Trakya Univ Tip Fak Derg 2009;26(4):297-305

Original Article / *Orijinal Çalışma - Araştırma*

A Bootstrap Confidence Interval for Skewness and Kurtosis and Properties of t-test in Small Samples from Normal Distribution

Normal Dağılımdan Alınan Küçük Örneklerde Çarpıklık ve Basıklık için Bootstrap Güven Aralığı ve t-testinin Özellikleri

Handan ANKARALI,[1] Ayşe CANAN YAZICI,[2] Seyit ANKARALI[3]

Departments of [1]Biostatistics and [3]Physiology, Medical Faculty of Zonguldak Karaelmas University, Zonguldak; Department of Biostatistics, Medical Faculty of Başkent University, Ankara

Submitted / Başvuru tarihi: *17.09.2008* ***Accepted / Kabul tarihi:*** *09.10.2008*

Objectives: We examined the test properties about mean and mean differences, sampling distributions properties of mean and standard deviation, especially when the sample size is lower than 10 and the variable has normal distribution in population. In addition, we aimed to construct a 95% bootstrap confidence interval for skewness and kurtosis values in various samples sizes.

Materials and Methods: In our simulation study, 10,000 samples with replacement were taken from the standard normal population. Various sample sizes were evaluated. Data which were obtained from an animal study were used for comparison of t-test's and Wilcoxon sign test's power in small sample.

Results: According to our results, sampling distribution of skewness coefficients has normal distribution; however, kurtosis coefficients have positively skewed distribution. A bootstrap confidence interval for these coefficients by using these sampling distribution can be used for normality test. Moreover, it was shown that test statistic for the mean has t-distribution in all studied sample sizes.

Conclusion: We can say that distribution shape of the variable plays a more important role than sample size in selection of appropriate test statistic.

Key words: Small sample size; bootstrap simulation; power of t-test; confidence interval; skewness and kurtosis.

Amaç: Normal dağılımdan alınmış ve özellikle örnek genişliği 10'dan küçük olan örneklerden hesaplanan ortalama ve standart sapmanın örnekleme dağılımları ve ortalamaya ait hipotez testinin özellikleri incelenerek yine bu örneklerde çarpıklık ve basıklık katsayılarına ait %95 ihtimalli bootstrap güven aralıklarının oluşturulması amaçlanmıştır.

Gereçler ve Yöntemler: Yapılan simülasyon çalışmasında standart normal dağılım gösteren popülasyondan geri iadeli olarak 10 000 örnek alınmıştır. Farklı örnek genişlikleri incelenmiştir. Küçük örnekte Wilcoxon işaret testi ve t-testinin güçlerini karşılaştırmalı olarak göstermek için bir hayvan deneyinden alınan veriler kullanılmıştır.

Bulgular: Elde edilen bulgular değerlendirildiğinde, çarpıklık katsayısının örnekleme dağılımı normal dağılım olarak bulunurken, basıklık katsayısının pozitif çarpık bir dağılımı olduğu görülmüştür. Bu katsayıların örnekleme dağılımları kullanılarak bulunan bootstrap güven aralığı verilerin normallik testinde kullanılabilir. Bunun ötesinde örnek genişliklerinin tamamında ortalamaya ait test istatistiğinin t-dağılımına sahip olduğu görülmüştür.

Sonuç: Uygun test istatistiğinin seçiminde verilerin dağılım şeklinin, örnek genişliğine göre daha önemli role sahip olduğu söylenebilir.

Anahtar sözcükler: Küçük örnekler; bootstrap simülasyon; t-testinin gücü; güven aralığı; çarpıklık ve basıklık katsayıları.

Presented at the XI. National Congress of Biostatistics, May 27-30, 2008, Malatya, Turkey (XI. Ulusal Biyoistatistik Kongresi'nde sunulmuştur, 27-30 Mayıs 2008, Malatya).

Correspondence (İletişim adresi): Dr. Handan Ankaralı. Düzce Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, 81620 Düzce. Tel: 0380- 542 14 16 / 4190 Fax (Faks): 0380 - 542 13 02 e-mail (e-posta): hankarali@yahoo.com

© Trakya Üniversitesi Tıp Fakültesi Dergisi. Ekin Tıbbi Yayıncılık tarafından basılmıştır. Her hakkı saklıdır.

© Medical Journal of Trakya University. Published by Ekin Medical Publishing. All rights reserved.

In recent years, medical literature has focused increasing attention on sample size requirements in medical research and peer-reviewed journals seriously look for the appropriateness of sample size in their manuscript review process. However, using sample size as small as possible has almost become a necessity due to financial, time saving and especially ethical reasons in clinical or experimental researches carried out on humans or animals. If a study has small sample, the significance of the results in reality (true differences) may not be detected and there are still many statistical problems, because the researchers or biostatisticians can not decide very definitely whether measurements have normal distribution or not. Some studies have suggested that some normality tests are powerful in small samples. Geary's skewness and kurtosis statistics and Shapiro-Wilk's statistic are the most frequently used tests for this purpose. Although some researches have recommended carrying out the normality test on the raw data sets collected from other researches which have used the same variables.[1]

In small samples, nonparametric median test and parametric t-test are most commonly used for one sample mean. In the literature, there are different opinions about the use of nonparametric tests.[1-8] These are:

(1) Nonparametric test should be used in hypothesis test without testing the distribution shape of studied variable when sample size is fewer than 25.

(2) Nonparametric test should be used when sample size is between 5 and 10.

(3) If studied variables have not normal distributions, nonparametric test should be used.

(4) Power of nonparametric tests is low in small sample size.

On the other hand, although some studies have reported that t-test is a robust test for all sample sizes, some others argued that this is invalid in small sample size. According to above mentioned points of view, general tendency is to use nonparametric tests in small samples, but nonparametric tests may also not give accurate results in our opinion. A more important point about this issue is determination of the distribution shape of data. In order to determine it, powerful tests should be used comparatively, and then an appropriate test of the mean should be selected.

Skewness and kurtosis coefficients define the distribution shape of the variable. When referring to the shape of probability distributions, skewness refers to asymmetry of the distribution. A distribution with an asymmetric tail extending out to the right is referred to as "positively skewed" while a distribution with an asymmetric tail extending out to the left is referred to as "negatively skewed".[9] Kurtosis is a measure of how flat the top of a symmetric distribution is when compared to a normal distribution of the same variance. Kurtosis is actually more influenced by scores in the tails of the distribution than scores in the center of a distribution.[10] Accordingly, it is often appropriate to describe a leptokurtic distribution as "fat in the tails" and a platykurtic distribution as "thin in the tails."

It has been shown that normality test can be performed using skewness and kurtosis coefficients.[8] A confidence interval to be defined for these coefficients will give a method which can be used in normality test of the data. In order to construct these intervals, it is required to define sampling distributions of both statistics in various sample sizes.

Aims of the study are; a) to obtain bootstrap confidence limits with 95% probability by generating sampling distributions of skewness and kurtosis values for using in normality test, b) to examine type I error and power of Student t-test with the aid of a simulation in various sample sizes taken from a normal population c) to test normality of a real data set by using this new confidence interval.

MATERIALS AND METHODS

Bootstrap Simulation Study

The bootstrap is a data-based simulation method for statistical inference, which involves repeatedly drawing random samples from the original

data, with replacement. It seeks to mimic, in an appropriate manner, the way the sample is collected from the population in the bootstrap samples from the observed data. The 'with replacement' means that any observation can be sampled more than once.[7]

In the first part of the bootstrap simulation study, 1000 samples, with replacement, that include different numbers of observations (ni) for each bootstrap sample have been taken from standard normal population which has μ=0 and σ=1, and included 1,000,000 observation (N). Sample sizes were selected as 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 50 and 100. Skewness and kurtosis statistics values were calculated from these samples, and then we obtained the sampling distributions of skewness and kurtosis. Bootstrap confidence intervals with 95% probability for skewness and kurtosis values have been constructed by using the sampling distributions.

In the next step of bootstrap simulation study, one sample mean and standard deviation values have been calculated from 10,000 samples, which have been taken from the above mentioned population, with replacement. Sample sizes were determined as 2, 4, 5, 8, 10, 50 and 100. Additionally, mean and standard deviation values were calculated from two independent samples, which are repeated 10,000 times. Type I error probability, type III error probability and power of t-test for both one sample mean and difference between two independent means were calculated. Power of t-test were determined for three different standardized effect sizes (Δ) of the outcome variable. These were Δ=1, 2, 3.[11]

Anderson-Darling test was used to test goodness of fit to any fixed distribution of the sampling distributions of mean, standard deviation, test statistic, skewness and kurtosis statistics calculated from samples.[12]

Nominal alpha level was accepted as 5%. FORTRAN IMSL, MINITAB (ver. 14.0) and EasyFit (ver. 4.0, Trial) was used in calculations.

Example Data

Lactating female mammals exhibit an aggressive behavior, called maternal aggression, to protect their pups towards male or female intruders. One of the most important measures of this behavior is the latency to first attack. In our experiment, the effects of nitric oxide (NO) inhibition via N-G Nitro-L-Arginine Methyl Esther (L-NAME), a nitric oxide synthase enzyme inhibitor, on maternal aggression were investigated. The control values for latency to first attack have been obtained by giving saline in the second day after delivery. In the third day, data resulting from L-NAME application were obtained. Total 10 Wistar rats were used. Previous studies conducted in animals from different species with different NO inhibition methods have shown that the NO inhibition reduces maternal aggression. It is probable that the results of this study are identical to previous studies. In the normality test of obtained data, Anderson-Darling and confidence intervals of skewness and kurtosis defined herein were used. Both paired t-test and Wilcoxon signed rank test were applied on data obtained from 10 animals. Moreover, total 45 samples which have two observations were taken from the data set that included 10 observations. In these samples, power of these two tests were calculated and compared.

RESULTS

Results of Simulation Study

Firstly, 1000 bootstrap samples with replacement were taken from the population and skewness and kurtosis values were calculated. Trimmed mean (TrMean), median and arithmetic mean values of sampling distribution of skewness values were similar to each other (Table 1). As sample size is increased, the values become closer to each other. Figure 1 shows sampling distribution of skewness values in bootstrap samples for some selected sample sizes. As seen, it was found that skewness values have normal distribution except sample size is 3 and there were no extreme values. When n is 3, skewness values had beta distribution, with 0.473 and 0.487 (a=–1.73 and b=1.73) parameters. For nominal α=0.05, α/2% and (1–α)/2% values of normal distribution were used in determination of confidence intervals with 95% for skewness

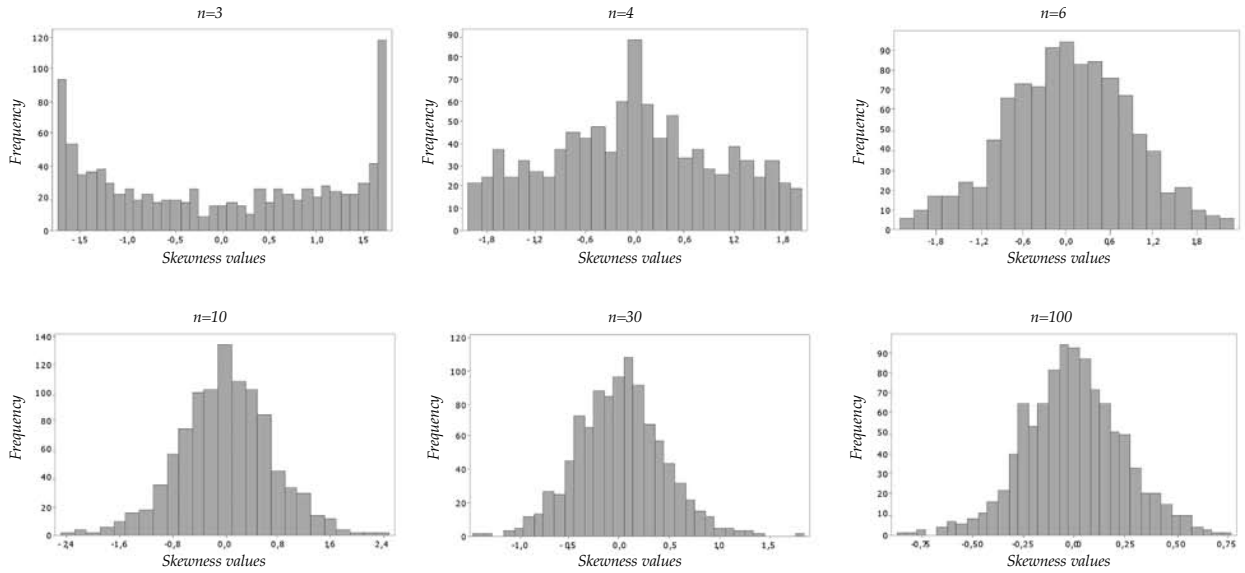

Fig. 1. *Examples of distributions of 1000 sample's skewness values in different sample size (n=3, 4, 6, 10, 30, 100).*

values of each sample size ($Z_{(\alpha)/2} = -1.96$ and $Z_{(1-\alpha)/2} = 1.96$). When skewness value calculated from any sample fall into the range of this confidence interval with 95%, it means that new sample skewness value is same as normal distribution.

Descriptive statistics for kurtosis values are given in Table 2. When sample size was larger than 5, sampling distribution of kurtosis values was found to be compatible with both generalized extreme value and Johnson SB distributions (Fig. 2). These two distributions have positive

Table 1. Descriptive statistics related to skewness values in bootstrap samples and confidence intervals with 95% probability

Sample size (n_i)	Mean of skewness	TrMean of skewness	Median of skewness	SE* of skewness	95% CI for skewness	
					Lower	Upper
3	-0.0249	-0.0276	-0.02	0.713	-1.42	1.37
4	-0.0209	-0.0223	0.00	0.509	-1.02	0.98
5	0.0104	0.0064	0.00	0.419	-0.81	0.83
6	0.0037	0.0082	0.005	0.355	-0.69	0.70
7	0.0127	0.0109	0.025	0.296	-0.57	0.59
8	-0.0047	-0.0007	-0.02	0.269	-0.53	0.52
9	0.0166	0.0242	0.04	0.242	-0.46	0.49
10	-0.0023	-0.0006	0.00	0.219	-0.43	0.43
12	-0.0003	0.0032	-0.01	0.192	-0.38	0.38
14	0.0104	0.0075	0.00	0.157	-0.30	0.32
16	-0.0120	-0.0138	0.00	0.137	-0.28	0.26
18	0.0007	-0.0010	0.00	0.123	-0.24	0.24
20	0.0041	0.0078	0.00	0.113	-0.22	0.23
25	-0.0299	-0.0288	-0.02	0.093	-0.21	0.15
30	0.0028	-0.0008	0.01	0.077	-0.15	0.15
50	-0.0158	-0.0140	0.00	0.048	-0.11	0.08
100	-0.0060	-0.0060	-0.01	0.024	-0.05	0.04

SE: Standard error.

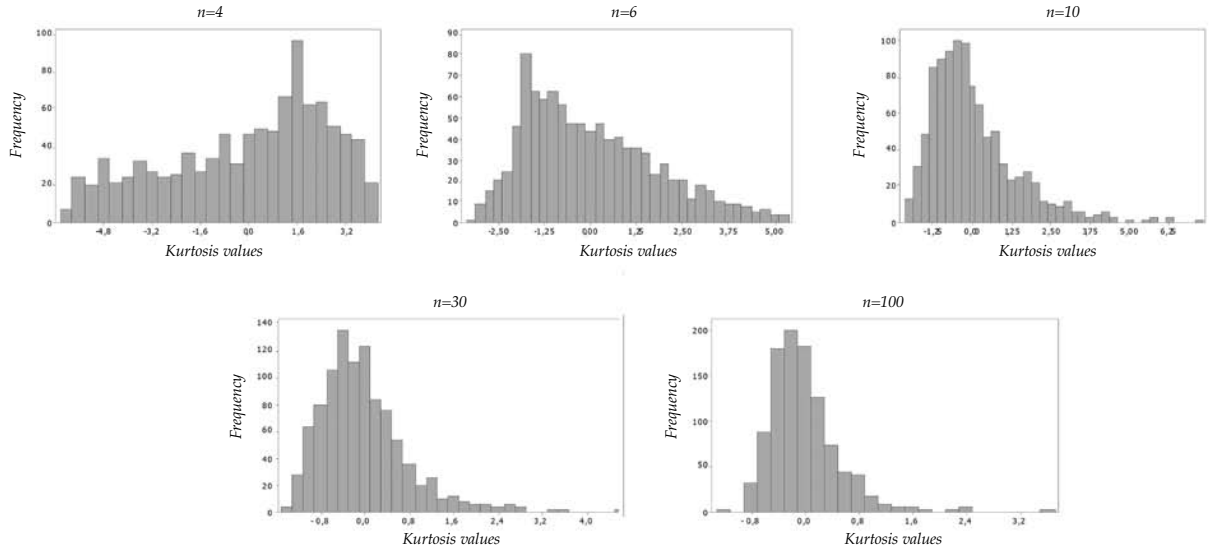

Fig. 2. *Examples of distributions of 1000 sample's kurtosis values in various sample size (n=4, 6, 10, 30, 100).*

skewness. For this reason, they are asymmetric and have negative or positive values. Because the sampling distribution of kurtosis was similar to these two distributions, confidence intervals for studied sample size were constructed according to these distributions. Limit values of distributions with $\alpha/2\%$ and $(1–\alpha/2)\%$ probability, (α=0.05) were determined as $G_{(1-\alpha/2)} = 3.68$ and $G_{(\alpha)/2} = –1.31$ for generalized extreme value distribution and $J_{(1-\alpha/2)} = 10.22$ and $J_{(\alpha)/2} = –0.955$ for Johnson SB distribution. If kurtosis value calculated from any sample fall into the range of this confidence interval with 95%, it means that new sample kurtosis value is same as normal distribution.

Table 2. Descriptive measurements related to distribution of kurtosis values in bootstrap samples and confidence intervals with 95% probability

Sample size (n_i)	Mean of Kurtosis	TrMean of Kurtosis	Median of Kurtosis	SE of Kurtosis	95% CI for Kurtosis according to JSB*		95% CI for Kurtosis according to GEV*	
					Lower	Upper	Lower	Upper
4	-0.0494	0.0406	0.565	1.327	-1.79	4.83	-1.32	13.51
5	0.1213	0.0797	0.160	0.897	-1.06	3.42	-0.73	9.28
6	0.0251	-0.0611	-0.305	0.737	-0.94	2.73	-0.68	7.56
7	-0.0236	-0.1220	-0.420	0.587	-0.79	2.13	-0.58	5.97
8	-0.0114	-0.1334	-0.410	0.549	-0.73	2.01	-0.53	5.60
9	-0.0287	-0.1432	-0.335	0.459	-0.63	1.66	-0.47	4.66
10	0.0234	-0.0856	-0.285	0.427	-0.53	1.59	-0.38	4.38
12	0.0768	-0.0346	-0.200	0.374	-0.41	1.45	-0.28	3.90
14	0.0121	-0.0778	-0.240	0.305	-0.38	1.13	-0.28	3.13
16	-0.0563	-0.1377	-0.290	0.250	-0.38	0.86	-0.29	2.49
18	0.0108	-0.0782	-0.205	0.238	-0.30	0.88	-0.22	2.44
20	-0.0079	-0.0838	-0.220	0.217	-0.29	0.79	-0.21	2.21
25	-0.0102	-0.0801	-0.230	0.182	-0.24	0.66	-0.18	1.85
30	-0.0266	-0.0888	-0.140	0.144	-0.21	0.50	-0.16	1.44
50	-0.0042	-0.0562	-0.120	0.095	-0.13	0.34	-0.09	0.97
100	-0.0202	-0.0553	-0.100	0.049	-0.08	0.16	-0.07	0.48

JSB : Johnson SB distribution; GEV: Generalized Extreme Value distribution.

Table 3. Actual Type I error probabilities and powers for t-test in various sample size for 10,000 bootstrap samples drawn from normal population with 1,000,000 observations

Sample size (n_i)	Distributions of statistics			Empirical (actual) alpha (%)	Power of t-test		
	Mean	Stdev	Test statistics		Δ=1	Δ=2	Δ=3
2	Normal	Pert, Weibull, Johnson SB, Beta	Near the Student-t	4.7	10.1	16.1	25.3
4	Normal	Normal	Near the Student-t	4.9	28.8	72.2	96.8
5	Normal	Normal	Near the Student-t	5.2	42.2	92.1	99.7
8	Normal	Normal	Student-t	4.6	67.2	99.7	100
10	Normal	Normal	Student-t	5.7	78.3	100	100
50	Normal	Normal	Student-t	5.1	100	100	100
100	Normal	Normal	Student-t	4.9	100	100	100

In bootstrap samples, mean, median and trimmed mean of kurtosis values were found to be quite different from each other, because the distribution includes extreme values in the right tail even if sample sizes are large. Sampling distribution of kurtosis values for some selected sample sizes are shown in Figure 2.

In the second part of the simulation study, sampling distribution of sample means had normal distribution in all studied sample sizes. On the contrary, sampling distribution of standard deviation in small samples (especially where n≤5) had positive skewness. In case the sample size is larger than 5, sampling distribution of standard deviations had approximately the normal distribution with a mean of '1'. When sample size is larger than 10, it had normal distribution.

Since the sampling distribution of standard deviation had positive skewness, where n≤5, some values of calculated test statistic were too high or too low. So distribution mean was zero and its shape was symmetric but extremely leptokurtic. The distribution of test statistics was compared to various continuous distributions by using goodness-of-fit tests and found that its shape was like t-distribution. When type I error probability of test statistic based on t-distribution was assessed, it has been determined that nominal alpha level (5% level) was preserved. Same results were found for hypothesis test about two independent mean differences. If empirical (actual) type I error was between 4% and 6% (i.e. 0.2 deviations from nominal alpha level) it means that there is no important deviation from the nominal alpha level (Table 1).[13,14]

When sample size was smaller than 5, power of the test was lower than expected. In order to increase the test power, Box-Cox transformation was applied to sample standard deviations at various lambda values and test statistic values were calculated again. Power of new test statistic values displayed a considerable level of increase, but type I and III error probabilities were larger than expected. When 'n' was larger than 5, t-test was sufficiently powerful (Table 3). Furthermore, type III error probability was found when sample size was "2 and 3" with Δ=1 and 2 and its value was very low (varying between 1/1000 and 6/1000).

Similar results were also obtained for the hypothesis test of two independent samples' mean differences. Illustrative simulation results for one sample are given in Table 3.

Results of Example Data

According to Anderson-Darling test, latency values in each group and differences had normal distribution (p values were larger than 0.05 for all). Furthermore, this normality test has been carried out by using the skewness and kurtosis values of these measurements according to our confidence interval with 95% given in Table 4. Consequently, confidence interval of skewness values for 10 observations were found by using Table 1 (lower=–0.43 and upper=0.43) and it has been observed that only the skewness values

Table 4. Results of example data

	Mean±SD	Skewness	Kurtosis	N
Control	310.5±140.6	0.496	-0.797	10
LNAME	887.1±504.4	-1.038	-1.211	10
DIFFERENCE	620.1±541.8	-1.063	-0.685	10

of control group fit the normal distribution. Confidence interval of kurtosis values according to both distributions was examined by using Table 2 and it has been observed that control group, L-NAME and difference between these two measurements fit normal distribution of kurtosis values.

When the data were analyzed by using paired t-test and Wilcoxon signed rank test, it was found that L-NAME significantly increases the latency to first attack in mothers (n=10; p values for paired t-test and WSR, 0.006 and 0.017 respectively). These results were similar to previous studies in the literature. If the same study was conducted with two subjects, what would be the power of both tests? To determine the test power, a total of 45 different samples with two subjects were taken from the original sample (n=10 subjects) and paired t-test and WSR test were applied. Statistically significant difference was found in 15 of 45 samples according to paired t-test, but no significant difference was found in WSR test. We can say that the power of WSR test is 0% when power of paired t-test is 15/45=33.3%. This result emphasizes that paired t-test is powerful than WSR. Descriptive statistics regarding data used in the study are given in Table 4.

DISCUSSION

The primary objective of all medical studies is to obtain reliable and rapid results with as fewer subjects as possible. The number of subjects are generally limited by financial possibilities and ethical rules in medical researches.[15,16] Health scientists frequently work with small groups of individuals, low incidence conditions, convenience samples, and limited funding. Thus, the assumption of large sample size is often violated by such studies using parametric statistical techniques.[17]

In this study, we were firstly defined a new bootstrap 95% confidence interval for both skewness and kurtosis values that can be used in normality test in small samples. If skewness and kurtosis values which were calculated by researchers from their study are within our 95% confidence interval, they can say that the studied variable has normal distribution. The results from these confidence intervals should be compared with other powerful normality test results which are used in small samples and then the most appropriate test should be selected.

In addition, characteristics of Student t-test for the one sample mean and difference between two independent sample means have been evaluated in small samples. According to our results, type I error probability of hypothesis test of one sample mean taken from a normal population preserves nominal alpha 5% level. In other words, t-test for mean in small samples is a robust test. A test is 'robust' if the actual significance level does not exceed 10% of the nominal significance level (i.e. less than or equal to 0.05±0.005 when the nominal significance level is 0.05).[18]

In addition, power of t-test was found low in sample sizes lower than 5. In other words, used statistical tests cannot detect a true difference between groups. For this reason, clinical/biological significance should also be taken into consideration with statistical results or the study should be repeated with larger sample especially if p values are a little larger than 0.05. It has been shown that sample size is directly related to researchers' ability to correctly reject a null hypothesis (power). As such, small sample sizes often reduce power and increase the chance of a type II error.

Some studies showed that even if assumptions of t-test are not valid, including the small samples, the testing power is higher than nonparametric alternatives. In a study carried out using Likert type data, two independent and dependent groups were compared and both parametric t-test and nonparametric alternatives were used, however, no significant difference was observed between them in terms of type

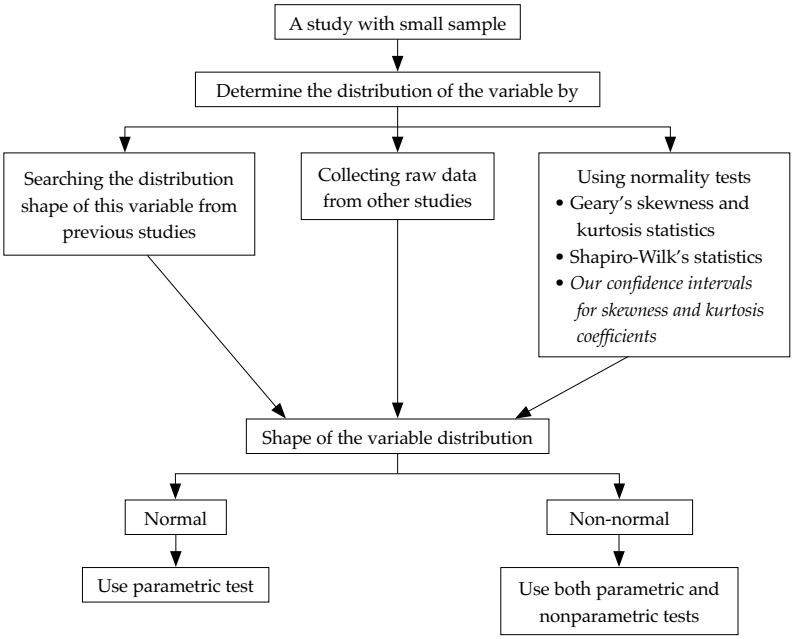

Fig. 3. *Summarized results of our study.*

I and II error probability. For variables whose distribution in population is unknown, it is recommended to evaluate parametric and nonparametric test results together.[7,19-22] In addition, in cases where assumptions related to distribution and variance homogeneity are all violated at the same time and sample size is too small, certain robust methods are suggested, but no firm balance has been established between type I error and testing power with these methods so far.[5]

Our results are summarized in Figure 3. We conclude that selection of suitable test statistic for the mean only based on sample size may give misleading results. The choice of parametric or nonparametric tests just depends on level of measurement, the researcher's knowledge of the variables' distribution in the population, and the shape of the distribution of the variable of interest. If in doubt, try using both parametric and nonparametric techniques in medical studies regardless of sample size. Researchers should try to determine the shape of the distribution of the studied variable especially in small samples. Nonparametric statistics are designed to be used when we know nothing about the distribution of the variable of interest, but parametric tests are more powerful than nonparametric tests for small samples when the studied variable has normal distribution in population according to our results. Defined confidence interval in this study is an alternative approach which can be used in small samples for normality test.

REFERENCES

1. Motulsky H, editor. Intuitive biostatistics. 1st ed. New York: Oxford University Press; 1995.
2. Meek GE, Ozgur C, Dunning K. Comparison of the t vs. Wilcoxon signed-rank test for likert scale data and small samples. J Mod Appl Stat Methods 2007;6:91-106.
3. Tomkins CC. An Introduction to non-parametric statistics for health scientists. Universty Alberta Health Sciences Journal 2006;3:20-6.
4. Sawilowsky SS, Blair RC. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychol Bull 1992;111:352-60.
5. Cribbie RA, Wilcox RR, Bewell C, Keselman HJ. Tests for treatment group equality when data are nonnormal and heteroscedastic. J Mod Appl Stat Methods 2007;6:117-32.
6. Zimmerman DW, Zumbo BD. The relative power of parametric and nonparametric statistical methods. In: Keren G, Lewis C, editors. A handbook for data analysis in the behavioral sciences: methodological issues. Hillsdale New Jersey: Lawrence Erlbaum Associates; 1993. p. 481-518.
7. Walters SJ. Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. Health Qual Life Outcomes 2004 ;2:26.

8. D'Agostino RB, Stephens MA, editors. Goodness-of-fit techniques. New York: M Dekker; 1986.
9. Hildebrand DK. Statistical thinking for behavioral scientists. Boston: Duxbury; 1986.
10. DeCarlo LT. On the meaning and use of kurtosis. Psychol Methods 1997;2:292-307.
11. Cohen J, editor. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale New Jersey: Lawrence Erlbaum Associates; 1988.
12. Yazici B, Yolacan S. A comparison of various tests of normality. J Stat Comput Simul 2007;77:175-83.
13. Stevens J, editor. Applied multivariate statistics for the social sciences. Hillsdale New Jersey: Lawrence Erlbaum Associates; 1986.
14. Kanyongo GY, Brooks GP, Blankson LK, Gocmen G. Reliability and statistical power: how measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. J Mod Appl Stat Methods 2007;6:81-90.
15. Halpern SD, Karlawish JH, Berlin JA. Re: "Ethics and sample size". Am J Epidemiol 2005;162:195-6.
16. Prentice R. Invited commentary: ethics and sample size--another view. Am J Epidemiol 2005;161:111-2.
17. Zar JH. Biostatistical analysis. 4th ed. Upper Saddle River New Jersey: Prentice Hall; 1999.
18. Sullivan LM, D'Agostino RB Sr. Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. Stat Med 2003;22:1317-34.
19. Heeren T, D'Agostino R. Robustness of the two independent samples t-test when applied to ordinal scaled data. Stat Med 1987;6:79-90.
20. Wampold BE, Drew CJ, editors. Theory and application of statistics. 1st ed. New York: McGraw-Hill Pub; 1990.
21. Siegel S, Castellan NJ Jr, editors. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill; 1988.
22. Blair RC, Higgins JJ. Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. Psychol Bull 1985;97:119-28.

View publication stats