STMATH 341, Chapter 10: Two Population Proportions

Example 1: A study found that of 549 participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths. Does this data suggest that the regular use of aspirin will decrease the incidence rate of colorectal cancer-specific deaths?

Inferences about the difference between p_1 and p_2.

Suppose we have two populations. Let p_1 be the proportion of individuals in population 1 with a certain characteristic, and let p_2 be the proportion of individuals in population 2 with the same characteristic. If we would like to compare these two proportions, we can use the following null and alternative hypotheses:

Null Hypothesis:	Alternative Hypothesis:
	$H_1: \ p_1 > p_2 \iff p_1 - p_2 > 0$ **OR**
$H_0: \ p_1 = p_2 \iff p_1 - p_2 = 0$	$H_1: \ p_1 < p_2 \iff p_1 - p_2 < 0$ **OR**
	$H_1: \ p_1 \neq p_2 \iff p_1 - p_2 \neq 0$

What is a good estimator for $p_1 - p_2$?

$$\begin{aligned}
n_1 &= \text{the size of the sample drawn from population 1} \\
x_1 &= \text{number of elements in the first sample with the characteristic} \\
\widehat{p}_1 &= \text{proportion of indvidiuals in the first sample with the characteristic}
\end{aligned}$$

$$\begin{aligned}
n_2 &= \text{the size of the sample drawn from population 2} \\
x_2 &= \text{number of elements in the second sample with the characteristic} \\
\widehat{p}_2 &= \text{proportion of indvidiuals in the second sample with the characteristic}
\end{aligned}$$

$\widehat{p}_1 - \widehat{p}_2$ is the natural estimator for $p_1 - p_2$. So we need to understand the sampling distribution of $\widehat{p}_1 - \widehat{p}_2$ in order to perform any inferences.

Recall, when we were just dealing with one sample, the standard deviation of $\widehat{p}$ is $\sigma_{\widehat{p}} = \sqrt{\dfrac{pq}{n}}$ where $q = 1 - p$, and so the variance is $\sigma^2_{\widehat{p}} = \dfrac{pq}{n}$

Just like with two means, $\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = \sigma^2_{\widehat{p}_1} + \sigma^2_{\widehat{p}_2} = \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}$ (where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$).

Mean and Standard Deviation of $\widehat{p}_1 - \widehat{p}_2$: For two large and independent samples, the sampling distribution of $\widehat{p}_1 - \widehat{p}_2$ is approximately normal with mean and standard deviation given by

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

$$\sigma_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Since we don't know p_1, q_1, p_2, and q_2, instead we'll use

$$s_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\widehat{p}_1 \widehat{q}_1}{n_1} + \frac{\widehat{p}_2 \widehat{q}_2}{n_2}}$$

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

$$(\widehat{p}_1 - \widehat{p}_2) \pm z\sqrt{\frac{\widehat{p}_1 \widehat{q}_1}{n_1} + \frac{\widehat{p}_2 \widehat{q}_2}{n_2}}$$

where z is obtained from the standard normal distribution for the given confidence level α.

Example 1: A study found that of 549 participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths. Does this data suggest that the regular use of aspirin will decrease the incidence rate of colorectal cancer-specific deaths?

(a) Find the 95% confidence interval for the difference in the proportion of cancer related deaths between the two groups.

(b) Check your calculations with $\texttt{STAT} \rightarrow \texttt{TESTS} \rightarrow \texttt{B} : \texttt{2} - \texttt{PropZInt}$.

When we perform hypothesis tests about $p_1 - p_2$, we assume the null hypothesis, that $p_1 = p_2$. If this is the case, then we can estimate the common value of $p_1 = p_2$ by

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \widehat{p}_1 + n_2 \widehat{p}_2}{n_1 + n_2}$$

This is just like assuming that samples 1 and 2 really came from the same population, and calculating the sample proportion for the combined sample.

This value $\overline{p}$ is called the **pooled sample proportion**, and it gives us a better estimate for $s_{\widehat{p}_1-\widehat{p}_2}$ when we're assuming the null hypothesis:

$$s_{\widehat{p}_1-\widehat{p}_2} = \sqrt{\overline{pq}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where $\overline{q} = 1 - \overline{p}$.

Test Statistic z for $\widehat{p}_1 - \widehat{p}_2$:

$$z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{s_{\widehat{p}_1-\widehat{p}_2}}$$

The value of $p_1 - p_2$ is substituted from H_0, and is usually 0.

Example 1 (cont.): Test at a 5% significance level whether the proportion of cancer-related deaths in the aspirin-using population is different than the proportion of deaths in the non-aspirin using population.

(d) State the null and alternative hypothesis.

(e) Compute $\overline{p}$, the pooled sample proportion, and $s_{\widehat{p}_1-\widehat{p}_2}$.

(f) Find the test statistic z.

(g) Find the p-value.

(h) Check your calculations with `STAT` $\to$ `TESTS` $\to$ `6 : 2 − PropZTest`.

(i) Make a decision:

REJECT H_0 DO NOT REJECT H_0

(j) State your conclusion.