8-Two-Sample Inferences

ENSY SILVER[1]

Saturday 5th September, 2020

[1]Thanks to my family, my friend and freedom.

1 Introduction

In the previous chapters, we have introduced that how to draw inference about parameters in a distribution, for example, to construct the confidence interval for μ in a normal distribution.

However, in most cases, wee are required to compare parameters in two different distributions, for example, if X and Y are normally distributed random variables, we need to test if $\mu_X = \mu_Y$.

In this chapter, we will introduce the two-sample inferences, with some special cases.

2 Testing $H_0 : \mu_X = \mu_Y$

This case is trivial, we introduce the theorem and omit the proof.

Theorem 2.1. *Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal distribution with mean μ_X and standard deviation σ and let $Y_1, Y_2, \cdots, Y_m$ be a random sample of size m from a normal distribution with mean μ_Y and standard deviation σ.*

Let S_X^2 and S_Y^2 be the two corresponding sample variances, and S_p^2 the pooled variance, where

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Then

$$T_{n+m-2} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has a Student t distribution with $n+m-2$ degrees of freedom.

If $H_0 : \mu_X = \mu_Y$ is true, then the variable $t = (\overline{X} - \overline{Y})/(S_p\sqrt{\frac{1}{n} + \frac{1}{m}})$ follows Student t distribution with $n+m-2$ degrees of freedom, which allows us to test H_0 at the α level of significance.

2.1 The Beehrens-Fisher problem

When the $\sigma_X \neq \sigma_Y$, the problem becomes complicated, which is called the **Beehrens-Fisher problem**. No exact solution is known, but a widely used approximation is based on the test statistic

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

B. L. Welch, a faculty member at UCL, in a 1938 Biometrika article showed that W is approximately distributed as a Student t random variable with degrees of freedom given by the nonintuitive expression

$$\frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^2}{n_1^2(n_1-1)} + \frac{\sigma_2^2}{n_2^2(n_2-1)}}$$

To understand Welch' s approximation, it helps to rewrite the random variable W as

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \div \frac{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

In this form, the numerator is a standard normal variable. Suppose there is a chi square random variable V with v degrees of freedom such that the square of the denominator is equal to V/v. Then the expression would indeed be a Student t variable with v degrees of freedom. However, in general, the denominator will not have exactly that distribution. The strategy, then, is to find an approximate equality for

$$\frac{S_X^2}{n} + \frac{S_Y^2}{m} = \left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)\frac{V}{v} \tag{1}$$

Here comes the peoblem, the textbook referred that the value of v can be deduced if the means and variance of both sides are equated in 1. However, the variance of chi square distribution with t degrees of freedom is $2t$, and the meean is t, and I can not get the result when I take the value into the equation **??**.

Now we move on, and let $\theta = \sigma_X^2/\sigma_Y^2$, then

$$v = \frac{\left(\theta + \frac{n}{m}\right)^2}{\frac{1}{n-1}\theta^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2} \tag{2}$$

Finally, we have the theorem for the case that $\sigma_X \neq \sigma_Y$.

Theorem 2.2. *Let $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_m$ be independent random samples from normal distributions with means μ_X and μ_Y, and standard deviations σ_X and σ_Y , respectively. Let*

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

Using $\hat{\theta} = S_X^2/S_Y^2$, take v to be the expression 2, rounded to the nearest integer. Then W has approximately a Student t distribution with v degrees of freedom.

3 Testing $H_0 : \sigma_X^2 = \sigma_Y^2$

Since S_X^2 and S_Y^2 are independent variables meet chi square distributions, the variable S_X/S_Y meet F distribution. Then, we can draw inference about S_X^2/S_Y^2.

4 Binomial data: testing $H_0 : p_X = p_Y$

Suppose that n Bernoulli trials related to treatment X have resulted in x successes, and that m (independent) Bernoulli trials related to treatment Y have yielded y successes.

The likelihood function can be writteen

$$L = P_X^x(1 - p_X)^{n-x} p_Y^y (1 - p_Y)^{m-y}$$

Setting the derivative of ln L with respect to p equal to 0 and solving for p gives a result that

$$p_e = \frac{x+y}{n+m}$$

The approach is to appeal to the central limit theorem and make the observation that

$$\frac{\frac{X}{n} - \frac{Y}{m} - E\left(\frac{X}{n} - \frac{Y}{m}\right)}{\sqrt{\operatorname{Var}\left(\frac{X}{n} - \frac{Y}{m}\right)}}$$

has an approximate standard normal distribution. Under H_0, of course

$$E\left(\frac{X}{n} - \frac{Y}{m}\right) = 0$$

and

$$\operatorname{Var}\left(\frac{X}{n} - \frac{Y}{m}\right) = \frac{p(1-p)}{n} + \frac{p(1-p)}{m} \tag{3}$$

Then, we can test 3 at the α level of significance.

5 Confidence intervals for the two-sample problem

The technique of testing hypothesis and build confidence intervals is the same, so we give the theorem directly.

Theorem 5.1. *Let $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_m$ be independent random samples drawn from normal distributions with means μ_X and μ_Y respectively, and with the same standard deviation, σ. Let s_p denote the data' s pooled standard deviation. A $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by*

$$\left(\overline{x} - \overline{y} - t_{\alpha/2,n+m-2} \cdot s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{x} - \overline{y} + t_{\alpha/2,n+m-2} \cdot s_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

Theorem 5.2. *Let $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_m$ be independent random samples drawn from normal distributions with standard deviations σ_X and σ_Y , respectively. A $100(1-\alpha)\%$ confidence interval for the variance ratio, σ_X^2/σ_Y^2, is given by*

$$\left(\frac{s_X^2}{s_Y^2} F_{\alpha/2,m-1,n-1}, \frac{s_X^2}{s_Y^2} F_{1-\alpha/2,m-1,n-1}\right)$$