Quantitative Observations

→ Why is quantitative data more complex than dichotomous data?

→ What are some of the statistics, besides the mean, that have sampling distributions?

Sampling Distribution of $\bar{y}$

→ _____ is used to estimate _____

→ What is the name of the distribution that describe the sampling variability of the sample mean?

→ Two Facts:

  1. Notation:

  2. Notation:
Why does the standard deviation of the sampling distribution of the mean get smaller as n gets larger?

Intuitively does this make sense?

Theorem 5:1 p.159

1.

2.

3.

a.

b.

Central Limit Theorem

The rule of thumb for “large” is

The more skewed the distribution, the ___________ n must be before the normal distribution is an adequate approximate of the sampling distribution of \( \bar{y} \).

Why does n need to be large for highly non-normal distributions (see question previous page)?
Example: Applets

http://statweb.calpoly.edu/chance/applets/senators/samplesenators.html
http://statweb.calpoly.edu/rottesen/dist/dist.html

Example: LA freeway commuters SBP (continued)

\[ \mu = 130 \]
\[ \sigma = 20 \]

Suppose we randomly sample 4 drivers.

→ Find \( \mu \)

→ Find \( \sigma \)

→ Visually

Suppose we randomly select 100 drivers.

→ Find \( \mu \)

→ Find \( \sigma \)

→ Visually
Suppose we want to find the probability that the mean of the 100 randomly selected drivers is more than 135 mmHg?

What is the standardizing formula that we use for the sampling distribution of the mean?

Why did the formula change from $Z = \frac{Y - \mu}{\sigma}$?

*Example:* LA freeway commuters SBP (cont')

Fill in the table:

<table>
<thead>
<tr>
<th>n</th>
<th>$P(125 &lt; \bar{Y} &lt; 135)$</th>
<th>$\sigma_{\bar{Y}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$P(-0.5 &lt; Z &lt; 0.5)$</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$P(-0.79 &lt; Z &lt; 0.79)$</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>$P(-1.12 &lt; Z &lt; 1.12)$</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>$P(-1.77 &lt; Z &lt; 1.77)$</td>
<td></td>
</tr>
</tbody>
</table>

What is this table telling us?
Notation

→ Fill in the table:

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling Distribution of $\bar{y}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other Aspects of Sampling Variability

→ There is also sampling variability in:

1.

2.

→ Overall: As $n$ gets large:

- $s$ approaches
- the shape of the sample will be close to
- the shape of the sampling distribution of $\bar{y}$ will approach

Statistical Estimation

→ Statistical estimation is a form of statistical inference in which we use the data to:

1.

2.
Example: A random sample of 45 residents in SLO was selected and IQ was determined for each one. Suppose the sample average was 110 and the sample variance was 100.

→ What do we know from this information?

→ 110 is an estimate of
→ 10 is an estimate of

Standard Error of the Mean

→ We know the discrepancy between \( \bar{y} \) and \( \mu \) from sampling error can be described by

→ What quantity measures the variability in the sampling distribution of \( \bar{y} \) ?

→ What is the problem with obtaining \( \sigma_\bar{y} \) from the data?

→ What seems like a good estimate for \( \sigma_\bar{y} \) ?

→ The standard error of the mean is

→ Notation: \( SE_\bar{y} = \)

→ Round SE to ____ significant digits

Example: SLO IQ (cont')

→ The standard error of the mean is calculated to be:

→ What does this mean?
Why do we expect \( \bar{y} \) to be within one SE of \( \mu \) most of the time?

If SE is \________we have a more precise estimate.

How does \( s \) affect reliability?

How does \( n \) affect reliability?

What is the difference between the standard deviation and the standard error?

SE is described “as the measure of precision for the estimate”, what estimate are we referring to?

As \( n \rightarrow \)
\( \bar{y} \rightarrow \)
\( s \rightarrow \)
\( SE \rightarrow \)

Example: SLO IQ (cont‘): Suppose the results of the 45 SLO residents were analyzed by gender.

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>( \bar{y} )</th>
<th>SE</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>5</td>
<td>117</td>
<td>6.40</td>
<td>14.3</td>
</tr>
<tr>
<td>Female</td>
<td>40</td>
<td>109</td>
<td>3.16</td>
<td>20.0</td>
</tr>
</tbody>
</table>

Which gender has a more reliable estimate of \( \mu \)? Why?
Of the plots above, which represents the sd and the SE?

Which plot describes the data better?

**Confidence Interval for \( \mu \)**

*Example: (Analogy from Cartoon Guide to Statistics)* Consider an archer shooting at a target. Suppose she hits the bulls eye (a 10 cm radius) 95% of the time. In other words, she misses the bulls eye one out of 20 arrows. Sitting behind the target is another person who can't see the bull's eye. The archer shoots a single arrow and it lands:

The person behind the target circles the arrow with a 10 cm radius circle, reasoning that with the archers 95% hit rate, the true center of bull's eye should be within part of that circle.

As she shoots more and more arrows, the person draws more and more circles and finally reasons that these circles will include the true center of the bull's eye 95% of the time.
Basic idea of a confidence interval. In this example:

\[ \mu \text{ is } \bar{y} \text{ is} \]

From the standard normal distribution we know:

\[ P(-1.96 < Z < 1.96) = \]

Proof

Formula: \[ \bar{y} \pm 1.96\left(\frac{s}{\sqrt{n}}\right) \]

Any problems with using this formula (previous page) with our data?

The T Distribution

What does T distribution depend on (that the normal distribution didn't)?

Why \( n - 1 \) df?

As \( n \rightarrow \infty \), a t distribution approaches a

Similarities between the t distribution and the normal distribution.

Differences between the t distribution and the normal distribution.
To use the table keep in mind:

1.
2.
3.

What do you do if you need a lower tail area?

**Using The T Distribution for CI's**

t_{0.025} is known as

Example: Suppose we wanted to find the “t multiplier” for a 95% confidence interval with 12 df

Notation and answer:

Why use the 0.025 column for a 95% CI? Draw a picture of the t distribution and the appropriate cut points and labels.

The t multiplier for a 95% CI when df = ∞ is 1.96. Does anything seem familiar about this number?

**Calculating a CI for \( \mu \)**

To calculate a 100(1 - \( \alpha \))% CI for \( \mu \)

1.
2.
3.
Formula: \[ \bar{y} \pm t(df)\frac{s}{\sqrt{n}} \]

→ Does the \( \frac{\alpha}{2} \) thing make sense? Explain.

→ Does the 100(1 - \( \alpha \)) thing make sense? Explain.

→ Show why \( \alpha \) is 0.05 for a 95% confidence interval.

**Application to Data**

**Example:** Suppose a researcher wants to examine CD4 counts for HIV(+) patients seen at his clinic. He randomly selects a sample of \( n = 25 \) HIV(+) patients and measures their CD4 levels (cells/uL). Suppose he obtains the following results:

**Descriptive Statistics: CD4**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>N*</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4</td>
<td>25</td>
<td>0</td>
<td>321.4</td>
<td>14.8</td>
<td>73.8</td>
<td>208.0</td>
<td>261.5</td>
<td>325.0</td>
<td>394.0</td>
<td>449.0</td>
</tr>
</tbody>
</table>

Calculate a 95% confidence interval for \( \mu \).

→ What do we know from the background information?

→ Calculate a 95% confidence interval for \( \mu \)

→ What does this interval mean?
What are the important points in your CI conclusion?
1.
2.
3.
4.
5.

Example: CD4 levels (cont')

Consider the following information:

The U.S. Government classification of AIDS has three official categories of CD4 counts:

asymptomatic = greater than or equal to 500 cells/uL
AIDS related complex (ARC) = 200-499 cells/uL
AIDS = less than 200 cells/uL

Now how can we interpret our CI?

What assumption were we making about the population when we calculated our interval?

How can we check this assumption?

Why do we even need to check this?

Does our data appear to be approximately normally distributed? How do you know, be specific about the appropriate statistics?
CI Interpretation

→ Identify the following CI interpretations as being correct or incorrect, and say why

\[ P(\text{the next sample will give a CI that contains } \mu) = 0.95 \]

\[ P(284 < \mu < 337) = 0.95 \]

The confidence level is a property of the method rather than of a particular interval

\[ \mu \]

Example: CD4 (cont’)

What if we calculate a 90% confidence interval for \( \mu \).

→ Without recalculating, will this interval be wider or narrower?

→ What changed?

→ 95%:

→ 90%:

→ How will we always know if the CI is wider or narrower?

As the confidence goes up the interval becomes ________

As the confidence goes down the interval becomes ________

→ What is the mathematical reason for this (above)?
What is a better solution for narrowing our CI’s? Why does this work?

There are two ways to decrease the width of a CI:

1.

2.

Relationship to the Sampling Distribution of \( \bar{y} \)

A CI will contain \( \mu \) for 95% of samples (in repeated sampling at 95% confidence).

One More Example

Example: A biologist obtained body weights of male reindeer from a herd during the seasonal round-up. He measured the weight of a random sample of 102 reindeer in the herd, and found the sample mean and standard deviation to be 54.78 kg and 8.83 kg, respectively. Suppose these data come from a normal distribution.

Calculate a 99% confidence interval.

Meaningful conclusion: