KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 • eISSN 2005-7563

More about the basic assumptions of t-test: normality and sample size

Tae Kyun Kim[1] and Jae Hong Park[2]

[1]*Department of Anesthesia and Pain Medicine, Pusan National University School of Medicine,* [2]*Department of Anesthesiology and Pain Medicine, Haeundae Paik Hospital, Inje University College of Medicine, Busan, Korea*

Most parametric tests start with the basic assumption on the distribution of populations. The conditions required to conduct the t-test include the measured values in ratio scale or interval scale, simple random extraction, normal distribution of data, appropriate sample size, and homogeneity of variance. The normality test is a kind of hypothesis test which has Type I and II errors, similar to the other hypothesis tests. It means that the sample size must influence the power of the normality test and its reliability. It is hard to find an established sample size for satisfying the power of the normality test. In the current article, the relationships between normality, power, and sample size were discussed. As the sample size decreased in the normality test, sufficient power was not guaranteed even with the same significance level. In the independent t-test, the change in power according to sample size and sample size ratio between groups was observed. When the sample size of one group was fixed and that of another group increased, power increased to some extent. However, it was not more efficient than increasing the sample sizes of both groups equally. To ensure the power in the normality test, sufficient sample size is required. The power is maximized when the sample size ratio between two groups is 1 : 1.

Keywords: Biostatistics; Normal distribution; Power; Probability; P value; Sample size; T-test.

Introduction

Science is based on probability. It is impossible to say with certainty whether the events observed in nature will occur or not. The same rule applies when testing hypotheses through research. Null hypotheses are basically assumed to be true. Fig. 1 presents two normal distribution probability density curves under the respective assumptions that the null hypothesis is true (left) or false (alternative hypothesis is true, right). They show the typical curves that approach but never reach the x-axis on both ends. Whether the assumption is true or not, the probability never becomes zero; this means that any result obtained from research always has a possibility of unreliability. It is always possible for conclusions from research to be wrong, and an appropriate hypothesis is required to conduct research as well as to reduce the risk of false conclusions.

If the null hypothesis is concluded to be true when the value is less than a specific point on the x-axis, and false when the value is greater, then the point is called the critical value. The probability curve of the null hypothesis partially exists on the right side of the critical value. This means that the null hypothesis is true, but as it exceeds the critical value, it is mistakenly thought to be false; this is called a Type I error. In contrast, even if the null hypothesis is false (orange probability distribution curve), because it exceeds the critical value, it is mistakenly accepted to be true; this is called a Type II error. The size of the error can be represented with a probability, which is the area under the curve lying outside the critical value. The probability of a Type I error is called the α or level of significance. The probability of a Type

Corresponding author: Jae Hong Park, M.D., Ph.D.
Department of Anesthesiology and Pain Medicine, Haeundae Paik Hospital, Inje University College of Medicine, 875 Haeun-daero, Haeundae-gu, Busan 48108, Korea
Tel: +82-51-797-0427, Fax: +82-51-797-0422
Email: H00150@paik.ac.kr
ORCID: https://orcid.org/0000-0003-0779-4483

Received: October 11, 2018.
Revised: March 6, 2019.
Accepted: March 25, 2019.

Korean J Anesthesiol 2019 August 72(4): 331-335
https://doi.org/10.4097/kja.d.18.00292

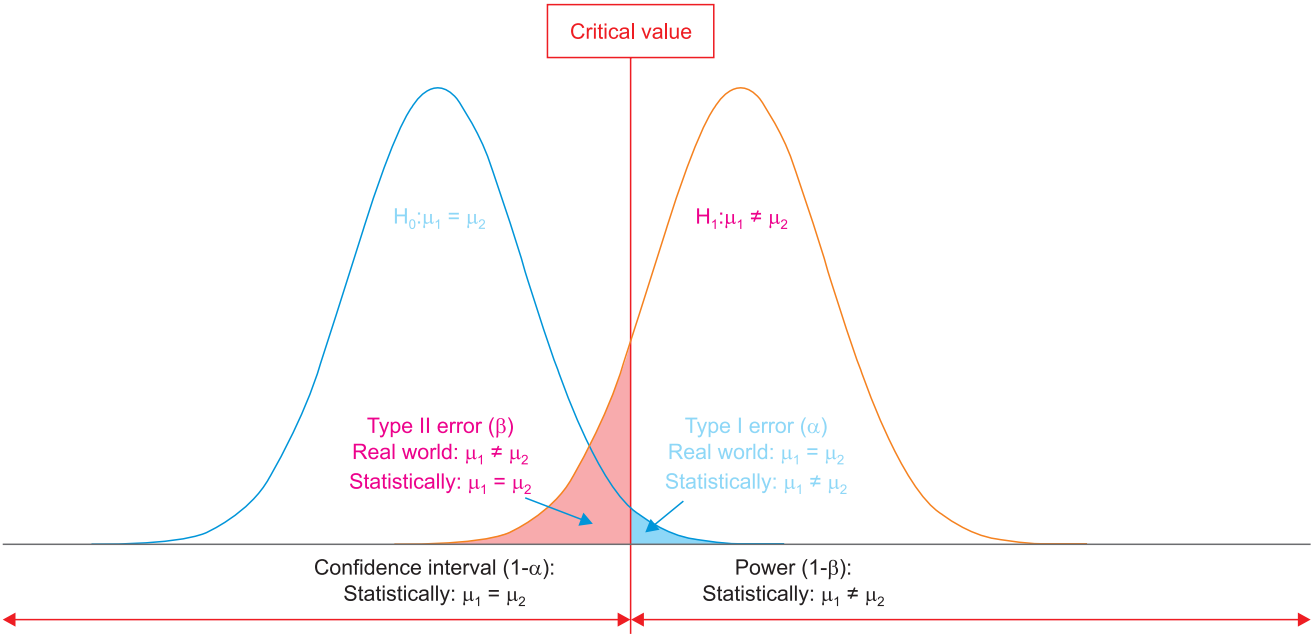

Fig. 1. Concept of hypothesis testing in independent t-test. H_0: null hypothesis, H_1: alternative hypothesis, μ_1 and μ_2: mean values of two groups.

II error is called β. Subtracting α from the probability of the null hypothesis becomes the confidence interval, and subtracting β from the probability of the alternative hypothesis becomes the power. Under the assumption that the null hypothesis is true, the P value—the probability of observing the test statistic of the data—can be obtained. In order to determine whether the hypothesis is true or false, it is necessary to confirm whether the observed event is statistically likely to occur under the assumption that the hypothesis is true. The P value is compared to a preset standard that determines the null hypothesis as false, which is the α. The α is the standard which has been agreed upon by researchers seeking the unknown truth, but the truth cannot be certain even if the P value is less than the α. The conclusion drawn from data analysis may not be the truth, and this is the error previously mentioned. Type I and Type II errors occur once we set a critical value; they show a pattern of trade-off between α and β, the probabilities of error. As can be seen in Fig. 1, the power means the probability of rejecting the null hypothesis when the alternative hypothesis is true; the power increases as the sample size increases.

The same rule applies to the normality test. The conditions required to conduct a t-test include the measured values in ratio scale or interval scale, simple random extraction, homogeneity of variance, appropriate sample size, and normal distribution of data. The normality assumption means that the collected data follows a normal distribution, which is essential for parametric assumption. Most statistical programs basically support the normality test, but the results only include P values and not the power of the normality test. Is it possible to conclude that the data follows a normal distribution if the P value is greater than or equal to α in the normality test? This article starts from this question and discusses the relationships between the sample size, normality, and power.

Normality

Statistical analysis methods based on acquired data are divided into parametric methods and nonparametric methods, according to the normality of the data. When the data satisfies the normality, it shows a probability distribution curve with the highest frequency of occurrence at the center, and the frequency decreases with distance from the center. The distance from the center of the curve makes it easier to statistically determine whether the data obtained is frequently observed. Since most of the data are gathered around the mean value, it reflects the nature of the group and gives information on whether there is a difference between groups and the magnitude of the difference. On the other hand, if the data does not follow the normal distribution, there is no guarantee that it is centered on the mean. Therefore, the comparison of characteristics between groups using the mean value is not possible. In this case, the nonparametric test is used, in which the observations are ranked or signed (e.g., + or −), and the sums are compared. However, the nonparametric test is somewhat less powerful than the parametric test [1]. Moreover, it is only possible to detect the difference between the values of groups but not to compare the magnitude

of these differences. Therefore, it is recommended that statistical analysis be performed using the parametric test if possible [1], and that the normality of the data be the first thing confirmed by the parametric test. The hypothesis in normality testing is as follows:

H_0: The data follows a normal distribution.
H_1: The data does not follow a normal distribution.

Thus, how many samples would be appropriate to assume normal distribution and to perform parametric tests?

According to the central limit theorem, the distribution of sample mean values tends to follow the normal distribution regardless of the population distribution if the sample size is large enough [2]. For this reason, there are some books which suggest that if the sample size per group is large enough, the t-test can be applied without the normality test. Strictly speaking, this is not true. Although the central limit theorem guarantees the normal distribution of the sample mean values, it does not guarantee the normal distribution of samples in the population. The purpose of the t-test is to compare certain characteristics representing groups, and the mean values become representative when the population has a normal distribution. This is the reason why satisfaction of the normality assumption is essential in the t-test. Therefore, even if the sample size is sufficient, it is recommended that the results of the normality test be checked first. Well-known methods of normality testing include the Shapiro–Wilks test and the Kolmogorov–Smirnov test. Hence, can the t-test be conducted with a very small sample size (e.g., 3) if the normality test is satisfied?

In the Shapiro–Wilks test, which is known as one of the most powerful normality tests, it is theoretically possible to perform the normality test with three samples [3,4]. However, even if the P value is greater than the significance level of 0.05, this does not automatically mean that the data follows a normal distribution. Type I and Type II errors occur in all hypothesis tests, which are detected using the significance levels and power. In general, statistical programs provide only a P value for the Type I error as a result of normality testing, and do not provide power for the Type II error. The power of the normality test indicates the ability to discriminate non-normal distributions from normal distributions. Since there is no formula that can calculate the power of the normality test directly, it is estimated by computer simulation. In the simulation, the computer repeatedly extracts samples of a certain size from the distribution to be tested, and tests whether the extracted samples have a normal distribution at a determined significance level. The power is the rate at which the null hypothesis is rejected from the data obtained through simulations repeated over several hundred times. If there are only three samples, it may be difficult to ensure that these are not normally distributed. Khan and Ahmad [4] reported the change of power according to sample sizes under different alternate non-normal distributions (Fig. 2). In fact, the types of distributions mentioned in the figure are not commonly observed in clinical studies, and are not essential to understand this figure. We do not explained in detail about that becauseit goes beyond our scope. The x-axis represents the number of samples extracted from each distribution type, and the y-axis represents the power of the normality test corresponding to the number of extracted samples. Fig. 2 shows that, although there are some degree of difference depending on the patterns of distribution, the power tends to decrease when the sample size decreases even if the significance level is fixed at 0.05. Therefore, in typical circumstances where the distribution pattern of the population is unknown, the normality test should be conducted with a sufficient sample size.

Sample size: Should the sample size ratio of the control group and the experimental group be 1 : 1?

When designing a study using the t-test, it is ideal to have the same sample size for the experimental and control groups [5]. However, as in cases of studying the therapeutic effects of drugs on rare diseases, there are cases in which it is difficult to secure

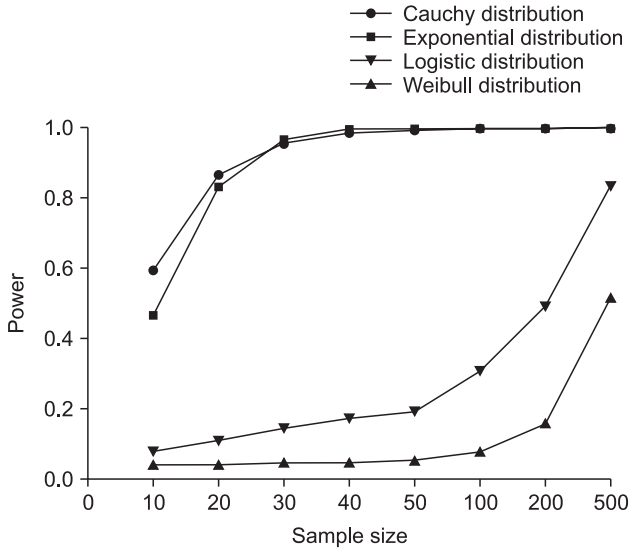

Fig. 2. Power results of Shapiro–Wilks test under different alternate non-normal distributions at α = 0.05. Power tends to decrease when the sample size decreases. Logistic distribution: alternate Logistic (Location = 0, Scale = 1) distribution, Weibull distribution: alternate Weibull (Scale = 2, Shape = 3) distribution (Modified from Khan RA, Ahmad F. Power Comparison of Various Normality Tests. Pakistan Journal of Statistics and Operation Research 2015; 11. Available from http://pjsor.com/index.php/pjsor/article/view/1082).

enough samples for the experimental group. It is well understood that increasing the sample size is a good way to improve both the significance level and the power. If so, would it be effective to increase only the sample size of control group without increasing that of the experimental group?

Assume that both sample groups follow the normal distribution and the variances are homogeneous. Assuming that sample group 1 is the experimental group and sample group 2 is the control group, the t- statistic can be expressed as [6]

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{(1+2)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

where $\overline{X}_1$ and $\overline{X}_2$ are the mean values of sample groups 1 and 2, n_1 and n_2 are the sample sizes of sample groups 1 and 2, and $S_{(1+2)}$ is the pooled standard deviation of the two sample groups. Assume that t_1 is a t-statistic when the sample sizes of the two groups are the same (n_1=n_2), and t_2 is a t-statistic when the sample size of the control group is two times larger than that of the experimental group ($2n_1$=n_2). In the equation, n_2 can be replaced by n_1 and summarized as follows:

$$t_1 = \sqrt{\frac{1}{2}} \cdot \frac{\overline{X}_1 - \overline{X}_2}{S_{(1+2)}\sqrt{\frac{1}{n_1}}},\ t_2 = \sqrt{\frac{2}{3}} \cdot \frac{\overline{X}_1 - \overline{X}_2}{S_{(1+2)}\sqrt{\frac{1}{n_1}}}$$

The above two equations can be concatenated so that t_2 can be expressed as a relation with t_1 as follows:

$$t_2 = \frac{2}{\sqrt{3}} t_1$$

When the sample size of the control group is doubled, the t-statistic increases by $\frac{2}{\sqrt{3}}$ times compared with the t-statistic when the sample size is not changed. This change results in the decrease of the P value. Therefore, if it is difficult to increase the sample size of the experimental group, it can be helpful to increase the sample size of the control group instead to improve the results. However, in such cases, relatively larger sample sizes are needed to obtain similar statistical results.

Table 1 shows the results of the minimum sample size calculation required to obtain a statistically significant result in the two-tailed independent t-test at a significance level of 0.05 and power of 0.8, when the ratios of sample sizes between groups are 1 : 1 and 1 : 2, respectively. Increasing the sample size of the control group can also reduce the size of the experimental group required to achieve the same statistical result. However, instead of reducing the sample size of the experimental group by 25%, the size of the control group should be increased by 50%, which involves additional effort, time, and cost. Power becomes maximum when the sample sizes of the two groups are the same. When the variances of the two groups to be compared are similar, the smallest sample size is acquired when those of the two groups are the same [5]. Table 2 shows the change of power in two-tailed independent t-tests under the same sample size but different sample size ratios. These results suggest that increasing the sample size of the control group cannot be used as a shortcut and should only be considered in unavoidable circumstances.

It should be noted that, when the sample size ratios between groups is different, more attention should be paid to the homogeneity of variance, one of the basic assumptions of the t-test. Clinical studies generally have small sample sizes. The smaller

Table 1. Minimum Sample Size Required to Obtain a Significant Result according to Different Sample Size Ratios in the Two-tailed Independent t-test

Tail(s)	Two	
Effect size d	0.5	
α err prob	0.05	
Power (1-β err prob)	0.8	
Sample size ratio between group (G1 : G2)	1 : 1	1 : 2
Noncentrality parameter δ	2.83	2.83
Critical t value	1.98	1.98
Degree of freedom	126	142
G1	64	48
G2	64	96
Total sample size	128	144
Actual power	0.80	0.80

α err prob: probability of Type I error, β err prob: probability of Type II error. Actual power: power acquired by statistical program after sample size calculation. Noncentrality parameter δ, G1: sample size group 1, G2: sample size group 2, critical t value and actual power were rounded to the third decimal place.

Table 2. Results of Post-hoc Power Analysis of Two-tailed Independent t-test under the Same Sample Size but Various Sample Size Ratios between Two Groups

Tail(s)	Two			
Effect size d	0.5			
α err prob	0.05			
Sample size group 1	80	100	120	140
Sample size group 2	80	60	40	20
Noncentrality parameter δ	3.16	3.06	2.74	2.09
Critical t value	1.98	1.98	1.98	1.98
Degree of freedom	158	158	158	158
Power (1-β err prob)	0.88	0.86	0.78	0.55

α err prob: probability of Type I error, β err prob: probability of Type II error. Noncentrality parameter δ, critical t value and power were rounded to the third decimal place.

the sample size, the greater the influence of the values of individual samples on variance. This variability becomes stable as the sample size increases. If the sample sizes of the groups are different, then this difference in variability may result in different variances.

Thus far, we have dealt with questions related to the basic assumptions of the t-test that can be found in the research design process. In the normality test, if the sample size is small, the power is not guaranteed. Therefore, it is necessary to secure a sufficient sample size. In order to maximize the power in the t-test, it is most efficient to increase the sample size of both groups equally. The above information may not be necessary for a researcher who has a sufficient sample size and conducts formal research. However, a deep understanding of the basic assumptions of the t-test will help you to understand a higher-level statistical analysis method. This will help you to take a leading role in the research process from research design to the statistical analysis and interpretation of results.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Author Contributions

Tae Kyun Kim (Conceptualization; Writing – review & editing)
Jae Hong Park (Visualization; Writing – original draft)

ORCID

Tae Kyun Kim, https://orcid.org/0000-0002-4790-896X
Jae Hong Park, https://orcid.org/0000-0003-0779-4483

References

1. Nahm FS. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean J Anesthesiol 2016; 69: 8-14.
2. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017; 70: 144-56.
3. PASS Sample Size Software Documentation | PASS Software Help [Internet]. NCSS Statistical Software [cited 2018 Nov 12]. Available from https://www.ncss.com/software/pass/pass-documentation/#Normality.
4. Khan RA, Ahmad F. Power Comparison of Various Normality Tests. Pak J Stat Oper Res 2015; 11: 331-45.
5. List JA, Sadoff S, Wagner M. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. Exp Econ 2011; 14: 439-57.
6. Kim TK. T test as a parametric statistic. Korean J Anesthesiol 2015; 68: 540-6.