10.1 Inferences for a Population Proportion

Confidence Interval

Recall the 3 rules from 7.3.1 for the sampling distribution of $\hat{p}$.

Thus, the CI for a population proportion p is

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

Assumptions: random sample, $n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$ (OR 5 or 10)

Note that n being large also allows for the standard error to use $\hat{p}$ since p is unknown.

Ex10.1) A survey of 1356 random adults asked them to pick out the funniest city name in a list. 923 chose "Keokuk", 74 chose "Walla Walla", and 359 chose "Seattle". Let p be the proportion of all adults who would have answered "Seattle" had they been polled. Construct a 95% confidence interval for p.

10.1.3 Sample Size Calculations

Consider the CI as $\hat{p} \pm E$, where $E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Recall that E is the margin of error. The width of the CI is $L = 2E$. Now, we still want to see how large a sample size is required; hence, we rearrange to

$$n \approx \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Round up n to next integer. Replace $\hat{p}$ by a prior estimate. If we don't have this information, how do we make n as large as possible? By choosing $\hat{p} = 0.5$, we maximize $\hat{p}(1-\hat{p})$ and get a conservative choice for n. This choice is most common. If, however, we expect $\hat{p}$ to be close to 0 or 1, say $\hat{p} \leq 0.1$, then we could set $\hat{p} = 0.1$ to obtain a smaller n. In this situation, though, we would usually want a smaller E.

Ex10.2) If you wish to conduct a poll so that the margin of error is at most 3 percentage points with 99% confidence, how large of a sample size is required?

Hypothesis Test
Assumptions: random sample, $np_0 \geq 15$ and $n(1 - p_0) \geq 15$ (OR 5 or 10)

H_0: $p = p_0$ H_A: $p \neq p_0$ (one-sided tests also possible)

Test statistic: $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

P-value: Find from $Z \sim N(0,1)$.

Conclusion: Similar to one-sample mean.

Ex10.3) Use the data from Ex10.1). Suppose p is 0.240. Does the sample disprove this claim? Test the claim using both approaches and $\alpha = 0.01$.

(Note that one-tailed tests with this example were discussed in class.)

10.2 Two Population Proportions for Large Samples

Properties of the Sampling Distribution of $\hat{p}_1 - \hat{p}_2$:

If the random samples on which $\hat{p}_1$ and $\hat{p}_2$ are based are selected independently of one another, then

1. $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ (OR $p_A - p_B$ by textbook)
2. $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$ and $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma^2_{\hat{p}_1 - \hat{p}_2}}$

3. If both n_1 & n_2 are large (if $n_1p_1 \geq 15$, $n_1(1-p_1) \geq 15$, $n_2p_2 \geq 15$, & $n_2(1-p_2) \geq 15$), then $\hat{p}_1$ and $\hat{p}_2$ each have a sampling distribution that is (approximately) normal. This implies that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is also (approximately) normal.

Assumptions:
1. Samples are independent random samples.
2. The 3rd property mentioned above holds when using $\hat{p}_1$ and $\hat{p}_2$ instead of p_1 and p_2.

Confidence Interval
The $(1-\alpha)100\%$ CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Substituting $\hat{p}_1$ and $\hat{p}_2$ is possible since we have approximate normality.

Hypothesis Test
We still don't have p_1 or p_2 here but we can use our assumption of equal proportions to help us. If $p_1 = p_2$, either sample should estimate p well. Ergo, a weighted average of $\hat{p}_1$ and $\hat{p}_2$ that gives more weight to the larger sample is used.

Pooled sample proportion: $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \overline{p}$

Test statistic: $z_0 = \dfrac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

The value of $p_1 - p_2$ is usually 0. This is because the above test statistic is not appropriate for values other than zero. Since the value of zero is the most likely in applications, there is no need to discuss other options here.

P-value: No different than how we calculated it in Section 10.1.

Conclusion: Reject/do not reject as in one-sample test; answer hypotheses/question posed.

(Examples done in class.)