ESTIMATING PROPORTIONS

- A **point estimate** is a single value used to approximate a population parameter. For example, the sample proportion is used as a point estimate for \( p \); which is the proportion of successes in the entire population.

- A **confidence interval** (CI) is a range of values used to estimate the true value of a population parameter. The confidence level \( 1 - \alpha \) gives the proportion of times that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times.
ESTIMATING PROPORTIONS

○ The margin of error \( E \) is the maximum likely difference (with probability \( 1 - \alpha \)) between the observed sample proportion \( \hat{p} \) and the true population proportion \( p \).

\[
E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ Variance of } \hat{p}
\]

○ Critical value

○ Then, the \((1 - \alpha)\) level CI is given by

\[
[\hat{p} - E, \hat{p} + E]
\]

○ Usually \((1 - \alpha)\) is 90%, 95%, 99%.

\[\hat{p} \text{ estimate of } p \text{ based in a sample}
\]

\[
\hat{p} = \frac{\text{successes}}{\text{total}}
\]
ESTIMATING PROPORTIONS

○ The critical value $z_{\alpha/2}$ can be obtained from a normal table (remember the normal approximation to the binomial) → How many standard deviations do I need to get the right coverage?

$z_{\alpha/2}$

$\alpha = 0.05 \implies z_{0.025} \approx 1.96$
ESTIMATING PROPORTIONS

- Also, if we want to determine the sample size for a given margin of error $E$ we can solve for $n$,

\[
 n = \frac{(z_{\alpha/2})^2 \hat{p}(1 - \hat{p})}{E^2} \quad \hat{p} \text{ known}
\]

\[
 n = \frac{(z_{\alpha/2})^2 0.25}{E^2} \quad \hat{p} \text{ unknown (upper bound)}
\]
**QUIZZ QUESTION**

- The drug Ziac is used to treat hypertension. In a clinical test, 3.2% of 221 Ziac users experienced dizziness. What would be a 98% confidence interval estimate of the proportion of Ziac users who experience dizziness?
  a) (0.004, 0.060).
  b) (0.010, 0.070).
  c) (0.002, 0.062).
  d) (0.011, 0.053).
  e) None of the above.

\[
\hat{p} = 0.032 \quad \alpha = 2\% \quad \sqrt{n} = 221 \\
2_{\alpha/2} = 2.33
\]

\[
E = 2.33 \times \sqrt{\frac{0.032 \times 0.968}{221}} = 0.0275
\]

CI(p) = (0.032 - 0.0275, 0.032 + 0.0275)
QUIZZ QUESTION

In the previous example, can we say that the probability that the interval contains the true proportion $p$ is 0.98?

a) Yes.
b) No.
c) Can't tell from the information provided.

The interval is the random quantity.
ESTIMATING A POPULATION MEAN (WHEN THE VARIANCE IS KNOWN)

- Margin of error:
  \[ E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \]
- Confidence interval:
  \[ (\bar{x} - E, \bar{x} + E) \]
- Sample size:
  \[ n = \left( \frac{z_{\alpha/2} \sigma}{E} \right)^2 \]

\[ x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2) \]

\[ \text{mean} \quad \text{standard deviation} \]
\[ \text{unknown} \quad (\text{known}) \]
MORE ON CONFIDENCE INTERVALS

- If we want a tighter interval we have 3 options:
  - Reduce the "noise" (variance of the observations).
  - Reduce the confidence level.
  - Increase the number of observations.

- Some ways are more effective than others: reducing the variance in half is equivalent to increasing the number of observations fourfold.
ESTIMATING A POPULATION MEAN
(WHEN THE VARIANCE IS KNOWN)

○ For $\alpha = 0.05$

\[
\left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)
\]

○ Remember that $\mu$ is a fixed quantity, while the confidence interval is random (sample dependent).
QUESTION

- Doctors collected the body temperatures of 106 men taken at 12 A.M. The mean of this sample is 98.20°F and its standard deviation is 0.62°F. What is the 95% level margin of error for the estimate of the mean?
  - 0.174
  - 0.065
  - 0.133
  - 0.118
  - None of the above.

\[ E = 2 \cdot \frac{\sigma}{\sqrt{n}} \]

\[ 2 \cdot \frac{0.62}{\sqrt{106}} = 0.118 \]

\[ 98.2 \pm 0.118 \]

\[ \text{CI}(\mu) = (98.2 - 0.118, \ 98.2 + 0.118) \]

\[ = (98.082, \ 98.318) \]
MORE ON CONFIDENCE INTERVALS

- If you answered e) to the previous questions, you are not completely wrong: The formula we provided before has \( \sigma \) on it, but we instead used \( s \). In other words

\[
\left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)
\]

assumes that we know \( \sigma \). Two questions arise:
- What can we do if \( \sigma \) is unknown?
- Are there any circumstances in which we can still use the formula for \( \sigma \) known when it is unknown?
THE STUDENT T DISTRIBUTION
The Student t distribution

- If the distribution of a population is essentially normal (approximately bell shaped) with mean \( \mu \), then the distribution of
  \[
  t = \frac{\bar{x} - \mu}{s/\sqrt{n}}
  \]
is essentially a Student t distribution with \( n - 1 \) degrees of freedom for all samples of size \( n \). The critical values are denoted \( t_{a/2}(n - 1) \).

Before
\[
\bar{x} = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]
\[\Rightarrow \] \[z_{5/2} = \bar{x} - \mu \Rightarrow \bar{x} = \mu + \frac{z_{5/2}}{s/\sqrt{n}}\]

Now
\[
t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \Rightarrow \bar{x} = \mu + t \frac{s}{\sqrt{n}}
\]
\[\downarrow \text{t distribution}\]
ESTIMATING A POPULATION MEAN
(WHEN THE VARIANCE IS UNKNOWN)

- If the observations are independent and identically distributed then we can use
  \[ t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \]
  to construct a confidence interval for \( \mu \) when \( \sigma \) is unknown.

- Margin of error: \( E = t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \)

- Confidence interval: \((\bar{x} - E, \bar{x} + E)\)
ESTIMATING A POPULATION MEAN
(WHEN THE VARIANCE IS UNKNOWN)

- These are the same formulas as before, but replacing $z_{\alpha/2}$ by $t_{\alpha/2}(n - 1)$.
- Note that, for large $n$, the t distribution is identical to the normal. In general, if $n > 30$, we use the formula for variance known even if it is being estimated from the sample.
CRITICAL T VALUES
<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>12.706</td>
<td>18.307</td>
<td>31.599</td>
<td>63.657</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>12.924</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>6.314</td>
<td>8.610</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>5.041</td>
<td>6.948</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.303</td>
<td>5.841</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.747</td>
<td>5.142</td>
</tr>
<tr>
<td>7</td>
<td>1.900</td>
<td>2.365</td>
<td>2.998</td>
<td>3.365</td>
<td>4.781</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.143</td>
<td>4.501</td>
</tr>
<tr>
<td>9</td>
<td>1.829</td>
<td>2.258</td>
<td>2.821</td>
<td>2.998</td>
<td>4.290</td>
</tr>
<tr>
<td>10</td>
<td>1.802</td>
<td>2.220</td>
<td>2.764</td>
<td>2.890</td>
<td>4.144</td>
</tr>
<tr>
<td>12</td>
<td>1.761</td>
<td>2.160</td>
<td>2.681</td>
<td>2.733</td>
<td>3.900</td>
</tr>
<tr>
<td>15</td>
<td>1.725</td>
<td>2.101</td>
<td>2.602</td>
<td>2.632</td>
<td>3.747</td>
</tr>
<tr>
<td>20</td>
<td>1.694</td>
<td>2.048</td>
<td>2.539</td>
<td>2.539</td>
<td>3.625</td>
</tr>
<tr>
<td>50</td>
<td>1.613</td>
<td>1.943</td>
<td>2.423</td>
<td>2.403</td>
<td>3.390</td>
</tr>
<tr>
<td>100</td>
<td>1.575</td>
<td>1.900</td>
<td>2.358</td>
<td>2.356</td>
<td>3.355</td>
</tr>
</tbody>
</table>
QUESTION

- In a classical 1908 paper by William Gosset, the following values were listed for the yields of head corn in pounds per acre. These values resulted from using regular seeds. What would be a 95% confidence interval estimate of the mean yield?
  - a) (1641.5, 2041.5)
  - b) (1611.3, 2071.7)
  - c) (1537.2, 2131.9)
  - d) (1672.1, 1945.6)
  - e) None of the above.

Data:
1903 1935 1910 2496 2108 1961 2060 1444 1612 1316 1511

\[
\begin{align*}
\bar{x} &= \frac{1903 + 1935 + 1910 + 2496 + \ldots + 1511}{11} \\
\bar{x} &= 1841.455 \\
S^2 &= \frac{1}{10} \left[ (1903 - 1841.455)^2 + (1935 - 1841.455)^2 + \ldots + (1511 - 1841.455)^2 \right] \\
S^2 &= 117.468.9 \implies S = 342.7373 \\
\text{To}_0.025 (10) &= 2.228
\end{align*}
\]
\[ E = t (n-1) \frac{s}{\sqrt{m}} = 2.228 \times \frac{342.3373}{230.2397} = 230.2397 \]

\[ CI(\mu) = (1841.455 - 230.2397, 1841.455 + 230.2397) = (1611.2, 2071.7) \]
ESTIMATING A POPULATION VARIANCE

- As before, we can use the variance of the sample

\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \]

- To construct confidence intervals for the variance we need to introduce the \( \chi^2 \) distribution. If we have an independent sample from a normal distribution, then

\[ \frac{(n - 1)s^2}{\sigma^2} \sim \chi^2(n - 1) \]
QUESTION

o Listed below are measured amounts of lead (in \( \mu g/m^3 \)) in the air. The EPA has established an air quality standard for lead: 1.5 \( \mu g/m^3 \). Use the given values to construct a 95% confidence interval for the estimate of the variance:

a) (1.195, 4.695).
b) (1.074, 4.772).
c) (1.344, 3.965).
d) (0.037, 2.782).
e) None of the above.

\[
\begin{align*}
\bar{x} &= 1.5383 \\
S^2 &= \frac{1}{5} \left( (5.4 - 1.5383)^2 + \ldots + (1.10 - 1.5383)^2 \right) \\
&= 3.664 \\
\chi^2_{0.025} &= 15.086, 0.831 \\
\chi^2_{0.175} &= 0.831, 15.086, 12.833 \\
CI(\sigma^2) &= \left( \frac{5 \times 3.664}{15.086}, \frac{5 \times 3.664}{12.833} \right) \\
&= (1.42, 22.04) \\
CI(\sigma^2) &= \left( \frac{1.42}{2.24}, 22.04 \right)
\end{align*}
\]
QUICK QUESTION

A sample of 40 smokers has a mean cotinine level of 172.5 ng/ml. Assuming that \( \sigma \) is known to be 119.5 ng/ml, find a 99% confidence interval estimate of the mean cotinine level of all smokers?

a) (141.4, 203.6)
b) (128.8, 217.3)
c) (137.1, 207.9)
d) (149.8, 195.2)
e) None of the above.

\[ n = 40 \]
\[ \bar{x} = 172.5 \]
\[ \sigma = 119.5 \]
\[ z_{0.005} = 2.575 \]

\[ C.I. (\mu) = (\bar{x} - z_{0.005} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.005} \frac{\sigma}{\sqrt{n}}) \]

\[ = (172.5 - 2.575 \frac{119.5}{\sqrt{40}}, 172.5 + 2.575 \frac{119.5}{\sqrt{40}}) \]

\[ = (123.84, 221.1536) \]