Two samples test and its package: the tscvh R package

R. Fajriyah [1]

[1] Department of Statistics, Universitas Islam Indonesia, Jogjakarta, Indonesia

Workshop
Yogyakarta, November 27, 2017

Outline

1. Motivation
   - Cross-Variance concept,[1] and [2]
   - Special case
2. The tscvh package
   - Main functions
   - Demo
3. Summary

Outline

Definition

Suppose we have two independent samples, X_i and Y_j; $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Their sample mean and variance are denoted by $\overline{X}, \overline{Y}$ and V_x, V_y. Let

$$V_x^* = \frac{\sum_{i=1}^{m}(X_i - \overline{Y})^2}{m-1} \text{ and } V_y^* = \frac{\sum_{i=1}^{n}(Y_i - \overline{X})^2}{n-1}$$

be the cross-variance for each sample X and Y respectively. The cross-variance sample of groups **X** and **Y** is defined as

$$\mathbf{T} = \frac{V_x^a + V_y^a}{2}, \tag{1}$$

where $V_x^a = \frac{V_x}{V_x^*}, \quad V_y^a = \frac{V_y}{V_y^*}.$

$T = \frac{V_x^a + V_y^a}{2}$ can be re-written as

$$\begin{aligned} T &= \frac{1}{2}\left[\frac{V_x}{V_x + \frac{m(\overline{Y}-\overline{X})^2}{m-1}} + \frac{V_y}{V_y + \frac{n(\overline{Y}-\overline{X})^2}{n-1}}\right] \\ &= \frac{V_x}{2V_x + 2\frac{m(\overline{Y}-\overline{X})^2}{m-1}} + \frac{V_y}{2V_y + 2\frac{n(\overline{Y}-\overline{X})^2}{n-1}} \\ &= Z_1 + Z_2 \end{aligned} \tag{2}$$

Assumptions

1. the sample sizes are equal
2. X_i and Y_i are i.i.d. normally distributed with unknown means and known variances σ_x^2, σ_y^2.

Equation (2) can be written as

$$T = Z_1 + Z_2 = \frac{U}{2U + 2abV} + \frac{S}{2S + 2bcV} \tag{3}$$

with

$$U = \frac{(n-1)V_x}{\sigma_x^2}, \quad S = \frac{(n-1)V_y}{\sigma_y^2}, \quad V = \frac{n(\overline{Y} - \overline{X})^2}{(\sigma_x^2 + \sigma_y^2)}$$

and

$$a = \frac{1}{\sigma_x^2}, \quad b = (\sigma_x^2 + \sigma_y^2), \quad c = \frac{1}{\sigma_y^2}.$$

To compute the distribution of T in Equation (3), it can be done by considering the fact that

1. U, V and S are independent
2. Z_1 and Z_2 are dependent

In this talk we will describe the first case, where we consider that U, V and S are independent.

Outline

1. Motivation
   - Cross-Variance concept,[1] and [2]
   - Special case
2. The tscvh package
   - Main functions
   - Demo
3. Summary

In the case of $\sigma_x^2 = \sigma_y^2$

$$T^* = \frac{U^*}{U^* + 4V^*} \tag{4}$$

where $U^* = \frac{(n-1)\left(V_x + V_y\right)}{\sigma_x^2}$ and $V^* = \frac{n\left(\overline{Y} - \overline{X}\right)^2}{2\sigma_x^2}$.

The pdf of T^* is derived from the ratio of linear combination of chi-square random variables [5]. First, let $Y = 1 + 4\frac{V^*}{U^*}$, where V^* is distributed $\chi^2_{(1)}$ and U^* is distributed $\chi^2_{2(n-1)}$. Second, the pdf of T^* is computed by taking $T^* = \frac{1}{Y}$.

Theorem

Let $\sigma_x^2 = \sigma_y^2$ *and the* V_x *and* V_y *is estimated by the pooled variance* $S_p^2 = \frac{V_x + V_y}{2}$, *then* $T*$ *in equation (4) has a pdf*

$$f_{T^*}(t^*) = \frac{4^{n-1} t^{n-2} (1 - t^*)^{\frac{1}{2}-1}}{B\left(\frac{1}{2}, n-1\right) (1 + 3t^*)^{n-\frac{1}{2}}}, \quad 0 \leq t^* \leq 1 \qquad (5)$$

about n

When the sample sizes are different

- max (n_1, n_2)
- average (n_1, n_2)

Outline

The functions I

There are 10 functions on the package:

1. **cumr**: This function is a cumulative probability density function for the Tstar random variable. It provides a cumulative probability density function value of the Tstar
2. **fr**: This function is a density function for the Tstar random variable. It provides a probability value of the Tstar
3. **quantr**: This function provides a value for the qth quantile of the Tstar
4. **tstf**: This function is a main function for doing the two-sided cross variance test. A test for testing the difference of mean between two samples
5. **tstf3c**: The two sample test call functions cross-variance

The functions II

6. **tstf3p**: The two sample test pooled functions cross-variance
7. **tstfa**: The two sample test function cross-variance with the option is average
8. **tstfma**: The two sample test function cross-variance with the option is maximum
9. **tstfmi**: The two sample test function cross-variance with the option is minimum
10. **tstfn**: The two sample test function cross-variance with the option is none

Outline

Install the package

- `tscvh_1.0.tar.gz`
- *install.packages(" tscvh")*
- `RCMDinstall~/Downloads/tscvh_1.0.tar.gz`

Use the package

- library(tscvh)
- x=c(35,40,15,21,46,28,48,16,30,32,48,31,22,39,25)
- y=c(2,27,31,1,19,1,3,1,2,3,2,2,3,37,2)
- tstf(d1=x,d2=y,alpha=0.05)
- u=c(26,21,22,26,19,22,26,25)
- v=c(18,23,21,20,20,29,20,16,20,26,21,25,17,18,19)
- tstf(d1=u,d2=v,alpha=0.05)

Summary

We have introduced

1. the cross-variance concept,
2. a new test based on the cross-variance,
3. the special case of the cross variance (as the proposed test)
4. the new probability density functions.
5. the package for the new test in case of homogeneit of variance
6. a demo on how to install and use it

Thank you.

Acknowledgement

The authors would like to thanks to

1. the Indonesian Government for its financial support of the Research Grant No. 012/DirDPPM/70/DPPM/ PascaDoktor- KEMENRISTEKDIKTI/IV/2017, under the Ministry of Research, Technology and Higher Education, Republic of Indonesia
2. Universitas Islam Indonesia, Jogjakarta, Indonesia
3. Prof. István Berkes, TU Graz, Austria

References I

- Fajriyah, R., *Micorarray Data Analysis: Background Correction and Diferentially Expressed Genes*, (2014), Dissertation. Graz, Austria: Graz University of Technology, 2014.
- Fajriyah, R., *Introducing a cross variance concept and its application*, (2016), 12^{th} ICMSA Proceeding, IEEE CFP 16 H58 ART, 2016.
- Gradshteyn, I. S. and Ryzhik, I. M., *Table of integrals, series, and prod- ucts. 7th edition*. New York: Academic Press, 2007.
- Rao, C. R., *Linear statistical inference and its applications. Second edition*. New York: Wiley, 1971.

References II

- Serge Provost and Edmund M. Rudiuk. The exact density function of the ratio of two dependent linear combinations of chi-square variables. *Annals of the Institute of Statistical Mathematics*, 46(3):557–571, 1994
- Welch, B. L., The generalization of ”student's” problem when several different population variances are involved. *Biometrika*, 34, 28–35, 1947.