This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
Section D

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 sd $\sigma$ the theoretical sampling distribution of sample means from all possible random samples of size $n$ is:

$$\sigma_{\bar{x}} = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$
Recap: CLT

- Technically this is true for “large $n$”: for this course, we’ll say $n > 60$; but when $n$ is smaller, sampling distribution is not quite normal, but follows a $t$-distribution
The t-distribution is the “fatter, flatter cousin” of the normal: t-distribution is uniquely defined by degrees of freedom

\[ \sigma_{\bar{x}} = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \]
Why the t?

- Basic idea: remember, the true SE(\(\bar{x}\)) is given by the formula
  \[ \sigma_{\bar{x}} = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \]

- But of course we don’t know \(\sigma\), and replace with \(s\) to estimate
  \[ SE(\bar{x}) = \frac{s}{\sqrt{n}} \]

- 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 \(\pm 2 \times SE(\bar{x})\) to get 95% coverage under the sampling distribution.
Underlying Assumptions

- How much bigger the $2$ 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 $n - 1$.

- One option: you can look up the correct number in a “t-table” or “t-distribution” with $n - 1$ degrees of freedom.

$$\bar{x} \pm t_{.95,n-1} \times \hat{SE}(\bar{x}) \Rightarrow$$

$$\bar{x} \pm t_{.95,n-1} \times \frac{s}{\sqrt{n}}$$
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>∞</td>
<td>1.960</td>
</tr>
</tbody>
</table>
Notes on the t-Correction

- You 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: $\overline{x}_{\text{change}} = -1.4 \text{ mmol/L}$
  
  $s_{\text{change}} = 0.55 \text{ mmol/L}$
Example

- 95% confidence interval for true mean change

\[ \bar{x} \pm t_{0.95,11} \times S\hat{E}(\bar{x}) \Rightarrow \]
\[ \bar{x} \pm 2.2 \times S\hat{E}(\bar{x}) \Rightarrow \]
\[ -1.4 \pm 2.2 \times \frac{0.55}{\sqrt{12}} \Rightarrow \]
\[ (-1.75, mmol/L, -1.05 \text{ mmol/L}) \]
Using Stata to Create Other CIs for a Mean

- The “cii” command,

```
cii 12 -1.4 .55
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std</th>
<th>. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>-1.4</td>
<td>.1587713</td>
<td>-1.749453</td>
<td>-1.050547</td>
</tr>
</tbody>
</table>