Ch. 19 – Two Population Proportions for Large Samples

Def'n: Two samples drawn from two populations are independent if the selection of one sample from one population does not affect the selection of the second sample from the second population. Otherwise, the samples are dependent.

Notation: Two samples require appropriate subscripts → p_1 and p_2, n_1 and n_2

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. $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$

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. Thus, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is also (approximately) normal.

Assumptions (for both the interval and the test):

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.

Hypotheses:

Although there are two population means (a.k.a. parameters) in our data structure and we consider them together as ONE parameter: $p_1 - p_2$. Thus, we have

H_0: $p_1 - p_2 = 0$ H_A: $p_1 - p_2 \neq 0$

Note that zero has a 'special' interpretation. Also, tests can be one-sided, so that will affect the hypotheses.

Test statistic:

We still don't have p_1 or p_2 here but we can use our assumption of equal proportions to give us a "better estimate". 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. The *pooled sample proportion* is

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2} = \overline{p}$$

Thus, an appropriate test statistic is

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

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 Ch. 17.

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

Ex19.1) Is the proportion of students who have watched *The Shawshank Redemption* different than the proportion of generic internet users? Answer the question by carrying out a full test at $\alpha = 0.02$ as well as constructing a 98% confidence interval.