An Empirical Comparison of Power of Two Independent Population Tests Under Different Underlined Distributions

Abstract

Determining whether sample differences in central tendency represent real differences in parent populations is a typical issue in applied research. If the conditions of normality, homogeneity of variance, and independence of errors are met, the t-test can be used for a two sample instance (two groups). However, the nonparametric equivalent is taken into account when these presumptions are violated. In order to determine which test is most effective and resilient to a certain distribution and sample size when samples are obtained from separate populations, the study compares the effectiveness and sensitivity of power of four test statistics.. These tests were examined under normal and some skew distributions at sample size of 5, 10, 15, 20, 25, 30, 40, 45, and 50 using simulation. The most effective test for a given distribution and sample size was chosen using the power of each test computed. The study found that when data are taken from a normal distribution and tested at small and large sample sizes, respectively, the t-test and Welch test have the highest power, while the Median is the most resistant to uniform and gamma, and the Man-Whitney test is the most reliable for exponential distributions.

Keywords: Simulation, Power, Normal, Gamma, Uniform, Exponential

1 Introduction

In research investigations, using statistical methods to examine the data gathered is given top priority. The primary focus has been on obtaining the pertinent information from the obtained data to address the study's intended objectives. Any researcher must be extremely careful to select the right tool for the job. Examining an analytical tool's resistance to assumptions' variation is another critical decision-making factor. To model the data effectively and make reliable inferences, numerous statistical methods have been created while taking these needs into account. To model the data effectively and make reliable inferences, numerous statistical methods have been created while taking these needs into account. During the process, it was discovered that only a small number of tools were dependent on various assumptions that were essential to their operation. A t-test, for instance, works effectively when data is gathered from a typical population. Prior to being utilized to test the hypothesis based on the population mean, this assumption must first be checked using the sample that was actually drawn. The conclusions made need not be accurate if the data do not support the assumption of normalcy but one nevertheless chooses to apply it (Srilakshminarayana, 2015).

Furthermore, a good alternative value compared to the null hypothesis may have low test power. The population may not always act regularly and may exhibit characteristics of the heavy tail phenomenon. Outliers or extremes are what produce heavy tail phenomena, hence the best model for the phenomenon must be selected. A statistical hypothesis is a premise or claim about one or more populations that may or may not be true. Typically, it is a statement about a set of traits in a population distribution. It is known as a hypothesis since it is uncertain of its veracity. In tests of substantial difference between two independent samples, a variety of approaches may be

employed, and each one may result in results that are considerably different. This implies that choosing the wrong test statistic may lead to an erroneous conclusion (Edith and Nkiru, 2016). To prevent the spread of misleading information, it is essential to properly investigate a select few methods for determining whether there is a significant difference between variables or subjects, in particular independent samples.

The relative effectiveness of the parametric t-test and the Wilcoxon Signed ranked test at various sample sizes on a pair and a single sample, respectively, has been studied by Akeyede et al. (2014) and Edith and Nkiru (2016). This study examines the effectiveness and sensitivity of four test statistics to establish which test is more useful in various situations where samples/populations are independent..

2 Related Empirical Studies

Scientists that conduct empirical investigations and contribute significantly to their respective fields of study rely on the results of their sample data. They check to see that the right sample strategy was applied when gathering the respondents' responses. The assumption of normalcy served as the foundation for the majority of parametric testing processes, and any departure from this assumption may have an impact on their ability to produce accurate results. This section examines many viewpoints on the assumption of normality and the effects that result from violating it that have been made by scientists, some of whom have created the methodologies. The normal distribution has played a significant role in the advancement of statistical methods since its inception (Abraham de Moivre, 1918), and it holds a prominent place in the evolution of statistics. In their book on the normal distribution, Patel and Read (1982) go into great length regarding the normal distribution and describe many of its key characteristics, approximations, and behaviors. It is suggested that a beginner who wants to learn about normal distribution start with this book before moving on to more sophisticated works that cover the topic. The t-test, one of the most significant and popular parametric tests, is heavily reliant on the supposition of normality.

Student (1908) assumes that the population from which the sample was drawn is normal and develops a t-distribution curve to represent sample means in his paper on the derivation of likely error of mean. Additionally, he notes that the results gained do not necessarily apply to groups that are known to not have a normal distribution. If one looks at population strict normalcy, the response is not always true. For the t-distribution to be used in the analysis, the population does not have to be strictly normal. If the population is around typical, that is enough. As a result, a researcher must make sure the sample is random. Fisher (1925) examined the uses of the t-distribution in his study. The paper provides the t-test statistic distribution for one sample and two sample problems under the supposition of normality. A t-distribution with the appropriate degrees of freedom can be used to analyze the behavior of the difference between two sample means. The variances of the two populations, however, have been considered to be equal. Welch (1947) investigated how the equality of variances assumption affects the behavior of the difference between two sample means. He suggests a degree of freedom correction to make the sample distribution of the test statistic the t-distribution once more. It has been assumed that the populations independently follow normal distribution in the case of two sample means. According to Welch (1938), the actual t-test for differences between two means with the sum of two sample sizes less two will work well when the variances are equal. Under unequal variances,

it does not perform effectively. When comparing differences in means of equal numbers of data, the t-test can still be utilized under mild departures from normality but not extreme departures, according to Bartlett (1935), who studies the impact of non-normality on t-distribution. Additional studies on the effectiveness of these tests when non-normality is present (Neyman and Pearson 1928).

Similar to this, Wilcoxon (1945) created an alternate method to a two sample (dependent) paired t-test that is now commonly known as the "Wilcoxon signed rank test." Kruskal-Wallis (1952) developed a test that can be used as an alternative to ANOVA, while Mann-Whitney (1947) provided a testing process as an alternative to the t-test (independent samples). The Friedman (1937) test can be applied if one is interested in evaluating the hypotheses on more than two related samples. The data are sorted according to magnitude in practically all non-parametric tests, and this ranking is also utilized to calculate the test statistic. Another significant feature of nonparametric methods is that for high sample sizes, the test statistic tends toward normality. The majority of statistical software has add-ons to calculate.

There aren't many discussions over whether to use a t-test or a Wilcoxon test depending on the test's power. In 1956, Hodges and Lehmann investigated the effectiveness of a few non-parametric alternatives to the t-test. Boneau (2020) contrasts the t-and test's U's power. Under several non-normal distributions, Blair and Higgins (1980) compared the power of the Wilcoxon test to that of the student t-test. The relative strength of the Mann-Whitney U test and the t-test is discussed by Blair and Higgins (2020). We compare the t-power test's to the Wilcoxon test's power for the data sets used in this study by taking into consideration a couple of these considerations. Nonparametric approaches may not yield accurate findings if even one of its presumptions is violated. Additionally, it is difficult to check the assumptions, such as symmetry, etc. The same problems that arise in cases of normality will also arise in this situation. The shortcomings of non-parametric tests have paved the way for the development of other, more reliable statistical techniques. They are referred to as "Robust Statistical Methods." These techniques include location estimates, interval creation, procedures for testing the relevance of location, one-sample and two-sample tests, etc. Due to their insensitivity to assumptions being violated, they are crucial un data analysis. Huber (1964) examines location estimation techniques and gives a variety of techniques. Additionally, Huber (2019) reviews various literature-based robust statistical techniques. These strategies, among other things, work effectively even when the data is contaminated. When there are numerous outliers and extreme observations in the data.

Therefore this study examined the relative efficiency of four different tests of sample means from independent populations under Normal, Uniform, Gamma and Exponential distributions. The tests considered are t-test, Welch’s t-test, Median, and Mann-Whitney U test.

3 Methodology

This study focuses on two parametric tests namely; Student t-test and Welch's *t*-test and two non-parametric tests Median and Mann-Whitney U tests of two independent samples. These tests were examined under normal and some skew distributions at sample size of 5, 10, 15, 20, 25, 30, 40, 45, and 50 through simulation. The power and Type I Error of each test were used to determine the best among the tests under a particular condition. The relative efficiency of these tests was examined through simulations in R statistical software.

3.1 Simulation Procedures and Analysis

For every sample size of 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50, which considered as small to large sample sizes respectively, random samples were simulated using the normal, uniform, gamma, and exponential distributions. The data were generated under the condition of equality of variances assumption from different aforementioned distributions and sample sizes. Data were also produced for variances with unequal variances using the same distributions and sample sizes. Each test approach was used on various sample sizes of data sets, and the test's power was examined in each case. Each distribution's means and variances—either the same means and the same variances, or separate means and variances from other family distributions—were used to simulate two samples from it simultaneously at each replication, resulting in independent samples. 1000 times were put through the process for each sample size selected.

It is frequently necessary to decide whether a test is significant or not. We can investigate two inquiries regarding the relative worth of statistical tests by converting the p-value into a binary decision:

i. What proportion of results that are noteworthy will a researcher wrongly consider to be unimportant?
ii. How many stated significant outcomes will actually turn out to be unimportant?

In fact, the number of times H_0 is rejected when it is true is tallied as a Type I error, and the number of times H_0 is accepted when it is true is tallied as the test's power for each statistic under consideration.

3.2 Student t-test

$H_0 : \mu_1 = \mu_2$

In a way, this hypothesis is about the mean difference because it is true if and only if $\mu_1 - \mu_2 = 0$. Therefore, we compute the sample means $\bar{x}_1$ and $\bar{x}_2$ from independent samples of size n_1 and n_2, respectively. The next step is to determine whether the difference between $\bar{x}_1$ and $\bar{x}_2$ is statistically significant. To determine whether there is a difference between the two populations' means

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \quad (1)$$

There are numerous names for this estimation. The phrase "the pooled unbiased estimator" as well as "Mean Square Within" and "Mean Square Error" have been used to refer to it. In particular, the formula for two groups is;

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

Note that the resulting t-statistic has $n_1 + n_2 - 2$ degrees of freedom

3.3 Welch's t-test

The two populations are thought to have normal distributions with equal variances for the purposes of the Student's t-test. Although the premise of normality, continuous data, and interval scales are preserved, Welch's t-test is made to handle unequal variances.

$$test\ statistic = \frac{estimate - hypothesised\ value}{standard\ \ error} \tag{3}$$

Our estimate is the difference between the two sample means when there are two independent samples. The hypothesized value is the difference between the two genuine population means that we assume to exist; this value is frequently zero (to test whether there is a difference or not between the means of the two populations).

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE} \tag{4}$$

We estimate our standard error in a Welch's two independent sample t-test using the sample standard deviations.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}} \tag{5}$$

The Welch-Satterthwaite equation is used to roughly estimate the degrees of freedom associated with these variance estimations.

The statistic's degree of freedom is determined as follow

$$df' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}} \tag{6}$$

3.4 Median Test

In a random sample of size n_j, where the two populations are measured at least on an ordinal scale, let x_i represent the i observation. The two samples are combined into one sample of size $n = \sum_{j=1}^{2} nj;\ j = 1,2 = n_1 + n_2$ observations, which is then, sorted either from the largest to the smallest or from the smallest to the largest in order to apply the two sample median test by ranks. In the absence of ties, any randomly chosen observation in the combined sample is equally likely to be greater or less than any other observation in the sample and, as a result, is equally likely to receive any one of the ranks assigned to the observations. This justifies the hypothesis of equal population medians. Let

$$R_j = \sum_{i=1}^{n_j} r_{ij} \tag{7}$$

be the average rank divided by the sum of the ranks given to the observations selected from population j for $j=1, 2$, and

$$\overline{r_{.j}} = \frac{R_j}{n_j} = \sum_{i=1}^{n_j} \frac{r_{ij}}{n_j} \tag{8}$$

The overall mean rank is

$$\overline{r} = \sum_{j=1}^{2} \frac{n_j \overline{r_{.j}}}{n.} = \sum_{j=1}^{2} \sum_{i=1}^{n_j} \frac{r_{ij}}{n}$$

That is, $\bar{r} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$ (9)

The total variance of all the ranks is

$$\sigma^2 = \frac{\sum_{j=1}^{2} \sum_{i=1}^{n_j} r_{ij}^2 - \left(\frac{\sum_{j=1}^{2} \sum_{i=1}^{n_j} r_{ij}}{n} \right)^2}{n-1}$$

$$\sigma^2 = \frac{\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{4}}{n-1} = \frac{n(n+1)}{12} \tag{10}$$

Currently, the total squared variation between the observed sample mean rank or treatment group mean rank and their overall mean rank $\bar{r}$ is

$$S_{ab}^2 = \sum_{j=1}^{2} n_j (\overline{r_{.j}} - \bar{r}) = \sum_{j=1}^{2} \frac{R_{.j}^2}{n_j} - \frac{n(n+1)^2}{4}$$

Now the quadratic form,

$$Q = \chi^2 = \frac{S_{ab}^2}{\sigma^2} = \frac{\sum_{j=1}^{2} \frac{R_{.j}^2}{n_j} - \frac{n(n+1)}{4}}{\frac{n(n+1)}{12}} \tag{11}$$

That is,

$$Q = \chi^2 = \frac{12}{n(n+n1)} \sum_{j=1}^{2} \frac{R_{.j}^2}{n_j} - 3(n+1) \tag{12}$$

has approximately a chi-square distribution with $k-1=2-1=1$ degree of freedom for sufficiently large n, and may be used to test the null hypothesis of equal population medians. The null hypothesis is rejected at α level of significance if $Q = \chi^2 \geq \chi^2_{1-\alpha,1}$ otherwise the null hypothesis is accepted. Note that equation 5 can be alternatively expressed as

has a chi-square distribution that, for sufficiently big n, has $k-1=2-1=1$ degree of freedom and can be used to test the null hypothesis that the population medians are equal. If the null hypothesis is accepted in all other cases, it is rejected at the level of significance if $Q = \chi^2 \geq \chi^2_{1-\alpha,1}$. Keep in mind that equation (5) can also be written as

$$S_{ab}^2 = \sum_{j=1}^{2} \frac{R_j^2}{n_j} - \frac{\left(\sum_{j=1}^{2} R_j^2\right)}{n} = \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} - \frac{(R_1+R_2)^2}{n_1+n_2}$$

Or when further simplified yields

$$S_{ab}^2 = \frac{(n_1R_2 - n_2R_1)^2}{n_1n_2(n_1+n_2)^2} \tag{13}$$

Hence, the test statistic of equation (6) can be equivalently written as:

$$Q = \chi^2 = \frac{12((n_1R_2 - n_2R_1)^2)}{n_1n_2(n_1+n_2)^2(n_1+n_2+1)} \tag{14}$$

If n_1 and n_2 are both at least (5), the test statistic of equations (6) and 8 is adequate and produces good results. Equation (8) further simplifies to $n_1 = n_2 = m$ if the two samples are equal.

$$Q = \chi^2 = \frac{12((n_1R_2 - n_2R_1)^2)}{m(m)^2(m+1)} \tag{15}$$

3.5 Mann-Whitney *U* test

The Mann-Whitney *U* test is a non-parametric test; therefore it makes no assumptions about how the scores will be distributed. The test's underlying presumptions are listed below.

i. The observations made by each group independently of the other.
ii. The results are ordered (i.e. one can at least say, if any two observations, which is the greater).

iii. The chance of an observation from one population *(X)* exceeding an observation from the second population *(Y)* equals the likelihood of an observation from *Y* exceeding an observation from *X* under the null hypothesis since the distributions of both groups are equal. In other words, there is asymmetry between populations in terms of the likelihood of a larger observation being drawn at random.

iv. The likelihood that an observation from one population *(X)* will surpass an observation from a different population *(Y)*, after ties are taken into account, is not equal to 0.5 under the alternative hypothesis. A one-sided test can also be used to express the alternative, as in the following example: $P(X > Y) + 0.5P(X = Y) > 0.5$.

$$U = n_1 n_2 - \frac{n_1(n_2+1)}{2} - \sum_{i=r_1+1}^{r_2} R_i \quad (16)$$

Where:

U =Mann-WhitneyU test

n_1 = sample size one

n_2 = Sample size two

R_i = Rank of the sample size

4 Results and Discussion

Data were simulated from normal and normal distributions of two different independent populations at different sample sizes. In the first instance, the variances of the two populations were set equal and each test was applied on the data generated in which power is recorded. The simulations and analyses were carried out with 1000 iterations and their average values were recorded for all categories in table 1 and 8 for equal and unequal variances respectively. The best test was determined by the power of the test and discussions were presented at the beneath of each table.

Table 1: Relative Frequency of Power of Test for Equal Variance ($\sigma_1^2 = \sigma_2^2$)from Normal Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	4.179×10^{-9}	6.145×10^{-9}	1.0000	0.0079
10	6.884×10^{-15}	1.186×10^{-14}	1.0000	1.083×10^{-5}
15	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	1.289×10^{-8}
20	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	1.451×10^{-11}
25	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	1.582×10^{-14}
30	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	$<2.2 \times 10^{-16}$
35	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	$<2.2 \times 10^{-16}$
40	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	$<2.2 \times 10^{-16}$
45	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	$<2.2 \times 10^{-16}$
50	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.0000	$<2.2 \times 10^{-16}$

Table 1 presents the result of power of different tests considered in this study at different sample sizes. It was observed that all the tests except median tests reject the wrong hypotheses fixed for

them and consequentially accept the right alternatives at various sample sizes. They indeed possess power of the tests. Specifically, t-test has the strongest rejection especially at a smaller sample sizes and therefore has the highest power but has closer values to Welch test as sample size is getting larger. However, Mann-Whitney test has the lower power compare with the two parametric tests at lower sample sizes but from sample sizes of 30 and above, it competes well with others, this is due to its values higher than them while Median test does not have the power in respect to normal distribution at all sample sizes

Table 2: Relative frequency of Power Test for Equal Variance ($\sigma_1^2 = \sigma_2^2$)from Uniform Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	0.0383	0.0669	0.0188	0.1508
10	4.171e-04	1.166e-03	9.766 e-06	1.083e-05
15	1.094e-06	1.634e-05	3.052e-08	9.025e-08
20	1.199e-06	1.172e-05	5.909 e-08	1.831e-06
25	1.751e-10	1.664e-08	7.749e-11	9.118e-11
30	5.967e-10	2.075e-08	2.974e-05	1.266e-10
35	8.024e-16	1.579e-12	2.088e-07	2.407e-12
40	3.275e-11	1.563e-09	2.114e-05	5.906e-09
45	9.526e-15	3.199e-12	4.667e-09	4.792e-13
50	1.124e-15	5.159e-13	2.231e-10	4.895e-13

In table 2, the results of analysis of data generated from uniform distribution for equal variances are shown. It is observed from the table that the two non-parametric tests have stronger rejection to the wrong hypothesis at $\alpha = 0.05$ than parametric counterparts at smaller sample sizes i.e from sample size of 5 to 25 where median is the best among them while. However, from sample size of 30 to 50, which can be classified as large sample sizes parametric t-test has the highest power compared with non-parametric tests.

Table 3: Relative frequency of Power Test for Equal Variance ($\sigma_1^2 = \sigma_2^2$)from Gamma Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	0.3087	0.3382	0.9688	0.0556
10	0.1656	0.1820	0.9990	0.0007
15	0.1868	0.1974	0.9990	0.0001
20	0.4002	0.4053	1.0000	5.024e-07
25	0.0982	0.1046	1.0000	6.972e-08
30	0.0510	0.0558	1.0000	3.567e-11
35	0.1051	0.1052	1.0000	3.118e-09
40	0.0269	0.0298	1.0000	5.77e-12
45	0.0021	0.0028	1.0000	<2.2e-16

50	0.0010	0.0013	1.0000	9.81e-11

The rejection of null hypothesis by Mann-Whitey in table 3 seems to be strongest at all categories of sample sizes and hence has the highest power. However, as the sample size increases, the two parametric tests seems to be better than median test when compare the power of their rejections of null hypotheses with the t-test has the best power in that category

Table 4: Relative frequency of Power of Test for Equal Variance ($\sigma_1^2 = \sigma_2^2$)from Exponential Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	0.1901	0.2244	0.5000	0.8413
10	0.0020	0.0050	0.0010	0.0007
15	0.0004	0.0013	0.0004	3.507e-06
20	0.0037	0.0060	0.0060	3.364e-05
25	4.768e-06	2.744e-05	2.98e-08	3.085e-12
30	3.69e-06	1.764e-05	2.887e-08	1.537e-11
35	7.782e-07	4.35e-06	1.836e-08	2.514e-13
40	2.428e-09	4.498e-08	7.467e-10	4.184e-14
45	2.141e-08	1.741e-07	4.667e-09	2.459e-12
50	6.8e-09	6.377e-08	2.231e-10	1.851e-13

Table 4 which shows the relative performance of the tests with respect to the data generated from exponential distribution of equal variances, the Mann-Whitney has the best power with strong rejection of wrong hypothesis followed by median test. The t-test is the best between the two parametric tests and the power of all the tests improve as sample size increases.

Table 5: Relative frequency of Power of Test for Unequal Variance ($\delta_1^2 \neq \delta_2^2$) from Normal Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	1.046×10^{-7}	1.008×10^{-7}	1	0.007937
10	3.356×10^{-11}	6.262×10^{-12}	1	1.083×10^{-5}
15	3.212×10^{-16}	3.089×10^{-16}	1	1.289×10^{-8}
20	2.432×10^{-16}	2.387×10^{-16}	1	1.451×10^{-11}
25	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1	1.582e-14
30	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1	$<2.2 \times 10^{-16}$
35	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1	$<2.2 \times 10^{-16}$
40	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1	$<2.2 \times 10^{-16}$
45	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1	$<2.2 \times 10^{-16}$
50	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1	$<2.2 \times 10^{-16}$

Table 5 shows the relative performance of the four tests, using power as a criterion for the assessment, when the variances of the two sample data generated from normal are not equal. It was observed that that the Welch's t- test has the strongest rejection value to the wrong null

hypothesis than other tests and indeed has the highest power. This is followed by the parametric test. However, as the sample size increases, their power becomes stronger and the t-, Welch's and Mann-Whitney U tests have similar power values.

Table 6: Relative frequency of Power of Test for Unequal Variance ($\delta_1^2 \neq \delta_2^2$) from Uniform Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	0.9179	0.9184	0.8125	0.6905
10	0.8551	0.8539	0.6230	0.6030
15	0.8654	0.8646	0.5000	0.4748
20	0.8762	0.8671	0.5881	0.4211
25	0.7699	0.7006	0.7878	0.7437
30	0.4151	0.4083	0.8998	0.1381
35	0.1339	0.1380	0.02048	0.01299
40	0.9669	0.9670	0.3179	0.0886
45	0.7794	0.7800	0.5000	0.4552
50	0.5046	0.5061	0.4439	0.4227

Table 6 shows that the two non-parametric tests have a stronger rejection to wrong null hypothesis and power at $\alpha = 0.05$ than parametric counterparts at various sample sizes, where the Mann-Whitney test is the best among them. As sample size is getting larger, the t-test has the weak power at various sample sizes and hence is the worst among the four tests followed by the Welch parametric test

Table 7: Relative frequency of Power of Test for Unequal Variance ($\delta_1^2 \neq \delta_2^2$) from Gamma Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	0.3116	0.3409	0.000968	0.05556
10	0.5254	0.5262	0.000599	0.00105
15	0.1983	0.2087	0.0002151	0.0002151
20	0.3307	0.3368	2.241e-08	9.249e-07
25	0.5343	0.5346	1.75e-12	1.75e-07
30	0.061	0.06585	<2.2e-16	4.126e-10
35	0.51	0.5121	<2.2e-16	4.503e-08
40	0.02781	0.03073	<2.2e-16	9.769e-11
45	0.002199	0.002896	<2.2e-16	<2.2e-16
50	0.3828	0.3848	<2.2e-16	9.878e-10

From table 7, the median test is the best among the four tests at all sample sizes for data simulated from Gamma distribution where variances are not equal followed by Man-Whitney U test. The least performing test in this case is the Welch test due to its weak rejection values. It also observed that Mann Whitney and median tests have the strongest wrong hypotheses, at 5%

level of significance, as sample sizes getting large especially from 30 upward, hence have they the highest power at that category.

Table 8: Relative frequency of Power of Test for Unequal Variance ($\delta_1^2 \neq \delta_2^2$) from Exponential Distribution

Sample Size	Test Statistics			
	T-test	Welch's t- test	Median test	Mann-Whitney U test
5	0.2875	0.3187	0.78658	0.87732
10	0.003012	0.00219	0.0007863	0.0008653
15	0.00014	0.00023	0.0005431	3.673e-06
20	0.00011	0.0004	0.005909	3.4523e-05
25	3.120e-06	2.744e-05	2.432e-08	2.112e-12
30	2.781e-06	2.064e-05	2.768e-12	1.307e-11
35	1.987e-07	2.001e-06	<2.2e-16	3.210e-13
40	1.098e-09	1.498e-08	<2.2e-16	4.329e-14
45	1.145e-09	1.241e-07	<2.2e-16	2.452e-13
50	1.044e-09	1.377e-08	<2.2e-16	3.145e-13

It was observed from table 8; that both non-parametric tests have stronger rejection values to wrong null hypothesis compared with the parametric counterpart from sample size of 5 to 50 which categorized them as the best with median test as the best. The performances of all tests increases as sample size increases.

5 Conclusion

This study revealed that the two parametric test are more efficient and perform better than nonparametric counterpart on data from normal distribution with the t-test as the best when variances of the two independent sample data are the same while Witch test is the best when variances are not equal, most especially at lower sample sizes. However as sample size increased from 30, the median test of non-parametric test compete with them. From the data generated from uniform and gamma distributions, Man-Whitney test is the best when variances are equal while median test is the best when variances are unequal at smaller sample sizes. As sample size increases the parametric the parametric tests seem to better than Mann-Whitney. It was also noted that the performance of the median test increases as sample size was increased based on the both criteria.

Furthermore, the Mann-Whitney has the best when data follows exponential distribution followed by median test. The t-test is the best between the two parametric tests and the power of all the tests improve as sample size increases. From sample size of 35 to 50, which can be classified as large sample sizes, median test is the best among the non-parametric tests.

6 Recommendation

The following recommendations are made from our findings;

i. Welch's t-test can be used at smaller sample sizes with equal variances while Median test for smaller sample sizes in normal distribution with unequal variance.
ii. Mann Whitney U test can be used for the skewed distributions when variances are equal.
iii. Median test can be used for all the distributions for unequal variance at smaller sample size under both criteria.
iv. This study also suggested for further study in considering other distributions like Cauchy, Weilbull, etc. and parametric and non-parametric statistics beyond two sample/populations

References

Abraham de Moivre, (1718). In Graltan-Guinness, I (ed), Landmark Writing in Western Mathematics 1640-1940, Amsterdam Elsevier pp. 105-120, ISBN 0-444-50871-6

Akeyede, I., Usman, M. and Chiawa, M. A. (2014).On Consistency and Limitation of paired t-test, Sign and Wilcoxon Sign Rank Test. IOSR Journal of Mathematics (IOSR-JM). Volume 10, Issue 1 Ver. IV. PP 01-06

Bartlett, M., S., (1935), The Effect of Non-Normality on the t Distribution", Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 31, No. 2, pp. 223 – 231

Blair, C., R., and Higgins, J., J., (2020). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distribution", Journal of Educational Statistics, Vol. 5, No. 4, pp. 309-335.

Blair, C., R., Higgins, J., J., and Smitley, W., D., S. (1980). On the relative power of the U and t tests, British Journal of Mathematical and Statistical Psychology, Vol. 33, pp. 114-130

Boneau, A., C., (2020). The effects of violations of assumptions underlying the t-test Psychological bulletin, Vol. 37, No. 1, pp. 49-64

Edith U. U. and Nkiru O. E. (2016). Comparison of Two Sample Tests Using Both Relative Efficiency and Power of Test. Open Journal of Statistics, 2016, 6, 331-345

Fisher, R., A., (1925). Applications of "students' distribution, Metron, Vol. 5, pp. 90-104

Friedman, M., (1937). The use of Rank to Avoid the Assumption of Normality Implicit in the Analysis of Variance. American Statistical Association 32, pp. 675-701

Hodges Jr, J., L., and E. L. Lehmann, E., L., (1956). The Efficiency of Some Nonparametric Competitors of the t-Test", The Annals of Mathematical Statistics, Vol. 27, No. 2, pp. 324-335.

Huber, P., J., (1964). Robust estimation of a location parameter, Annals of Mathematical Statistics, Vol. 35, pp. 73-101

Huber, P., J., (2019). Robust statistics: A review, Annals of Mathematical Statistics, Vol. 43, pp. 1041-1067

Neyman, J.; Pearson, E. S., (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika. 20(1/2); 175-240

Patel, J., K., and Read C., B. (1982). Handbook of the normal distribution', Marcel Dekker, INC

Srilakshminarayana, G. (2015). On Importance of Normality Assumption in Using a T-Test: One Sample and Two Sample Cases. Proceedings of the International Symposium on Emerging Trends in Social Science Research IS15 Chennai Symposium

Student (1908). The Probable error of mean, Biometrika, Vol. 6, No. 1, pp. 1-25.

Welch, B., L., (1938). The significance of the difference between two means when the population variances are unequal, Biometrika, Vol. 29, No. ¾, pp. 350-362.

Welch, B., L., (1945). The Theoretical Basis of Psychotherapy* Psychoanalysis, Behaviorism, and Gestalt psychology. American Journal of Orthopsychiatry. Vol. 15, Issue 2

Welch, B., L., (1947). The Generalization of `Student's' Problem when Several Different Population Variances are involved, Biometrika, Vol. 34, No. ½, pp. 28-35