# Stat 401, section 9.4 Difference Between Population Proportions

notes by Tim Pilachowski

In section 9.4, we will continue comparing two groups, but now, instead of looking at means, we'll be considering proportions. (See Lectures 6.1c, 7.2c and 8.4 [8th edition section 8.3].)

Recall that a sample proportion is calculated as $\hat{p} = \dfrac{\text{number of successes}}{\text{total number tested}} = \dfrac{X}{m}$.

When comparing population proportions, we'll be considering the parameter $p_1 - p_2$.

Proposition:

Let $\hat{p}_1 = \dfrac{X}{m}$ and $\hat{p}_2 = \dfrac{Y}{n}$ where $X \sim \text{Bin}(m,\ p_1)$ and $Y \sim \text{Bin}(n,\ p_2)$ with $X$ and $Y$ independent variables.

Then $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of the parameter $p_1 - p_2$, and $V(\hat{p}_1 - \hat{p}_2) = \dfrac{p_1 q_1}{m} + \dfrac{p_2 q_2}{n}$.

notes on the proof:

$$E(X) = mp_1, \quad E(Y) = np_2$$

$$\begin{aligned} E(\hat{p}_1 - \hat{p}_2) &= E\left(\frac{X}{m} - \frac{Y}{n}\right) \\ &= \frac{1}{m}E(X) - \frac{1}{n}E(Y) \\ &= \frac{1}{m}mp_1 - \frac{1}{n}np_2 \\ &= p_1 - p_2 \end{aligned}$$

$$V(X) = mp_1q_1, \quad V(Y) = np_2q_2$$

$$\begin{aligned} V(\hat{p}_1 - \hat{p}_2) &= V\left(\frac{X}{m} - \frac{Y}{n}\right) \\ &= \frac{1}{m^2}V(X) + \frac{1}{n^2}V(Y) \\ &= \frac{1}{m^2}mp_1q_1 + \frac{1}{n^2}np_2q_2 \\ &= \frac{p_1q_1}{m} + \frac{p_2q_2}{n} \end{aligned}$$

In this class, we'll only be considering cases where $mp_1 \geq 10,\ mq_1 \geq 10,\ np_2 \geq 10,\ nq_2 \geq 10$, that is, a large sample case for which both $\hat{p}_1 = \dfrac{X}{m}$ and $\hat{p}_2 = \dfrac{Y}{n}$ have approximately normal distributions. In particular, the standard error of $p_1 - p_2$ can be estimated using sample statistics, $\sigma(p_1 - p_2) \approx \sigma\left(\dfrac{X}{m} - \dfrac{Y}{n}\right) = \sqrt{\dfrac{\hat{p}_1\hat{q}_1}{m} + \dfrac{\hat{p}_2\hat{q}_2}{n}}$.

Provided we have these conditions, a (100 – $\alpha$)% confidence interval is given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}.$$

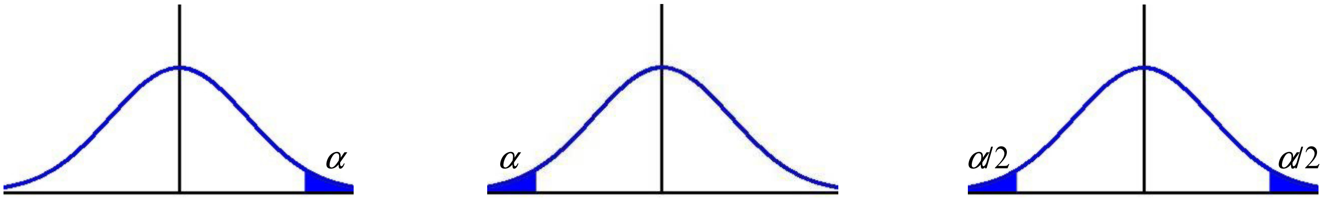
Example A – confidence interval. A news organization polls voters and asks, "Do you intend to vote for incumbent Senator Phillip E. Buster in the upcoming election?" Pollsters record the following numbers.

| | Yes | Undecided | No |
|---|---|---|---|
| Male | 72 | 11 | 38 |
| Female | 82 | 12 | 35 |

Construct a 90% confidence interval for the difference in the proportion of male voters who would vote for Senator Buster vs. the proportion of female voters who would vote for Senator Buster.

For a hypothesis test of the difference between population proportions, we have a further step we need to take.

Unlike the hypothesis test for a difference between means, for which both cases $\Delta_0 = 0$ and $\Delta_0 \neq 0$ used the same test statistic, a hypothesis test of the difference between population proportions for which $\Delta_0 \neq 0$ is significantly beyond the scope of this class. Instead, we'll focus solely on a hypothesis test for which we'll have $H_0 : p_1 - p_2 = 0$. (This matches practice for the vast majority of actual situations.)



| | | |
|---|---|---|
| $H_0 : p_1 - p_2 \leq 0$ | $H_0 : p_1 - p_2 \geq 0$ | $H_0 : p_1 - p_2 = 0$ |
| $H_a : p_1 - p_2 > 0$ | $H_a : p_1 - p_2 < 0$ | $H_a : p_1 - p_2 \neq 0$ |

The *P*-value will be $p = 2(1 - \Phi(|z|))$ for a two-tailed test, and $p = 1 - \Phi(|z|) = \Phi(-|z|)$ for a one-tailed test. If $p \leq \alpha$, reject $H_0$. Otherwise, fail to reject $H_0$. State the conclusion in the context of the problem.

However, the null hypothesis assumption $p_1 - p_2 = 0$ simplifies our hypothesis test process because it implies $p_1 = p_2$. That is, we can consider the separate samples of size *m* and *n* as being a single sample of size *m* + *n* from a single population with proportion *p*.

The estimator for this proportion *p* is thus $\hat{p} = \dfrac{X + Y}{m + n} = \dfrac{m * \hat{p}_1 + n * \hat{p}_2}{m + n}$.

For a large-sample hypothesis test of the difference between population proportions, the test statistic will be

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}.$$

Example A – hypothesis test. A news organization polls voters and asks, “Do you intend to vote for incumbent Senator Phillip E. Buster in the upcoming election?” Pollsters record the following numbers.

| | Yes | Undecided | No |
|---|---|---|---|
| Male | 72 | 11 | 38 |
| Female | 82 | 12 | 35 |

Conduct a hypothesis test ($\alpha = 0.10$) of the assertion that the proportion of male voters who would vote for Senator Buster is less than the proportion of female voters who would vote for Senator Buster.

When the situation doesn’t meet the large-sample rule of thumb (that is, at least one of $mp_1,\ mp_2,\ np_2,\ nq_2$ is less than 10), there are several methods which may be used, with some disagreement among statisticians over which is the best. Fortunately for us, the discussion is beyond the requirements of this class.

As with the hypothesis tests of the difference between means, we won’t be covering the probability of a Type II error $\beta$ and determination of sample size in this class.