now want to develop tools that allow us to determine how confident we are of the sample estimates being representative of the underlying population.

- We know that, on average, $\bar{X}$ is equal to $\mu$.
- We want some way to express how confident we are that a given $\bar{X}$ is near the actual $\mu$ of the population.
- We do this by constructing a confidence interval, which is some range around $\bar{X}$ that most probably contains $\mu$. 

Think of this process as one of changing hats

- We first put on our statistician hat to study distributions in the abstract.
  - We start with some given distribution with mean $\mu$ and standard deviation $\sigma$.
  - We then discover that the mean of a sample of size $n$ will be distributed $N(\mu, \sigma/\sqrt{n})$.
- This is like a controlled experiment; we get to choose the initial distribution ourselves.

Using this result, we now put on our practitioner's hat.

- As a researcher, we have some data, but no idea what is the real parent distribution.
- Say we have a sample of size $n$ from a distribution with standard deviation $\sigma$.
  - This sample happens to have mean $\bar{X}$.
  - Then our best guess is that the parent distribution has mean $\bar{X}$ as well.
Confidence Interval (cont.)

Definitions
- A confidence interval is constructed around a point estimate (e.g., \( \bar{X} \)), and it is stated that this interval is likely to contain the corresponding population parameter (e.g., \( \mu \)).
- Two components:
  - The standard error is a measure of how much error there is in the sampling process.
  - The level of confidence attached to the interval.
- The confidence level associated with a confidence interval states how much confidence we have that the interval contains the true population parameter.
  - The confidence level is denoted by \((1 - \alpha)100\%\)
  - Common values are 90%, 95% and 99%
  - Corresponding \(\alpha\)-levels are .10, .05, and .01.

Confidence Interval: \(\sigma\) known (cont.)

Constructing a 95% Confidence Interval
- Graph
  - First, we know from the central limit theorem that the sample mean \(\bar{X}\) is distributed normally, with mean \(\mu\) and standard error \(SE = \frac{\sigma}{\sqrt{n}}\).

Second, we determine how confident we want to be in our estimate of \(\mu\).
- Defining how confident you want to be is called the \(\alpha\)-level.
  - A 95% confidence interval has an associated \(\alpha\)-level of .05.
  - We find a range under the curve with area of 0.95.
  - If we are concerned with both higher and lower values, then the relevant range will have \(\alpha/2\) probability in each tail.

\[ \alpha = (1 - 0.95) = .05 \text{ or 5\% level} \]

Third, we find an interval around \(\bar{X}\) that contains 95% of the area under the curve
- The actual interval is \([1.96 \times SE]\) on either side of the sample mean.
- We then know that 95% of the time, this interval will contain \(\mu\).
- This interval is defined by:

\[ [\bar{X} - 1.96 \times SE, \bar{X} + 1.96 \times SE] \]
What's the probability that the population mean $\mu$ will fall within the interval $\pm 1.96 \times SE$?

Now, let's take this interval of size $[-1.96 \times SE, 1.96 \times SE]$ and use it as a measuring rod.

In general:
- We know from the 68-95-99.7 rule that a 95% confidence interval will be about 2 standard deviations on either side of $\bar{X}$.
- To be precise, from the z-table, we find the z-value associated with a .025 probability is 1.96.
- If we take a random sample of size $n$ from the population,
  - 95% of the time the population mean will be within the range:

$$\bar{X} \pm (Z_{0.025} \times \frac{\sigma}{\sqrt{n}}) \leq \mu \leq \bar{X} + (Z_{0.025} \times \frac{\sigma}{\sqrt{n}})$$

$$\mu = \pm Z_{0.025} \times SE$$

Example: Calculating a 95% confidence interval
- Say we sample 180 people and see how many times they ate at a fast-food restaurant in a given week.
  - Sample size $n=180$
  - The sample has a mean of 0.82, and
  - The population standard deviation $\sigma$ is 0.48.
- Calculate the 95% confidence interval for these data.
Confidence Interval (cont.)

Answer:

Step 1: Calculate SE = \(
\frac{48}{\sqrt{180}} = 0.036
\)

Step 2: Calculate Margin of Error = \( z_{0.025} \ times \ SE = 1.96 \times 0.036 = 0.071 \)

Step 3: Calculate Confidence Interval = 0.82 ± 0.07, or [.75 < \( \mu \) < .89]

Example 2: Calculating a 90% confidence interval

A random sample of 16 observations was drawn from a normal population with

- Standard deviation, \( \sigma = 6 \), and
- Sample mean \( \mu = 25 \).

Find a 90% (\( \alpha = .10 \)) confidence interval for the population mean.

First, find \( Z_{0.05/2} \) in the standard normal tables:

\( Z_{0.05/2} = 1.64 \)

Second, calculate the 90% confidence interval

\[ \mu = \bar{x} \pm z_{0.05/2} \times \frac{\sigma}{\sqrt{n}} \]

\[ \mu = 25 \pm 1.64 \times \frac{6}{\sqrt{16}} \]

\[ SE = 1.5 \]

\[ \mu = 25 \pm 1.64 \times 1.5 = 25 \pm 2.46 \]

\[ 22.53 < \mu < 27.46 \]

90% of the time, the mean lies within this range.

What if we wanted to be 99% of the time sure that the mean falls within the interval?

Select \( \alpha \)-level: \( Z_{0.005} = 2.58 \)

Calculate margin of error:

\[ 2.58 \times 1.5 = 3.87 \]

Calculate Confidence Interval:

\[ 25 \pm 3.87 \] or \[ 21.13 \leq \mu \leq 28.87 \]

What happens when we move from a 90% to a 99% confidence interval?

The range gets larger.
Confidence Interval (cont.)

Why 95%?
- It is standard to accept that our estimate will be wrong 1 out of 20 times.
- We could reduce the possibility of error, of course, by making the interval larger.
- Increasing the interval, however, makes our estimates less precise.
  - That is, the margin of error increases
  \[ z_{\alpha/2} \times \text{SE} \]
  - Trade off precision for the probability that the true mean lies in a given range.

Confidence Interval: \( \sigma \) Unknown

Confidence Intervals when \( \sigma \) is unknown
- We have been calculating confidence intervals assuming that we know the population standard deviation \( \sigma \).
- Of course, in most cases, we are not only uncertain of the mean \( \mu \), but also of the underlying variance of the parent population.
- When this is the case, we must estimate \( \sigma \).
  - The best estimate of \( \sigma \) is the sample standard deviation \( s \):
  \[
s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}
\]
  - This introduces a new source of error that must be taken into account.

Confidence Interval: \( \sigma \) Unknown (cont.)

Characteristics of a Student-t distribution
- Shape the student t-distribution

The t-distribution changes shape as the sample size gets larger, and in the limit it becomes identical to the normal.
- When to use t-distribution
  - \( \sigma \) is unknown
  - Sample size \( n \) is small (\( n < 30 \))

Confidence Interval: \( \sigma \) Unknown (cont.)

Constructing Confidence Intervals using t-Distribution
- 95% confidence interval is: \( \bar{x} \pm t_{0.025} \frac{s}{\sqrt{n}} \)
  Where:
  \[
t_{0.025} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}
\]
  The size of the confidence interval changes as sample size changes.
Confidence Interval: \( \sigma \) Unknown (cont.)

- Using t-tables
  - Given a sample size \( n \), what is the critical value to get 95% of the area under the curve?
  - Step 1: Find Degrees of Freedom
    - Degrees of freedom is the amount of information used to calculate the standard deviation, \( s \).
    - We denote it as d.f. = \( n-1 \)
  - Step 2: Look up in the t-table
    - Now we go down the side of the table to the degrees of freedom and across to the appropriate t-value.
    - That's the cutoff value that gives you area of .025 in each tail, leaving 95% under the middle of the curve.

- Application:
  - Suppose we have sample size \( n=15 \) and \( t_{0.025} \).
  - What is the critical value? 2.13

Confidence Interval: \( \sigma \) Unknown (cont.)

- Comparison to the normal distribution
  - As d.f. gets large the shape of the curve tends toward a normal distribution.
  - As \( n \) get larger, \( t_{0.025} \) gets closer and closer to 1.96 and with infinite degrees of freedom, it equals 1.96.
  - As the sample size grows, the difference between the t and the normal distribution disappears.
  - Look back at the standard normal tables...

Confidence Interval: \( \sigma \) Unknown (cont.)

- Example:
  - Four students had grades on a test of 64, 66, 89, 77. Calculate a 95% confidence interval for the class average.
  - Mean \( \bar{X} = \frac{(64 + 66 + 89 + 77)}{4} = 74 \)
  - Sample Variance \( s^2 = \frac{(64 - 74)^2 + (66 - 74)^2 + (89 - 74)^2 + (77 - 74)^2}{3} \)
  - \( = 132.7 \)
  - Sample Standard Deviation \( s = \sqrt{132.7} = 11.52 \)
Confidence Interval:  
σ Unknown (cont.)

Answer:
- Calculate Margin of Error:
  \[ \text{SE} = \frac{s}{\sqrt{n}} \]
  d.f. = 3  
  \[ t_{0.025} \times \text{SE} = 3.18 \times 5.76 = 18 \]
- Calculate confidence interval:
  \[ \mu = \pm 18 = 56 < \mu < 92 \]

Not very precise with a sample of size 4.

Confidence Interval:  
Differences of Means (cont.)

Population Variance Known (σ-known)
- We are interested in estimating the value (\(\mu_1 - \mu_2\)) by the sample means, using (\(\bar{X}_1 - \bar{X}_2\)).
- Take samples of the size \(n_1\) and \(n_2\) from the two populations.
- Estimate the differences in two population means.
- To tell how accurate these estimates are, we can construct the familiar confidence interval around their difference:
  \[ (\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{0.025} \left( \frac{s_1}{\sqrt{n_1}} + \frac{s_2}{\sqrt{n_2}} \right) \]
- This holds if the sample size is large and we know both \(\sigma_1\) and \(\sigma_2\).

Confidence Interval:  
Differences of Means

- We can use these same techniques to address a number of different questions.
- For example, we may wish to determine if two populations (e.g., men and women) have the same mean (e.g., salary).
- Other examples:
  - How two sections of the same class did on an exam.
  - The comparative effectiveness of two drugs in treating the same disease.

Population Variance Unknown (σ-unknown)
- If, as usual, we do not know \(\sigma_1\) and \(\sigma_2\), then we use the sample standard deviations instead.
- When the variances of populations are not equal (\(s_1 \neq s_2\)):
  \[ (\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{0.025} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]
- Example:
  - Test scores of two classes where one is from an inner city school and the other is from an affluent suburb.
Confidence Interval: Pooled Sample Variances (cont.)

- **Pooled Sample Variances**, \( s^2_p = s^2 \) (\( \sigma^2 \) is unknown)
  - If both samples come from the same population (e.g., test scores for two classes in the same school), we can assume that they have the same population variance, \( s^2_p \).
  - where \( s^2_p = \frac{\sum (x_i - \bar{x}_i)^2 + \sum (x_j - \bar{x}_j)^2}{(n_1 - 1) + (n_2 - 1)} \)
  - 95% Confidence Interval
    \[ (\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2} \sqrt{s^2_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \]
  - The **degrees of freedom** are \((n_1 - 1) + (n_2 - 1)\), or \((n_1 + n_2 - 2)\).

**Example:**
- Two classes from the same school take a test. Calculate the 95% confidence interval for the difference between the two class means.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>71</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>53</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>77</td>
</tr>
</tbody>
</table>

**Observation**

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>71</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>53</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>77</td>
</tr>
</tbody>
</table>

**Sum** 296 180

**Mean** 74 60

Confidence Interval: Pooled Sample Variances (cont.)

**Step 1:** Calculate sample estimates
- \( \bar{x}_1 = 74; \bar{x}_2 = 60 \)
- \( \bar{x}_1 - \bar{x}_2 = 14 \)
- \( n_1 = 4; n_2 = 3 \)

\[ s^2_p = \frac{\sum (x_i - \bar{x}_i)^2 + \sum (x_j - \bar{x}_j)^2}{(n_1 - 1) + (n_2 - 1)} \]

\[ s^2_p = \frac{[(64 - 74)^2 + (66 - 74)^2 + (89 - 74)^2 + (77 - 74)^2] + [56 - 60]^2 + (71 - 60)^2 + (53 - 60)^2}{(4 - 1) + (3 - 1)} \]

\[ s^2_p = \frac{[100 + 64 + 625 + 9] + [16 + 9 + 9]}{2} \]

\[ s^2_p = \frac{1725}{3} = 575 \]

\[ s^2_p = 575 \]

**Step 2:** Calculate standard error

\[ SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 10.8 \sqrt{\frac{1}{4} + \frac{1}{3}} = 8.26 \]

**Step 3:** Calculate 95% Confidence Interval

\[ t_{d.f. = 5, 0.025} = 2.57 \]

\[ (\bar{x}_1 - \bar{x}_2) \pm t_{d.f. = 5, 0.025} \cdot SE = 14 \pm 21 \]

\[ 7 \leq (\bar{x}_1 - \bar{x}_2) \leq 21 \]

Even though the first class looks like they are doing better, we can’t ignore the possibility that Class 2 is doing better than Class 1.
**Confidence Interval:**

**Matched Samples**

- **Definition**
  - Matched samples are ones where you take a single individual and measure him or her at two different points and then calculate the difference.

- **Advantage**
  - One advantage of matched samples is that it reduces the variance because it allows the experimenter to control for many other variables which may influence the outcome.

**Calculating a 95% Confidence Interval (cont.)**

For each individual we can calculate their difference $D$ from one time to the next.

- We then use these $D$'s as the data set to estimate $\Delta$, the population difference.
- The sample mean of the differences will be denoted $\bar{D}$.
- The standard error will just be: $SE = \frac{s_D}{\sqrt{n}}$.

- Use the $t$-distribution to construct 95% confidence interval:
  
  $$\Delta = \bar{D} \pm t_{0.025} \frac{s_D}{\sqrt{n}}$$

**Example:**

**Student** | $X_1$ (Fall) | $X_2$ (Spring) | $D = X_1 - X_2$
---|---|---|---
Trimble | 64 | 57 | 7
Wilde | 66 | 57 | 9
Giannos | 89 | 73 | 16
Ames | 77 | 65 | 12
Sum | 296 | 252 | 44
Mean | 74 | 63 | 11

- **95% Confidence Interval**
  
  $$\Delta = \bar{D} \pm t_{0.025} \frac{s_D}{\sqrt{n}}$$

$S_D = \sqrt{(74-63)^2 + (63-63)^2 + (9-8)^2 + (16-16)^2 + (11-11)^2} = 15.233$ \Rightarrow \ $SE = \frac{s_D}{\sqrt{n}} = 1.96$

- Notice that the standard error is much smaller than in our unmatched pairs of equal sample size.

**Example:**

- Just before the 1996 presidential election, a Gallup poll of about 1500 voters showed 840 for Clinton and 660 for Dole.

- Calculate the 95% confidence interval for the population proportion $\pi$ of Clinton supporters.
  
  - $n = 1500$
  - Sample proportion $P$: $P = \frac{840}{1500} = 0.56$

  That is, in our sample of 1500 individuals, 840 people responded that they preferred Clinton to Dole.
Confidence Interval:

Proportions

Create a 95% confidence interval:
- where \( \pi \) and \( P \) are the population and sample proportions, respectively, and \( n \) is the sample size.

\[
\pi = P \pm \frac{1.96}{\sqrt{n}} \sqrt{\frac{\pi(1-\pi)}{n}}
\]

\( \pi = .56 \pm 1.96 \sqrt{\frac{.56(1-.56)}{1500}} \)

That is, with 95% confidence, the proportion of voters for Clinton in the whole population was between 53% and 59%.

Variance of Binomial Dist.

In general, the variance is the expected value of \((x-\mu)^2\)
- Take a binomial with \( P(x=1) = \pi \)
  \( P(x=0) = 1-\pi \)
- Mean \( \mu = \pi * 1 + (1-\pi) * 0 = \pi \)
- Variance = \( \pi * (1 - \pi)^2 + (1 - \pi) * (0 - \pi)^2 \)

\[
= (1-\pi)[\pi(1-\pi)] + \pi^2[\pi(1-\pi)]
= [\pi(1-\pi)]
\]