Chapter 8: Comparing Two Independent Populations
Part 3: Comparing two population proportions

The concepts in this section are covered in section 10.3 of Ott and Longnecker.

1 An example

Consider the following example. Does handedness differ according to sex? A sample of $n_M = 54$ males and $n_F = 21$ females was taken, and each person was asked to indicate which was their dominant hand. The data are as follows:

Female: 12 left, 9 right
Male: 23 left, 31 right

If we let π_{FL} be the proportion of females that are left-handed, and π_{ML} be the proportion of males that are left-handed, then our hypotheses could be expressed as:

$H_0 : \pi_{FL} - \pi_{ML} = 0$
$H_A : \pi_{FL} - \pi_{ML} \neq 0$

The hypotheses could equally well be expressed in terms of proportion that are right-handed. The form of the hypotheses suggests that the difference of the sample proportions, $\hat{\pi}_{FL} - \hat{\pi}_{ML}$ would be a natural choice of test statistic. Provided the sample sizes are large enough, it can be shown that:

$$\hat{\pi}_{FL} - \hat{\pi}_{ML} \mathrel{\dot\sim} N(\pi_{FL} - \pi_{ML},\ \tfrac{\pi_{FL}(1-\pi_{FL})}{n_F} + \tfrac{\pi_{ML}(1-\pi_{ML})}{n_M}).$$

This expression can be derived using the fact that the numbers of left-handed people in each population are distributed as binomials, and by using rules of expectation and variance and the CLT.

It may be worthwhile to review the expectation and variance of a binomial here, and perhaps go through some steps of pulling the expectation and variance through the expression.

This is a general result. However, for the purposes of testing, we are primarily concerned with the distribution of the test statistic under the null hypothesis. When the null is true, $\pi_{FL} = \pi_{ML} = \pi_L$, and the expression simplifies:

$$\hat{\pi}_{FL} - \hat{\pi}_{ML} \mathrel{\dot\sim} N(0,\ \pi_L(1-\pi_L)(\tfrac{1}{n_F} + \tfrac{1}{n_M}).$$

The common proportion π_L is unknown, but can be estimated using a weighted average of the two individual sample proportions:

$$\hat{\pi}_L = \tfrac{\hat{\pi}_{FL} n_F + \hat{\pi}_{ML} n_M}{n_F + n_M}.$$

Usually, the test statistic is given in standardized form:

$$\frac{\hat{\pi}_{FL}-\hat{\pi}_{ML}}{\sqrt{\hat{\pi}_L(1-\hat{\pi}_L)(\frac{1}{n_F}+\frac{1}{n_M})}} \dot{\sim} N(0,1).$$

Then p-values can be computed using the standard normal.

So, how large does the sample size need to be? For testing purposes, what is required is that $\pi_L n_F$, $(1-\pi_L)n_F$, $\pi_L n_M$, and $(1-\pi_L)n_M$ are all greater than 5. We can use $\hat{\pi}_L$ as an estimate of π_L for the purposes of evaluating this.

Let's now do the test using our data. $\hat{\pi}_{FL} = 0.571$, $\hat{\pi}_{ML} = 0.426$, and $\hat{\pi}_L = \frac{12+23}{21+54} = 0.467$. All of the requirements for large sample size are met. Our test statistic is thus:

$$\frac{0.571-0.426-0}{\sqrt{0.467(1-0.467)(\frac{1}{21}+\frac{1}{54})}} = 1.13.$$

Comparing this to a standard normal, we find $p - value = 0.258$. Thus we conclude that there is not enough evidence to say that males and females have a different proportion of left-handed individuals.

2 Recap

The data consists of separate samples from two populations, label them 1 and 2. Let:
π_1 = true proportion in population 1
π_2 = true proportion in population 2
n_1 = sample size taken from population 1
n_2 = sample size taken from population 2

We wish to test:
$H_0 : \pi_1 - \pi_2 = 0$ vs. $H_A : \pi_1 - \pi_2 \neq 0$

When the null is true, $\pi_1 = \pi_2 = \pi$. The unknown π can be estimated using a weighted average of the two individual sample proportions:

$$\hat{\pi} = \frac{\hat{\pi}_1 n_1 + \hat{\pi}_2 n_2}{n_1+n_2}.$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are the sample proportions as computed from the two samples. If based on our prior knowledge we are willing to assume:

- All of the data points are independent, both within and between populations
- The sample sizes are large enough (πn_1, $(1-\pi)n_1$, πn_2, and $(1-\pi)n_2$ are all greater than 5).

Then the test statistic is:

$$\frac{\hat{\pi}_1-\hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1}+\frac{1}{n_2})}} \dot{\sim} N(0,1).$$

Calculate the p-value and compare to the given significance level α.