J. Chosun Natural Sci.
Vol. 12, No. 3 (2019) pp. 91 – 99
https://doi.org/10.13160/ricns.2019.12.3.91

Effective Sample Sizes for the Test of Mean Differences Based on Homogeneity Test

Sunyeong Heo†

Abstract

Many researchers in various study fields use the two sample $t-$test to confirm their treatment effects. The two sample $t-$test is generally used for small samples, and assumes that two independent random samples are selected from normal populations, and the population variances are unknown. Researchers often conduct $F-$test, the test of equality of variances, before testing the treatment effects, and the test statistic or confidence interval for the two sample $t-$test has two formats according to whether the variances are equal or not. Researchers using the two sample $t-$test often want to know how large sample sizes they need to get reliable test results. This research gives some guidelines for sample sizes to them through simulation works. The simulation had run for normal populations with the different ratios of two variances for different sample sizes (≤ 30). The simulation results are as follows. First, if one has no idea equality of variances but he/she can assume the difference is moderate, it is safe to use sample size at least 20 in terms of the nominal level of significance. Second, the power of $F-$test for the equality of variances is very low when the sample sizes are small (<30) even though the ratio of two variances is equal to 2. Third, the sample sizes at least 10 for the two sample $t-$ test are recommendable in terms of the nominal level of significance and the error limit.

Keywords : Homogeneity, Power of Test, Sample Sizes, Test Reliability, Two Sample $t-$ test

1. Introduction

Many researchers in various study fields, such as medical sciences, food and nutrition sciences, educational sciences, and so on, use the two samples $t-$test to find out if there are any treatment effects in their experiments. The two sample $t-$test is generally used to test treatment effects for small samples.

In 1908, Gosset[1] published a paper with his nickname, Student (Student, 1906), In the paper, he proved that when a random sample is selected from normal population, the sample mean and sample variance are uncorrelated each other, and the distribution of difference between sample mean and population mean divided by sample standard deviation follows t-distribution. He is not the first person who proved the independence between sample mean and sample variance, and drove $t-$ distribution (Pfanzagl and Sheynin, 1996)[2]. However, his paper has been wildly known before other authors works had known to people.

The $t-$distribution is generally used for the inferences of population means under small sample sizes when population distributions are normal and population variances are unknown.

For the two samples $t-$test, the data structure is as follows

(1) $X_1, X_2, \cdots, X_m$ consists a random sample of size m from $N(\mu_1, \sigma_1^2)$, and (μ_1, σ_1^2) are unknown

(2) $Y_1, Y_2, \cdots, Y_n$ consists a random sample of a size n from $N(\mu_2, \sigma_2^2)$, and (μ_2, σ_2^2) are unknown,

(3) $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ are independent.

If sample sizes m and n are large (as a rule of thumb, greater than 30), by the central limit theorem the sample

Department of statistics, Changwon National University, Changwon

†Corresponding author : syheo@changwon.ac.kr
(Received : August 21, 2019, Revised : September 5, 2019,
Accepted : September 12, 2019)

distribution of the difference between two sample means, $\overline{X}-\overline{Y}$, follows approximately normal distribution, and the inferences about population mean differences $\mu_1-\mu_2$ are conducted using normal distribution.

However, when m and n are small (≤ 30), the inferences about $\mu_1-\mu_2$ are conducted based on the $t-$distribution, and the test using $t-$distribution is called the two sample $t-$test. There are two situations on the two sample $t-$test; one is the case that population variances can be assumed to be equal, and the other is the case that the population variances can be assumed to be unequal.

When σ_1^2, σ_2^2 are unknown but one can assume $\sigma_1^2=\sigma_2^2=\sigma^2$,

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$

where

$S_p^2=\left[(m-1)S_1^2+(n-1)S_2^2\right]/(m+n-2)$ and S_1^2, S_2^2 are the sample variances of $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$, respectively. Here, $t(m+n-2)$ denotes the $t-$distribution with $m+n-2$ degrees of freedom. And, a $(1-\alpha)100\%$ confidence interval about $\mu_1-\mu_2$ is given by

$$(\overline{X}-\overline{Y})\pm t_{m+n-2}(\alpha/2)\cdot S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \qquad (1)$$

where $t_{m+n-2}(\alpha/2)$ is the upper $(\alpha/2)100$ th percentile of $t(m+n-2)$.

On the other hand, when one can assume that the unknown variances are not equal, $\sigma_1^2\neq\sigma_2^2$,

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}}\sim t(\phi)$$

where ϕ is the degrees of freedom suggested by Satterthwaite (1946)[3] and defined by

$$\phi=\frac{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1}+\frac{(s_2^2/n)^2}{n-1}}$$

where s_1^2, s_2^2 are the values of the sample variances from random samples (Park, 2003)[4]. And a $(1-\alpha)100\%$ confidence interval for $\mu_1-\mu_2$ is given by

$$(\overline{X}-\overline{Y})\pm t_{\phi}(\alpha/2)\cdot\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}} \qquad (2)$$

(Casella and Berger,1990; freund, 1992)[5,6].

Almost every statistical method is established under some underlying assumptions like the two samples $t-$test. If a method is not much affected from the violation of underlying assumptions, and then the method is said as being robust (The Korean Statistical Association, 1987)[7].

Many researchers using the two sample $t-$test are interested a lot how large sample sizes they need to obtain reliable test results. In theory, they can conduct the two sample $t-$test for any sample sizes generally larger than 2 if the populations sampled have normal distributions. However, when the sample sizes are too small, the error limits of the test are too large, and so the values of the test results as information become too low.

The purpose of this research is to give some guidelines about sample sizes for researchers to conduct the two samples $t-$test. To achieve this purpose, I ran simulations using R version 3.5.2. For each selected parameter, I repeated the same calculation $100,000$ times. This research will be presented in the following order.

First, the proportions of the 95% confidence intervals including the true $\mu_1-\mu_2$ are calculated for selected sample sizes (≤ 30) when the population variances are equal or unequal. Here, the confidence intervals are obtained using the equation (1) regardless of whether the variances are equal or not.

Second, the proportions of rejecting $H_0: \sigma_1^2=\sigma_2^2$ under the 95% confidence level are calculated for the same sample sizes and $r=\sigma_1/\sigma_2$ as the above first step.

Third, the proportions of the 95% confidence intervals including the true $\mu_1-\mu_2$ are calculated according to the test results of the equality of variances, for the same sample sizes and $r=\sigma_1/\sigma_2$ as the above two steps.

2. Effective Sample Sizes for Two Sample *t*-test

For simulation work, I had chosen two normal

populations, $N(0,1)$ and $N(0,1/r^2)$. The reason I chose these populations will be given at the end of this section. And I selected two independent random samples of size n each from these populations. and calculated the 95% confidence interval about $\mu_1-\mu_2$. I repeated this calculation 100,000 times for each selected (n,r), and calculated the proportions of the confidence intervals including $\mu_1-\mu_2=0$.

In this simulation setting, $\mu_1-\mu_2=0$ is true, and so it is the most desirable that the proportions of the 95% confidence intervals including $\mu_1-\mu_2=0$ are at least 0.95 for all selected (n,r). As for variances, $r=1$ means that two population variances are equal, and $r\neq 1$ means that the variances are unequal. For the simulation, I have selected $r=0.5,0.6,0.7\,0.8,0.9,1.0$, For example, $r=0.5$ means that one variance is 2 times larger then the other. Since two sample $t-$test is generally used for small samples, I have chosen sample sizes $n=3,5,10,15,20,30$.

2.1 Effective Sample Sizes when the Inequality of Population Variances Ignored

<Table 1> shows the proportions of the 95% confidence intervals including the true $\mu_1-\mu_2=0$ for all different (n,r). Here, the confidence intervals are obtained using the equation (1) which assumes $\sigma_1^2=\sigma_2^2$. Note that $r<1$ means that the population variances are not equal.

<Table 1> shows that when $r=1$, for all selected n the proportions are very close to 0.95. But, as r decreases, for small n the proportions are getting smaller from 0.95 more and more. However, when $n=30$, the proportions are almost 0.95 for all r.

So, if there is no reliable information about variances and one wants to use the equation (1), it is safe to choose large sample sizes (≥ 30)

2.2 Effective Sample Size for the test of Homogeneity Between Two Population Variances

Before conducting the two sample $t-$test, it is common to test the equality of population variances. <Table 2> shows proportions of rejecting H_0 out of 100,000 replicated test about $H_0:\sigma_1^2=\sigma_2^2$ against $H_1:\sigma_1^2\neq\sigma_2^2$ for selected (n,r). Here, I tested theses hypotheses using $F-$ test about the ratio of two variances. I do not mention details about $F-$test here. You can refer to statistical textbooks for the $F-$test, e.g., Casella and Berger (1990) and Freund (1992).

When $r=1$, the proportions are close to 0.05 for all selected n, and it is what we expect. However, when $r\neq 1$, we expect the proportion close to 0.95 for all (n,r) combinations, but it is much smaller than we expect for almost every (n,r). But, when n is getting larger and r is getting smaller, the proportions are getting closer to 0.95. Especially, when $r=0.5$ and $n=30$, the proportion is a little larger then 0.95.

When $n=3$, even though $r=0.5$ the proportion is only 0.101, which means that only 10.100 times out of 100,000 tests rejecting $H_0:\sigma_1^2=\sigma_2^2$ when $H_1:\sigma_1^2\neq\sigma_2^2$ is true.

From <Table 2>, we can know that the power of $F-$ test is very low when the sample sizes are small, even though the ratios of two variances are small like $r=0.5$. So, when one tests the equality of two variances with small samples less than 30, he/she needs to be careful to accept the test results.

2.3 Effect of Sample Size on the Reliability of Two Sample *t*-test

<Table 3> shows the proportions of the 95% confidence intervals including the true $\mu_1-\mu_2=0$. For each pair of selected random samples, the equality of variances was tested. And, as the result of the test, when $H_0:\sigma_1^2=\sigma_2^2$ was accepted, the 95% confidence interval was calculated using the equation (1), and when $H_0:\sigma_1^2=\sigma_2^2$ was rejected, the 95% confidence interval was calculated using the equation (2). I repeated these calculations 100,000 times for each selected (n,r) combination, and the results are given in <Table 3>

For example, when $r=0.8$, $n=3$, $H_0:\sigma_1^2=\sigma_2^2$ was rejected only 5,593 times out of 100,000 replicated tests. For the samples rejected the equality of variances, the 95% confidence intervals about $\mu_1-\mu_2$ were calculated using the equation (2). And the 98.5% of 5,593 confidence intervals included the true $\mu_1-\mu_2=0$. On the other hand, for 94,407 pairs of samples which accepted $H_0:\sigma_1^2=\sigma_2^2$, the 95% confidence intervals were calculated using the equation (1), and the 94.9% of the 94,407 confidence intervals included the true $\mu_1-\mu_2=0$.

Note that, if $r=1$, then accepting $H_0:\sigma_1^2=\sigma_2^2$ is right decision, and if $r<1$, then rejecting $H_0:\sigma_1^2=\sigma_2^2$ is right decision.

In <Table 3>, when $r=1$ but the true $H_0:\sigma_1^2=\sigma_2^2$ was rejected, the proportions of the confidence intervals including the true $\mu_1-\mu_2=0$ are larger than 0.95 for all selected n, but the proportion decreases as the sample size increases, On the other hand, when $r=1$ and the true $H_0:\sigma_1^2=\sigma_2^2$ accepted, the proportions of confidence intervals including the true $\mu_1-\mu_2=0$ are almost equal to 0.95, the nominal level of significance, for all selected sample sizes.

When $r<1$, for all selected r the proportions of confidence intervals including the true $\mu_1-\mu_2=0$ decreases as the sample size increases regardless of whether $H_0:\sigma_1^2=\sigma_2^2$ is rejected or not.

When $r<1$ and $H_0:\sigma_1^2=\sigma_2^2$ is rejected, the proportions of confidence intervals including the true $\mu_1-\mu_2=0$ are larger then 0.95 for all selected sample sizes. When $r<1$ but $H_0:\sigma_1^2=\sigma_2^2$ is accepted, the proportion of confidence intervals including the true $\mu_1-\mu_2=0$ decreases as the sample size increases and r decreases, and the proportions are getting much smaller than the nominal level of significance 0.95 for large n and small r, for example $n=30$ and $r=0.5$. However, we should note that when $r<1$ the proportions of accepting the true $H_0:\sigma_1^2\neq\sigma_2^2$ were very small when both r and n are small. We can refer to <Table 2>.

In <Table 3>, the last column "total" gives the proportion of the total confidence intervals including the true $\mu_1-\mu_2=0$, from both rejecting and accepting $H_0:\sigma_1^2=\sigma_2^2$, divided by $100{,}000$. For example, when $r=0.8$ and $n=3$, out of $5{,}593$ pairs of samples which rejected $H_0:\sigma_1^2=\sigma_2^2$. the $5{,}510$ pairs of samples gave the confidence intervals including the true $\mu_1-\mu_2=0$, and out of $94{,}407$ pairs of samples which accepted $H_0:\sigma_1^2=\sigma_2^2$, the $89{,}592$ pairs of samples gave the confidence intervals including the true $\mu_1-\mu_2=0$. So, the total confidence interval including the true $\mu_1-\mu_2=0$ are $95{,}102$ pairs of samples, and the proportion of confidence intervals including the true $\mu_1-\mu_2=0$ is about 95.1%.

From the last column of <Table 3>, we can know that on average the 95% confidence intervals about $\mu_1-\mu_2$ satisfy the nominal level of significance for all sample sizes and r. However, we need to pay attention to which in <Table 2> when sample sizes are small, the $F-$test results are very poor.

In <Table 4>, I added the error limits from the 95% confidence interval for $\mu_1-\mu_2$ and the relative sizes of error limits to <Table 3>. The error limit is defined as a half length of confidence interval. For the relative size of error limit, I considered the error limits when $n=30$ as 1.

For example, when $r=0.8$ and $n=3$, out of $100{,}000$ replicated tests only $5{,}593$ times $H_0:\sigma_1^2=\sigma_2^2$ were rejected, and the error limits are calculated for the $5{,}593$ pairs of samples using the equation (2) and the average of $5{,}593$ error limits is 3.813. Since I considered the average error limit of $n=30$ as 1, when $r=0.8$, the relative size of error limit of $n=3$ is $3.813/0.601\approx 6.3$. That is, the length of confidence interval when $n=3$ is about 12.6 times larger than $n=30$, on average.

<Fig. 1 > shows the relative sizes of error limits for $n=3,5,10,30$ by sizes of r.

From <Table 3> and <Table 4>, we can see that when $H_0:\sigma_1^2=\sigma_2^2$ rejected, for all $r(\leq 1)$ the proportions of including the true $\mu_1-\mu_2=0$ are larger than 0.95, the nominal level of significance, and the proportion decreases as the sample size increases. But the error limits increases a lot as n decreases. Especially when $n=3$ or 5, the relative sizes of error limit are more than 6 times or 3 times larger than $n=30$.

When $r<1$ but $H_0:\sigma_1^2=\sigma_2^2$ accepted, the proportions of including the true $\mu_1-\mu_2=0$ are decreasing as the sample sizes are increasing. If r is small, the proportions are getting much smaller than 0.95 for all n, and the proportions are the least when $n=30$. However, when $r=0.9$ or 0.8, the proportions are close to 0.95 for all n, but the relative sizes of error limits are about 4 or 3 times larger when $n=3$ or 5 than when $n=30$.

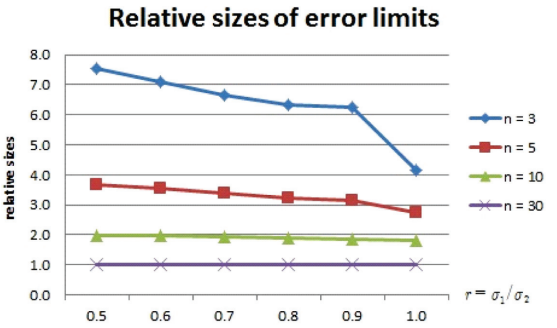

Fig. 1. The relative sizes of error limits for different sizes of n and r.

In conclusions, on average the two sample $t-$test satisfies the nominal level of significance rate for all selected (n, r). However, if one conducts $F-$test for equality of variances (homogeneity test), and according to the test results he/she chooses a test between (1) and (2). the chance he/she obtains the test results satisfying the nominal level of significance is high when sample sizes are small and the differences between two variances are small. But when sample sizes are small like 3 or 5, the error limit are large, and so the reliability of the test results is reduced. It is recommendable to choose the sample sizes at least 10 in terms of the nominal level of significance and the error limit.

2.4 Generalization

This simulation results can be generalized for any samples from normal populations with $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Let the ratio of two population standard deviation be $r = \sigma_1/\sigma_2$, then the populations can be rewritten as $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2/r^2)$.

When $X \sim N(\mu_1, \sigma^2)$, the normalized variable has standard normal distribution, that is,

$$Z_x = \frac{X-\mu_1}{\sigma} \sim N(0, 1).$$

Now, for $Y \sim N(\mu_2, \sigma^2/r^2)$ let us transform Y with the same μ_1 and σ as X, such as

$$Z_y = \frac{Y-\mu_1}{\sigma} = \frac{Y-\mu_2}{\sigma} + \frac{\mu_2-\mu_1}{\sigma} = \frac{1}{r}\frac{Y-\mu_2}{\sigma/r} + \frac{\mu_2-\mu_1}{\sigma},$$

and then since $Y \sim N(\mu_2, \sigma^2/r^2)$, in the first term of the right-hand side

$$Z = \frac{Y-\mu_2}{\sigma/r} \sim N(0, 1).$$

Therefore,

$$Z_y = \frac{1}{r} Z + \frac{\mu_2-\mu_1}{\sigma}$$

and r, μ_1, μ_2, σ are constants, and so

Table 1. Proportions of the confidence intervals including the true $\mu_1 - \mu_2 = 0$ out of $100{,}000$ replication when the equation (1) was used for different sizes of $r(=\sigma_1/\sigma_2)$ and sample sizes (Normal population and 95% confidence interval).

r	σ_1^2/σ_2^2	Sample sizes					
		3	5	10	15	20	30
1	1	0.951	0.951	0.950	0.949	0.950	0.951
0.9	0.81	0.950	0.950	0.951	0.949	0.950	0.951
0.8	0.64	0.949	0.949	0.950	0.948	0.949	0.951
0.7	0.49	0.948	0.947	0.949	0.947	0.949	0.950
0.6	0.36	0.945	0.945	0.948	0.947	0.948	0.950
0.5	0.25	0.939	0.942	0.946	0.945	0.948	0.949

Table 2. Proportions of rejecting $H_0 : \sigma_1^2 = \sigma_2^2$ out of $100{,}000$ replicated tests for different sizes of r and sample sizes (Normal population, 95% level of significance).

r	σ_1^2/σ_2^2	Sample sizes					
		3	5	10	15	20	30
1	1	0.050	0.049	0.049	0.050	0.051	0.050
0.9	0.81	0.052	0.053	0.059	0.066	0.072	0.084
0.8	0.64	0.056	0.065	0.094	0.124	0.154	0.214
0.7	0.49	0.064	0.087	0.166	0.245	0.323	0.468
0.6	0.36	0.077	0.130	0.294	0.450	0.579	0.771
0.5	0.25	0.101	0.208	0.496	0.704	0.835	0.957

Table 3. Proportions of rejecting $H_0 : \sigma_1^2 = \sigma_2^2$, and confidence intervals including $\mu_1 - \mu_2 = 0$ (true) out of 100,000 replication by rejecting or accepting $H_0 : \sigma_1^2 = \sigma_2^2$, for different sizes of r and sample sizes (Normal population, 95% level of significance)

r	Sample size	Proportion of rejecting $H_0 : \sigma_1^2 = \sigma_2^2$	Proportion of including $\mu_1 - \mu_2 = 0$		
			when $H_0 : \sigma_1^2 = \sigma_2^2$ rejected	when $H_0 : \sigma_1^2 = \sigma_2^2$ accepted	Total
1	3	0.050	0.985	0.951	0.953
	5	0.049	0.973	0.950	0.952
	10	0.049	0.957	0.950	0.951
	15	0.050	0.954	0.949	0.949
	20	0.051	0.953	0.950	0.950
	30	0.050	0.954	0.951	0.951
0.9	3	0.052	0.986	0.951	0.952
	5	0.053	0.974	0.950	0.951
	10	0.059	0.958	0.950	0.951
	15	0.066	0.959	0.948	0.949
	20	0.072	0.955	0.950	0.950
	30	0.084	0.958	0.950	0.951
0.8	3	0.056	0.985	0.949	0.951
	5	0.065	0.977	0.948	0.950
	10	0.094	0.965	0.949	0.950
	15	0.124	0.960	0.947	0.949
	20	0.154	0.960	0.948	0.950
	30	0.214	0.958	0.949	0.951
0.7	3	0.064	0.986	0.947	0.949
	5	0.087	0.982	0.946	0.949
	10	0.166	0.969	0.946	0.949
	15	0.245	0.963	0.944	0.949
	20	0.323	0.960	0.945	0.950
	30	0.468	0.957	0.945	0.951
0.6	3	0.077	0.990	0.943	0.947
	5	0.130	0.983	0.942	0.947
	10	0.294	0.970	0.940	0.949
	15	0.450	0.962	0.938	0.949
	20	0.579	0.958	0.939	0.950
	30	0.771	0.955	0.936	0.951
0.5	3	0.101	0.990	0.936	0.942
	5	0.208	0.985	0.934	0.944
	10	0.496	0.969	0.928	0.949
	15	0.704	0.960	0.924	0.949
	20	0.835	0.955	0.922	0.950
	30	0.957	0.952	0.914	0.950

Table 4. Proportions of confidence intervals including $\mu_1 - \mu_2 = 0$ (true), error limits, relative sizes of error limit comparing to when $n = 30$, by rejecting or accepting $H_0 : \sigma_1^2 = \sigma_2^2$, for different sizes of r and sample sizes (Normal population, 95% level of significance)

r	Sample size	when $H_0 : \sigma_1^2 = \sigma_2^2$ rejected			when $H_0 : \sigma_1^2 = \sigma_2^2$ accepted		
		Proportion of including $\mu_1 - \mu_2 = 0$	Error limit	Relative size of error limit	Proportion of including $\mu_1 - \mu_2 = 0$	Error limit	Relative size of error limit
1	3	0.985	3.237	6.3	0.951	2.134	4.1
	5	0.973	1.626	3.2	0.950	1.414	2.7
	10	0.957	0.957	1.9	0.950	0.927	1.8
	15	0.954	0.754	1.5	0.949	0.741	1.4
	20	0.953	0.642	1.2	0.950	0.636	1.2
	30	0.954	0.516	1.0	0.951	0.515	1.0
0.9	3	0.986	3.459	6.2	0.951	2.254	4.1
	5	0.974	1.744	3.1	0.950	1.492	2.7
	10	0.958	1.029	1.9	0.950	0.979	1.8
	15	0.959	0.811	1.5	0.948	0.783	1.4
	20	0.955	0.691	1.2	0.950	0.671	1.2
	30	0.958	0.554	1.0	0.950	0.543	1.0
0.8	3	0.985	3.813	6.3	0.949	2.405	4.2
	5	0.977	1.942	3.2	0.948	1.592	2.8
	10	0.965	1.140	1.9	0.949	1.043	1.8
	15	0.960	0.890	1.5	0.947	0.834	1.4
	20	0.960	0.752	1.3	0.948	0.715	1.2
	30	0.958	0.601	1.0	0.949	0.578	1.0
0.7	3	0.986	4.349	6.6	0.947	2.603	4.2
	5	0.982	2.219	3.4	0.946	1.719	2.8
	10	0.969	1.273	1.9	0.946	1.124	1.8
	15	0.963	0.981	1.5	0.944	0.897	1.4
	20	0.960	0.825	1.3	0.945	0.767	1.2
	30	0.957	0.654	1.0	0.945	0.619	1.0
0.6	3	0.990	5.129	7.1	0.943	2.867	4.3
	5	0.983	2.567	3.5	0.942	1.887	2.8
	10	0.970	1.434	2.0	0.940	1.225	1.8
	15	0.962	1.092	1.5	0.938	0.972	1.5
	20	0.958	0.915	1.3	0.939	0.829	1.2
	30	0.955	0.723	1.0	0.936	0.665	1.0
0.5	3	0.990	6.210	7.5	0.936	3.237	4.5
	5	0.985	3.021	3.7	0.934	2.108	3.0
	10	0.969	1.639	2.0	0.928	1.346	1.9
	15	0.960	1.241	1.5	0.924	1.059	1.5
	20	0.955	1.038	1.3	0.922	0.897	1.3
	30	0.952	0.824	1.0	0.914	0.714	1.0

$$Z_y = \frac{Y-\mu_1}{\sigma} \sim N\left(\frac{\mu_2-\mu_1}{\sigma}, \frac{1}{r^2}\right).$$

When $\mu_1 = \mu_2 = \mu$,

$$Z_x = \frac{X-\mu}{\sigma} \sim N(0,1)$$
$$Z_y = \frac{Y-\mu}{\sigma} \sim N(0, 1/r^2).$$

The distributions of Z_x, Z_y are the same as the distributions of the simulation setting in the beginning of Section 2.

The r can be estimated by the ratio of sample standard deviations from two independent samples, that is, $\hat{r} = s_x/s_y$ where s_x, s_y are the values of sample standard deviations of two independent random samples, and based on this r and <Table 3> and <Table 4> researchers can select their sample sizes.

3. Conclusion

Many researcher in various study fields need to proof their treatment effects, for example, the benefits of newly developed medicine, the effectiveness of new training method, the comparison of two kinds of breads baked with different ingredients's component ratios, and so on. In these kinds of research, researchers often have to make decision based on small sample sizes. In the situations, they generally use the two sample $t-$test. And they need to know how large sample sizes they need to get reliable test results.

This research gives guidelines for sample sizes to them through simulation results. I ran simulation $100{,}000$ times for each combination of selected ratios of two population standard deviation, denoted r, and sample sizes n. I have chosen $r = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$ and $n = 3, 5, 10, 15, 20, 30$. The $r = 1$ means that two population variances are equal. Through the simulation results, we find that

first, if there is no reliable information about variances and one wants to use the equation (1) which assumes equal variances, it is safe to choose large sample sizes (≥ 30) in terms of achieving the nominal level of significance. One can refer to <Table 1> for details.

Second, researchers are generally conducting $F-$test, the test of the equality of variances, before testing the differences between means. The power of $F-$ test is very low when the sample sizes are small, even though the differences between two variances are large. So, when one tests the equality of two variances with small samples less than 30, he/she needs to be careful to accept the test results. One can refer to <Table 2> for details.

Third, the two sample $t-$test satisfies the nominal level of significance for all selected (n, r). However, if one conducts $F-$test for the equality of variances, and according to the test results he/she chooses a test between the equations (1) and (2), the chance he/she achieves the test results satisfying the nominal level of significance is high when sample sizes small and the difference between variances are small. But when sample sizes are small like 3 or 5, the error limit are very large, and so the reliability of the test results is reduced. Therefore, I recommend to choose the sample sizes at least 10 to get the reliable test results in terms of the nominal level of significance and the error limit. One can refer to <Table 3> and <Table 4> for details.

When a researcher needs to determine sample size for the test of their treatment effect, this simulation results can be good references for deciding sample sizes.

Acknowledgements

This research was supported by Changwon National University in 2019-2020.

References

[1] Student, "The probable error of mean", Biometrika, Vol. 6, pp. 1-25, 1908.

[2] J. Pfanzagl and O. Sheynin, "Studies in the history of probability and statistics XLIV, A forerunner of the $t-$distribution", Biometrika, Vol. 83, pp. 891-898, 1996.

[3] F. E. Satterthwaite, "An approximate distribution of estimates of variance components", Biometrics Bullein, Vol. 2, pp. 110-114, 1946.

[4] S. H. Park, Design of Experiments (2nd), Minyeongsa, Seoul, 2003.

[5] G. Casella, and R. L., Berger, Statistical Inference, Brooks/Cole Publishing Company, 1990.
[6] J. E. Freund, Mathematical Statistics (5th ed), Prentice-Hall International, Inc., New Jersey, California, 1992.
[7] The Korean Statistical Society, Statistical Glossary, Freeacademy Inc, Seoul, 1987.