IOP Conf. Series: Materials Science and Engineering **300** (2018) 012030 doi:10.1088/1757-899X/300/1/012030

The Robustness of Two Independent Samples t Test Using Monte Carlo Simulation with R

[1]Prana Ugiana Gio, [2]Elly Rosmaini

[1,2] Department of Mathematics, University of Sumatera Utara

Email: gioprana89@gmail.com, ellyrosmaini@gmail.com

Abstract. Two independent samples t test is a parametric statistical method that have several assumptions underlying this method. The assumption applied in the use of two independent samples t test is normal population and equal variance. These assumptions will affect accuracy of the result if the underlying assumptions satisfied or not. Many people use the t statistic to compare two samples even when the underlying assumptions are in doubt. This paper is intended as an introductory article that gives brief overview about the robustness or sensitivity of this popular test statistic with respect to changes in the assumptions. One way to investigate the robustness or sensitivity of two independent samples t test using Monte Carlo simulation with R.

1. Introduction

Suppose given two independent samples, $x_1, x_2, x_3, \dots, x_m$ and $y_1, y_2, y_3, \dots, y_n$. Note that m and n are respectively the number of element in x and y sample. In this case wishes to test the hypothesis that mean of the x population are equal with mean of the y population that mathematically can be stated

$$H_0: \mu_x = \mu_y. \qquad (1)$$

Let $\bar{x}$ and $\bar{y}$ denote the mean of x and y sample, with standard deviation for each sample s_x and s_y. The standard test to test (1) is based on the t statistic

$$T = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}, \qquad (2)$$

with s_p denotes pooled standard deviation that can be calculated

$$s_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}}. \qquad (3)$$

Under the hypothesis H_0 (Walpole et al., 2007; Albert, 2009; Mann, 2013), the test statistic T has t distribution with $m + n - 2$ degree of freedom, when:

- Both the x and y samples are independent random samples from normal distributions.
- The standard deviation of x and y populations are equals, $\sigma_x = \sigma_y$

Suppose the significance level of the test is stated at α. The hypothesis H_0 will be rejected when

$$|T| \geq t_{n+m-2,\frac{\alpha}{2}}, \quad (4)$$

where $t_{df,\alpha}$ is the $(1-\alpha)$ quantile of a t random variable with df degrees of freedom (Albert, 2009).

If the underlying assumptions are satisfied, namely x population and y population are normal distribution and have equal variances, then the level of significance of t test will be stated at α (Albert, 2009). But, in practice, many people use t statistic to compare two samples even when the underlying assumption are in doubt (Albert, 2009). Therefore, one of the interesting problem is to investigate sensitivity and robustness of this popular statistic test respect to the change of the underlying assumptions using Monte Carlo simulation with R.

2. Probability Distribution

2.1. Normal Distribution

A random variable X with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

is normal random variable with parameters μ where $-\infty < \mu < \infty$ and $\sigma > 0$ (Montgomery and Runger, 2014). Note that μ is location parameter and σ is scale parameter.

2.2. t Distribution

Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. The random variable

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (6)$$

has a t distribution with $n-1$ degrees of freedom (Montgomery and Runger, 2014). The t probability density function is

$$f(t) = \frac{\Gamma\left[\frac{(k+1)}{2}\right]}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \times \frac{1}{\left[\left(\frac{t^2}{k}\right)+1\right]^{\frac{(k+1)}{2}}} \quad ; -\infty < t < \infty \quad (7)$$

where k is the number of degrees of freedom. The mean and variance of t distribution are zero and $\frac{k}{k-2}$ (for $k > 2$), respectively (Montgomery and Runger, 2014).

2.3. Exponential Distribution

L Montgomery and Runger (2014) a continuous random variable X with probability density function

$$f(x) = \lambda e^{-\lambda x} \quad ; 0 \leq x < \infty \quad (8)$$

is exponential random variable with parameter $\lambda > 0$. Note that λ is also called rate parameter. If the random variable X has an exponential distribution with parameter λ,

$$\mu = \frac{1}{\lambda} \quad and \quad \sigma^2 = \frac{1}{\lambda^2}. \quad (9)$$

3. Monte Carlo Simulation

To apply the Monte Carol method, the analyst constructs a mathematical model that simulates a real system. A large number of random sampling of the model is applied yielding a large number of random samples of output results from the model. The origin began in the 1940s by three scientists, John von Neumann, Stanislaw Ulam and Nicholas Metropolis who were employed on a secret assignment in the Los Alamos National Laboratory, while working on a nuclear weapon project called the Manhattan Project. The Manhattan team formulated a model of a system they were studying that included input variables, and a series of algorithms that were too complicated to analytically solve (Thomopoulos, 2013).

The method is based on running the model many times as in random sampling. For each sample, random variates are generated on each input variable, computations are run through the model yielding random outcomes on each output variable. Since each input is random, the outcomes are random. In the same way, they generated thousands of such samples and achieved thousands of outcomes for each output variable. In order to carry out this method, a large stream of random numbers were needed (Thomopoulos, 2013).

The Monte Carlo method proved to be successful and was an important instrument in the Manhattan Project. After the War, during the 1940s, the method was continually in use and became a prominent tool in the development of the hydrogen bomb. The Rand Corporation and the U.S. Air Force were two of the top organizations that were funding and circulating information on the use of the Monte Carlo method. Soon, applications started popping up in all sorts of situations in business, engineering, science and finance (Thomopoulos, 2013).

4. R Programming Language

R is a scripting language for statistical data manipulation and analysis. A lot of new functions are contributed by users, many of whom are prominent statisticians (Matloff, 2011). R is a very suitable platform for writing a simulation algorithm. One can generate random samples from a wide variety of probability distributions, and R has an extensive set of data analysis capabilities for summarizing and graphing the simulation output (Albert, 2009). In this paper, simple R function has been designed to investigate the robustness or sensitivity of two independent samples t test (t statistic) using Monte Carlo simulation with R. with respect to changes in the assumptions.

5. Programming a Monte Carlo Simulation

One way to investigate the robustness or sensitivity of two independent samples t test (t statistic) using Monte Carlo simulation (Albert, 2009). In this paper use R for writing a simulation algorithm. The following is algorithm to do the Monte Carlo simulation to investigate the robustness or sensitivity of two independent samples t test (t statistic).

- First of all is generated random sample, namely $x_1, x_2, x_3, \dots, x_m$ for x sample and $y_1, y_2, y_3, \dots, y_n$ for y sample. The random sample is generated from various of probability distribution. In this paper, the size for each sample is 10 ($m = n = 10$).
- Next, compute t statistic from x and y sample. The formula to compute t statistic can be seen in (2).
- Determine t critical value. The t critical value is depended on degrees of freedom and stated significance level α.
- The t statistic is compared with t critical value to determine whether H_0 is rejected or no.

In this paper, step 1–4 is repeated 10000 times. Albert (2009) the estimates the true significance level can be computed

$$\hat{\alpha}^T = \frac{Number\ of\ rejections\ of\ H_0}{N} \qquad (10)$$

6. Sensitivity of t Statistic of Two Independent Samples t Test Respect to The Change of Assumption Using Monte Carlo Simulation

Table 1 is presented true significance level of the t-test computed by Monte Carlo experiment. Based on Table 1, the random sample is generated from various kind of probability distribution, namely normal, t, and exponential distribution. Based on the result in Table 1, when first population and second population are normal distribution with the equal mean and spread (N(0,1) and N(0,1)), the inaccuracy of the estimated true significance level respect to the stated significance level is 0,51%. But there is an increase the innacuracy, namely 17,09%, when the spread is significantly different (N(0,1) and N(0,10)). When first population and second population are t distribution with the equal spread ($t_{df=3}$ and $t_{df=3}$), the inaccuracy of the estimated true significance level respect to the stated significance level is 5,12%. But there is an decrease the innacuracy, namely 3,34%, when the spread is different ($t_{df=3}$ and $t_{df=3}$). The inaccuracy of the estimated true significance level respect to the stated significance level is 4,68%, when first population and second population are exponential distribution with the same rate (exp(5) and exp(5)).

Table 1. True Significance Levels of the T-Test Computed by Monte Carlo Experiments

Simulation	N(0,1) & N(0,1)	N(0,1) & N(0,10)	$t_{df=3}$ & $t_{df=3}$	$t_{df=3}$ & $t_{df=25}$	Exp(5) and Exp(5)	N(10,2) and Exp(1/10)	N(1,1) and Exp(1)
1	0.0986	0.114	0.0952	0.0939	0.0948	0.1568	0.0998
2	0.1001	0.1194	0.0896	0.0952	0.0923	0.1473	0.1096
3	0.1021	0.1182	0.093	0.0956	0.0938	0.1514	0.1061
4	0.1044	0.1174	0.0974	0.0995	0.0957	0.151	0.1088
5	0.0986	0.1178	0.0929	0.0946	0.0995	0.1506	0.1043
6	0.1015	0.1131	0.0972	0.093	0.0929	0.1577	0.1047
7	0.1014	0.1178	0.0963	0.1014	0.0983	0.1489	0.1096
8	0.098	0.112	0.0954	0.1014	0.0948	0.1536	0.0997
9	0.099	0.1254	0.0971	0.0977	0.0949	0.1546	0.1083
10	0.1	0.1109	0.0945	0.0975	0.095	0.1518	0.1049
11	0.1039	0.1137	0.1023	0.0926	0.0957	0.1499	0.1013
12	0.103	0.1188	0.092	0.0953	0.0912	0.1585	0.1070
13	0.0981	0.1205	0.0994	0.1023	0.0936	0.1489	0.1083
14	0.093	0.1162	0.0858	0.0948	0.0967	0.1514	0.1066
15	0.1012	0.1148	0.0945	0.095	0.098	0.1532	0.1083
16	0.0998	0.1178	0.0937	0.0985	0.0949	0.1501	0.0997
17	0.1036	0.1109	0.096	0.0945	0.1041	0.1579	0.1087
18	0.0955	0.113	0.0953	0.0978	0.0995	0.1543	0.1094
19	0.0995	0.1145	0.1	0.0943	0.0966	0.1591	0.1069
20	0.1022	0.1207	0.0957	0.0962	0.0942	0.1507	0.1076
21	0.1001	0.1135	0.0952	0.1018	0.0955	0.144	0.1071
22	0.1036	0.1166	0.0966	0.0954	0.0949	0.1605	0.1094
23	0.1021	0.1214	0.0917	0.0966	0.0931	0.1522	0.1032
24	0.1055	0.1189	0.0943	0.1006	0.0937	0.1542	0.1075

25	0.0976	0.1225	0.094	0.0954	0.0954	0.1536	0.1053
26	0.1	0.1172	0.0945	0.0953	0.093	0.1587	0.1068
27	0.0997	0.1214	0.0923	0.092	0.0936	0.1588	0.1102
28	0.0975	0.122	0.0963	0.0952	0.0933	0.1527	0.1055
29	0.1029	0.1167	0.094	0.0975	0.0944	0.1541	0.1037
30	0.1028	0.1156	0.0942	0.0988	0.0961	0.1576	0.1037
Average	0.10051	0.11709	0.09488	0.09665667	0.0953166	0.15347	0.1060
Inaccuracy (%)	0.51	17.09	5.12	3.34333333	4.6833333	53.47	6.066666

The last, first population is normal distribution and second population is exponential population, with equal mean but different spread, the inaccuracy of the estimated true significance level respect to the stated significance level is 53,47%. Note that, normal and exponential distribution substantially have different shape. When the first population is normal distribution and second population is exponential population, with equal mean and spread, the inaccuracy of the estimated true significance level respect to the stated significance level is 6,06%.

7. Conclusion

Based on the result of the investigation, the inaccuracy of the estimated true significance level respect to the stated significance level is lower than 10% when first population and second population have equal mean and spread, eventhough substantially have different shape of distribution. In this case, the stated significant level is 10%. The inaccuracy of the estimated true significance level respect to the stated significance level is higher than 10% when first population and second population have equal mean but significantly different spread. But it does not hold for t population. It means that, when first population and second population are t distribution, the inaccuracy of the estimated true significance level respect to the stated significance level is lower than 10%, eventhough have big difference of degrees of freedom.

References

[1] Albert, J. 2009. *Bayesian Computation with R, 2nd Edition*. New York: Springer.
[2] Mann, P. S. 2013. *Introductory Statistics, 8th Edition*. United States of America: John Wiley & Sons.
[3] Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. 2007. *Probabilty and Statistics for Engineers and Scientist, 8th Edition*. Prentice Hall.
[4] Montgomery, D. C. and Runger, G. C. 2014. *Applied Statistics and Probability for Engineers, 6th Edition.* John Wiley and Sons.
[5] Thomopoulos, N. T. 2013. *Essentials of Monte Carlo Simulation, Statistical Methods for Building Simulation Models*. New York: Springer.
[6] Matloff, N. 2011. *The Art of R Programming, A Tour of Statistical Software Design*. San Fransisco: No Starch Press.