ESTIMATING PROPORTIONS

- A *point estimate* is a single value used to approximate a population parameter. For example, the sample proportion $\hat{p}$ 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.

\[ p = \frac{\text{# of successes}}{\text{# of trials}} \]
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 $p$.

$$E = \frac{z_{\alpha/2}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

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

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

- Usually $(1 - \alpha)$ is 90%, 95%, 99%.
ESTIMATING PROPORTIONS

- The critical value $z_{a/2}$ 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?

\[
\alpha = 5\% = 0.05 \Rightarrow z_{0.025} = 1.96
\]

\[
\alpha = 10\% = 0.10 \Rightarrow z_{0.05} = 1.645
\]
QUESTION

- A scientist conducts a genetics experiment by crossing green and yellow peas. Of the 580 peas obtained, 26.2% were yellow. In this case, what is the 95% confidence interval for the proportion of yellow peas?
  a) [0.231, 0.292].
  b) [0.219, 0.304].
  x  [0.226, 0.298].
  d) [0.272, 0.357].
  e) None of the above.

\[
\hat{p} = 0.262 \quad n = 580 \quad Z_{0.025} = 1.96
\]

\[
E = Z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \sqrt{\frac{0.262 \times 0.738}{580}}
\]

\[
E = 0.036 \quad \Rightarrow CI = [0.262 - 0.036, 0.262 + 0.036] = [0.226, 0.298]
\]
**QUESTION**

- Mendel's theory states that the proportion of yellow peas should be 25%. Does the previous result support Mendel's theory?
  1. Yes.
  2. No.
  3. The previous results do not give me any information to make this decision.
ESTIMATING PROPORTIONS

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

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

$$n = \frac{(z_{a/2})^2 0.25}{E^2} \quad \hat{p} \text{ unknown (upper bound)}$$
QUESTION

After reading about the results from the previous experiment, another scientist is planning to replicate the results. What is the minimum number of samples he should collect if he wants a 95\% margin of error of less than 1\%?

(a) 5000.
(b) 7000.
(c) 4000.
(d) 8000.
(e) 11000.

\[ E = 0.01 \]
\[ d = 0.05 \Rightarrow 2 \sqrt{d} = 1.96 \]
\[ P = \frac{1}{4} \]
\[ n = \frac{(1.96)^2 \times 0.25 \times 0.75}{(0.01)^2} = 7200 \]

\( \downarrow \)

round up (not down)
INTERPRETING CONFIDENCE INTERVALS

○ Tempting (but wrong): "There is a 95% chance that the true value of \( p \) will fall between 0.226 and 0.298."

○ Correct: "We are 95% confident that the interval from 0.226 to 0.298 actually does contain the true value of \( p \)."

○ *This is not only semantics:* The first one implies that \( p \) is random, the second one implies that the confidence interval is random.
INTERPRETING CONFIDENCE INTERVALS

- Note that the confidence interval was constructed based on the sampling distribution of the observed proportion.
- Therefore, the natural interpretation is that, if we were to obtain 100 samples and compute the confidence interval for each one of these intervals, about $(1 - \alpha\%)$ of them will contain the true proportion (and $\alpha\%$ of them will not!).
- When we report the confidence interval, we are trusting that we are in one of those $(1 - \alpha\%)$ cases (therefore, we report our "confidence" on the interval actually containing the value).
INTERPRETING CONFIDENCE INTERVALS

o 90% CI for 20 samples from Bin(600, 0.25).

<table>
<thead>
<tr>
<th>X</th>
<th>Lower Lim</th>
<th>Upper Lim</th>
</tr>
</thead>
<tbody>
<tr>
<td>169</td>
<td>0.251</td>
<td>0.312</td>
</tr>
<tr>
<td>139</td>
<td>0.203</td>
<td>0.260</td>
</tr>
<tr>
<td>147</td>
<td>0.216</td>
<td>0.274</td>
</tr>
<tr>
<td>167</td>
<td>0.248</td>
<td>0.308</td>
</tr>
<tr>
<td>150</td>
<td>0.221</td>
<td>0.279</td>
</tr>
<tr>
<td>159</td>
<td>0.235</td>
<td>0.295</td>
</tr>
<tr>
<td>140</td>
<td>0.205</td>
<td>0.262</td>
</tr>
<tr>
<td>150</td>
<td>0.221</td>
<td>0.279</td>
</tr>
<tr>
<td>139</td>
<td>0.203</td>
<td>0.260</td>
</tr>
<tr>
<td>149</td>
<td>0.219</td>
<td>0.277</td>
</tr>
<tr>
<td>166</td>
<td>0.247</td>
<td>0.307</td>
</tr>
<tr>
<td>152</td>
<td>0.224</td>
<td>0.283</td>
</tr>
<tr>
<td>152</td>
<td>0.224</td>
<td>0.283</td>
</tr>
<tr>
<td>153</td>
<td>0.226</td>
<td>0.284</td>
</tr>
<tr>
<td>138</td>
<td>0.202</td>
<td>0.258</td>
</tr>
<tr>
<td>143</td>
<td>0.210</td>
<td>0.267</td>
</tr>
<tr>
<td>169</td>
<td>0.251</td>
<td>0.312</td>
</tr>
<tr>
<td>154</td>
<td>0.227</td>
<td>0.286</td>
</tr>
<tr>
<td>149</td>
<td>0.219</td>
<td>0.277</td>
</tr>
<tr>
<td>153</td>
<td>0.226</td>
<td>0.284</td>
</tr>
</tbody>
</table>
INTERPRETING CONFIDENCE INTERVALS

- The CI is a random interval and has 95% chance of covering $p$ before a sample is chosen. Once a sample is drawn, and we compute the CI, it either contains $p$ or not, there is no probability involved!!!
MULTIPLE CHOICE. Choose the one alternative that best completes the statement or answers the question.

Identify the type of observational study (cross-sectional, retrospective, prospective).
1) The principal of Harrod High School plans to obtain data by interviewing students who graduate from Harrod High School to determine what percentage eventually graduate from college.
   A) Cross-Sectional B) Retrospective X) Prospective D) None of these

Identify which of these types of sampling is used: random, systematic, convenience, stratified, or cluster.
2) An education researcher randomly selects 48 middle schools and interviews all the teachers at each school.
   A) Convenience B) Stratified C) Cluster D) Random E) Systematic

Determine which score corresponds to the higher relative position.
3) Which score has a better relative position: a score of 72.4 on a test for which \( \bar{x} = 58 \) and \( s = 8 \), or a score of 306.8 on a test for which \( \bar{x} = 215 \) and \( s = 51 \)?
   A) A score of 306.8 B) 306.8 C) Both scores have the same relative position D) A score of 72.4

Solve the problem.
4) Suppose that replacement times for washing machines are normally distributed with a mean of 9.7 years and a standard deviation of 1.7 years. Find the replacement time that separates the top 18% from the bottom 82%.
   \( \bar{x} = 9.7 \) \( s = 1.7 \)
   A) 10.0 years B) 8.1 years C) 11.3 years D) 10.6 years

Find the minimum sample size you should use to assure that your estimate of \( \hat{p} \) will be within the required margin of error around the population \( p \).
5) Margin of error: 0.04; confidence level: 90%; from a prior study, \( \hat{p} \) is estimated by 0.25.
   A) 318 B) 282 C) 13 D) 954

SHORT ANSWER. Write the word or phrase that best completes each statement or answers the question.

Solve the problem.
6) A cat has a litter of 7 kittens. Find the probability that exactly 5 of the little furballs are female. Assume that male and female births are equally likely.

7) The Columbia Power Company experiences power failures with a mean of \( \mu = 0.210 \) per day. Find the probability that there are exactly two power failures in a particular day.
8) Among students at one college are 3806 women and 3022 men. The following table provides relative-frequency distributions for subject major for males and females at the college.

<table>
<thead>
<tr>
<th>Major</th>
<th>Relative frequency for women</th>
<th>Relative frequency for men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humanities</td>
<td>0.182</td>
<td>0.165</td>
</tr>
<tr>
<td>Science</td>
<td>0.291</td>
<td>0.336</td>
</tr>
<tr>
<td>Social Science</td>
<td>0.124</td>
<td>0.365</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>0.134</td>
</tr>
</tbody>
</table>

A student is selected at random from the college. Determine the probability that the student selected is female given that he or she is a Humanities major.

9) The annual precipitation amounts in a certain mountain range are normally distributed with a mean of 92 inches, and a standard deviation of 12 inches. What is the probability that the mean annual precipitation during 36 randomly picked years will be less than 94.8 inches?

10) The data shows the round trip mileage that 43 randomly selected professors and students drive to school each day. Use a frequency polygon to compare the results and determine whether there appears to be any significant difference between the two groups.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency (professors)</th>
<th>Score</th>
<th>Frequency (students)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-14</td>
<td>2</td>
<td>10-14</td>
<td>0</td>
</tr>
<tr>
<td>15-19</td>
<td>5</td>
<td>15-19</td>
<td>6</td>
</tr>
<tr>
<td>20-24</td>
<td>13</td>
<td>20-24</td>
<td>9</td>
</tr>
<tr>
<td>25-29</td>
<td>17</td>
<td>25-29</td>
<td>21</td>
</tr>
<tr>
<td>30-34</td>
<td>6</td>
<td>30-34</td>
<td>7</td>
</tr>
</tbody>
</table>
\[ P(X > q) = 0.18 \implies P(X \leq q) = 0.82. \]

2-score associated with 0.82 = 0.92

\[ q = 9.7 + 1.7 \times 0.915 = 11.3 \]

\[ \mu + \sigma \times 2 \cdot \text{score} \]

\[ n = \frac{(2\sqrt{p})^2 \hat{p}(1-\hat{p})}{E^2} = \frac{(1.64)^2 \times \frac{1}{4} \times \frac{3}{4}}{(0.04)^2} = 315. \]
$$X = \# \text{ of female kittens (out of 7)}$$

$$X \sim \text{Bin}(7, \frac{1}{2})$$

$$\Pr(X = 5) = \frac{7!}{5! \cdot 2!} \left(\frac{1}{2}\right)^5 \left(1 - \frac{1}{2}\right)^2$$

$$= \frac{7 \times 6 \times 5!}{5! \cdot 2} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^2$$

$$= \frac{21}{2} \left(\frac{1}{2}\right)^7 = \frac{21}{128}$$

---

$$X = \# \text{ of power failures (in a day)}$$

$$X \sim \text{Poisson}(0.210)$$

$$\Pr(X = 2) = e^{-0.21} \frac{(0.21)^2}{2!} = 0.0818$$
\[ P_r(F) = \frac{3806}{6828} \quad P_r(F^c) = \frac{3022}{6828} \]

H: Humanities major
S_c: Science major
S_s: Social Sciences major
O: Other major

0.165 = P_r(H | M)
0.336 = P_r(S_c | M)
0.182 = P_r(H | F)

\[
P_r(F | H) = \frac{P_r(H | F) P_r(F)}{P_r(H | F) P_r(F) + P_r(H | M) P_r(M)}
\]

\[
= \frac{0.182 \times \frac{3806}{6828}}{0.182 \times \frac{3806}{6828} + 0.165 \times \frac{3022}{6828}}
\]
\( X_i \sim N(92, 12) \) annual
\[
\bar{X} \sim N(92, \frac{12}{361}) \Rightarrow \bar{X} \sim N(92, 2)
\]
\[\Pr(\bar{X} < 94.8) = \Pr\left( z < \frac{94.8 - 92}{2} \right)\]
\[= \Pr(z < 1.4) = 0.9192\]
1) C
2) C
3) B
4) C
5) A
6) 0.16406
7) 0.018
8) 0.581
9) 0.9192
10) There does not appear to be a significant difference.

Dotted: Professor
Solid: Student