Chapter 8
Statistical Intervals for a Single Sample

Part 1: Confidence intervals (CI) for population mean $\mu$

Section 8.1:
CI for $\mu$ when $\sigma^2$ known & drawing from normal distribution

Section 8.1.2:
Sample size calculation for estimating $\mu$ with specified error, $\sigma^2$ known

Section 8.2:
CI for $\mu$ when $\sigma^2$ unknown & drawing from normal distribution
We end the last chapter with a phrase: “Moving beyond point estimates”

Point estimates are a good start, but we should also give the client some idea of the confidence in our estimate.

More data gives more information. We will have more confidence in an estimate for $\mu$ from an $n = 50$ sample, than an estimate from an $n = 3$ sample.

The confidence in an estimate is related to the size (or width) of such an interval.
We use the observed $\bar{x}$ as the point estimate for $\mu$.

We provide a two-sided CI for $\mu$ as a ‘window’ or interval for which we are fairly confident the unknown population mean $\mu$ lies.

$\bar{x}$ will be at the center of our two-sided CIs

$[\bar{x} - \text{cushion}, \bar{x} + \text{cushion}]$

For example, suppose $\bar{x} = 8$ and our cushion is 3
We want to have high confidence that our interval contains $\mu$.

How do we choose this ±‘cushion’ so that we have high confidence that it contains $\mu$? Or the length of our interval?

We use the behavior (or probability distribution) of $\bar{X}$...

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

for any sample size $n$

to form our CI in such a way that we can say something very powerful, like...

“We are 95% confident that the true mean $\mu$ falls in this interval.”
Right now, we are estimating $\mu$ and we say that we know $\sigma^2$. Perhaps not terribly realistic, but we will loosen this up later...

- Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from a normal distribution with $X_i \sim N(\mu, \sigma^2)$ for all $i$, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- Using this probability distribution, we have

$$P(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95$$

$$P(-z_{0.025} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{0.025}) = 0.95$$

where $z_{0.025}$ is the $97.5^{th}$ percentile of the standard normal (next slide).
Confidence Interval for $\mu$ - Normal parent population, known $\sigma^2$

$$P(-z_{0.025} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{0.025}) = 0.95$$

- Manipulating what’s inside the parentheses give us the Upper and Lower end-points for our 95% CI for $\mu$...

$$P(\bar{X} - z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{0.025} \frac{\sigma}{\sqrt{n}}) = 0.95$$

**NOTATION:** $z_{0.025}$ is the $z$-value such that 97.5% of the distribution is below and 2.5% is above it (an upper tail $z$-value).
We can state the lower and upper end-points of the 95% CI for $\mu$ from a random sample of size $n$ drawn from a normally distributed population with variance $\sigma^2$ and sample mean of $\bar{x}$ as:

Lower end-point (L) = $\bar{x} - z_{0.025} \frac{\sigma}{\sqrt{n}}$

Upper end-point (U) = $\bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}$

NOTE: $\bar{x}$ lies in the center of the 2-sided confidence interval.

95% CI for $\mu$ when $\sigma^2$ known and drawing from a normally distributed population:

$\bar{x} - z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}$

Or...

$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$
Example (Fill weights of boxes)

The sample mean for the fill weights of 100 boxes is \( \bar{x} = 12.050 \). The population variance of the fill weights is known to be \((0.100)^2\). Find a 95% confidence interval for the population mean \( \mu \) fill weight of the boxes.

**ANS:**

\[
L = \bar{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} = 12.050 - 1.96 \cdot \frac{0.100}{\sqrt{100}} = 12.030.
\]

\[
U = \bar{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} = 12.050 + 1.96 \cdot \frac{0.100}{\sqrt{100}} = 12.070.
\]

The 95% confidence interval for \( \mu \) is \([12.030, 12.070]\).

We are 95% confident that the true parameter value lies in this interval.

NOTE: Because \( \sigma^2 \) was very small and \( n \) was fairly large, we have a very narrow confidence interval for \( \mu \) (which is good).
Confidence Interval for $\mu$ - Normal parent population, known $\sigma^2$

CI for any choice of confidence level, or $100(1-\alpha)\%$ confidence

- The confidence level of choice is stated as $100(1 - \alpha)\%$.
  
  For a 95% confidence interval, $\alpha = 0.05$.
  For an 80% confidence interval, $\alpha = 0.20$.

- We can re-write the earlier $Z$ probability as
  
  $$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

  and this leads to the $100(1 - \alpha)\%$ confidence interval for $\mu$

  $$P(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

- In a two-sided confidence interval, the $\alpha$ amount is split between the two tails, thus we see $\alpha/2$ or specifically, $z_{\alpha/2}$ in the formula.
Confidence Interval for $\mu$ - Normal parent population, known $\sigma^2$

- **100(1-$\alpha$)% Confidence interval on the mean, variance known**
  
  If $\bar{x}$ is the sample mean of a random sample of size $n$ from a normal population with known variance $\sigma^2$, a 100(1 − $\alpha$)% confidence interval for $\mu$ is given by

  \[
  \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
  \]

  where $z_{\alpha/2}$ represents the z-value from the standard normal distribution with $\alpha/2$ in the upper tail (e.g. if $\alpha = .05$, $z_{\alpha/2} = z_{.025} = 1.96$).

- **Commonly used z scores**

<table>
<thead>
<tr>
<th>Conf. Level</th>
<th>$\alpha$</th>
<th>$\alpha/2$</th>
<th>$z_{\alpha/2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.10</td>
<td>0.05</td>
<td>1.645</td>
</tr>
<tr>
<td>95%</td>
<td>0.05</td>
<td>0.025</td>
<td>1.96</td>
</tr>
<tr>
<td>99%</td>
<td>0.01</td>
<td>0.005</td>
<td>2.576</td>
</tr>
</tbody>
</table>
Example (Fill weights of boxes)

The sample mean for the fill weights of 100 boxes is \( \bar{x} = 12.050 \). The population variance of the fill weights is known to be \((0.100)^2\). Find a 80% confidence interval for the population mean \( \mu \) fill weight of the boxes.

**ANS:**

\[
L = \bar{x} - z_{0.10} \cdot \frac{\sigma}{\sqrt{n}} = 12.050 - 1.28 \cdot \frac{0.100}{\sqrt{100}} = 12.037.
\]

\[
U = \bar{x} + z_{0.10} \cdot \frac{\sigma}{\sqrt{n}} = 12.050 + 1.28 \cdot \frac{0.100}{\sqrt{100}} = 12.063.
\]

The 80% confidence interval for \( \mu \) is \([12.037, 12.063]\).

We are 80% confident that the true parameter value lies in this interval.

**NOTE:** Because \( \sigma^2 \) was very small and \( n \) was fairly large, we have a very narrow confidence interval for \( \mu \) (which is good).
Confidence Interval for $\mu$ - Normal parent population, known $\sigma^2$

- Compare the 80% and 95% confidence intervals:

  The 80% confidence interval for $\mu$ is $[12.037, 12.063]$
  (The width of this interval is 0.026)

  The 95% confidence interval for $\mu$ is $[12.030, 12.070]$
  (The width of this interval is 0.040)

- The 95% CI is wider... i.e. All else being held constant, if you want to be more confident you capture $\mu$, you’ll have to make your net bigger.
Confidence Interval for $\mu$ - Normal parent population, known $\sigma^2$

- Looking at the form of the confidence interval:

$$
\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}
$$

- **Sample mean** based on confidence level changes for different % CI.

- **Multiplier value** based on $\sigma$ and sample size.

- **Standard error of the sample mean** changes.
Confidence Interval for $\mu$ - Normal parent population, known $\sigma^2$

- More narrow CIs are desirable.
- How can this be achieved?
  \[
  \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}
  \]
  Sample multiplier value
  mean based on based on confidence level sample size
- Increase your sample size (Good idea if possible)
- Decrease $\sigma$? Not an option, it’s fixed by original distribution
- Decrease your confidence level? (Not a great idea. You reduce the CI width, but you’re less likely to capture $\mu$)
Confidence Interval Interpretation

- Once the confidence interval is formed (based on observed $\bar{x}$), it either does or does not contain the fixed unknown value $\mu$

- For example, the 95% CI for box fill weights was: [12.030, 12.070]

  and the true population mean either is or isn’t in this interval.

- The confidence interval level arises based on the randomness of the interval. BEFORE we collect the data, the CI is a random interval and it could take on many different values due to the randomness of $\bar{X}$.
For a 95% CI, we are 95% confident that the true $\mu$ lies in the interval. This statement of confidence reflects the following...

If we repeated this process 100 times (i.e. collect a sample, compute $\bar{x}$, compute the CI), 95 out of 100 times we will capture the true $\mu$ on average, in the long run.

The confidence relates to the method used to calculate the CI. We don’t know if our CI captured $\mu$ or not ($\mu$ is unknown), but using the same method, 95 out of 100 times I’ll get it (on average).

See confidence interval applet website:
http://www.rossmanchance.com/applets/ConfSim.html
Simulating Confidence Intervals

Method
- Means
- Normal
- z with sigma

Parameters:
- \( \mu \) = 20
- \( \sigma \) = 10.0
- \( n \) = 100

Intervals: 100

Confidence level: 95%

Intervals containing \( \mu \):
95 / 100 = 95.0%

Running Total:
95 / 100 = 95.0%

Buttons:
- Sample
- Recalculate
- Sort
- Reset
Sample Size Calculation for $\mu$

- The length of the CI is a measure of **precision** of estimation.

- Precision is related to sample size $n$. Higher precision coincides with a larger sample size (all else being held constant).

- What sample size should you choose? (when you CAN choose)
  
  Let $E$ be the error in estimating $\mu$, distance of observed $\bar{x}$ from target.

  $$E = |\bar{x} - \mu|$$

  Other books may state this error $E$ as the ‘**Margin of Error**’.

- Choose a sample size that gives you a pre-specified level of precision.
Sample Size Calculation for $\mu$

- Choose $n$ to provide a certain bound on the error $E$ with confidence $100(1 - \alpha)$.

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Cl half-width or $E$

- Pre-specified error: $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \implies n = \left( \frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$

- Sample size for estimating $\mu$ with $100(1 - \alpha)$% confidence and error $E$:

$$n = \left( \frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$
Sample Size Calculation for $\mu$

**Example (The fill weight example)**

In the fill weight example, how many boxes must be sampled to obtain a 99% confidence interval of full width 0.024 oz.? (i.e. $E = 0.012$)

**ANS:** $\sigma = 0.100$ from before, and we want 99% CI, so $\alpha = 0.01$ and $z_{0.005} = 2.576$.

Error $E$ is set at 0.012 (half-width of CI).

$$n = \left( \frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2 = \left( \frac{2.576 \times 0.1}{0.012} \right)^2 = 460.8$$

We can’t sample a fraction of a box, so we **round-up** to ensure our confidence level is at least 99%, thus the required **sample size** is $n = 461$.

**NOTE:** Read sample size problems closely to determine if they are giving precision as a half-width of a CI which is $E$ (the cushion up or down), or the full width of the CI which is $2E$. 
Occasionally, you may be interested in finding a bound for $\mu$ on only one side.

- A 100$(1 - \alpha)$% upper-confidence bound for $\mu$ is
  \[ \mu \leq \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \]
  and this gives an interval $(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}})$. (an upper bound on $\mu$)

- A 100$(1 - \alpha)$% lower-confidence bound for $\mu$ is
  \[ \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \]
  and this gives an interval $(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$. (a lower bound on $\mu$)
Confidence Interval for $\mu$ - Normal parent pop’n, unknown $\sigma^2$

- What if we don’t know $\sigma$? Can I just plug-in my estimator for $\sigma$ (or $s$) and again have the same 95% CI?

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

This was a 95% CI for $\mu$ when $\sigma$ was known

$$\bar{x} - z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}$$

Is this a 95% CI for $\mu$?

$$\bar{x} - z_{0.025} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{s}{\sqrt{n}}$$

- HINT: This feels like cheating. If I don’t know $\sigma$, I must have more uncertainty in trying to capture $\mu$ than when I do know $\sigma$.

So, how do we incorporate this extra uncertainty (for not knowing $\sigma$)?
The answer comes from the \( t \)-distribution.

The 95% CI for \( \mu \) when \( \sigma \) unknown

\[
\bar{x} - t_{0.025,n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{0.025,n-1} \frac{s}{\sqrt{n}}
\]

where...

- \( t_{0.025,n-1} \) is the 97.5\(^{th}\) percentile of the \( t \)-distribution with \( n - 1 \) degrees of freedom (next slide).

- \( s \) is the sample standard deviation \( s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \)

- \( n \) is the number of observations (the sample size)
Connection to the $Z$-distribution...

- The $Z$ random variable follows a $N(0, 1)$ distribution

\[ Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \]

- A $T$ random variable follows a $t$-distribution

\[ T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \]

where $t_{n-1}$ is a $t$-distribution with $n - 1$ degrees of freedom.
What does the $t$-distribution look like?

There is only one $Z$-distribution, but there are many $t$-distributions (distinguished by their degrees of freedom $df$ as $t_{df}$). They look a lot like the $N(0, 1)$, except they have heavier tails.

For estimating a single parameter $\mu$, the degrees of freedom is $n - 1$.

The heaviness of the tails depends on the degrees of freedom (the subscript on the $t$), so it depends on the sample size $n$.

Differing $t$-distributions are shown below with $df = k$. 
Confidence Interval for $\mu$ - Normal parent pop’n, unknown $\sigma^2$

- For a large sample size $n$, $df = n - 1$ is very large, and the $t_{n-1}$ looks just like the $N(0, 1)$.

- So, $Z \sim N(0, 1)$ is the limiting distribution for $t_{n-1}$ as $n \to \infty$. 

![Graph showing the distribution of $t_{n-1}$ with increasing $k$ values and approaching $N(0, 1)$ as $k$ increases to infinity]
**Confidence Interval for \( \mu \) - Normal parent pop’n, unknown \( \sigma^2 \)**

- **100(1-\( \alpha \))% Confidence interval for mean, variance unknown**
  If \( \bar{x} \) is the sample mean and \( s \) is the sample standard deviation of a random sample of size \( n \) from a normal population, a 100(1 - \( \alpha \))% confidence interval for \( \mu \) is given by

\[
\bar{x} - t_{\alpha/2,df} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,df} \frac{s}{\sqrt{n}}
\]

- **How do I get the \( t_{\alpha/2,n-1} \) value?** (next slide)
Confidence Interval for $\mu$ - Normal parent pop’n, unknown $\sigma^2$

- How do I get the $t_{\alpha/2, n-1}$ value? Similar to getting a $z$-value.

- A $t$-table can be found in your book Appendix A, Table V, page A-11.
- When $\alpha = 0.05$ (for 95% CI) and the sample size is $n = 10$, 
  $$t_{\alpha/2, n-1} = t_{0.025,9}$$
- This is the $t$-value for a $t_9$ distribution with 2.5% above and 97.5% below. Looking at the table...
  $$t_{0.025,9} = 2.262$$
Confidence Interval for $\mu$ - Normal parent pop’n, unknown $\sigma^2$

Example (CI for $\mu$ using $t$-distribution)

Suppose a sample of size $n = 10$ is taken from a normal population and $\bar{x} = 8.94$ and $s = 4.3$. Construct a 95% CI for the population mean.

Upper end-point:
$$\bar{x} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} = \bar{x} + t_{0.025, 9} \cdot \frac{s}{\sqrt{n}} = 8.94 + 2.262 \left( \frac{4.3}{\sqrt{10}} \right) = 10.02$$

Lower end-point:
$$\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} = 8.94 - 2.262 \left( \frac{4.3}{\sqrt{10}} \right) = 5.86$$

The 95% confidence interval for $\mu$ is $[5.86, 10.02]$.

We are 95% confident that the true mean $\mu$ is between 5.86 and 10.02.
Confidence Interval for $\mu$ - Normal parent pop’n, unknown $\sigma^2$

Normality assumption for these $t$-based confidence intervals:

- When $\sigma^2$ is unknown and we have a rather small sample, we need the parent population to be normally distributed (or nearly normal) to truly achieve our $100(1 - \alpha)\%$ confidence level.
  - After we collect our data, we can check this assumption of normality by creating a *normal probability plot* (recall section 6.7).

- If the data are not normally distributed, we have to use a different approach. Something that doesn’t depend on this normality assumption, such methods are called *nonparametric methods* (which we won’t cover in this class).
$100(1 - \alpha)\%$ Confidence Interval for $\mu$

**RULE OF THUMB**

- When $\sigma$ is known, use
  \[
  \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
  \]

- When $\sigma$ is unknown, use
  \[
  \bar{x} - t_{\alpha/2,df} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,df} \frac{s}{\sqrt{n}}
  \]

When $n$ is REALLY LARGE ($n > 60$) a 95% CI for $\mu$ can be
\[
\bar{x} - z_{0.025} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{s}{\sqrt{n}}
\]

**NOTE:** At $n = 60$ the $Z$-table and $t_{60}$-table are very very similar.
But just use the rule of thumb,
which says $s$ goes with $t$ and $\sigma$ goes with $z$. 