Sampling Variability and Confidence Intervals

John McGready
Department of Biostatistics, Bloomberg School of Public Health

Lecture Topics

- Sampling distribution of a sample mean
- Variability in the sampling distribution
- Standard error of the mean
- Standard error vs. standard deviation
- Confidence intervals for the population mean
- Sampling distribution of a sample proportion
- Standard error for a proportion
- Confidence intervals for a population proportion

Section A

The Random Sampling Behavior of a Sample Mean Across Multiple Random Samples
Random Sample

- When a sample is randomly selected from a population, it is called a random sample.
  - Technically speaking, values in a random sample are representative of the distribution of the values in the population sample, regardless of size.
- In a simple random sample, each individual in the population has an equal chance of being chosen for the sample.
- Random sampling helps control systematic bias.
- But even with random sampling, there is still sampling variability or error.

Sampling Variability of a Sample Statistic

- If we repeatedly choose samples from the same population, a statistic (like a sample mean or sample proportion) will take different values in different samples.
- If the statistic does not change much if you repeated the study (you get the similar answers each time), then it is fairly reliable (not a lot of variability).
  - How much variability there is from sample to sample is a measure of precision.

Example: Hospital Length of Stay

- Consider the following data on a population of all patients discharged from a major US teaching hospital in year 2005.
- Assume the population distribution is given by the following:

  - Population mean ($\mu$) = 5.0 days
  - Population sd ($\sigma$) = 6.9 days
Example 2: Hospital Length of Stay
- Boxplot presentation

Population Distribution of LOS Data, US Teaching Hospital
All Patients Discharged in Year 2005

25th percentile: 1.0 days
50th percentile: 3.0 days
75th percentile: 6.0 days

Example 2: Hospital Length of Stay
- Suppose we had all the time in the world
- We decide to do a set of experiments
- We are going to take 500 separate random samples from this population of patients, each with 20 subjects
- For each of the 500 samples, we will plot a histogram of the sample LOS values, and record the sample mean and sample standard deviation
- Ready, set, go…

Random Samples
- Sample 1: \( n = 20 \)
  \[ \bar{x}_{LOS} = 6.6 \text{ days} \]
  \[ s_{LOS} = 9.5 \text{ days} \]
- Sample 2: \( n = 20 \)
  \[ \bar{x}_{LOS} = 4.8 \text{ days} \]
  \[ s_{LOS} = 4.2 \text{ days} \]
Example 2: Hospital Length of Stay

- So we did this 500 times: now let’s look at a histogram of the 500 sample means

![Histogram of 500 Sample Means](image)

Example 2: Hospital Length of Stay

- Suppose we had all the time in the world again
- We decide to do one more experiment
- We are going to take 500 separate random samples from this population of me, each with 50 subjects
- For each of the 500 samples, we will plot a histogram of the sample LOS values, and record the sample mean and sample standard deviation
- Ready, set, go…

Random Samples

- Sample 1: $n = 50$
  - $\bar{X}_{LOS} = 3.3$ days
  - $S_{LOS} = 3.1$ days

- Sample 2: $n = 50$
  - $\bar{X}_{LOS} = 4.7$ days
  - $S_{LOS} = 5.1$ days
Distribution of Sample Means

- So we did this 500 times: now let’s look at a histogram of the 500 sample means

Example 2: Hospital Length of Stay

- Suppose we had all the time in the world again
- We decide to do one more experiment
- We are going to take 500 separate random samples from this population of me, each with 100 subjects
- For each of the 500 samples, we will plot a histogram of the sample BP values, and record the sample mean and sample standard deviation
- Ready, set, go...

Random Samples

- Sample 1: \( n = 100 \)
- Sample 2: \( n = 100 \)

\[
\bar{X}_{LOS} = 5.8 \text{ days} \\
S_{LOS} = 9.7 \text{ days}
\]

\[
\bar{X}_{LOS} = 4.5 \text{ days} \\
S_{LOS} = 6.5 \text{ days}
\]
Histogram of 500 Sample Means

- So we did this 500 times: now let’s look at a histogram of the 500 sample means

Distribution of Sample Means

- Let’s Review The Results

Example 2: Hospital Length of Stay

- Population distribution of individual LOS values for population of patients: Right skewed
- Population mean 5.05 days: Population sd = 6.90 days
- Results from 500 random samples:

<table>
<thead>
<tr>
<th>Sample Sizes</th>
<th>Means of 500 Sample Means</th>
<th>SD of 500 Sample Means</th>
<th>Shape of Distribution of 500 Sample Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=20</td>
<td>5.05 days</td>
<td>1.49 days</td>
<td>Approx normal</td>
</tr>
<tr>
<td>n=50</td>
<td>5.04 days</td>
<td>1.00 days</td>
<td>Approx normal</td>
</tr>
<tr>
<td>n=100</td>
<td>5.08 days</td>
<td>0.70 days</td>
<td>Approx normal</td>
</tr>
</tbody>
</table>

Example 2: Hospital Length of Stay

- Let’s Review The Results

Example 2: Hospital Length of Stay

- Let’s Review The Results
Summary

- What did we see across the two examples (BP of men, LOS for teaching hospital patients)?
- A couple of trends:
  - Distributions of sample means tended to be approximately normal (symmetric, bell shaped) even when original, individual level data was not (LOS)
  - Variability in sample mean values decreased as size of sample each mean based upon increased
  - Distributions of sample means centered at true, population mean

Clarification

- Variation in sample mean values tied to size of each sample selected in our exercise: NOT the number of samples

Sampling Distribution of the Sample Mean

- In the previous section we reviewed the results of simulations that resulted in estimates of what’s formally called the sampling distribution of a sample mean
- The sampling distribution of a sample mean is a theoretical probability distribution: it describes the distribution of all sample means from all possible random samples of the same size taken from a population
Sampling Distribution of the Sample Mean

- In real research it is impossible to estimate the sampling distribution of a sample mean by actually taking multiple random samples from the same population - no research would ever happen if a study needed to be repeated multiple times to understand this sampling behavior
- Simulations are useful to illustrate a concept, but not to highlight a practical approach!
- Luckily, there is some mathematical machinery that generalizes some of the patterns we saw in the simulation results

The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful results
  - The sampling distribution of sample means based on all samples of same size \( n \) is approximately normal, regardless of the distribution of the original (individual level) data in the population/samples
  - The mean of all sample means in the sampling distribution is the true mean of the population from which the samples were taken, \( \mu \)
  - The standard deviation in the sample means of size \( n \) is equal to \( \frac{\sigma}{\sqrt{n}} \) (standard deviation of individual values)

Recap: CLT

- So the CLT tells us the following:
  - When taking a random sample of continuous measures of size \( n \) from a population with true mean \( \mu \) the theoretical sampling distribution of sample means from all possible random samples of size \( n \) is:

\[
\text{Normal Distribution} \quad \mu, \quad \frac{\sigma}{\sqrt{n}}
\]

\[
\text{Standard Error (SE)} = \frac{\sigma}{\sqrt{n}}
\]
CLT: So What?

- So what good is this info?
  Well using the properties of the normal curve, this shows that for most random samples we can take (95%), the sample mean $\bar{x}$ will fall within 2 SE's of the true mean $\mu$:

  $\mu$ - 2SE($\sigma$) ...

- So AGAIN what good is this info?
  We are going to take a single sample of size $n$ and get one $\bar{x}$. So we won't know $\mu$ and if we did know $\mu$ why would we care about the distribution of estimates of $\mu$ from imperfect subsets of the population?

- CLT: So What?
  We are going to take a single sample of size $n$ and get one $\bar{x}$. But for most (95%) of the random samples we can get, our $\bar{x}$ will fall within +/- 2SEs of $\mu$. 

  $\bar{x}$ - $\mu$
We are going to take a single sample of size \( n \) and get one \( \bar{x} \). So if we start at \( \bar{x} \) and go 2SEs in either direction, the interval created will contain \( \mu \) most (95 out of 100) of the time.

**CLT: So What?**

A confidence interval for the population mean \( \mu \) is given by \( \bar{x} \pm 2SE(\bar{x}) \). What is interpretation of a confidence interval?

**Estimating a Confidence Interval**

- Such interval is a called a 95% confidence interval for the population mean \( \mu \).
- Interval given by \( \bar{x} \pm 2SE(\bar{x}) \).
- Interpretation of a confidence interval?

A 95% confidence interval (CI)

- Layperson's: Range of "plausible" values for true mean
  - Researcher never can observe true mean \( \mu \)
  - \( \bar{x} \) is the best estimate based on a single sample
  - The 95% CI starts with this best estimate, and additionally recognizes uncertainty in this quantity
- Technical: were 100 random samples of size \( n \) taken from the same population, and 95% confidence intervals computed using each of these 100 samples, 95 of the 100 intervals would contain the values of true mean \( \mu \) within the endpoints
Technical Interpretation

- One hundred 95% confidence intervals from 100 random samples of size n=50

Notes on Confidence Intervals

- Random sampling error
  - Confidence interval only accounts for random sampling error—not other systematic sources of error or bias

Semantic: Standard Deviation vs. Standard Error

- The term “standard deviation” refers to the variability in individual observations in a single sample (s) or population
- The standard error of the mean is also a measure of standard deviation: but not of individual values, rather variation in multiple sample means computed on multiple random samples of the same size, taken from the same population
Section B

Estimating Confidence Intervals for the Mean of a Population Based on a Single Sample of Size \( n \): Some Examples

### Estimating a 95% Confidence Interval

- In last section we defined a 95% confidence interval for the population mean \( \mu \).
- Interval given by \( \bar{x} \pm 2SE(\bar{x}) \).
- Problem: how to get \( SE(\bar{x}) \).
  - Can estimate by formula: \( SE(\bar{x}) = \frac{s}{\sqrt{n}} \), where \( s \) is the standard deviation of the sample values.
- Estimated 95% CI for \( \mu \) based on a single sample of size \( n \):
  \[ \bar{x} \pm 2 \times \frac{s}{\sqrt{n}} \]

### Example 1

- Suppose we had blood pressure measurements collected from a random samples of 100 Hopkins students collected in September 2008. We wish to use the results of the sample to estimate a 95% CI for the mean blood pressure of all Hopkins students.
- Results:
  - \( \bar{x} = 123.4 \) mm Hg;
  - \( s = 13.7 \) mm Hg
- \( SE(\bar{x}) = \frac{13}{\sqrt{100}} = 1.3 \) mm Hg
- So a 95% CI for the true mean BP of all Hopkins Students:
  - \( 123.4 \pm 2 \times 1.3 \rightarrow (120.8 \text{ mmHg}, 126.0 \text{ mmHg}) \)
Example 2

- Data from the National Medical Expenditures Survey (1987): U.S. Based Survey Administered by the Centers for Disease Control (CDC)

Some Results:

<table>
<thead>
<tr>
<th>Smoking History</th>
<th>No Smoking History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean 1987 Expenditures (US $)</td>
<td>2,260</td>
</tr>
<tr>
<td>SD (US $)</td>
<td>4,850</td>
</tr>
<tr>
<td>N</td>
<td>6,564</td>
</tr>
</tbody>
</table>

Example 2

- 95% CIs For 1987 medical expenditures by smoking history

Smoking History: $2,260 ± 2 x $4,850 / $5,016 = $2,260 ± $2,600 → ($1,950, $2,380)

No smoking History: $2,080 ± 2 x $4,600 / $5,016 = $2,080 ± $2,080 → ($1,950, $2,210)

Example 3

- Effect of Lower Targets for Blood Pressure and LDL Cholesterol on Atherosclerosis in Diabetes: The SANDS Randomized Trial

**Objective** To compare progression of subclinical atherosclerosis in adults with type 2 diabetes treated to reach aggressive targets of low-density lipoprotein cholesterol (LDL-C) of 70 mg/dL or lower and systolic blood pressure (SBP) of 115 mm Hg or lower vs standard targets of LDL-C of 100 mg/dL or lower and SBP of 130 mm Hg or lower.

---

1 Howard B et al., Effect of Lower Targets for Blood Pressure and LDL Cholesterol on Atherosclerosis in Diabetes: The SANDS Randomized Trial, Journal of the American Medical Association 299, no. 14 (2008)
Example 3

- **Design, Setting, and Participants** A randomized, open-label, blinded-to-end point, 3-year trial from April 2003-July 2007 at 4 clinical centers in Oklahoma, Arizona, and South Dakota. Participants were 499 American Indian men and women aged 40 years or older with type 2 diabetes and no prior CVD events.
- **Interventions** Participants were randomized to aggressive (n=252) vs standard (n=247) treatment groups with stepped treatment algorithms defined for both.

Example 3

- **Results** Mean target LDL-C and SBP levels for both groups were reached and maintained. Mean (95% confidence interval) levels for LDL-C in the last 12 months were 72 (69-75) and 104 (101-106) mg/dL, and SBP levels were 117 (115-118) and 129 (128-130) mm Hg in the aggressive vs. standard groups, respectively.

Example 3

- **Lots of 95% CIs!**

![Table 3: Effectiveness of Target LDL-C in the Aggressive vs. Standard Groups](image)
FYI: True Confessions Biostat Style: What We Mean by Approximately Normal and What Happens to the Sampling Distribution of the Sample Mean with Small \( n \)

Recap: CLT

- So the CLT tells us the following:
  When taking a random sample of continuous measures of size \( n \) from a population with true mean \( \mu \) and true standard deviation \( \sigma \) the theoretical sampling distribution of sample means from all possible random samples of size \( n \) is:

\[
\sigma_x = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}
\]

Recap: CLT

- Technically this is true for “large \( n \)”: for this course, we'll say \( n > 60 \); but when \( n \) is smaller, sampling distribution not quite normal, but follows a \( t \)-distribution

\[
\sigma_x = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}
\]
The t-distribution is the "fatter, flatter cousin" of the normal: t-distribution uniquely defined by degrees of freedom.

Why the t?

- Basic idea: remember, the true SE( ) is given by the formula
  \[ \sigma_x = SE(\mu) = \frac{\sigma}{\sqrt{n}} \]
- But of course we don’t know \( \sigma \), and replace with \( s \) to estimate
  \[ \hat{SE}(\mu) = \frac{s}{\sqrt{n}} \]
- In small samples, there is a lot of sampling variability in \( s \) as well; so this estimate is less precise.
- To account for this additional uncertainty, we have to go slightly more than \( \pm 2 \times SE(\mu) \) to get 95% coverage under the sampling distribution.

Underlying Assumptions

- How much bigger the \( z \) needs to be depends on the sample size.
- You can look up the correct number in a "t-table" or "t-distribution" with \( n-1 \) degrees of freedom.
The t-distribution

- So if we have a smaller sample size, we will have to go out more than 2 SEs to achieve 95% confidence
- How many standard errors we need to go depends on the degrees of freedom—this is linked to sample size
- The appropriate degrees of freedom are \( n - 1 \)

\[
\bar{x} \pm t_{\text{df}, 0.025} \times SE(\bar{x}) \\
\bar{x} \pm \frac{t_{\text{df}, 0.025} \times s}{\sqrt{n}}
\]

Notes on the t-Correction

- The particular t-table gives the number of SEs needed to cut off 95% under the sampling distribution

<table>
<thead>
<tr>
<th>df</th>
<th>t</th>
<th>df</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.706</td>
<td>12</td>
<td>2.179</td>
</tr>
<tr>
<td>2</td>
<td>4.303</td>
<td>13</td>
<td>2.160</td>
</tr>
<tr>
<td>3</td>
<td>3.182</td>
<td>14</td>
<td>2.145</td>
</tr>
<tr>
<td>4</td>
<td>2.776</td>
<td>15</td>
<td>2.131</td>
</tr>
<tr>
<td>5</td>
<td>2.571</td>
<td>20</td>
<td>2.086</td>
</tr>
<tr>
<td>6</td>
<td>2.447</td>
<td>25</td>
<td>2.060</td>
</tr>
<tr>
<td>7</td>
<td>2.365</td>
<td>30</td>
<td>2.042</td>
</tr>
<tr>
<td>8</td>
<td>2.360</td>
<td>40</td>
<td>2.021</td>
</tr>
<tr>
<td>9</td>
<td>2.262</td>
<td>60</td>
<td>2.000</td>
</tr>
<tr>
<td>10</td>
<td>2.228</td>
<td>120</td>
<td>1.980</td>
</tr>
<tr>
<td>11</td>
<td>2.201</td>
<td>\infty</td>
<td>1.960</td>
</tr>
</tbody>
</table>

Notes on the t-Correction

- Can easily find a t-table for other cutoffs (90%, 99%) in any stats text or by searching the internet
- Also, using the cii command takes care of this little detail
- The point is not to spend a lot of time looking up t-values: more important is a basic understanding of why slightly more needs to be added to the sample mean in smaller samples to get a valid 95% CI
- The interpretation of the 95% CI (or any other level) is the same as discussed before
Example

- Small study on response to treatment among 12 patients with hyperlipidemia (high LDL cholesterol) given a treatment
- Change in cholesterol post-pre treatment computed for each of the 12 patients
- Results:

\[
\bar{x}_{\text{change}} = -1.4 \text{ mmol/L} \\
S_{\text{change}} = 0.55 \text{ mmol/L}
\]

Example

- 95% confidence interval for true mean change

\[
\bar{x} \pm t_{0.025,11} \times \hat{SE}(\bar{x}) \Rightarrow \\
\bar{x} \pm 2.2 \times \hat{SE}(\bar{x}) \Rightarrow \\
-1.4 \pm 2.2 \times \frac{0.55}{\sqrt{12}} \Rightarrow \\
(-1.75, \text{mmol/L}, -1.05 \text{ mmol/L})
\]

Section D

The Sample Proportion as a Summary Measure for Binary Outcomes and the CLT
Proportions (p)

- Proportion of individuals with health insurance
- Proportion of patients who became infected
- Proportion of patients who are cured
- Proportion of individuals who are hypertensive
- Proportion of individuals positive on a blood test
- Proportion of adverse drug reactions
- Proportion of premature infants who survive

For each individual in the study, we record a binary outcome (Yes/No; Success/Failure) rather than a continuous measurement.

Compute a sample proportion, \( \hat{p} \) (pronounced “p-hat”), by taking observed number of “yes’s” divided by total sample size.

This is the key summary measure for binary data, analogous to a mean for continuous data.

There is a formula for the standard deviation of a proportion, but the quantity lacks the “physical interpretability” that it has for continuous data.

Example 1

- Proportion of dialysis patients with national insurance in 12 countries (only six shown...)

Exhibit 1
Descriptive Measures of the Prevalent Cross-Sectional Patient Sample, Dialysis Patients in Twelve Countries, 2002-2003

Example: Canada:

\[ \hat{p} = \frac{400}{503} = 0.796 \]

1 Hirth R et al., Out-Of-Pocket Spending And Medication Adherence Among Dialysis Patients In Twelve Countries, Health Affairs 27, no. 1 (2008)
Example 2

- Maternal/Infant Transmission of HIV
  HIV-infection status was known for 363 births (180 in the zidovudine (AZT) group and 183 in the placebo group). Thirteen infants in the zidovudine group and 40 in the placebo group were HIV-infected.

\[ \hat{p}_{\text{AZT}} = \frac{13}{180} = 0.07 = 7\% \]

\[ \hat{p}_{\text{PLAC}} = \frac{40}{183} = 0.22 = 22\% \]

---

Proportions (p)

- What is the sampling behavior of a sample proportion?
- In other words, how do sample proportions, estimated from random samples of the same size from the same population, behave?

---

The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful results.
  - The sampling distribution of sample proportions based on all samples of same size \( n \) is approximately normal.
  - The mean of all sample proportions in the sampling distribution is the true mean of the population from which the samples were taken, \( \bar{p} \).
  - The standard deviation in the sample proportions of size \( n \) is called the standard error of the sample proportion and sometimes written as \( SE(\hat{p}) \).

---

1 Spector S et al., A Controlled Trial of Intravenous Immune Globulin for the Prevention of Serious Bacterial Infections in Children Receiving Zidovudine for Advanced Human Immunodeficiency Virus Infection, New England Journal of Medicine 331, no. 18 (1994)
We are going to take a single sample of size \( n \) and get one \( \hat{p} \). But for most (95%) of the random samples we can get, our \( \hat{p} \) will fall within +/- 2SEs of \( p \).

### Estimating a Confidence Interval

- Such an interval is called a 95% confidence interval for the population proportion \( p \).
- Interval estimated given by \( \hat{p} \pm 2SE(\hat{p}) \).
- Problem: how to estimate \( SE(\hat{p}) \).

Can estimate via following formula:

\[
SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}
\]

- Estimated 95% CI for based on a single sample of size \( n \):

\[
\hat{p} \pm 2 \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}
\]
Example 1

- Proportion of dialysis patients with national insurance in 12 countries (only six shown..)

**Example: France:**

\[ \hat{p} = \frac{219}{481} = .46 \]

- Estimated confidence interval

\[ \hat{p} \pm 2 \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} \]

\[ .46 \pm 2 \times \sqrt{\frac{.46 \times (1-.46)}{481}} \]

\[ .46 \pm .023 \]

\[ .46 \pm .046 \]

\[ (.414, .505) = (.41, .51) \]

\[ \rightarrow 41\% \text{ to } 51\% \]

Example 2

- Maternal/Infant Transmission of HIV

HIV-infection status was known for 363 births (180 in the zidovudine (AZT) group and 183 in the placebo group). Thirteen infants in the zidovudine group and 40 in the placebo group were HIV-infected.

\[ \hat{p}_{AZT} = \frac{13}{180} = 0.07 = 7\% \]

\[ \hat{p}_{PLAC} = \frac{40}{183} = 0.22 = 22\% \]
Example 2

- Estimated confidence interval for transmission percentage in the placebo group

\[ \hat{p} \pm 2 \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \]

\[ .22 \pm 2 \times \sqrt{\frac{.22 \times (1 - .22)}{183}} \]

\[ .22 \pm 2 \times .031 \]

\[ .46 \pm .062 \]

\[ (.158, .282) = (.16, .28) \]

\[ \rightarrow 16\% \text{ to } 28\% \]

Notes on 95% Confidence Interval for Proportion

- Sometimes \( \pm 2 \times SE(\hat{p}) \) is called
  - 95% error bound
  - Margin of error