Econ 325/327
Notes on Two Population Proportions
By Hiro Kasahara

Estimator and Confidence Interval

We are often interested in a comparison of two population proportions. For example, in the study of lung cancer and smoking, Doll and Hill compared the proportion of smokers between patients with lung-cancer and patients with other diseases. In such a case, one may be interested in testing the hypothesis that the proportion of smokers is different between lung-cancer patients and other patients.

Suppose that $\{X_1, X_2, ..., X_{n_x}\}$ is a random sample, where X_i takes a value of zero or one with probability $1 - p_x$ and p_x, respectively. Suppose also that $\{Y_1, Y_2, ..., Y_{n_y}\}$ is a random sample, where Y_i takes a value of zero or one with probability $1 - p_y$ and p_y, respectively. Our concern is with the population difference $p_x - p_y$. In the example of lung cancer, p_x represents the proportion of smokers in population for the patients with lung-cancer while p_y represents the proportion of smokers in population for the patients with other diseases.

Because $\hat{p}_x := \bar{X} = (1/n_x)\sum_{i=1}^{n_x} X_i$ and $\hat{p}_y := \bar{Y} = (1/n_y)\sum_{i=1}^{n_y} Y_i$ provide the estimators of p_x and p_y, respectively, we may consider the estimator of the difference between p_x and p_y defined as $\hat{p}_x - \hat{p}_y$.

The expected value of $\hat{p}_x - \hat{p}_y$ is

$$E[\hat{p}_x - \hat{p}_y] = E[\hat{p}_x] - E[\hat{p}_y] = p_x - p_y$$

so that $\hat{p}_x - \hat{p}_y$ is an unbiased estimator of $p_x - p_y$.

The variance of $\hat{p}_x - \hat{p}_y$ is

$$\mathrm{Var}(\hat{p}_x - \hat{p}_y) = \mathrm{Var}(\hat{p}_x) + \mathrm{Var}(\hat{p}_y) - 2\mathrm{Cov}(\hat{p}_x, \hat{p}_y) = \frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y},$$

where the last line uses $\mathrm{Cov}(\hat{p}_x, \hat{p}_y) = 0$ because of the random sampling assumption (and hence $\hat{p}_x$ and $\hat{p}_y$ are independent).

Define a standardized version of $\hat{p}_x - \hat{p}_y$ as

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - E[\hat{p}_x - \hat{p}_y]}{\sqrt{\mathrm{Var}(\hat{p}_x - \hat{p}_y)}} = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}.$$

Because $\hat{p}_x$ and $\hat{p}_y$ converges to p_x and p_y in probability as the sample sizes n_x and n_y go to infinity, we may apply the Central Limit Theorem to show that the distribution of the random variable Z is approximated by a standard normal distribution when n_x and n_y are sufficiently large. Therefore,

$$\Pr\left(-z_{\alpha/2} \leq \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\mathrm{Var}(\hat{p}_x - \hat{p}_y)}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$\Leftrightarrow \Pr\left((\hat{p}_x - \hat{p}_y) - z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{p}_x - \hat{p}_y)} \leq p_x - p_y \leq (\hat{p}_x - \hat{p}_y) + z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{p}_x - \hat{p}_y)}\right) = 1 - \alpha,$$

where $\sqrt{\text{Var}(\hat{p}_x - \hat{p}_y)} = \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$ and $z_{\alpha/2}$ is the critical value such that $P(z_{\alpha/2} > Z) = 1 - \alpha$ when $Z \sim N(0,1)$.

Therefore, $100(1-\alpha)$ percent confidence interval for the difference between population proportions in large samples is given by

$$(\hat{p}_x - \hat{p}_y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}. \tag{1}$$

For example, if $1 - \alpha = 0.95$, then $z_{\alpha/2} = z_{0.025} = 1.96$.

Example 1 (Lung Cancer and Smoking) *Table 1 presents a fraction of smokers among lung-cancer patients, denoted by p_x, and a fraction of smokers among patients with other diseases, denoted by p_y. What is the 95 percent confidence interval for the difference between population proportions p_x and p_y?*

Denote the sample proportion of smokers among lung-cancer patients by $\hat{p}_x$ and denote the sample proportion of smokers among patients with other diseases by $\hat{p}_y$. In large sample, $\frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}} \sim N(0,1)$ by the Central Limit Theorem, where n_x and n_y are the sample size to compute $\hat{p}_x$ and $\hat{p}_y$. The 95 percent confidence interval can be constructed as (1) with $z_{\alpha/2} = z_{0.025} = 1.96$.

In this example, we have $\hat{p}_x = 0.995$, $\hat{p}_y = 0.955$, and $n_x = n_y = 1357$. Then, $\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = 0.00595$ and the 95 percent confidence interval is given as

$$[LB, UB] = [0.040 - 1.96 \times 0.00595, 0.040 + 1.96 \times 0.00595] = [0.028, 0.051].$$

Table 1: Lung Cancer and Smoking

Disease Group	No. of Non-Smokers	No. of Smokers
Men:		
1357 lung-cancer patients	7 (0.5%)	1350 (99.5%)
1357 patients with other diseases	61 (4.5%)	1296 (95.5%)

Notes: Computed from Table V of Doll and Hill (1952).

Hypothesis test when the sample size n is large

Consider the null hypothesis that

$$H_0 : p_x - p_y \leq 0 \quad \text{against} \quad H_1 : p_x - p_y > 0.$$

If this null hypothesis is rejected, then we have statistical evidence that the population proportion of the first sample with variable X is larger than that of the second sample with variable Y.

Consider the estimator of $p_x - p_y$, i.e., $\hat{p}_x - \hat{p}_y$. We are interested in deriving the distribution of $\hat{p}_x - \hat{p}_y$ when the null hypothesis is true. When the null hypothesis is true at $p_0 = p_x = p_y$, we have $p_x - p_y = 0$. Therefore, with n_x and n_y sufficiently large, we may approximate the distribution of $\hat{p}_x - \hat{p}_y$ when the null hypothesis is true by $N(0, \frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y})$. Therefore, to test H_0 at the α percent level, we consider a test statistic:

$$Z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}},$$

where $\hat{p}_0$ represents the estimate of the population proportion when the null hypothesis is true, i.e., $p_0 = p_x = p_y$.

We reject the null hypothesis $H_0 : p_x - p_y \leq 0$ at the α percent level if

$$Z \geq z_\alpha,$$

where $z_\alpha = \Phi(\alpha)$, given the standard normal cdf $\Phi(\cdot)$. For example, $z_{0.025} = \Phi(0.025) = 1.96$.

Example 2 (Mommograms) *Table 2 reports the two-way table of results of mammograms taken on 10,000 women with tumors (either benign or malignant) taken from page 509 of Bennett, Briggs, and Triola (2000). We are interested in evaluating if this test is useful to detect if one has malignant cancer or not. Assume that 1570 women with tumors are randomly sampled from a population of men with tumors whose test result is positive and that 8430 women with tumors are randomly sampled from a population of women with tumors whose test result is negative. Denote the population proportion of malignant tumors among the population of women with tumors whose test result is positive by p_x and that among the population of women with malignant tumors whose test result is negative by p_y. We have sample sizes of $n_x = 1570$ and $n_y = 8430$.*

We may test the null hypothesis that

$$H_0 : p_x - p_y \leq 0 \quad \textit{against} \quad H_1 > 0$$

If this null hypothesis is rejected, then we have statistical evidence that mammograms is useful for detecting malignant tumor.

To test H_0 at the 5 percent level, we consider a test statistic:

$$Z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}.$$

where

$$\hat{p}_x = \frac{85}{1570}, \quad \hat{p}_y = \frac{15}{8430}, \quad \textit{and} \quad \hat{p}_0 = \frac{100}{10000},$$

which leads to $Z = 19.14$. Note that $\hat{p}_0$ represents the estimate of the population proportion of women with malignant tumors when the null hypothesis is true, i.e., $p_0 = p_x = p_y$.

We reject the null hypothesis $H_0 : p_x - p_y \leq 0$ at the 5 percent level if

$$Z \geq z_{0.05} = 1.64.$$

Because $Z = 19.14$*, we reject the null hypothesis and concludes that there is statistical evidence that those who have positive test is more likely to have malignant tumor than those who have negative test.*

The p-value of test is to find the value of α *such that* $z_\alpha = 19.14$*. In standard normal distribution table* $z_\alpha = 3.39$ *when* $\alpha = 0.0003$*. So, we may conclude that the p-value is smaller than* 0.0003*, i.e., the hypothesis will be rejected even at the 0.03 percent significance level.*

Table 2: Two-way table of results of tests on 10,000 patients with Tumors

	Malignant (cancer)	Benign (no cancer)	Total
Test Positive	85	1485	1570
Test Negative	15	8415	8430
Total	100	9900	10000

Notes: From Table 10.19 of Bennett, Briggs, and Triola (2000).