Two Sample Proportion Tests in Stata

Before delving into two-way associations using contingency (two-by-two) tables, we first examine the structure of the two-sample test of proportions, using the normal approximation to the binomial.

Exercises:

1. How might we define a test statistic for comparing two proportions? Specifically, we would like to test the hypothesis that $H_0 : p_1 = p_0$ versus the alternative that $p_1 \neq p_0$ at the $\alpha = 0.05$ level. How does this test compare to the two-sample mean test for normally distributed data from last week?

   Recall the two-sample t-test for equal variances:

   Assume $X_1 \sim N(\mu_1, \sigma^2)$, and the sample mean of multiple realizations of X_1 is $\bar{x}_1$ and sample standard deviation is s_1; and $X_2 \sim N(\mu_2, \sigma^2)$, and the sample mean of multiple realizations of X_2 is $\bar{x}_2$ and sample standard deviation is s_2.

   To test $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$, our test statistic for the two-sample t-test with equal variances was:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2}$$

   Remember: the variance is independent of the mean for normally distributed data. For binomial data, the variance is a function of the mean.

   For binomial data:

   - Assume $X_1 \sim Binomial(n_1, p_1)$ and $X_0 \sim Binomial(n_0, p_0)$.
   - Define $\hat{p}_1 = X_1/n_1$ and $\hat{p}_0 = X_0/n_0$.
   - Using the Central Limit Theorem, we know that $\hat{p}_1 \sim N(p_1, p_1(1 - p_1)/n_1)$ and $\hat{p}_0 \sim N(p_0, p_0(1 - p_0)/n_0)$.
   - Under the null hypothesis that $p_1 = p_0$, $\hat{p}_1 - \hat{p}_0 \sim N(0, V)$, where $V = p(1 - p)\left(\frac{1}{n_1} + \frac{1}{n_0}\right)$ and $p = \frac{X_1+X_0}{n_1+n_0}$.
   - Therefore, a natural test statistic for testing $H_0 : p_1 = p_0, H_A : p_1 \neq p_0$ is:

$$\frac{\hat{p}_1 - \hat{p}_0}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}} \overset{H_0}{\sim} N(0, 1)$$

   For binomial data, the structure of the test statistic is similar to the two-sample t-test with equal variances, because, under the null, the variances are equal in both groups.

2. Let p_1/p_0 denote the proportion of CA residents below/above the federal poverty level who visited the doctor at least once in the past year. Test the hypothesis that $p_1 = p_0$ versus the alternative that $p_1 \neq p_0$ at the $\alpha = 0.05$ level. What do you conclude? Report a 95% CI along with your results.

   - What test are you using? Is normality reasonable?

     ```
     tabulate doctor
     ```

     Check that $n_1 p > 5, n_1(1-p) > 5, n_0 p > 5, n_0(1-p) > 5$, where $p = 0.804$.

     Two-group proportion test in Stata

     ```
     . prtest doctor, by(poverty)
     ```

   - What is the value of your test statistic?
     $Z = 2.3$
   - What is the distribution of your test statistic?
     $Z \sim N(0,1)$
   - What is the p-value of your test?
     $p = 0.024$
   - Do you reject or not reject the null hypothesis?
     Reject H_0
   - What do you conclude?
     There is evidence in the data that individuals in CA who are below poverty are less likely to go to the doctor.

3. Based on these data, you decide to conduct an intervention among those below the poverty line. You randomize individuals to intervention or no intervention. Suppose you power your study to detect a 15% risk difference with 90% power, assuming the proportion in the control group would equal the estimated proportion among those below poverty (70%) in this study. What sample size would you need, with equal numbers of individuals per arm, if you plan to conduct your test at the $\alpha = 0.05$?

```
. sampsi 0.7 0.85, power(0.9) alpha(0.05)
```