Extensions of the Mann-Whitney U Test

Submitted in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science: Honours in Statistics

Jinlei Xu
Supervisor: Dr. Edward Susko
Dalhousie University
Nov 23th, 2023

Contents

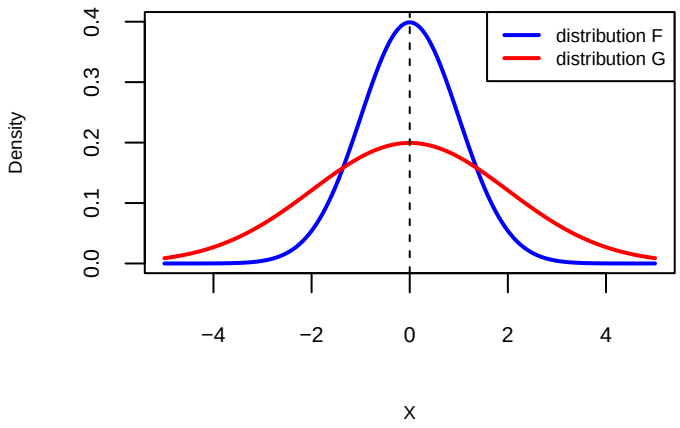
1 Introduction

Statistical hypothesis testing stands as an essential tool for deriving meaningful insights from data, particularly when comparing two populations. Among the array of statistical methods, the Mann-Whitney test(Lehmann & Abrera, 1975) has established itself as a valuable instrument for testing the equality of distributions between two populations. This paper embarks on an in-depth exploration of Mann-Whitney tests, delving into proportions, confidence intervals, simulation studies, and real-life applications. It also considers the use of the Mann-Whitney test statistics as an effect size and the construction of confidence interval in this case.

The Mann-Whitney test aim to comparing two populations, guided by the null hypothesis (H_0) positing the identical distribution of two populations. By define two population X and Y, the Mann-Whitney test statistic is the proportion of pairs of observation for which $y > x$. The statistics provides a estimation of the probability that population $Y > X$ for a randomly selected pairs of population each from X and Y which we denote as θ, $\theta = P[Y > X]$.

The Mann-Whitney statistics can also be considered as an effect size, the size of effect defined by proportion. Consider a hypothetical example where we have two samples about testing patients blood pressure. We define X as the treatment group that we give treatment to patient and Y as the control group that we do nothing. Then we calculated the Mann-Whitney statistics value as 0.73. The interpretation of this value implies that 73% of the time, the blood pressure treatment proves more effective than the control group which is a significant indicator of treatment efficacy. This stands in stark contrast to another hypothetical scenarios where Mann-Whitney statistics hover around 0.5, signifying a lack of substantial effectiveness. When presented with Mann-Whitney statistics like 0.73, it becomes crucial to acknowledge the infeasibility of definitively asserting the treatment's effectiveness. From a statistical standpoint, the conclusion is confined to supporting evidence for the rejection of the hypothesis that the treatment is not effective.

In traditional Mann-Whitney hypothesis, there is an assumption of identical distributions for two populations under the null hypothesis. That implies that we are assuming $\theta := P[Y > X] = \frac{1}{2}$ initially(Lehmann & Abrera, 1975). But generally, we compare between distributions. Define a distribution for X as F and distribution for Y as G, a $\theta = P[Y > X] = \frac{1}{2}$ does not directly implies the distribution of two population is the same in which we describe as $F = G$. Then, we are not expected that $F = G$ as used in derived Mann-Whitney test.



In light of these considerations, this paper endeavors to investigate extensions of the Mann-Whitney test methodology by take out to generate confidence intervals for the effect size(Newcombe, 2005). This extension seeks to offer a more nuanced and informative comprehension of the treatment's impact, recognizing the inherent uncertainties in statistical inference and delivering a comprehensive assessment of observed effects.

2 Traditional Mann-Whitney test

In a research context where the comparison of two groups is central, a foundational hypothesis is established, distinguishing the treatment group with a size of n and the control group with a size of m. The designation "Receive treatment: n" characterizes the group undergoing a specific treatment, while "Control group: m" identifies the group not receiving the treatment. The null hypothesis, denoted as H_0, posits that the ranking of observations remains uninfluenced by the treatment effect. In contrast, the alternative hypothesis, denoted as H_A, asserts that the ranking is indeed affected by the treatment.

We initially consider Mann-Whitney in case that there are no ties that would arise with count data, meaning that there are no identical values in both populations. However, when tied observations are present, the Mann-Whitney test adjusts its approach. Ties introduce the possibility of $x_i = y_j$, which can impact the calculation of the test statistic. We will delve into the implications of ties on the test statistic shortly. In scenarios without tied observations, our consideration is focused solely on instances where $y_j > x_i$ for the computation of the test statistics.

2.1 Wilcoxon Statistics

It turns out that the Mann-Whitney test is equivalent to the Wilcoxon rank-sum test as we will discuss in this section. Specifically, under the null hypothesis H_0, the order statistics is defined as $X_{(1)}, X_{(2)}, ..., X_{(n)}$, with $X_{(1)} < X_{(2)} < ... < X_{(n)}$ within the treatment group. Similarly, for the alternative hypothesis H_A, the order statistics are denoted as $Y_{(1)}, Y_{(2)} ..., Y_{(m)}$ within the control group, maintaining a consistent order structure $Y_{(1)} < Y_{(2)} < ... < Y_{(m)}$. Let the pooled values be $z_1 = x_1, ..., z_n = x_n, z_{n+1} = y_1, ..., z_{n+m} = y_m$. Thus, the order statistics for z would be $z_{(1)} < ... < z_{(n+m)}$. The rank R_j of $y_{(j)}$ as the

position of order value in $z_{(1)} < ... < z_{(n+m)}$ such that $R_j = i$ if and only if $y_j = z_{(i)}$.

In this scenario, if the treatment has no influence on the observed effect, the order statistics for the treatment and control groups, denoted by X and Y respectively, should exhibit exchangeability. Consequently, under the null hypothesis, the sums of ranks over all y_j should have a known expectation and variance as we will show. The Wilcoxon rank-sum test involves computing the statistic W_s, defined as the sum of the ranks assigned to the elements in set Y:

$$W_s = \sum_{j=1}^{m} R_j$$

(Lehmann & Abrera, 1975)

2.2 Mann-Whitney U Statistics

In the Mann-Whitney U test, the null hypothesis (H_0) states the distribution between two populations, denoted as F and G, have the relationship $H_0 : F = G$. This assertion is implies as $P[y > x] = \frac{1}{2}$. Conversely, the one-sided alternative hypothesis (H_A) often proposes that $P[y > x] > \frac{1}{2}$. Meanwhile, the two-sided alternative hypothesis (H_A) asserts that $P[y > x] \neq \frac{1}{2}$.

To quantify this hypothesis testing procedure, the Mann-Whitney test statistic (U) is employed, defined as the proportion when y_j is greater than x_i for all observations i and j. Which the count of y_j is greater than x_i denote as sum of indicator that $y > x$. The calculation involves iterating over all pairs of observations (n from population x and m from population y), and the resulting statistic is normalized by nm. Mathematically, this is expressed as:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{I\{y_j > x_i\}}{nm}$$

(Lehmann & Abrera, 1975)

Under the assumption of the null hypothesis (H_0), the expected value of U is equivalent to the proportion $P_{H_0}[y_j > x_i]$, which, as stipulated by H_0, is precisely $\frac{1}{2}$.

2.3 Comparison of statistics

The Wilcoxon rank-sum test is essentially another name for the Mann-Whitney U test. Both tests compare two independent groups and assess whether their distributions are equivalent or if one group tends to have higher ranks than the other.

Defining the Mann-Whitney null hypothesis $H_0 : F = G$, and the Wilcoxon rank-sum test with $y_1, y_2, ..., y_m$ and $x_1, x_2, ..., x_n$ denoted as x_i, y_j, respectively. We define the test statistic W_s as the sum of ranks of y_j, in $z_{(1)} < ... < z_{(n+m)}$, which is pooled into one distribution such that $y_j = z_{(i)}$. Thus, when the x_i, y_j values are pooled, the rank of y_j in the pooled sample is the number of y_i less than or equal to y_j and plus the number of x_i less

than y_j. The Wilcoxon test statistics related to the indicators used in the Mann-Whitney test statistics is:

$$R_j = \sum_{i=1}^{n} I\{y_j > x_i\} + \sum_{i=1}^{m} I\{y_j \geq y_i\}$$

Thus,

$$\sum_{j=1}^{m} R_j = \sum_{j=1}^{m}\sum_{i=1}^{n} I\{y_j > x_i\} + \sum_{j=1}^{m}\sum_{k=1}^{m} I\{y_j \geq y_k\}$$

For this equation, the rank is constructed by two parts: the counts of x_i smaller than y_j in the z-pooled distribution and the y_j itself, and the counts of y before y_j in this pooled distribution. For the first summation, it is just the count of $y_j > x_i$ for the pooled distribution, so it is $\sum_{j=1}^{m}\sum_{i=1}^{n} I\{y_j > x_i\}$.

Then, the second summation considers all pairs of (j,k) in this pooled sample, we can write it as $\sum_{j=1}^{m}\sum_{k=1}^{m} I\{y_{(j)} \geq y_{(k)}\}$. Assuming definition of $y_{(j)}$ through $y_{(1)} < ... < y_{(j)}$, we have $\sum_{k=1}^{m} I\{y_{(j)} \geq y_{(k)}\} = j$. Then, the second summation can be inferred as $\sum_{j=1}^{m}\sum_{k=1}^{m} I\{y_{(j)} \geq y_{(k)}\} = \sum_{j=1}^{m} j$. Based on the sum of natural sequences, the second summation can be $\frac{m(m+1)}{2}$. Thus, the statistics for the Whicoxon rank-sum test W_s can also be written as:

$$\sum_{j=1}^{m} R_j = \sum_{j=1}^{m}\sum_{i=1}^{n} I\{y_j > x_i\} + \frac{m(m+1)}{2}$$

.

Since the test statistics W_s for the sum of ranks of x_i, y_j is $W_s = \sum_{j=1}^{m} R_j$, then we can define the relationship between the statistics for the Mann-Whitney test U and the Wilcoxon rank-sum test W_s:

$$U = \frac{W_s - \frac{m(m+1)}{2}}{nm}$$

Thus, we can easily infer about the normal relationship between U and W_s, especially the mean relationship between U and W_s. Previously, under null hypothesis, we define that $E[I\{y_j > x_i\}] = P_{H_0}[y_j > x_i] = \frac{1}{2}$ given by Mann-Whitney U test. It means that the mean of Mann-Whitney U test is $\frac{1}{2}$. And we also given the relationship between statistics for the Mann-Whitney U test and the Wilcoxon rank-sum test, then we can easily refer the mean of Wilcoxon rank sum test under null hypothesis that $F = G$ as given

$$\begin{aligned} E(W_s) &= E(U) + E\left(\tfrac{m(m+1)}{2}\right) \\ &= \tfrac{1}{2} * nm + \tfrac{m(m+1)}{2} \\ &= \tfrac{m(n+m+1)}{2} \end{aligned}$$

3 Mann-Whitney Effect Size and CI

Given the assumption of identical distributions for two populations under the null hypothesis, our objective is to augment the Mann-Whitney test by extending it to produce confidence intervals for the effect size. In order to consider confidence interval, we first need to figure out the variance between Mann-Whitney and Wilcoxon test.

3.1 Variance

Within the context of the traditional Mann-Whitney test, first, we considered in cases that no tied observation of $x_i = y_j$ exist in this Mann-Whitney test. The effect size θ are determined as $\theta = P[y_j > x_i] = E[I\{y_j > x_i\}]$ because the expected value of an indicator random variable for an event is just the probability of that event. Then, the variance of traditional Mann-Whitney test statistics is defined as

$$\mathrm{Var}(U) = \frac{1}{n^2m^2}\mathrm{Cov}\left(\sum_i\sum_j I\{y_j \geq x_i\}, \sum_{i'}\sum_{j'} I\{y_{j'} \geq x_{i'}\}\right)$$

by the definition of covariance. Utilizing the bilinearity properties of covariance, we rearrange the position of covariance and the summations as

$$\mathrm{Var}(U) = \frac{1}{n^2m^2}\sum_{i,j}\sum_{i',j'}\mathrm{Cov}(I\{y_j \geq x_i\}, I\{y_{j'} \geq x_{i'}\}) \qquad \ldots(1)$$

We consider three cases for i, j, i', and j', excluding terms with $j \neq j'$ and $i \neq i'$ because they give a covariance contribution equal to $\emptyset$. The three cases are:

Case 1. When $i = i'$ and $j = j'$,

$$\begin{aligned}
\mathrm{Cov}(I\{y_j \geq x_i\}, I\{y'_j \geq x'_i\}) &= E[I\{y_j > x_i\}^2] - E[I\{y_j > x_i\}]^2 && \text{(covariance definition)}\\
&= E[I\{y_j > x_i\}^2] - \theta^2 && (\theta \text{ definition})\\
&= E[I\{y_j > x_i\}] - \theta^2 && \text{(natural of indicator variable)}\\
&= \theta(1-\theta)
\end{aligned}$$

Since there are m and n be the summation in (i', j'), (i, j), the contribution to (1) from these terms is

$$m \cdot n \cdot \theta(1-\theta)$$

Case 2. When $i = i'$ and $j \neq j'$,

$$\begin{aligned}
\mathrm{Cov}(I\{y_j \geq x_i\}, I\{y_{j'} \geq x_{i'}\}) &= E[I\{y_j > x_i\}I\{y_{j'} > x_i\}] - E[I\{y_j > x_i\}]E[I\{y_{j'} > x_i\}]\\
&= E[I\{y_j > x_i\}I\{y_{j'} > x_i\}] - \theta^2
\end{aligned}$$

Because the $j \neq j'$, we have $n-1$ i' for case 2. Since there are $n \cdot m(m-1)$ many covariance be the summation in (i', j'), (i, j) in this case, then contribution to (1) from these terms is

$$n \cdot m \cdot (m-1) \left(E[I\{y_j > x_i\} I\{y'_j > x_i\}] - \theta^2 \right)$$

Case 3. When $i \neq i'$ and $j = j'$, Similarly, the $i \neq i'$, so we have $n-1$ i' for case 3. The $\text{Cov}(I\{y_j \geq x_i\}, I\{y_{j'} \geq x_{i'}\})$ is equal to$E[I\{y_j > x_i\} I\{y_j > x_{i'}\}] - \theta^2$ and there are $m \cdot n(n-1)$ many covariance be the summation in (i', j'), (i, j). Then, the contribution to (1) for these terms is

$$m \cdot n \cdot (n-1) \left(E[I\{y_j > x_i\} I\{y_j > x_{i'}\}] - \theta^2 \right)$$

Thus, after analysis of each cases, we can have the variance of traditional Mann-Whitney test statistics be

$$\begin{aligned} Var(U) &= \tfrac{m \cdot n \cdot \theta(1-\theta)}{n^2 m^2} + \tfrac{n \cdot m \cdot (m-1)}{n^2 m^2} \left(E[I\{y_j > x_i\} I\{y_{j'} > x_i\}] - \theta^2 \right) \\ &\quad + \tfrac{m \cdot n \cdot (n-1)}{n^2 m^2} \left(E[I\{y_j > x_i\} I\{y_j > x_{i'}\}] - \theta^2 \right) \end{aligned} \quad \ldots (2)$$

3.2 Traditional Mann-Whitney Variance

We now consider variance calculations for the traditional Mann-Whitney test under the null hypotthesis condition $H_0 : F = G$. Under null hypothesis that $F = G$ we assume $\theta = P[Y_j > X_i] = \frac{1}{2}$ for case 1 (when $i = i'$ and $j = j'$) for null hypothesis

$$\begin{aligned} \theta(1-\theta) &= \tfrac{1}{2}(1 - \tfrac{1}{2}) \\ &= \tfrac{1}{4} \end{aligned}$$

For case 2 (when $i = i'$ and $j \neq j'$),

$$\begin{aligned} E[I\{Y_j > X_i\} I\{Y_{j'} > X_i\}] &= P[Y_j > X_i, Y_{j'} > X_i] \\ &= P[F(Y_j) > F(X_i), F(Y_{j'}) > F(X_i)] \end{aligned}$$

Let $U_j = F(Y_j), U_{j'} = F(Y_{j'})$, and $U_i = F(X_i)$. Given if $Y \sim F$ then $U = F(Y) \sim U(0,1)$ under H_0, both $X_i, Y_j, Y_{j'} \sim F$. Then $U_i, U_j, U_{j'} \sim U(0,1)$. So $P[Y_j > X_i, Y_{j'} > X_i] = P[min(Y_{j'}, Y_j) \geq X_i)]$ where $X_i, Y_j, Y_{j'}$ independent in $U(0,1)$.

$$P[Y_j > X_i, Y_{j'} > X_i] = P[Y_j > Y_{j'} > X_i] + P[Y_{j'} > Y_j > X_i]$$

we have two sub cases for case 2.

First, $Y_j > Y_{j'} > X_i$ for this integral. For the uniform distribution, $p(y_{j'}, x_i, y_j) = 1$ on the interval $[0, 1]$. Then,

$$\begin{aligned} P[Y_j > Y_{j'} > X_i] &= \int_0^1 \int_0^{y_j} \int_0^{y_{j'}} dx_i dy_{j'} dy_j \\ &= \int_0^1 \int_0^{y_j} y'_j dy_{j'} dy_j \\ &= \int_0^1 \frac{y_j^2}{2} dy_j \\ &= \frac{1}{6} \end{aligned}$$

By symmetry, $P[Y_{j'} > Y_j > X_i] = P[Y_j > Y_{j'} > X_i] = \frac{1}{6}$.

Thus, the covariance in case 2 $E[I\{Y_j > X_i\}I\{Y_{j'} > X_i\}] - \theta^2$ in null hypothesis with θ assume $\theta = \frac{1}{2}$ is $\frac{1}{6} + \frac{1}{6} - \frac{1}{2}^2 = \frac{1}{12}$ for case 2.

A similar argument applies in Case 3 and gives

$$E[I\{Y_j > X_i\}I\{Y_j > X_{i'}\}] - \theta^2 = 1/12$$

Substituting in (2), for $Var(U)$ in null hypothesis, we have

$$\begin{aligned} \hat{Var}(U) &= \frac{m \cdot n \cdot \theta(1-\theta)}{n^2 m^2} + \frac{n \cdot m \cdot (m-1)}{n^2 m^2} \left(E[I\{y_j > x_i\}I\{y_{j'} > x_i\}] - \theta^2\right) \\ &\quad + \frac{m \cdot n \cdot (n-1)}{n^2 m^2} \left(E[I\{y_j > x_i\}I\{y_j > x_{i'}\}] - \theta^2\right) \\ &= \frac{m \cdot n \cdot \frac{1}{4}}{n^2 m^2} + \frac{n \cdot m \cdot (m-1)}{n^2 m^2} \left(\frac{1}{12}\right) + \frac{m \cdot n \cdot (n-1)}{n^2 m^2} \left(\frac{1}{12}\right) \\ &= \frac{m+n+1}{12mn} \end{aligned}$$

3.3 Hanley–McNeil Approximation

Confidence limits for the effect size $\theta = P[y_j > x_i] = E[I\{y_j > x_i\}]$ are determined using the Hanley–McNeil Wald method(Hanley & McNeil, 1982). According to Hanley–McNeil Wald method, assuming the values $m^* = m - 1$ and $n^* = n - 1$ adjust for sample size constraints(Newcombe, 2005), we have variance

$$V_1 = \{\hat{\theta}(1-\hat{\theta}) + m^*(\hat{Q}_1 - \hat{\theta}^2) + n^*(\hat{Q}_2 - \hat{\theta}^2)\}/mn$$

(Newcombe, 2005).

In the case that we are interested in U as an effect size we can no longer assume that $F = G$ and the terms in (2) need to be approximated. The Hanley-McNeli Wald method approximate the Case 1 expression by $\hat{\theta}(1-\hat{\theta})$ where $\hat{\theta} = U$. The Case 2 expression is approximated by $\hat{Q}_1 - \hat{\theta}^2$ and the Case 3 expression by $\hat{Q}_2 - \hat{\theta}^2$. The $\hat{Q}_1$ and $\hat{Q}_2$ are expressed as following:

$$\begin{aligned} \hat{Q}_1 &= \sum_i \sum_j \sum_{j' \neq j} I\{y_j > x_i\}I\{y_{j'} > x_i\}/[m(m-1)n] \\ &= \sum_i \sum_j I\{y_j > x_i\}/mn \sum_{j' \neq j} I\{y_{j'} > x_i\}/(m-1) \end{aligned}$$

and for $\hat{Q}_2$,

$$\begin{aligned}\hat{Q}_2 &= \sum_j \sum_i \sum_{i' \neq i} I\{y_j > x_i\} I\{y_j > x_{i'}\}/[n(n-1)m] \\ &= \sum_j \sum_i I\{y_j > x_i\}/nm \sum_{i' \neq i} I\{y_j > x_{i'}\}/(n-1)\end{aligned}$$

Therefore, we have $\hat{Q}_1 \approx E[I\{y_j > x_i\} I\{y_{j'} > x_i\}]$ and $\hat{Q}_2 \approx E[I\{y_j > x_i\} I\{y_j > x_{i'}\}]$ when there is no ties. With these approximations we get

$$\begin{aligned}V_1 &= \{\hat{\theta}(1-\hat{\theta}) + (m-1)(\hat{Q}_1 - \hat{\theta}^2) + (n-1)(\hat{Q}_2 - \hat{\theta}^2)\}/mn \\ &= (\hat{\theta}(1-\hat{\theta}) + (m-1)(E[I\{y_j > x_i\} I\{y_{j'} > x_i\}] - \hat{\theta}^2) \\ &\quad + (n-1)(E[I\{y_j > x_i\} I\{y_j > x_{i'}\}] - \hat{\theta}^2))/mn\end{aligned}$$

Under the null hypothesis $F = G$, substituting $\theta = 1/2$ and the expressions for the covariances obtained above, we get the expression

$$\begin{aligned}V_1 &= \left(\tfrac{1}{4} + (m-1) \cdot \tfrac{1}{12} + (n-1) \cdot \tfrac{1}{12})\right)/mn \\ &= \tfrac{(n+m+1)}{12nm}\end{aligned}$$

Thus, $V_1 = \frac{(n+m+1)}{12nm} = Var(U)$. And through calculating confidence interval by normal approximation, we have interval with $\hat{\theta} \pm z_{\alpha/2}\sqrt{V_1}$. $z_{\alpha/2}$ has $(1-\alpha) \times 100\%$ confidence interval from a normal $N(0,1)$ (e.g., $z = 1.959964$ for a 95% confidence interval), ensures the desired level of confidence in the computed limits.

3.4 Duality

We are also talking about why we can turn this extension study from test to confidence interval. We are explore the concept of 1-1 correspondence, also known as duality, between statistical tests and confidence intervals. The framework establishes a seamless connection between hypothesis testing and interval estimation.

Consider a confidence interval construction procedure, denoted as $[L(x), U(x)]$, where $(1-a) \times 100\%$ is the confidence level. This procedure sets the stage for defining a corresponding statistical test. Defining a hypothesis test for $H_0 : \theta = \theta_0$, we set the indicator variable $\phi(x)$ indicate rejection of the null hypothesis as $\phi(x) = 1$. $\phi(x) = 1$ also means that if θ_0 is not within the interval $[L(x), U(x)]$. On the contrary, if θ_0 falls within the interval $[L(x), U(x)]$, $\phi(x) = 0$, signifying a failure to reject the null hypothesis.

The probability under the null hypothesis of rejecting H_0 is given by $P_{H_0}[\text{reject } H_0] = 1 - P_\theta[\theta \text{ inside interval [L(X), U(X)]}]$. Notably, this probability equates to the chosen significance level $(1-a)$, rendering the test an alpha-level test.

The defined test, with its corresponding confidence interval, embodies an alpha-level test, providing a direct link between the critical regions of the test and the confidence interval boundaries.

3.5 Ties

In the presence of ties the Mann-Whitney U statistic definition replaces $I\{y_j > x_i\}$ with $I\{Y_j > X_i\} + I\{Y_j = X_i\}]/2$. This alteration influences the practical computation of both the Mann-Whitney test statistic and its associated variance.

When list the presence of $I\{y_j \geq x_i\}$ within pooled two datasets refer by $y_1, y_2, \ldots, y_m$ and $x_1, x_2, \ldots, x_n$ denoted as x_i, y_j, it should adapt the calculation to accommodate tied values. This adjustment acknowledges that $P\{Y_j \geq X_i\} = 1 \cdot P\{Y_j > X_i\} + 1/2 \cdot P\{Y_j = X_i\}$(Lehmann & Abrera, 1975)., encapsulating both scenarios of y_j being greater than or equal to x_i. So the effect sizeθ from $\theta = P[Y > X]$ to $\theta = P[Y > X] + P[Y = X]/2$. Under null hypothesis $H_0 : F = G$, the effect size become with ties that $\theta = \frac{1}{2}$.

This corresponding adjustment not only reflects a method refinement in the context of tied observations but also ensures a more accurate representation of the relationships between y_j and x_i in the Mann-Whitney U test. Importantly, this approach proves to be more applicable in real-world data analyses where tied values may be prevalent.

4 Real Data Analysis

In this section, we illustrate the practical application of both the traditional Mann-Whitney test and the extended Mann-Whitney confidence interval using real-world data. Our analysis addresses scenarios with both tied and untied observations in the samples, considering two categorical variables that influence the distribution of the data.

4.1 Weight of Baby

The study focuses on the dependent variable, child birth weight, and its relationship with the categorical explanatory variable of parental smoking status (smoker or non-smoker). We formulate the hypothesis that child birth weight for non-smokers exceeds that of smokers (marked as 1), contrasting with the null hypothesis suggesting equal child birth weights for both groups (marked as 0).(Perktold, 2010)

Using the `wilcox.test()` method and calculating the Mann-Whitney confidence interval through $\hat{\theta} \pm z_{\alpha/2}\sqrt{V1}$, we obtain a Wilcoxon test p-value of almost zero, along with a Mann-Whitney confidence interval of $[0.64, 0.66]$ at a 95% confidence level. This leads to the rejection of the null hypothesis, suggesting a significant difference in the distribution of child birth weight between non-smokers and smokers. This decision is corroborated by the Wilcoxon test, where the p-value is below the 0.05 significance level, and the confidence interval does not encompass one half.

To validate these results, a t-test was performed, yielding a p-value almost zero. The convergence of results across different statistical tests fortifies the rejection of the null hypothesis, indicating a substantial difference in child birth weight distribution between non-smokers and smokers. Additionally, by comparing the mean child birth weights for both groups (a smoker has a mean baby weight of 3137.66 and a non-smoker has a mean baby weight

of 3412.91), it is apparent that non-smokers tend to have higher child birth weights than smokers, providing valuable practical advice.

Subsequently, we explore another categorical variable: whether the child is the first for the mother (marked as 1) or not (marked as 0).(Perktold, 2010) We assume for the null hypothesis that first-born and non-first-born children share the same distribution of child birth weight.

Again utilizing the *wilcox.test*() method and computing the Mann-Whitney confidence interval, we obtain a Wilcoxon test p-value of 1.145e-05 and a Mann-Whitney confidence interval of $[0.52, 0.55]$ at a 95% confidence level. This leads to the rejection of the null hypothesis that first-born and non-first-born children share the same distribution of child birth weight. Similar to the previous case, the decision is based on both the Wilcoxon test with significant value and the non-overlapping confidence interval with one half.

To validate these results, a t-test was conducted, resulting in a p-value of still almost zero. Although it indicated a small difference in child birth weight distribution between first-born and non-first-born children, further examines the mean child birth weights with mean of first-born baby weight 3329.60 and mean of not-first-born baby weight 3386.68. It reveals a slight advantage for non-first-born children. This insight offers practical advice, suggesting that non-first-born children tend to have higher birth weights compared to their first-born counterparts.

4.2 Tonsil Size

The Mann-Whitney confidence interval and Mann-Whitney test can be applied even in scenarios with tied data values. In this context, we examine data from Table 1 provided by Holmes and Williams (1954), where 1398 children aged 0-15 years are categorized based on their relative tonsil size. The table presents information on tonsil size among carriers and non-carriers of Streptococcus pyogenes.(McCullagh, 1980)

	Present but not enlarged	Enlarged	Greatly enlarged	Total
Carriers	19	29	24	72
Non-carriers	497	560	269	1326
Total	516	589	293	1398

Table 1: Tonsil size of carriers and non-carriers of Streptococcus pyogenes

After applying the *wilcox.test*() method and calculating the Mann-Whitney confidence interval, we obtained a Wilcoxon test p-value of 0.01 and a Mann-Whitney confidence interval of $[0.61, 0.63]$ at a 95% confidence level. These results lead to the rejection of the null hypothesis, indicating that carriers and non-carriers have different tonsil size distributions, and this dissimilarity may influence tonsil enlargement.

To validate these findings, a t-test was conducted, resulting in a p-value of 0.01. Although it suggests a slight difference in the distribution of present and enlarged tonsil size between carriers and non-carriers, a closer examination of the means reveals an advantage for carriers

(mean for carrier tonsil size: 1.07; mean for non-carrier tonsil size: 0.83). This insight offers practical advice, suggesting that carriers of Streptococcus pyogenes tend to have higher tonsil size compared to non-carriers.

4.3 Vision Quality

Exploring the potential impact of vision quality on individuals, we categorize the subjects into two groups: women (marked as 1) and men (marked as 0). Our null hypothesis assumes that the level of vision quality effect is the same for both women and men. Vision quality is further categorized into four levels: 1. Highest, 2. Relative high, 3. Relatively low, and 4. Lowest. The data table(Table 2) are shown below(McCullagh, 1980):

		Vision Quality			
		Highest	2	3	Lowest
Men	Total	1053	782	893	3242
	Women	1976	2256	2456	789
Total		3029	3038	3349	7477

Table 2: Quality of right eye vision in men and women

Utilizing the Wilcoxon rank-sum test to compare vision quality between women and men, we obtain a p-value of 0.3347. The test results do not provide sufficient evidence to reject the null hypothesis, indicating that there is no significant difference in the true location between women and men concerning vision quality.

Subsequently, a Two Sample t-test is conducted on the same data, resulting in a t-statistic of 0.341 and a p-value of 0.733. These results do not support the alternative hypothesis, suggesting no true difference in means between women and men. The 95% Mann-Whitney confidence interval is from 0.50 to 0.51, indicating that the true difference likely falls within this range. Although the sample estimates show a slightly lower mean vision quality for men 2.27 compared to women 2.28, the Wilcoxon test, t-test, and Mann-Whitney confidence interval collectively suggest that vision quality does not significantly differ between genders.

5 Simulation Study

The Mann-Whitney confidence interval and test are fundamental tools in non-parametric statistics, particularly for comparing two independent samples. To bolster the credibility of these techniques, a comprehensive simulation study is undertaken. This study is motivated by the need to assess the performance of the Mann-Whitney methods across diverse scenarios and under various assumptions.

5.1 Two standard Normal Population with Common Mean and Variance

In this section, we conduct a simulation study to compare the Type I error and power of the Mann-Whitney test, t-test and Mann-Whitney confidence interval in two same normal

population. We are expecting to see a type I error close to significant level and a power that is relatively small.

First compare Type I error under the null hypothesis (H_0 true: $\theta = 0$). The simulation involves two populations with the same standard normal distribution, characterized by a mean of 0 and a standard deviation of 1 of simulated normal distribution.

The simulation is conducted under the assumption that the null hypothesis is true ($\theta = 0$). Three statistical tests, the Mann-Whitney test, the t-test and the confidence interval, are employed to assess the Type I error is the percentage of rejecting in null hypothesis. The simulation is repeated 1000 times for sample sizes of 50 and 100, providing a robust evaluation of the tests' performance.

	Sample Size	
Test Type	$n = 50$	$n = 100$
Mann-Whitney	0.061	0.049
t-test	0.063	0.055
Mann-Whitney CI	0.071	0.062

Table 3: Type I error for Mann-Whitney test, t-test, and Mann-Whitney CI for Common Mean and Variance

From the results (Table 3), it is evident that under the null hypothesis ($\theta = 0$), All of the Mann-Whitney test, the t-test and the Mann-Whitney CI exhibit similar proportions of rejecting the null hypothesis. And as the simulated sample size become bigger, the error between three proportions are lower.

The simulation study provides valuable insight into the behavior of the Mann-Whitney test and t-test under the null hypothesis that they are all close to the significant level 0.05. The similarity in rejection proportions underscores the robustness of these tests when applied to populations with identical distributions and supports the notion that, under these conditions, all tests perform similarly in terms of type I error.

Then, we extend the simulation study to compare the power of the Mann-Whitney test, the t-test and Mann-Whiney CI under the alternative hypothesis. The simulation involves two populations with the same standard deviation (1) but different means. The null hypothesis (H_0 true: $\theta = 0$) is compared against the alternative hypothesis (H_a: $\theta = 0.2$) in the simulated normal distributions.

The simulation is conducted under the assumption of different means in the alternative hypothesis scenario. The first simulated group has a mean of 0, while the second simulated group has a mean of 0.2. The simulation is repeated 1000 times for sample sizes of 50 and 100.

The output of the simulation, detailing the proportion of times each test rejects the null hypothesis across the 1000 simulations, is presented below:

From Table 4, the power—indicating the proportion of instances where each test correctly rejects the null hypothesis—shows variability, ranging from 0.16 to 0.29. All three tests

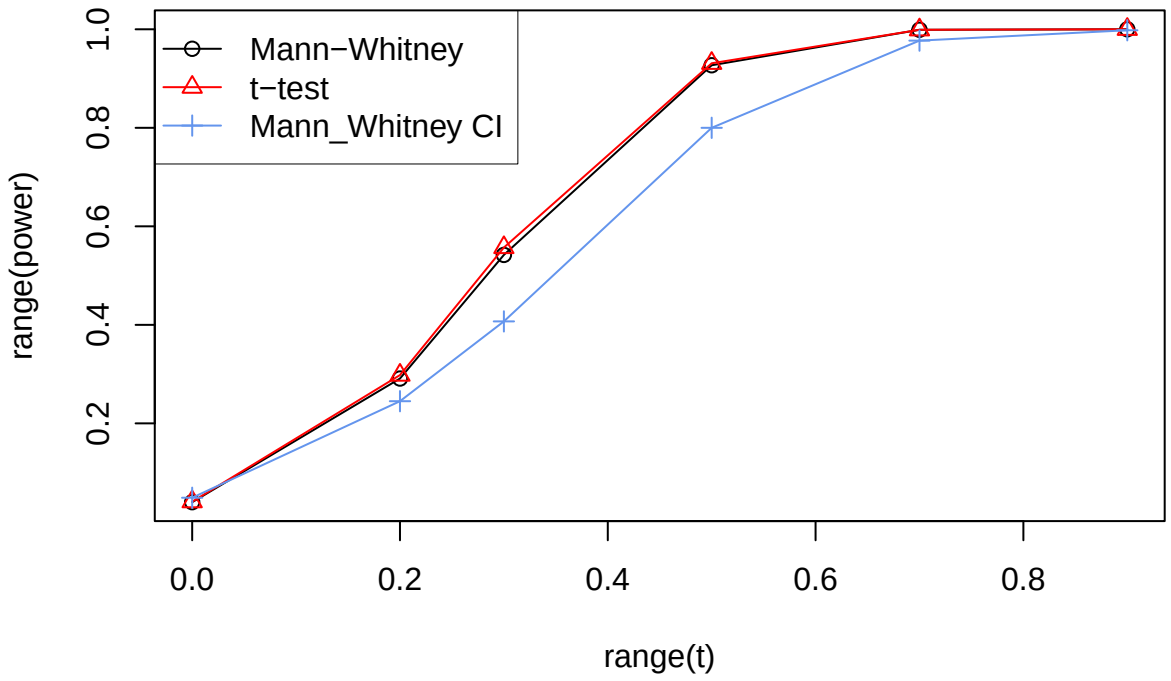
Test Type	Sample Size	
	$n = 50$	$n = 100$
Mann-Whitney	0.167	0.288
t-test	0.18	0.291
Mann-Whitney CI	0.163	0.214

Table 4: 0.2 Power for Mann-Whitney test, t-test, and Mann-Whitney confidence interval for Common Mean and Variance

perform similarly in terms of power, each surpassing the significance level of 0.05. Notably, the traditional Mann-Whitney test exhibits a relatively higher difference in power between sample sizes 50 and 100. In contrast, the Mann-Whitney CI demonstrates a relatively stable and lower difference in power between sample sizes 50 and 100.

This pattern suggests that, while all tests demonstrate sufficient power to reject the null hypothesis, the traditional Mann-Whitney test's sensitivity to changes in sample size merits consideration in experimental design. The consistent and stable performance of the Mann-Whitney CI across varying sample sizes indicates its robustness and dependable ability to detect differences. Consequently, the Mann-Whitney CI appears particularly well-suited for this scenario.

To deepen our exploration of the power associated with the Mann-Whitney test, t-test, and Mann-Whitney confidence interval, refer to the graph below. In this analysis, we keep the first normal distribution with a mean of 0 and a standard deviation of 1 constant. The mean of the second distribution varies across values of 0, 0.2, 0.3, 0.5, 0.7, and 0.9, while the standard deviation remains constant at 1. It is noteworthy that when the mean is set to 0, the power of the three tests aligns with the Type I error.



As the mean of the second distribution deviates from 0, specifically at 0.2, 0.3, 0.5, 0.7, and 0.9, both the traditional Mann-Whitney U test and the t-test exhibit nearly identical behaviors across these mean values. Simultaneously, the Mann-Whitney confidence interval tends to have a relatively lower value of power. This observation suggests that, under these conditions, the Mann-Whitney U test and the t-test provide comparable outcomes, while the Mann-Whitney confidence interval tends to be more conservative in terms of power.

In summary, all three tests maintain an expected Type I error that closely aligns with the chosen significance level of 0.05. In terms of power, both the traditional Mann-Whitney test and the t-test demonstrate comparable outcomes, while the Mann-Whitney confidence interval falls slightly short of achieving the anticipated power under the alternative hypothesis. Importantly, the Mann-Whitney CI exhibits a slightly higher rate of Type II errors compared to the other two tests. This suggests that in scenarios involving an alternative hypothesis, the traditional Mann-Whitney test and the t-test may provide more reliable and comparable results, while the Mann-Whitney CI is somewhat more susceptible to Type II errors.

5.2 Two standard Normal Population with Common Mean but Different Variance

In this section, we delve into a scenario where we have two standard normal populations with a common mean but different variances. Our focus is on conducting a simulation study to explore the type I error and power of the Mann-Whitney U test, t-test, and Mann-Whitney confidence interval under such circumstances. We anticipate observing variations in the outcomes of the three tests due to the differences in variance.

First, we start our exploration by examining the type I error rates for the Mann-Whitney test, t-test, and Mann-Whitney confidence interval under the condition of different variances within the normal distribution. Specifically, we generate simulated data with group 1 and group 2 having the same mean (0) but different variances (1 for group 1 and 25 for group 2). Our simulation involves 1000 iterated times for these data, and we calculate the proportion of times the null hypothesis is rejected when it is type I error with mean is the same.

	Sample Size	
Test Type	$n = 50$	$n = 100$
Mann-Whitney	0.063	0.083
t-test	0.039	0.052
Mann-Whitney CI	0.053	0.058

Table 5: Type I error for Mann-Whitney test, t-test, and Mann-Whitney CI for Different Variance

From the results (Table 5) we observed, this pattern continue emerges: as the sample size increases, the proportion of a type I error tends to go up. Among the three tests we looked at, Mann-Whitney CI consistently shows the most consistent results across different sample sizes. On the other hand, the traditional Mann-Whitney test is noticeably affected by changes in sample size.

Our initial expectation was that the type I error rate would closely match our chosen significance level of 0.05, which corresponds to a 95% confidence interval. In this context, both the t-test and Mann-Whitney CI performed well, producing results that align closely with the expected significance level. However, the traditional Mann-Whitney test produced type I error rates (0.063 and 0.083) that surpassed the anticipated 0.05 level.

This deviation from the expected significance level supports our hypothesis that the traditional Mann-Whitney test is significantly influenced by variations in the distribution's variance. This influence hinders its ability to reliably show that $F = G$ when the means are equal but the variances different. In contrast, the consistent performance of Mann-Whitney CI suggests its resilience to such variations, making it a more dependable choice in situations involving differing variances.

Next, we turn our attention to power analysis. We explore the ability of the Mann-Whitney test, t-test, and Mann-Whitney confidence interval to accurately reject the null hypothesis in the presence of standard normal data exhibiting distinct means and variances (group 1 mean = 0, variance = 1; group 2 mean = 0.5, variance = 25). This analysis involves conducting 1000 simulations, same as the approach employed previously.

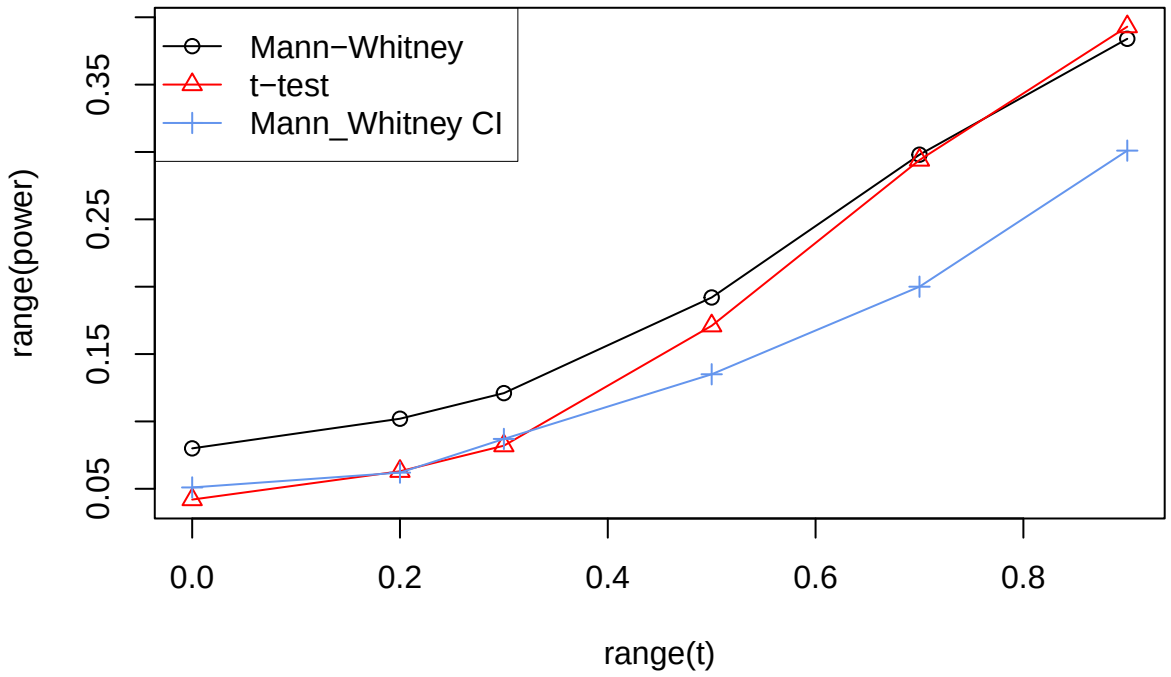
	Sample Size	
Test Type	$n = 50$	$n = 100$
Mann-Whitney	0.121	0.192
t-test	0.082	0.179
Mann-Whitney CI	0.082	0.144

Table 6: 0.5 Power for Mann-Whitney test, t-test, and Mann-Whitney CI for Different Variance

From the observed data (Table 6), the trend of an increasing proportion of power with larger sample sizes persists. Among the three tests examined, Mann-Whitney CI continues to demonstrate the most consistent results across different sample sizes, albeit with a more pronounced increase of around 0.06. In contrast, the t-test is notably influenced by sample size, with a difference of close to 0.1 affected by the changing size.

We expected that the power rate might deviate from our chosen significance level of 0.05, corresponding to a 95% confidence interval. In this context, all three tests produce results that deviate from the expected significance level, aligning with our hypothesis. This departure supports our hypothesis that the proportion of rejecting the null hypothesis for all three tests is significantly influenced by variations in the distribution's variance.

To explore deeper into the power of the Mann-Whitney test, t-test, and Mann-Whitney confidence interval, we aim to visualize these relationships through a graph. In this analysis, we keep the first normal distribution with a mean of 0 and a standard deviation of 1 constant. However, we vary the mean of the second distribution, exploring values of 0, 0.2, 0.3, 0.5, 0.7, and 0.9, while maintaining a standard deviation of 5. Notably, when the mean is 0, the power of the three tests aligns with the Type I error.



We maintain the characteristics of the first normal distribution with a mean of 0 and a standard deviation of 1. The second distribution's mean varies from 0 to 0.2, 0.3, 0.5, 0.7, and 0.9, while the standard deviation changes to 5. Specifically, when the mean is 0, the Type I error rates for the t-test and Mann-Whitney confidence interval are comparable, both close to 0.05. In contrast, the traditional Mann-Whitney U test exhibits a Type I error rate closer to 0.09 under these conditions. This discrepancy suggests that, in these specific conditions, the traditional Mann-Whitney U test may be more likely to falsely rejecting the null hypothesis.

As the mean of the second distribution deviates from 0, distinct shifts in power dynamics among the Mann-Whitney U test, t-test, and Mann-Whitney confidence interval become evident. Specifically, at mean values of 0.2 and 0.3, both the Mann-Whitney confidence interval and the t-test demonstrate nearly identical power. As the mean of the second distribution increases to 0.7 or 0.9, the traditional Mann-Whitney test and the t-test start to exhibit identical power.

For a mean of 0.5 in the second distribution, the power hierarchy among the three distributions, from largest to lowest, is Mann-Whitney test, t-test, and Mann-Whitney confidence interval. This observation suggests that the Mann-Whitney test is more sensitive to shifts in the mean, especially when it increases. The t-test also responds to mean shifts and appears to have more reliable power as the mean increases. Conversely, the Mann-Whitney confidence interval exhibits less sensitivity to changes in the mean, particularly in the context of varying standard deviations.

In summary, both the Mann-Whitney confidence interval and t-test maintain an expected Type I error that closely aligns with the chosen significance level of 0.05. However, the traditional Mann-Whitney confidence interval deviates from the expected Type I error rate due to differences in variance, implying that the t-test and Mann-Whitney CI are more re-

liable for rejecting the null hypothesis. Concerning power, all three tests are influenced by changes in variance, resulting in a higher rate of Type II errors compared to the previous graph. Specifically, both the traditional Mann-Whitney test and the t-test demonstrate relatively comparable outcomes, while the Mann-Whitney confidence interval exhibits slightly lower power under the alternative hypothesis. This suggests that in scenarios involving an alternative hypothesis, the traditional Mann-Whitney and the t-test continue to provide more reliable and comparable results than the Mann-Whitney confidence interval. However, compared to the previous graph, it is evident that all three tests are less reliable in reject an false null hypothesis when there are variations in variance.

5.3 Double Exponential Distribution

In this section, we simulate the double exponential distribution (Laplace distribution), which is defined by location and scale parameters. The scale parameter is set to a default value of 1 to avoid heavy tails, while the location parameter (θ) is systematically varied to examine its effects on both Type I error and power during simulation. The double exponential distribution are form as the following probability density function $p(x;\theta) = \frac{1}{2}e^{-|x-\theta|}$. Despite its symmetry resembling the normal distribution, this section delves into the deviations observed in the location parameter, particularly concerning the t-test. The null hypothesis is defined as $F = G$, indicating equality between two populations, while the alternative hypothesis posits the presence of two distinct double exponential distributions, $F \neq G$.

First, the provided Table 7 presents Type I error rates same null hypothesis tests applied to two similar populations with a double exponential distribution. The simulations are conducted for sample sizes of 50 and 100, both with a location parameter of 0 and a scale parameter of 1. The simulations are repeated 1000 times to detect the proportion of Type I error in this double exponential scenario.

	Sample Size	
Test Type	$n = 50$	$n = 100$
Mann-Whitney	0.052	0.035
t-test	0.044	0.033
Mann-Whitney CI	0.076	0.043

Table 7: Type I error for Mann-Whitney test, t-test, and Mann-Whitney CI double exponential

In the obtained results, the Type I error rates for the Mann-Whitney test, t-test, and Mann-Whitney confidence interval (CI) all hover around the anticipated significance level of 0.05. Unexpectedly, the t-test demonstrates effectiveness despite the simulation's initial assumptions. Both the t-test and Mann-Whitney U test exhibit comparable Type I error rates, contradicting the predicted behavior.

The unexpected efficacy of the t-test prompts a more in-depth examination of the simulation assumptions and their interaction with the location parameter of the double exponential distribution. This unanticipated outcome underscores the need for further exploration into the

location parameter influencing the t-test's performance in the context of double exponential distribution.

Additionally, the Mann-Whitney CI, particularly for smaller sample sizes, exhibits a slightly elevated Type I error rate of 0.076. This finding suggests a sensitivity of the Mann-Whitney CI to sample size variations. Conversely, the Mann-Whitney CI appears to be more reliable for larger sample sizes. This observation underscores the importance of considering sample size implications when utilizing Mann-Whitney confidence intervals in the context of a double exponential distribution.

We investigate the power of different statistical tests using the data presented in Table 8, which focuses on two populations with a double exponential distribution. The simulations involve sample sizes of 50 and 100, with a location parameter of 0 for the first population and a location parameter of 0.3 for the second population. Both populations have a scale parameter set to 1. The simulations are iterated 1000 times to assess the power of these tests in the context of a double exponential scenario as the proportion of the test reject the null hypothesis.

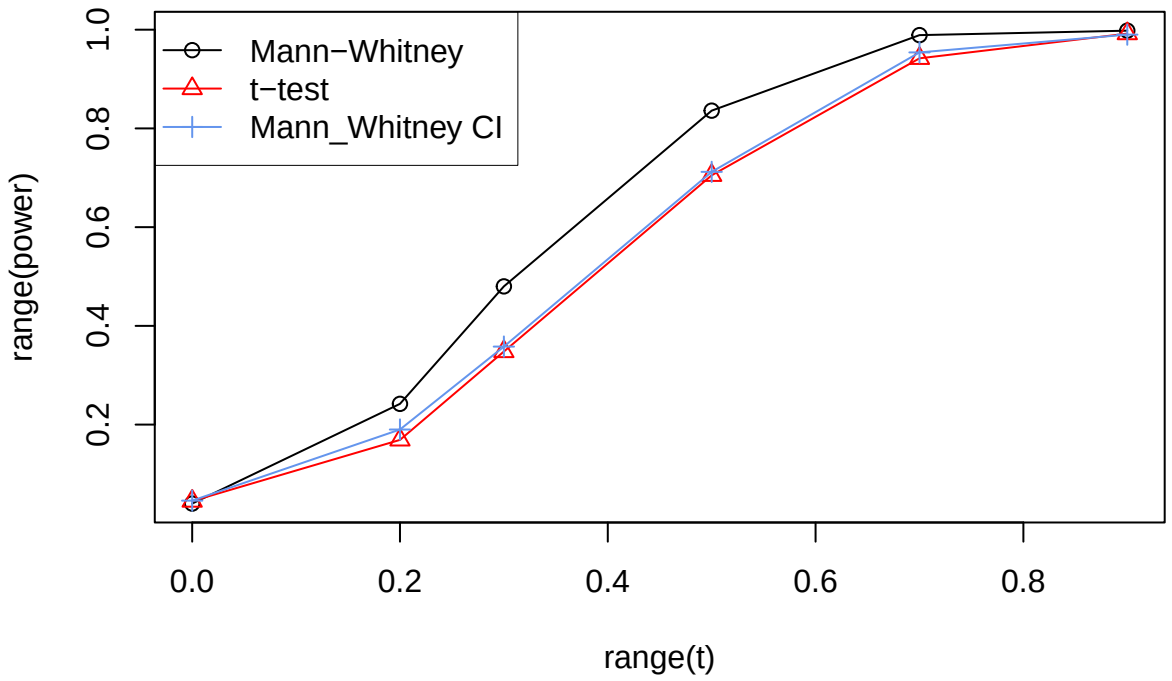
	Sample Size	
Test Type	$n = 50$	$n = 100$
Mann-Whitney	0.229	0.464
t-test	0.16	0.345
Mann-Whitney CI	0.196	0.373

Table 8: 0.3 Power for Mann-Whitney test, t-test, and Mann-Whitney CI double exponential

The table data is quite clear: when we compare it to a similar table simulating two normal distributions, the Mann-Whitney test, t-test, and Mann-Whitney CI show a stronger ability to detect differences. Notably, the Mann-Whitney test perform with a maximum power of 0.46 among three test, indicating its effectiveness in identifying true effects when the distribution has a location parameter of 0.3.

However, the t-test behaves unexpectedly in this scenario, displaying a power of 0.16 for a sample size of 50 and 0.35 for a sample size of 100—numbers lower than anticipated, especially when contrasted with the other two tests. To delve deeper into understanding why the t-test is less effective with a double exponential distribution, we will conduct alternative test simulations and visualize the results through graphical representations.

In our continued exploration, we delve deeper into the power of the Mann-Whitney test, t-test, and Mann-Whitney confidence interval using graphical analysis. We maintain the setting of two double exponential distributions, where the first distribution has a constant scale parameter of 1 and a fixed location parameter of 0. However, for the second distribution, we systematically vary the location parameter, examining values of 0, 0.2, 0.3, 0.5, 0.7, and 0.9, while keeping the scale parameter constant at 1 and repeated simulate 1000 times as usual. When the location parameter is set to 0, the power of the three tests aligns with the Type I error for the null hypothesis.



From the graph listed above, we can clearly seen that the type I error for Mann-Whitney test, t-test, Mann-Whitney confidence interval is close to 0 or at least around 0.05 significant level.

Contrary to expectations, yet aligning with our earlier table analysis, it is evident that the Mann-Whitney test, t-test, and Mann-Whitney CI demonstrate higher power values when compared to their counterparts in the simulation of two normal distributions. Specifically, the Mann-Whitney test consistently exhibits greater power across various alternative tests with location parameters set at 0.2, 0.3, 0.5, 0.7, and 0.9. The enhanced power levels, especially in the Mann-Whitney test, suggest its robust ability to detect true effects and its overall effectiveness when compared to the other two tests in the context of these two double exponential distributions.

While the t-test's power is challenging to distinguish from that of the Mann-Whitney confidence interval, it appears to be less efficient among the three tests. However, this doesn't imply that the t-test is ineffective in this scenario. In fact, the graphical representation proves to be more efficient compared to scenarios involving normal distributions with varying variances (population 1 variance is 1, and population 2 variance is 25). In summary, both the Mann-Whitney confidence interval and t-test demonstrate relatively effective power in the context of double exponential distributions. However, when considering different variances for normal populations, these two tests exhibit reduced efficiency, suggesting that the t-test may struggle when the scale parameter of the double exponential distribution varies. Consequently, in this scenario, the Mann-Whitney test emerges as the most efficient option.

References

(1): Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

(2): Lehmann, E. L. & Abrera, H. J. M. (1975). Nonparametrics: Statistical Methods Based On Ranks. Holden-Day, Inc.

(3): McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

(4): Newcombe, R. G. (2005). Confidence intervals for an effect size measure based on the Mann–whitney statistic. part 2: Asymptotic methods and evaluation. Statistics in Medicine, 25(4), 559–573. https://doi.org/10.1002/sim.2324

(5): Perktold, J. (2010). Treatment effects under conditional independence. Treatment effects under conditional independence - statsmodels 0.15.0 (+132). https://www.statsmodels.org/dev/examples/notebooks/generated/treatment_effect.html#Create-TreatmentEffect-instance-and-compute-ipw