A Comparison of Parametric and Nonparametric Tests of Location for Non-normal Data

Fandi Chang, Jasmine Wang, Lin Zou, Yezhi Pan

December 10, 2020

1 Introduction

Biomedical data does not always satisfy the normality assumption in parametric tests for comparison of the mean. The normality assumption suggests that the sample data were drawn from a normally distributed population, and thus, should also be normally distributed. A normal distribution is symmetric with respect to the mean of the distribution, where the mean, median, and mode are equal. For example, in the case of the two sample test of independence, the normality assumption suggests that each independent sample is normally distributed. Parametric tests should not be implemented when there is deviation from normality for the two independent samples because the type I error rate would be inflated. Specifically, parametric tests are robust in terms of type I error, so as the distribution of the groups grow apart, the type I error increases.

Therefore, checking whether or not the normality assumption holds for the data is critical for determining the appropriate parametric or non-parametric test of location. Assessing normality can be accomplished through implementing statistical tests, such as the Kolomogorov-Smirnov or Shapiro-Wilk tests, in conjunction with graphical assessments, such as histograms and box-plots. The skewness and kurtosis of the data can also be computed and compared to that of a normal distribution. When assessing normality, it is important to consider the sample size. That is, when the sample size is small, we cannot conclude that the data is normal due to the uncertainty involved in the lack of information. For studies with small sample sizes, normality should be controlled prior to the analysis in order to avoid violations of assumptions.

Some of the most commonly used statistical tests for comparison of the mean are the t-test, ANOVA, and Mann-Whitney U test. The t-test and ANOVA lie under the branch of parametric statistics, which means that the tests are based on a particular distribution. For these two tests, the distribution would be a normal distribution. The assumption of equal variance between the groups of interest can also be specified. Contrarily, the Mann-Whitney U test lies under the branch of non-parametric statistics, which means that the test does not assume any distribution in particular. Instead of comparing the mean, this test involves computation using rank statistics. However, the results from the t-test and Mann-Whitney U test do not always agree when implemented on diseases with high variability, such as the influenza or the novel COVID-19.

Therefore, our primary research question is to determine ways in which we could account for non-normality in bio-medical data when implementing statistical tests for the comparison of location. In our manuscript, we compared the results of the t-test and Mann-Whitney U test under different conditions of skewness and kurtosis in simulated data.

2 Methodology

In our project, we will use skewness and kurtosis to check normality for our data, which is more straightforward and provide more flexibility and hence is commonly used in the literature.

2.1 Skewness

Skewness is a measure of checking symmetry, or lack of symmetry. We say a data set is symmetric if when we plot it as a histogram, the left of the center point looks the same as the right of the center point. If we have data$X_1, X_2, ..., X_N$and the formula to compute the skewness is:

$$g = \frac{\sum_{i=1}^{N}(X_i - \bar{X})/N}{s^3}$$

In this equation, s is the sample standard deviation computed with N degree of freedom, N is the number of data points, and$\bar{X}$is the mean of the data set. For normal distribution, the skewness value is zero. For non-normal data, if the skewness value is negative, it indicates that the data are skewed left, which means that the left tail is relatively longer than the right tail. And if the value is positive, the data are skewed right, that on the contrary, the right tail is longer than the left tail. In our project, we only will discuss the situation of defining non-normal data set with positive skewness value.

2.2 Kurtosis

Kurtosis is used to measure whether the data are heavy-tailed or light-tailed compared to normal distribution. If a data set has high kurtosis, it tends to have heavy tails or outlier issues, and if it has low kurtosis value, it tends to have light tails and lack of outliers. Similarly, if we have univariate data $X_1, X_2, ..., X_N$, the formula for kurtosis is:

$$\frac{\sum_{i=1}^{N}(X_i - \bar{X})^4/N}{s^4}$$

or

$$\frac{\sum_{i=1}^{N}(X_i - \bar{X})^4/N}{s^4} - 3$$

Similarly, in these two formulas, s represents the standard deviation with N degree of freedom, $\bar{X}$ represents the mean of the data and N is the number of data points. For the first formula, the kurtosis value for standard normal distribution is 3. In order to make it easier to analyze, we can also write it as the form of the second, where in this case, the kurtosis value for standard normal distribution is 0, and positive number indicates a heavy-tailed distribution and negative number indicates a light-tailed distribution. In this project, we building non-normal data set, only positive kurtosis values are considered.

2.3 Simulation Design

To implement the simulation process, we generate data and perform analysis in R.

2.3.1 Data Generation

The first step of the procedure is to generate normal and non-normal data for a given sample size and a given normal/non-normal proportion. We will run 500 replications for each condition and study design factor. The first design factor, sample size, is set to 60, 100, 300, and 1000, as they represent datasets from small to large (Orcan, 2020). With sample sizes given, we generate normal data with *rnorm* and split the dataset into two groups: one group contains the original normal data, and the second group will be transformed into non-normal data using the Fleishman polynomial transformation (Fleishman, 1978):

$$Y = a + bX + cX^2 + dX^3$$

where X is a normally distributed random variable with zero mean and variance 1, and Y represents the linear combination of the first three powers of X. The key is to determine the coefficients a, b, c and d which can help the distribution of Y to have expected results of the moments of the first four orders: mean, variance, skewness, and kurtosis. Fleishman's power transformation method is very commonly used to simulate non-normality. We use *fleishman.coef* function in the NonBinNor R library to generate the coefficients with the given skewness and kurtosis level.

The second design factor is the percentage of normal group in the sample, which is set to 25%, 50%, 75% and 100%. The remaining 75%,50%, 25%, or 0% of the data will be transformed into data with given skewness and kurtosis. As suggested by Orcan, we choose three combinations of skewness and kurtosis that represent slight non-normality to severe non-normality as shown in table 1 from condition 2 to 4 (Orcan, 2020):

Table 1: Skewness and Kurtosis Conditions

Condition	1	2	3	4
Skewness	normal (0)	1	1.5	1.75
Kurtosis	normal (0)	1	2.5	5

Now that we have two independent groups of data, we will merge the two groups to get one big dataset in which the grouping variable is available for data analysis purpose But before merging data and carrying out the sample tests, we assume equal variance in our data. The non-normal group of the data generated from the Fleishman coefficients are expected to have standard deviation and mean close to the standard normal data where we generated the coefficients from, which is zero and one, respectively. However, there are deviation in skewness and kurtosis in the transformed data and multiple simulations may be required to ensure a good fit of data (Bendayan et al, 2013). Therefore, in our study, the assumption of equal variance is loosely held. The reason why we need such assumption is because research shows that type I error probabilities of both t-tests and u-tests will substantially increase if non-normal distributions, especially skewed distributions, have heterogeneous variances(Zimmerman, 2004). Hence we want to control for equal variance to avoid the potential type 1 error fluctuation caused by unequal variances, so that we can make conclusion that discrepancy in mean comparison test results is associated with non-normality in the data.

2.3.2 Data Analysis

As mentioned above, we will perform both independent sample tests and dependent sample tests to record the discrepancies of the test results with each combination of study design factors and non-normal conditions. For parametric independent sample tests, the null hypothesis is that the sample mean of the non-normal group is the sample as mean of the normal group. With the equal variance assumption, pooled variance t-test is used instead of the default unequal variance t-test in the R function *t.test*. In pooled variance t-tests, the t-statistics is computed as

$$t = \frac{\bar{x_1} - \bar{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s_p is the pooled standard deviation and has the expression:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

. And the degree of freedom when comparing the t-stats to critical t-values in the t-distribution is:

$$df = n_1 + n_2 - 2$$

On the other hand, unequal variance t-tests have t-statistics computed as

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with the Welch approximation as the degree of freedom as equal variance is not assumed:

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2}$$

. For non-parametric tests, we use Mann-Whitney U test with the function *wilcox.test* which also requires the assumption of equal variance and tests the null-hypothesis that for randomly selected values X and Y from two populations$P(X) > P(Y) = P(X) < P(Y)$. After performing both tests, we count the number of replications where the t-test and u-test give out same results, which is when both tests reject the null hypothesis or both tests fail to reject the null at the 0.05 alpha level. For dependent sample tests, we merge the normal and non-normal datasets and test the null hypothesis that the sample mean of the data set is equal to 0.

3 Results

3.1 One Sample Test Results

Based on the simulation process illustrated above, the results of one sample test were given in Figure 1. The condition of skewness represented for unique combination of skewness and kurtosis mentioned above.

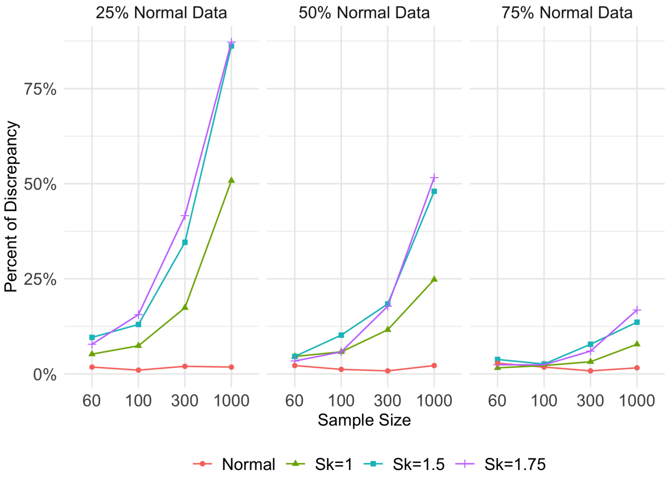

Figure 1: Discrepancy between t-test and U-test for One-Sample tests

Figure 1 showed the discrepancy between one-sample t-test with equal variance and Mann-Whitney U test for sample mixed with different proportion of normal data. Take the first panel as an example, it showed how discrepancy changed when sample size and skewness changed for data contained 25% normal data. More specifically, for a fixed proportion of normal data, the discrepancy increased as skewness increased from 1 to 1.75. For example, in 25% normal data with sample size equaled 1000, the discrepancy increased from 50.8% to 87.2% as skewness increased from 1 to 1.75. It is also worth noting that under the sample size was 1000, the largest skewness group had the highest percentage of discrepancy, even though this was not the case for other sample sizes. Furthermore, under the normally distributed data, the discrepancy did not vary so much and the maximum was around 2.8%. In other words, the t-test and U-test returned the same results at least 98.2% of the times regardless of sample size.

In addition, the discrepancy was positively related to the sample size (see Figure 2). Examining four panels in Figure 2, when the proportion of normal data and skewness were fixed, the dissimilarities between the tests became more prominent as the sample size increased. For example, when skewness was 1.75 and the proportion of normal data was 50%, the dissimilarity increased from 3.4% to 51.6% when the sample size increased from 60 to 1000.

Besides the influence of skewness and sample size, the discrepancy was also dependent on the percent of skewed data. Compare among three panels in Figure 1, the magnitude of discrepancy decreased as the percentage of normal data increased. For example, when the sample size was 300 and skewness was 1.5, the dissimilarity decreased from 43.6% to 7.8% as the proportion of normal data increased from 25% to 75%.

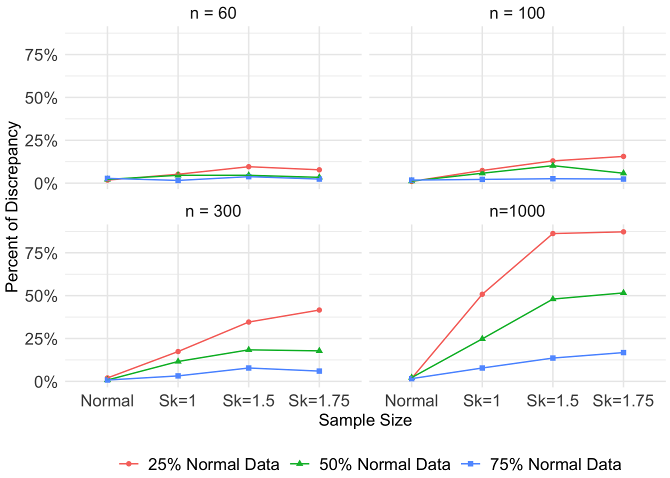

Figure 2: Discrepancy of One-Sample Tests for Different Sample Sizes

In summary, two tests on one sample returned different results in terms of p-values when data was skewed. More specifically, this discrepancy increased as sample size, skewness of data, and percentage of unnormal data increased.

3.2 Two Samples Test Results

Now, we switch the focus to the results of two-sample test which are given in Figure 3.

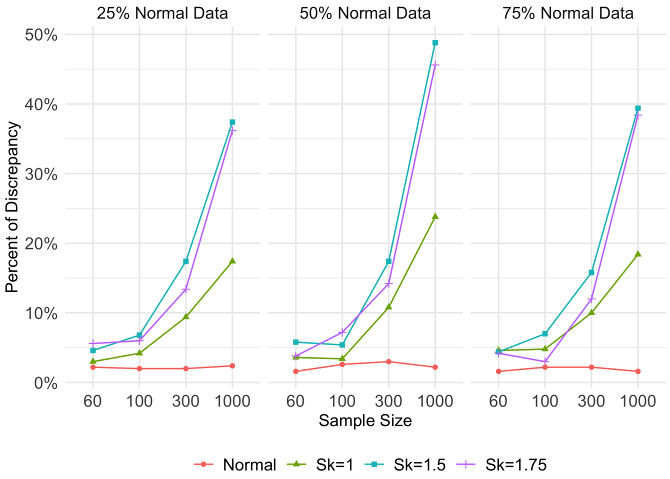

Figure 3: Discrepancy between t-test and U-test for Two-Sample Tests

Similar to Figure 1, Figure 3 showed the discrepancy between two independent sample t-test and Mann-Whitney U test with the same percentage of normal data. Take the second panel as an example, it showed how discrepancy changed when sample size and skewness changed for data contained 50% normal data. The effect of skewness was similar to the one-sample test results except for the group with skewness was 1.75. In other words, for a fixed percentage of normal data, the discrepancy increased as skewness increased from 1 to 1.5. For example, in the data with 50% normal data with sample size equaled 300, the discrepancy increased from 10.8% to 17.4% when skewness increased from 1 to 1.5. In addition, under the normally distributed data, the discrepancy did not vary so much and the maximum was around 3.0%. Thus, we concluded that the t-test and U-test returned the same results at least 97.0% of the times regardless of sample size.

On the other hand, the discrepancy under skewness of 1.75 was less stable and the change of discrepancy was inconsistent with other skewness groups. For example, under 75% normal data, the discrepancy under skewness was 1.75 started at a similar place of other skewness groups, then it remained lower than the skewness of 1.5 group.

The effect of sample size still held in two-sample tests. That is to say, the discrepancy became more prominent when the sample size increased. For example, when fix skewness was 1 and the proportion of normal data was 25%, the dissimilarity increased from 3.0% to 17.4% when the sample size increased from 60 to 1000.

The percentage of normal data did not have much effect on the discrepancy of independent samples tests. Compare among the three panels in Figure 3, the magnitude of discrepancy did not change significantly as the proportion of normal data increased. It should be emphasized the patterns of the discrepancy between 25% and 75% normal data were very similar. On the other hand, the dissimilarities under 50% normal data were the greatest when the sample size was 1000. The results of independent tests were given in Table 2 in detail.

Table 2: Discrepancy Values (%) between t-test and U-test for Independent Sample Tests

Sample Size	% of Normality	Normal	Skewness = 1	Skewness = 1.5	Skewness = 1.75
60	25	2.2	3.0	4.6	5.6
	50	1.6	3.6	5.8	3.8
	75	1.6	4.6	4.4	4.2
100	25	2.0	4.2	6.8	6.0
	50	2.6	3.4	5.4	7.2
	75	2.2	4.8	7.0	3.0
300	25	2.0	9.4	17.4	13.4
	50	3.0	10.8	17.4	14.2
	75	2.2	10.0	15.8	12.0
1,000	25	2.4	17.4	37.4	36.2
	50	2.2	23.8	48.8	45.6
	75	1.6	18.4	39.4	38.4

Table 2 showed the average discrepancy values between the t-test and U-test. In summary, two tests gave different results in terms of p-values when data is skewed. In addition, this discrepancy increased as sample sizes increased. The skewness of data (skewness = 1 and 1.5) was also positively correlated with the dissimilarity. However, the proportion of normal data did not change the discrepancy much.

4 Discussion

This simple simulation found that the discrepancy between t-test with equal variance and Mann-Whitney U test were associated with sample sizes, percentage of normal data, and skewness of samples. In both one-sample and two-samples tests, sample sizes were positively related to the dissimilarity of test results. The percentage of normal data was inversely related to the discrepancy in one sample tests case, but this relationship did not hold for two-samples tests. The skewness of 1 and 1.5 were positively related to the discrepancy, while the skewness of 1.75 does not follow this pattern strictly in both tests.

It could be caused by the property of algorithms used to generate samples. As illustrated above, for 500 replications, the sample was generated at once and stored in a matrix. In other words, a $n \times 500$ matrix was generated and each column represented one sample. Another potential way to generate the sample is to generate a vector of length n for each time and repeat this process 500 times. Compared to this potential method, the method used in this simulation involved less randomness. As a result, the final discrepancies based on those samples may not be representative. Another possible explanation is that the sample size of

1000 is not large enough. There is no clear cut on how large a sample should be to be considered as a large sample. It could be better if a larger sample size such as 10,000 can be used in the simulation. In addition, we made the assumption that two samples had equal variance. To get more accurate results, future studies can test if equal variance assumption holds for samples. If the equal variance assumption is violated, t-test with unequal variance and Kolmogorov-Smirnov Test can be utilized instead.

Building upon this simulation, it is important to test for normality assumption before utilizing any tests and models. The violation of assumptions could lead to inaccurate and even invalid test results. In practice, many biostatisticians will use parametric tests intensively, such as t-test and analysis of variance, to decide if a treatment or drug is effective by comparing the mean of different groups. It is critical for researchers to check on the normality to choose the correct tests and models. If any of the assumptions is violated, the non-parametric methods can be used to get a more reliable conclusion without making any assumptions on the distribution of the data.

Furthermore, since the simulation showed that the skewness and kurtosis, percentage of normal data and sample sizes can affect results of t-test and U-test, it is risky to use skewness and kurtosis alone to test normality. Instead, there are variety of ways to test normality such as Shapiro-Wilk test and histogram. The conclusion could be more reliable and stable if different methods are used to check normality.

5 Conclusion

Researchers must carefully check the assumptions of statistical tests and models before using them. Otherwise, the results will be inaccurate and the conclusion will be invalid. Consequently, this may lead to a waste of time and limited research funds. If the normality assumption is violated, nonparametric methods should be used, as they do not rely on strict assumptions of the underlying data. In our study, the t-test with the assumption of equal variance and the Mann-Whitney U test gave different results in terms of p-values when the data was skewed. These discrepancies are evident under different conditions of skewness, sample size, and percentage of normal data present. Hence, it is naive to only use skewness and kurtosis to test normality. In addition to these measurements, the results from statistical tests, such as the Kolomogorov-Smirnov or Shapiro-Wilk tests, should be used in conjunction with graphical assessments, such as histograms and boxplots, in order to obtain a comprehensive understanding of the data and to make a more reliable and stable conclusion of normality. We suspect that the anomaly of the discrepancy between the t-test and U test for skewness 1.75 was caused by the simplicity in our data generation procedure, which provides less randomness compared to generating a vector of length n and repeating this process 500 times. Consequently, the discrepancy values for these particular samples may not be representative. We recommend future studies to generate 500 vectors of length n and to assess the discrepancy using $n = 10,000$ to obtain more stable results. Furthermore, we made the assumption of equal variance for the t-test, but this assumption may not hold in other cases. Future studies should test if the equal variance assumption holds for their samples. We recommend using the t-test with unequal variance and the Kolomogorov-Smirnov test when the equal variance assumption is violated.

Reference

Bendayan, R., Arnau, J., Blanca, M. J., & Bono, R. (2013). Comparison of the procedures of Fleishman and Ramberg et al. for generating non-normal data in simulation studies. Anales De Psicología / Annals of Psychology, 30(1), 364-371. https://doi.org/10.6018/analesps.30.1.135911

Fleishman, A.I. (1978). A method for simulating non-normal distributions. Psychometrika 43, 521–532. https://doi.org/10.1007/BF02293811

Orcan, F. (2020). Parametric or Non-parametric: Skewness to Test Normality for Mean Comparison . International Journal of Assessment Tools in Education , 7 (2) , 255-265 . DOI: 10.21449/ijate.656077

Zimmerman, Donald W. (2004). Inflation of Type I Error Rates by Unequal Variances Associated with Parametric, Nonparametric, and Rank-Transformation Tests. https://files.eric.ed.gov/fulltext/EJ848306.pdf