CHAPTER 8: Inference for a Single Proportion

We will now turn to looking at statistical inference on population proportions.

Examples
- How common is it for students at Purdue to fail a class? Out of a sample of 200 students, 50 of them have failed a class. That is 25% of them have failed a class. Based on these data, what can we say about all students at Purdue?
- What proportion of people in Lafayette receive a flu shot each year. From an SRS of 50 people living in Lafayette, 8 of them received a flu shot. What can we say about all people in Lafayette?

Note, we are interested in estimating the unknown proportion $p$ from a population. The statistic that estimates that parameter $p$ is the sample proportion $\hat{p}$.

Sampling Distribution of a Sample Proportion
Choose an SRS of size $n$ from a large population with population proportion $p$ having some characteristic of interest. It is usual to call whatever characteristic we are studying a “success.” Let $\hat{p}$ be the sample proportion of success,

$$\hat{p} = \frac{\text{count of successes in the sample}}{n}$$

Then:
- The sampling distribution of $\hat{p}$ is approximately normal and is closer to a normal distribution when the sample size $n$ is large. (ie: $np \geq 10$ and $(n(1-p)) \geq 10$.)
- The mean of the sampling distribution is $p$.
- The standard deviation of the sampling distribution is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Probabilities related to $\hat{p}$:
We can use the techniques from chapter 5 with proportions. Since $\hat{p}$ is approximately normal, we can find probabilities related to $\hat{p}$ by converting $\hat{p}$ to a Z-score. $Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$.

Large-Sample Confidence Interval for a Population Proportion
Choose an SRS of size $n$ from a large population with unknown proportion $p$ of successes.
- The sample proportion is $\hat{p} = \frac{X}{n}$ where $X$ is the total number of successes in the sample.
- The standard error of $\hat{p}$ is $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- An approximate level C confidence interval for $p$ is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Use this interval when the number of successes and the number of failures are both greater than or equal to 15 and the confidence level is 90%, 95%, or 99%.
Since p is a proportion, it takes on values between 0 and 1. If your CI has an upper bound of greater than one, just round the upper bound to one. If you get a lower bound that is negative, change it to 0.

**Example 1:** (Moore and McCabe 4th edition)
1. When trying to hire managers and executives, companies sometimes verify the academic credentials described by the applicants. One company that performs these checks summarized its findings for a six-month period. Of the 84 applicants whose credentials were checked, 15 lied about having a degree.

   a. Give an estimate for the proportion of applicants who lied about having a degree and give the estimate for the standard error of \( \hat{p} \).

   b. Consider these data to be a random sample of credentials from a large collection of similar applicants. Give a 95% confidence interval for the true proportion of applicants who lie about having a degree.

   c. Are the results of your confidence interval valid?

**Large-Sample Significance Test for a Population Proportion**
- State the Null and Alternative hypothesis.
  \[ H_0 : p = p_0 \]
  \[ H_a : p > p_0, H_a : p < p_0 \text{ or } H_a : p \neq p_0 \]
- Draw an SRS of size \( n \) from a large population with unknown proportion \( p \) of successes.
- Compute the test statistic: \( z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \) (Note: \( \hat{p} = \frac{X}{n} \)).
- Calculate the \( p \)-value.
  In terms of a standard normal random variable \( Z \), the approximate \( P \)-value for a test of \( H_0 \) against
  \[ H_a : p > p_0 \text{ is } P(Z > z_0) \]
  \[ H_a : p < p_0 \text{ is } P(Z < z_0) \]
  \[ H_a : p \neq p_0 \text{ is } 2P(Z > |z_0|) \]

It is fine to use the Large-Sample Significance Test for a Population Proportion if the expected number of successes (\( np_0 \)) and the expected number of failures (\( n(1-p_0) \) are both greater than 10. If this is not met, or if the population is less than 10 times as large as the sample, other procedures should be used.
- Compare the \( P \)-value to the \( \alpha \)-level.
  If \( P \)-value \( \leq \alpha \), then reject \( H_0 \) (significant results).
  If \( P \)-value \( \geq \alpha \), then fail to reject \( H_0 \) (non-significant results).
- State conclusions in terms of the problem.
  Reject (Do not reject) \( H_0 \). There is (is not enough) evidence that the population proportion of _______ is greater than/less than/not equal to _______.
Example 1 (continued):
2. Once again consider the data to be a random sample of credentials from a larger collection of similar applicants. In the past the company found that 15% of its applicants lied about having degrees. Has the proportion of applicants that have lied about having degrees increased? Use $\alpha = 0.05$.
   a. State the null and alternative hypotheses to answer this question.

   b. Calculate the $z$-statistic and $P$-value.

   c. What are your conclusions?

   d. Is your hypothesis test legitimate?

Sample Size for Desired Margin of Error: (page 579)
The level C confidence interval for a proportion $p$ will have a margin of error approximately equal to a specified value $m$ when the sample size satisfies

$$n = \left( \frac{z^*}{m} \right)^2 p^*(1 - p^*)$$

Here $z^*$ is the critical value for confidence C, and $p^*$ is a guessed value for the proportion of successes in the future sample.

Example 2:
1. You want to estimate the proportion of students at your college or university who are employed for 10 or more hours per week while classes are in session. You plan to present your results by a 95% confidence interval. Using the guessed value $p^* = 0.40$, find the sample size required if the interval is to have an approximate margin of error of $m = 0.06$. 
Comparing Two Proportions

Assumptions for comparing two Proportions:
- The data consist of the two **independent** SRS’s
- The two SRS’s are large.

Typically we want to compare two proportions by giving a confidence interval for \( p_1 - p_2 \), or by testing the hypothesis of no difference, \( H_0 : p_1 - p_2 = 0 \) or \( H_0 : p_1 = p_2 \).

**Confidence Intervals for Comparing Two Proportions:**
Choose an SRS of size \( n_1 \) from a large population having proportion \( p_1 \) of successes and an independent SRS of size \( n_2 \) from another population having proportion \( p_2 \) of successes. An **approximate level C confidence interval** for \( p_1 - p_2 \) is:

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE_D
\]

Where \( \hat{p}_1 = \frac{X_1}{n_1} \) and \( \hat{p}_2 = \frac{X_2}{n_2} \)

And \( SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \)

and \( z^* \) is the value for the standard Normal density curve with area \( C \) between \( -z^* \) and \( z^* \). The **margin of error** is \( m = \pm z^* \times SE_D \)

Use this method when the number of successes and the number of failures in both samples are all at least 10 and the confidence level is 90%, 95%, or 99%.

**Example 3:** (Moore and McCabe 4th edition)
1. Is lying about credentials by job applicants changing? In example 1, we looked at the proportion of applicants who lied about having a degree in a six-month period. To see if there is a change over time, we can compare that period with the following six months. Here are the data:

<table>
<thead>
<tr>
<th>Period</th>
<th>n</th>
<th>X(lied)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>106</td>
<td>21</td>
</tr>
</tbody>
</table>

Use a 95% confidence interval to address the question of interest.
Significance Tests for Comparing Two Proportions: (page 592)

- State the Null and Alternative hypothesis.
  
  To test the hypotheses
  
  $H_0 : p_1 = p_2$ against
  
  $H_a : p_1 - p_2 < 0$, $H_a : p_1 - p_2 > 0$ or $H_a : p_1 - p_2 \neq 0$

- Find the test statistic:
  
  Compute the $z$ statistic:
  
  $z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{D_E}}$

  where $\hat{p}_1 = \frac{X_1}{n_1}$ and $\hat{p}_2 = \frac{X_2}{n_2}$

  and where the pooled standard error is
  
  $SE_{D_E} = \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$

  and where
  
  $\hat{p} = \frac{X_1 + X_2}{(n_1 + n_2)}$

- Calculate the $P$-value
  
  In terms of a standard normal random variable $Z$, the $P$-value for the test of $H_0$ against
  
  $H_a : p_1 - p_2 > 0$ is $P(Z \geq z)$
  
  $H_a : p_1 - p_2 < 0$ is $P(Z \leq z)$
  
  $H_a : p_1 - p_2 \neq 0$ is $2P(Z \geq |z|)$

- Compare the $P$-value to the $\alpha$-level.
  
  If $P$-value $\leq \alpha$, then reject $H_0$ (significant results)

  If $P$-value $\geq \alpha$, then fail to reject $H_0$ (non-significant results)

- State conclusions in terms of the problem.

Note: This can be used if the number of successes and the number of failures in each sample is at least 5.

Example 3 (continued):
2. Data on the proportion of applicants who lied about having a degree in two consecutive six-month periods are given in the previous example. Formulate appropriate null and alternative hypotheses that can be addressed with these data, carry out the significance test, and summarize the results.