TITLE:

STUDYING THE ASYMPTOTIC DISTRIBUTION AND CONSISTENCY OF THE MANN-WHITNEY U STATISTIC

Submitted by: Anupriya Halder

Registration No.: A01-2112-0786-17

Roll No.: 0475

Supervisor: Prof. Pallabi Ghosh

Declaration: I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Anupriya Halder

(Signature)

ABSTRACT

The Mann-Whitney U test (Mann and Whitney, 1947) is based on the idea that the particular pattern exhibited when the X and Y random variables are arranged together in increasing order of magnitude provides information about the relationship between their populations. A sample pattern of arrangement where most of the Y's are greater than most of the X's, or vice versa, or both, would be evidence against a random mixing and thus tend to discredit the null hypothesis (H_0)[1] of identical distribution. The Mann-Whitney U test statistic is defined as the number of times a Y precedes an X in the combined ordered arrangement of the two independent random samples drawn from continuous distributions.

The purpose of this paper is to study the asymptotic distribution and investigate the consistency of the Mann-Whitney U statistic. To examine the asymptotic distribution of U statistic (i.e., U test), it is inspected whether the asymptotic distribution of the Mann-Whitney U Statistic converges to normality when samples are drawn from any continuous population not known beforehand (say, Exponential). To examine the consistency of U statistic, it is then verified whether the estimated probability of the value of Mann-Whitney U Statistic exceeding a cut off value U_α (where U_α is ∋ P_{H_o} [U ≥ U_α] = α, α being the size[4] of the test) converges to α when the samples are drawn from the same continuous population. If the two samples are drawn from populations with different location parameters, it is observed whether the above estimated probability (which is basically the Power[5]) goes to 1 as we increase the difference between the location parameters of the two populations and hence it is verified that the test becomes more and more consistent[6].

TABLE OF CONTENTS

Page Number

❖ INTRODUCTION TO MANN-WHITNEY U TEST:

Under the assumption that the two groups X and Y are independently distributed with DFs F_x and F_y both continuous but not necessarily known, Gurevich (2009), the Mann-Whitney U statistic is broadly used to interpret whether there are differences in the distribution of the two groups or differences in their locations. The Mann-Whitney test was developed as a test of stochastic equality, Mann (1947).

In statistics, the Mann-Whitney U test is a non-parametric test[17] of the null hypothesis that it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second population, that is equivalent to another non-parametric test, the Wilcoxon rank-sum test though the test statistics are different. For the Wilcoxon rank-sum test, we assume that there are two continuous distributions (one of X and other of Y) which differ only in their locations (medians). The distribution of X has median θ_1 and that of Y median θ_2. We set $\Delta= \theta_1 - \theta_2$ and this, known as the shift, is the parameter of interest.

This test can be used to investigate whether two *independent* samples were selected from populations having the same distribution. We assume that the two samples are drawn from continuous distributions, so that the possibility $X_i = Y_j$ for some i and j need not be considered. A similar nonparametric test used on *dependent* samples is the Wilcoxon Signed Rank Test[7]. Although Mann and Whitney developed the Mann–Whitney *U* test under the assumption of continuous responses with the alternative hypothesis being that one distribution is stochastically greater than the other, there are many other ways to formulate the null and alternative hypotheses[2] such that the Mann–Whitney *U* test will give a valid test.

A very general formulation is to assume that:

- All the observations from both groups are independent of each other,
- The responses are ordinal (i.e., one can at least say, of any two observations, which is the greater),
- Under the null hypothesis H_0, the distributions of both populations are equal.
- The alternative hypothesis H_1 is that the distributions are not equal.

Under more strict assumptions than the general formulation above, e.g., if the responses are assumed to be continuous and the alternative is restricted to a shift in location, i.e., $F_1(x) = F_2(x + \delta)$, we can interpret a significant Mann–Whitney U test as showing a difference in medians or means. Under this location shift assumption, we can also interpret the Mann–Whitney U test as assessing whether the Hodges–Lehman estimate[8] of the difference in central tendency between the two populations differs from zero. The Hodges–Lehmann estimate for this two-sample problem is the median of all possible differences between an observation in the first sample and an observation in the second sample.

- The Mann–Whitney U test / Wilcoxon rank-sum test is not the same as the Wilcoxon *signed*-rank test, although both are nonparametric and involve summation of ranks. The Mann–Whitney U test is applied to independent samples. The Wilcoxon signed-rank test is applied to matched or dependent samples.
- The asymptotic distribution of Mann-Whitney remains of interest and Ferge (1994) suggested a number of tests and evaluated the asymptotic significance

level, and compared the asymptotic power for some of these tests. When normality holds, Mann-Whitney test has an asymptotic efficiency[9] of about $3/\pi$ when compared to t-test[10], Lehamn(1999).

In reporting the results of a Mann-Whitney U test, it is important to state:

- A measure of the central tendencies of the two groups (means or medians; since the Mann–Whitney *U* test is an ordinal test, medians are usually recommended).
- The value of *U*
- The sample sizes
- The significance level.

Only independence and continuous distributions need to be assumed to test the null hypothesis of identical populations. The test is simple to use for any size samples, and tables of the exact null distribution are widely available. The large-sample approximation is quite adequate for most practical purposes, and corrections for ties can be incorporated in the test statistic. The test has been found to perform particularly well as a test for equal means (or medians), since it is especially sensitive to differences in location. In order to reduce the generality of the null hypothesis in this way, however, we must feel that we can legitimately assume that the populations are identical except possibly for their locations. A particular advantage of the test procedure in this case is that it can be adapted to confidence interval estimation of the difference in location.

❑ History:

In a single paper in 1945, Frank -Wilcoxon proposed both the one-sample signed rank and the two-sample rank sum test, in a test of significance with a point null-hypothesis against its complementary alternative (that is, equal versus not equal). However, he only tabulated a few points for the equal-sample size case in that paper (though in a later paper he gave larger tables). A thorough analysis of the statistic, which included a recurrence allowing the computation of tail probabilities for arbitrary sample sizes and tables for sample sizes of eight or less appeared in the article by Henry Mann and his student Donald Ransom Whitney in 1947. This article discussed alternative hypotheses, including a stochastic ordering (where the cumulative distribution functions satisfied the point-wise inequality $F_X(t) < F_Y(t)$). This paper also computed the first four moments and established the limiting normality of the statistic under the null hypothesis, that it is asymptotically distribution-free.

- **Related test statistics**:

Kendall's tau: The Mann–Whitney *U* test is related to a number of other non-parametric statistical procedures. For example, it is equivalent to Kendall's tau correlation coefficient[12] if one of the variables is binary (that is, it can only take two values).

ρ statistic*:* A statistic called ρ that is linearly related to *U* and widely used in studies of categorization (discrimination learning involving concepts), and elsewhere is calculated by dividing *U* by its maximum value for the given sample sizes, which is simply $n_1 \times n_2$. ρ is thus a non-parametric measure of the overlap between two distributions; it can take values between 0 and 1 .

❖ **Test Procedure:**

Let X and Y be two independent RVs having continuous distributions given by the DFs F_x and F_y respectively. Here, we assume that the two distributions may differ only with respect to their locations, i.e., we may have $\mathbf{F_y(x) = F_x(x-\theta)}$ for all x and some $\theta \neq 0$. Under this assumption, to test

$$\mathbf{H_0: F_y(x) = F_x(x)} \ \forall \ x \text{ against } \mathbf{H_1: Not\ H_0}$$

is equivalent to test $\mathbf{H_0 : \theta = 0}$ against $\mathbf{H_1 : Not\ H_0}$.

Let $\mathbf{X_1, X_2, ..., X_{n_1}}$ and $\mathbf{Y_1, Y_2, ..., Y_{n_2}}$ be two random samples drawn independently from the distributions of X and Y respectively. We combine the two sets of observations and arrange them in ascending order of magnitude. We then assign ranks as 1 to $n(=n_1+n_2)$ from the smallest to the largest , where there are groups of tied values, assign a rank equal to the midpoint of unadjusted rankings(e.g., the ranks of (3,5,5,5,5,8) are (1,3.5,3.5,3.5,3.5,6) instead of the unadjusted rank (1,2,3,4,5,6)).

Let $R_1, R_2, ..., R_{n_1}$ be the ranks of the first sample observations and $R_{n_1+1}, R_{n_1+2}, ..., R_{n_1+n_2}$ be the ranks of the second sample observations.

Let us define an indicator function:

$$\phi(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i < Y_j \\ 0, & \text{otherwise} \end{cases}, \quad i = 1(1)n_1,\ j = 1(1)n_2$$

Then the **test statistic** is defined as below:

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

U is the number of times an $\boldsymbol{X_i}$ precedes a $\boldsymbol{Y_j}$ among all $\boldsymbol{(X_i, Y_j)}$ pairs. The logical region for the one-sided alternative that the Y's are stochastically larger than the X's, H_1: $F_y(x) \leq F_x(x)$ would clearly be large values of U.

If we denote by U′ the number of $\boldsymbol{(X_i, Y_j)}$ pairs for which $\boldsymbol{X_i > Y_j}$, then assuming no X=Y ties, we have $U+U' = n_1n_2$. When H_1: Δ (as defined earlier) > 0 is true, U will tend to be larger than U′. Hence we reject $H_0 : \Delta = 0$ in favour of H_1 for large values of U or, equivalently, for small values of U′. Similarly, for testing $H_0 : \Delta = 0$ against H_1: $\Delta < 0$, we reject H_0 for small values of U, while for the two-sided alternative we reject H_0 for small values of min(U,U′). The selection of U, U′ or min(U,U′) depends on the type of alternative. If the computed value of the appropriate U is less than or equal to the tabulated value, then we reject the null hypothesis at the stated level of significance.

The test rule can also be stated as the following:

Suppose here we want to test **H_0: θ=0 against H_1: θ>0**. Under H_1: θ>0, it implies that Y is stochastically larger than X, i.e., $F_y(x) < F_x(x)$. Hence, under H_1, the second sample observations will tend to be higher than the first sample observations and consequently ϕ will tend to take the value 1 more frequently and hence 'U' will tend to be large under H_1. Hence, a too large value of 'U' will indicate departure from the null hypothesis H_0. Therefore, <u>we will be rejecting H_0: θ=0 against H_1: θ > 0 at level α if and only if</u>

$$\mathbf{U_{obs} \geq U_\alpha}$$

where, U_α is such that $P_{H_0}[U \geq U_\alpha] = \alpha$, $\mathbf{U_{obs}}$ is the observed value of the test statistic.

❖ **Relation to other tests:**

- **Comparison to Student's *t*-test:**

When the populations are assumed to differ only in location, the Mann-Whitney U test is directly comparable with Student's *t-test* for means. The asymptotic relative efficiency (ARE)[9] of U relative to *t* is never less than 0.864, and if the populations are normal, the ARE is quite high, at $3/\pi = 0.9550$. The Mann-Whitney U test performs better than the *t* test for some non-normal distributions. For example, The ARE is 1.50 for the double exponential distribution and 1.09 for the logistic distribution, which are both heavy-tailed distributions. The Mann–Whitney U test is preferable to the t-test when the data are ordinal but not interval scaled, in which case the spacing between adjacent values of the scale cannot be assumed to be constant. For distributions sufficiently far from normal and for sufficiently large sample sizes, the Mann–Whitney U test is considerably more efficient than the t-test. This comparison in efficiency, however, should be interpreted with caution, as Mann-Whitney and the t-test do not test the same quantities. If, for example, a difference of group means is of primary interest, Mann-Whitney is not an appropriate test.

- **Robustness:**

As it compares the sums of ranks, the Mann–Whitney U test is less likely than the t-test to spuriously indicate significance because of the presence of outliers. However, the Mann-Whitney U test may have worse type I error[3] control when data are both heteroscedastic[11] and non-normal.

❖ The asymptotic Distribution of U-Statistic:

We know that, $U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, Y_j)$

We also know that, $\mathbf{U = W -} \frac{n_2(n_2+1)}{2}$, where **W** is the sum of the ranks of the second sample observations.

Now, under H_0, the (n_1+n_2) observations can be considered as a single random sample drawn from a single continuous population. So, under H_0, the second sample ranks can be taken as any subset of the $n(= n_1+n_2)$ elements $\{1,2,..., n\}$. Hence, under H_0, the second sample ranks may be considered as a Simple Random Sample of size n_2 drawn without replacement[13] from the first n natural numbers 1,2,..,n . Thus under H_0, R_{n_1+j} can take any value α with probability (1/n), α=1(1) n.

Therefore, $\mathbf{P[R_{n_1+j} = \alpha] = 1/n, \forall \alpha=1(1)n}$

and $\mathbf{P[R_{n_1+j} = \alpha, R_{n_1+j'} = \alpha'] = \frac{1}{n(n-1)}, \alpha\neq\alpha', j\neq j', \alpha,\alpha'=1(1)n, j,j'= 1(1)n_2}$

But, $\mathbf{E_{H_0}[R_{n_1+j}] = \sum_{\alpha=1}^{n} \alpha\, P[R_{n_1+j} = \alpha]}$

$$= \frac{1}{n}\sum_{\alpha=1}^{n} \alpha = \frac{n+1}{2}$$

Thus, $\mathbf{E_{H_0}[W] = \sum_{j=1}^{n_2} E_{H_0}[R_{n_1+j}] = \frac{n_2(n+1)}{2}}$

Therefore, $\mathbf{E_{H_0}[U] = E_{H_0}[W - \frac{n_2(n_2+1)}{2}]}$

$$= \mathbf{E_{H_0}[W] -} \frac{n_2(n_2+1)}{2}$$

$$= \frac{n_2(n+1)}{2} - \frac{n_2(n_2+1)}{2}$$

$$= \frac{n_2\,(n+1-n_2-1)}{2} = \frac{n_1 n_2}{2}. \quad \text{(Since, } n_1+n_2=n\text{)}$$

& $\mathbf{Var_{H_0}(U) = Var_{H_0}(W) = Var(\sum_{j=1}^{n_2} R_{n_1+j})}$

$$= \sum_{j=1}^{n_2} Var(R_{n_1+j}) + \sum_{\substack{j=1 \\ j \neq j'}}^{n_2} \sum_{j'=1}^{n_2} Cov(R_{n_1+j}, R_{n_1+j'})$$

But, $\mathbf{Var(R_{n_1+j}) = E[R^2_{n_1+j}] - [E(R_{n_1+j})]^2}$

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

& $\mathbf{Cov(Rn_1+j, Rn_1+j') = \sum_{\substack{\alpha=1 \\ \alpha \neq \alpha'}}^{n} \sum_{\alpha'=1}^{n} [\alpha - E(R_{n_1+j})][\alpha' - E(R_{n_1+j'})]\, P[R_{n_1+j} = \alpha, R_{n_1+j'} = \alpha']}$

$$= \frac{1}{n(n-1)} \sum_{\substack{\alpha=1 \\ \alpha \neq \alpha'}}^{n} \sum_{\alpha'=1}^{n} \left(\alpha - \frac{n+1}{2}\right)\left(\alpha' - \frac{n+1}{2}\right)$$

Consider,

$$\left[\sum_{\alpha=1}^{n} \left(\alpha - \frac{n+1}{2}\right)\right]^2 = \sum_{\alpha=1}^{n} \left(\alpha - \frac{n+1}{2}\right)^2 + \sum_{\substack{\alpha=1 \\ \alpha \neq \alpha'}}^{n} \sum_{\alpha'=1}^{n} \left(\alpha - \frac{n+1}{2}\right)\left(\alpha' - \frac{n+1}{2}\right)$$

or, $\mathbf{0 = nVar(R_{n_1+j}) + \sum_{\substack{\alpha=1 \\ \alpha \neq \alpha'}}^{n} \sum_{\alpha'=1}^{n} \left(\alpha - \frac{n+1}{2}\right)\left(\alpha' - \frac{n+1}{2}\right)}$

or, $\mathbf{\sum_{\substack{\alpha=1 \\ \alpha \neq \alpha'}}^{n} \sum_{\alpha'=1}^{n} \left(\alpha - \frac{n+1}{2}\right)\left(\alpha' - \frac{n+1}{2}\right) = -nVar(R_{n_1+j})}$

Thus, $\mathbf{Cov(R_{n_1+j}, R_{n_1+j'}) = \frac{1}{n(n-1)}[-nVar(R_{n_1+j})] = -\frac{n+1}{12}}$

Therefore, $\mathbf{Var_{H_0}(U) = Var_{H_0}(W)}$

$$= \frac{n_2(n^2-1)}{12} - \frac{n_2(n_2-1)(n+1)}{12} = \frac{n_1 n_2(n+1)}{12}.$$

Under the null hypothesis, the test statistic is given by,

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \,\tilde{a}\, \mathbf{N(0, 1)}$$

❖ **Examination of the convergence of the asymptotic distribution of the Mann-Whitney U-Statistic to normality:**

Now it is required to examine whether the asymptotic distribution of U-Statistic truly converges to normality if the samples are drawn from any continuous population (whose distribution is not known in advance) say, for example Exponential, Gamma, Beta, Normal Distribution etc. For this purpose, at first two random samples of equal sizes are drawn from the same population (say, Exponential (mean=2)) and the value of the U-Statistic is calculated(taking simulation number, R=100) . Then the Shapiro-Wilk Test for normality is performed separately for each of the sample sizes n=5, 10, 15, 20. This procedure is repeated for different continuous populations and the respective Frequency curves[14] are being obtained for each case. In this context the Shapiro-Wilk Test for normality is described below.

- Shapiro-Wilk Test For Normality: The Shapiro–Wilk test is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk. The Shapiro–Wilk test tests the null hypothesis that a sample $x_1,\ldots, x_n$ came from a normally distributed population. The test statistic is

$$\boldsymbol{W} = \frac{\left(\sum_{i=1}^{n} \boldsymbol{a_i x_{(i)}}\right)^2}{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}$$

where $x_{(i)}$ is the i^{th} order statistic, $\overline{x} = (x_1+x_2+\ldots+x_n)/n$ is the sample mean,

The coefficients $\boldsymbol{a_i}$ are given by

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{C}$$

where C is a vector norm:

$C = ||V^{-1} m|| = (m^T V^{-1} V^{-1} m)^{1/2}$ and the vector m, $m=(m_1,\ldots,m_n)^T$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal; finally, V is the covariance matrix of those normal order statistics.

Interpretation:

The null-hypothesis of this test is that the population is normally distributed. Thus, if the *p* value[16] is less than the chosen α level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the *p* value is greater than the chosen α level, then the null hypothesis that the data came from a normally distributed population is accepted.

Monte Carlo simulation has found that Shapiro–Wilk has the best power for a given significance, followed closely by Anderson–Darling when comparing the Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests.

- The results obtained are as follows:

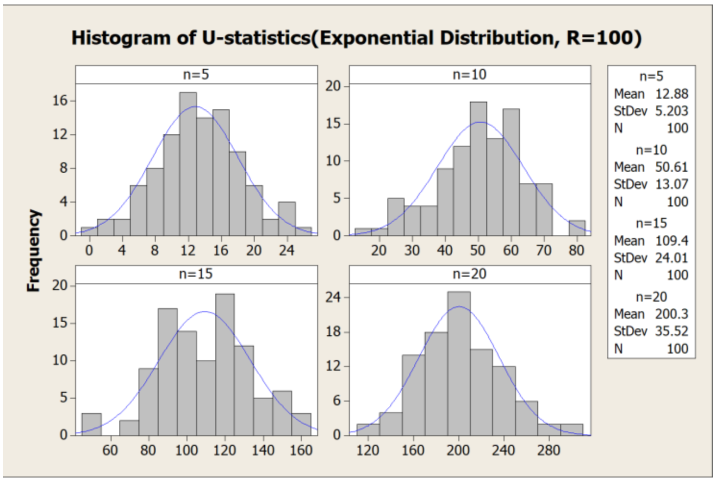
Sample Size(n)		n=5	n=10	n=15	n=20
Exponential Distribution	p-value	**0.82800**	**0.21900**	**0.42590**	**0.45570**
	W	**0.99209**	**0.98282**	**0.98682**	**0.98727**
Gamma Distribution	p-value	**0.03772**	**0.55720**	**0.39150**	**0.60250**
	W	**0.97301**	**0.98865**	**0.98629**	**0.98923**
Beta Distribution	p-value	**0.18900**	**0.51250**	**0.80160**	**0.45430**
	W	**0.98198**	**0.98806**	**0.99173**	**0.98725**
Normal Distribution	p-value	**0.03060**	**0.69430**	**0.49030**	**0.05152**
	W	**0.97184**	**0.99037**	**0.98775**	**0.97474**

- Comment:

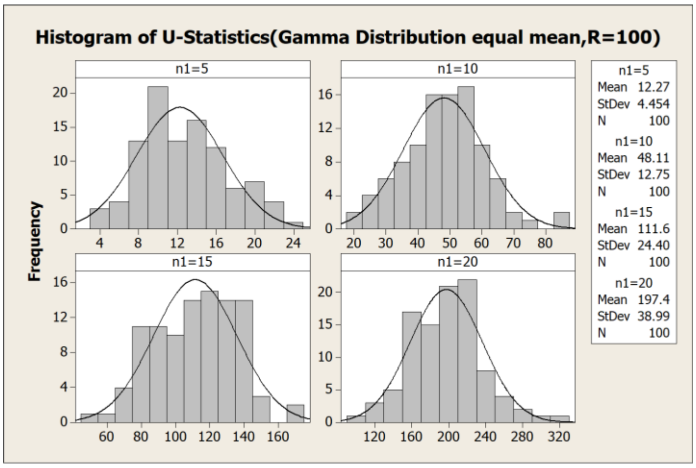
  Here α=0.05. In all the above cases, except for n=5 in Gamma Distribution and in Normal Distribution, it is observed that p-value > 0.05= α. Only for the case of n=5 in Gamma Distribution and in Normal Distribution p-value < α. Hence, the null hypothesis that the population is normally distributed is accepted for each case except for n=5 in Gamma Distribution and in Normal Distribution.

- The corresponding frequency curves (for n=5,10,15,20) for each distribution are given below.

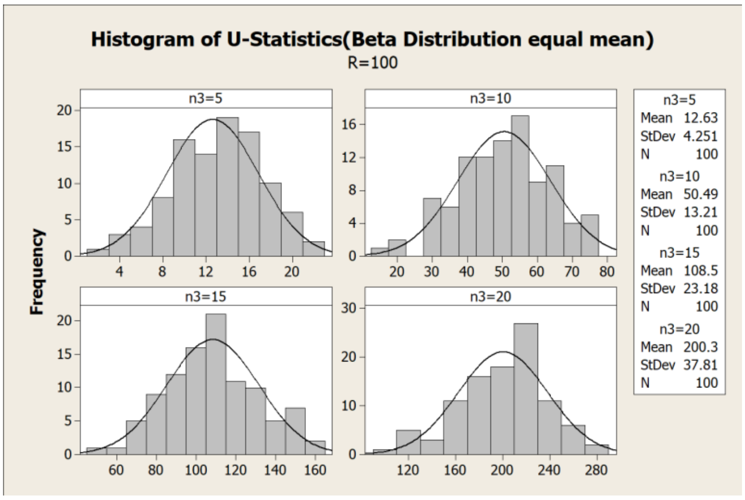
<u>For Exponential Distribution with equal means:</u>



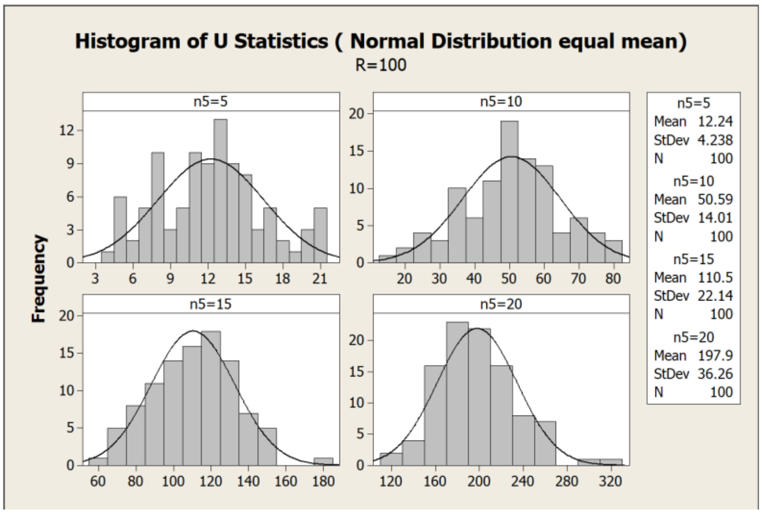
<u>For Gamma Distribution with equal means:</u>



For Beta Distribution with equal means:



For Normal Distribution with equal means:



Then, again two random samples are drawn but now from populations with different location parameters (say, Exp(2) and Exp(4)) and similarly the value of U-statistic is calculated followed by which the Shapiro-Wilk Test for normality is conducted for each case as before and the corresponding frequency curves are obtained.

- The results obtained are as follows:

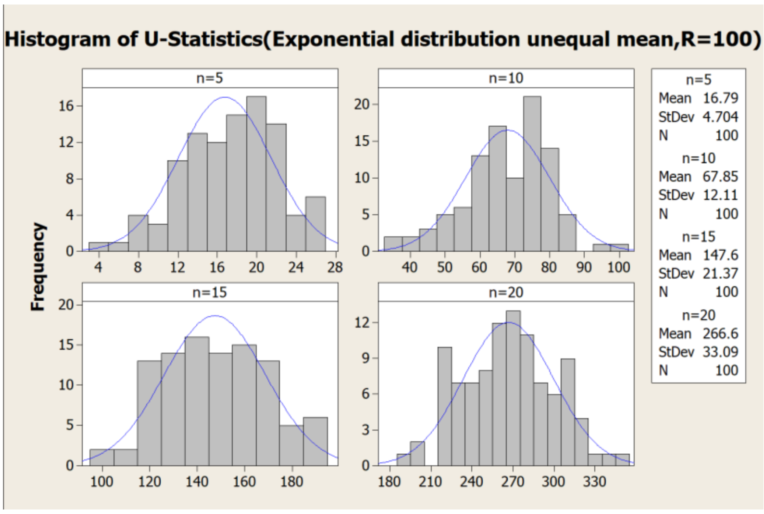
Sample Size(n)		n=5	n=10	n=15	n=20
Exponential Distribution	p-value	**0.08403**	**0.05821**	**0.34300**	**0.87360**
	W	**0.97745**	**0.97542**	**0.98547**	**0.99277**
Gamma Distribution	p-value	**$1.196*10^{-5}$**	**0.10300**	**0.40100**	**0.48660**
	W	**0.91866**	**0.97858**	**0.98644**	**0.98770**
Beta Distribution	p-value	**0.12600**	**0.04681**	**0.10450**	**0.37150**
	W	**0.97970**	**0.97421**	**0.97866**	**0.98596**
Normal Distribution	p-value	**0.50740**	**0.38440**	**0.14570**	**0.36970**
	W	**0.98799**	**0.98617**	**0.98051**	**0.98593**

- Comment:

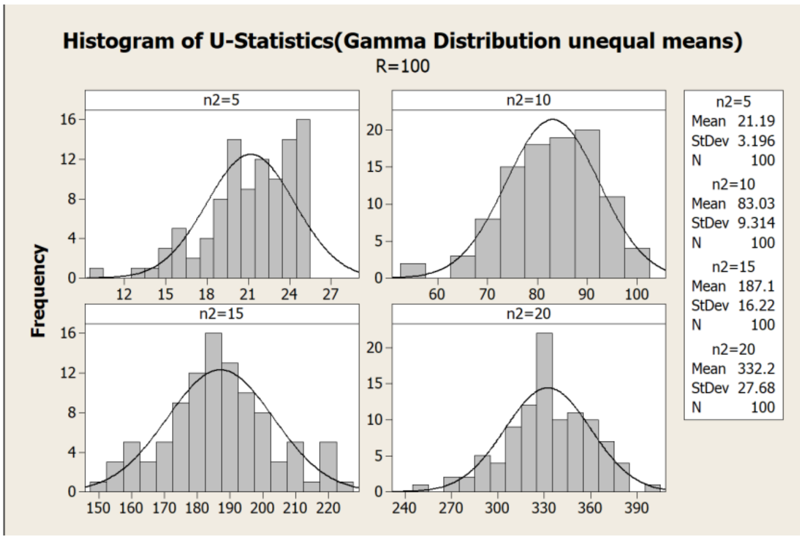
Here also α is taken as 0.05. For n=5 in Gamma Distribution and n=10 in Beta Distribution p-value < α. For all the cases other than the above two, the p-values are observed to be greater than α. Hence, the null hypothesis that the population is normally distributed is accepted in each case except for the above mentioned cases.

- The corresponding frequency curves (for n=5,10,15,20) for each distribution are given below.

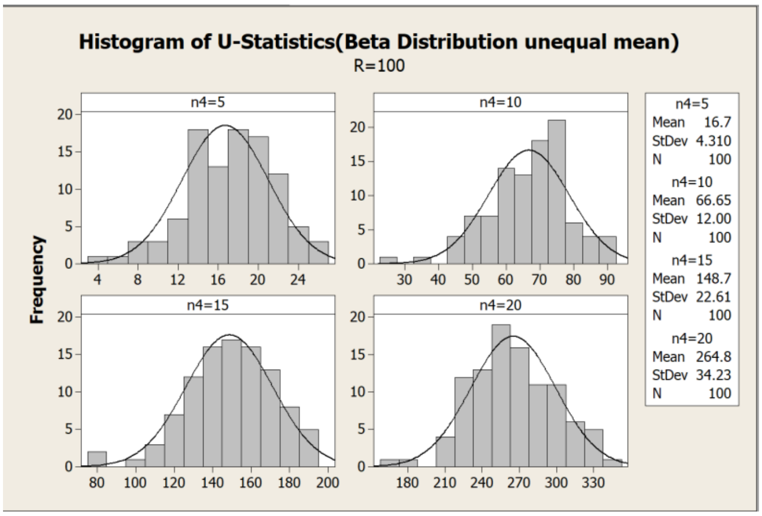
For Exponential Distribution with unequal means:



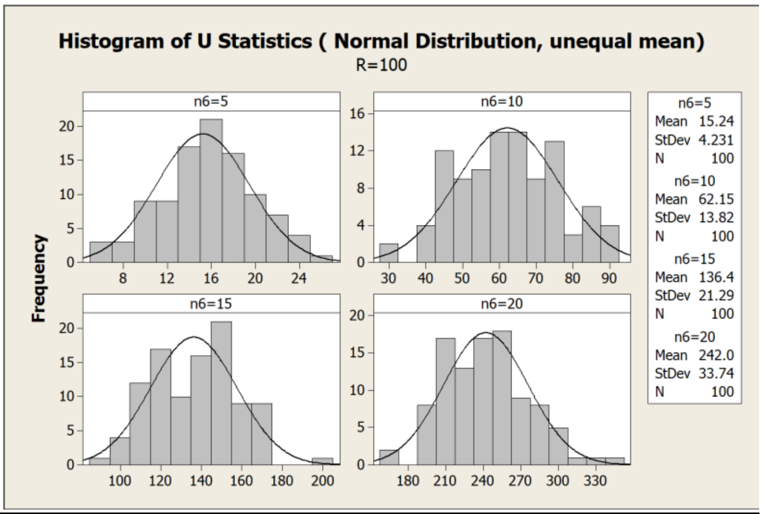
For Gamma Distribution with unequal means:



For Beta Distribution with unequal means:



For Normal Distribution with unequal means:



- Interpretation: From the above discussion and the graphs it is established that the distribution of U statistic converges to normality as the sample sizes increase irrespective of the fact that the random samples, used in the test, have been drawn from any continuous population. The asymptotic distribution of U Statistic is hereafter used to obtain the cut off values U_α and hence in the study of the consistency of Mann-Whitney U test.

❖ **Examination of the consistency of U test:**

The purpose of this paper, as mentioned in the abstract, is to investigate whether the U test becomes more and more consistent as the sample size is increased, i.e., here it is desired to calculate the cut off values U_α for different sample sizes followed by determining the corresponding powers and examining whether they converge to unity with the increase in the sample size. This is because; Consistency of a test for a specified alternative is defined as if the power of the test, when that alternative is true, approaches 1 as the sample size approaches infinity.

- ***To find*** U_α from the asymptotic distribution of U-statistic, which is given by,

$$\left.\begin{aligned} E(U) &= n_1 n_2/2 \\ V(U) &= n_1 n_2 (n+1)/12 \end{aligned}\right\} \text{for different values of n(we here consider } n=n_1=n_2)$$

with α taken as 0.05.

The values of U_α are found as:

$$U_\alpha = \begin{cases} 28.27211, & \text{for } n=6 \\ 71.75937, & \text{for } n=10 \\ 152.156, & \text{for } n=15 \\ 260.8077, & \text{for } n=20 \end{cases}$$

(Note that U_α is the upper α point of a $N(\mu, s)$ distribution , here $\alpha=0.05$

and $\mu= n_1n_2/2$, $s =\sqrt{\{n_1 n_2 (n+1)/12\}}$.)

- ***To get the desired estimated probabilities,***

Now, after the values of U_α for different values of n are obtained, two random samples of size n (i.e., equal sample size) are drawn from different continuous populations. The value of U-Statistic has been calculated for each case with simulation number R=100000. Further the corresponding estimated probabilities of the observed values of the U-Statistic exceeding U_α are obtained. We run the R code[18] repeatedly for n=6, 10, 15, 20 and also separately for different continuous populations (changing the code accordingly) namely Normal, Exponential, Gamma and Beta Distributions respectively. Here the two populations from which the two random samples are drawn are considered to have same variance or dispersion parameter.

- The values of the desired estimates of $P[U \geq U_\alpha]$ (i.e., p) for each sample size (n) and all the above mentioned continuous distributions are tabulated separately for each distribution below.

❑ **X~N(0,1), Y~N(μ,1):**

Sample size(n)	n=6	n=10	n=15	n=20
μ=0.00	**0.04559**	**0.05247**	**0.04834**	**0.05139**
μ=0.10	**0.06230**	**0.07939**	**0.08092**	**0.09139**
μ=0.20	**0.08435**	**0.11471**	**0.12788**	**0.15183**
μ=0.50	**0.18236**	**0.28244**	**0.36321**	**0.45104**
μ=0.80	**0.33245**	**0.52042**	**0.66349**	**0.78414**
μ=0.90	**0.38964**	**0.60289**	**0.75184**	**0.86214**
μ=0.95	**0.42036**	**0.64213**	**0.79079**	**0.89210**
μ=0.98	**0.43863**	**0.66431**	**0.81251**	**0.90771**
μ=1.00	**0.45049**	**0.67912**	**0.82626**	**0.91740**
μ=1.20	**0.57184**	**0.81042**	**0.92640**	**0.97683**
μ=1.50	**0.73988**	**0.93316**	**0.98658**	**0.99812**

- **X ~Exponential(2), Y ~Exponential(mean):**

Sample size(n)	n=6	n=10	n=15	n=20
Mean=2.000	0.04651	0.05255	0.04955	0.05229
Mean=2.222	0.06287	0.07661	0.07992	0.08934
Mean=2.500	0.08558	0.11327	0.12804	0.15116
Mean=3.333	0.16564	0.24894	0.31337	0.39389
Mean=4.000	0.23457	0.36472	0.47022	0.58075
Mean=5.000	0.33525	0.52186	0.66116	0.78240
Mean=5.263	0.36063	0.55774	0.70214	0.81771
Mean=10.000	0.67106	0.88664	0.96889	0.99264
Mean=11.111	0.71426	0.91521	0.98079	0.99630
Mean=20.000	0.88502	0.98623	0.99913	0.99992
Mean=25.000	0.92269	0.99377	0.99975	0.99998
Mean=40.000	0.96725	0.99878	0.99999	1.00000
Mean=50.000	0.97873	0.99953	1.00000	1.00000
Mean=75.000	0.99053	0.99993	1.00000	1.00000
Mean=80.000	0.99156	0.99995	1.00000	1.00000
Mean=90.000	0.99368	0.99998	1.00000	1.00000
Mean=100.000	0.99473	0.99999	1.00000	1.00000

❑ **X ~Gamma(0.1,1), Y ~Gamma(α,1);**

Sample size(n)	n=6	n=10	n=15	n=20
α=0.100	0.04522	0.05356	0.04848	0.05180
α=0.150	0.13253	0.19607	0.23917	0.29828
α=0.180	0.19728	0.30514	0.38869	0.48764
α=0.200	0.24278	0.37880	0.48757	0.59870
α=0.240	0.32840	0.51275	0.65271	0.77074
α=0.285	0.41909	0.63637	0.78527	0.88642
α=0.300	0.44936	0.67204	0.81660	0.91081
α=0.330	0.50047	0.73137	0.86922	0.94518
α=0.360	0.54732	0.78262	0.90766	0.96570
α=0.390	0.59355	0.82384	0.93544	0.97970
α=0.410	0.61854	0.84605	0.94739	0.98497
α=0.440	0.65687	0.87577	0.96254	0.99096
α=0.496	0.71341	0.91518	0.97978	0.99628
α=0.520	0.73550	0.92752	0.98505	0.99754
α=0.550	0.75861	0.94140	0.98955	0.99860
α=0.640	0.81961	0.96712	0.99581	0.99963
α=0.700	0.84977	0.97858	0.99805	0.99986
α=0.780	0.88383	0.98690	0.99906	0.99995
α=0.800	0.89052	0.98839	0.99913	0.99996
α=0.830	0.89911	0.98997	0.99951	1.00000

❑ <u>**X ~Beta(0.1,1) Y ~Beta(α,1);**</u>

Sample size(n)	n=6	n=10	n=15	n=20
α=0.100	0.04555	0.05310	0.04909	0.05104
α=0.130	0.09369	0.12901	0.14585	0.17700
α=0.160	0.15274	0.22724	0.28201	0.35251
α=0.210	0.25658	0.40169	0.51466	0.63058
α=0.240	0.31634	0.49383	0.63149	0.75119
α=0.270	0.37140	0.57742	0.72423	0.83710
α=0.310	0.43871	0.66767	0.81231	0.90774
α=0.350	0.50376	0.73539	0.87379	0.94770
α=0.370	0.53252	0.76488	0.89642	0.95993
α=0.420	0.59270	0.82525	0.93489	0.98019
α=0.480	0.65319	0.87411	0.96230	0.99070
α=0.534	0.69850	0.90618	0.97680	0.99517
α=0.600	0.74422	0.93308	0.98666	0.99805
α=0.667	0.78169	0.95079	0.99218	0.99899
α=0.730	0.81055	0.96315	0.99488	0.99950
α=0.775	0.82739	0.96932	0.99621	0.99966
α=0.825	0.84333	0.97512	0.99735	0.99981
α=0.850	0.85081	0.97704	0.99794	0.99986
α=0.895	0.86346	0.98015	0.99841	0.99992
α=0.911	0.86656	0.98127	0.99846	0.99988
α=0.956	0.87561	0.98434	0.99865	0.99993
α=1.050	0.89587	0.98811	0.99922	0.99997
α=1.650	0.95426	0.99772	0.99995	1.00000

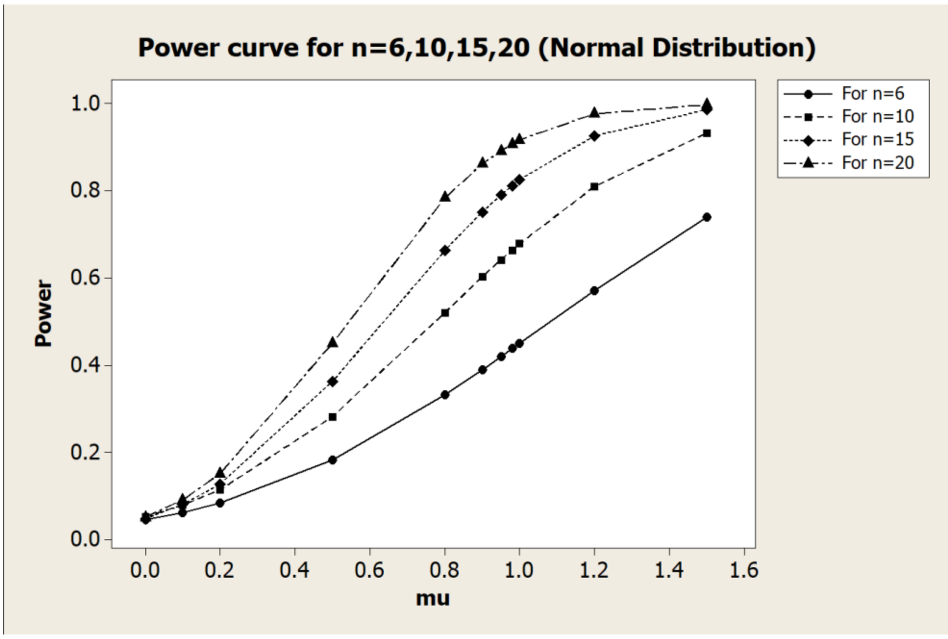
❖ OBSERVATIONS:

- It is observed that when the two samples are drawn from the same continuous population (i.e., populations with the same location parameter) the estimated probabilities of the value of U statistic exceeding U_α (α being the size) tend to α=0.05.

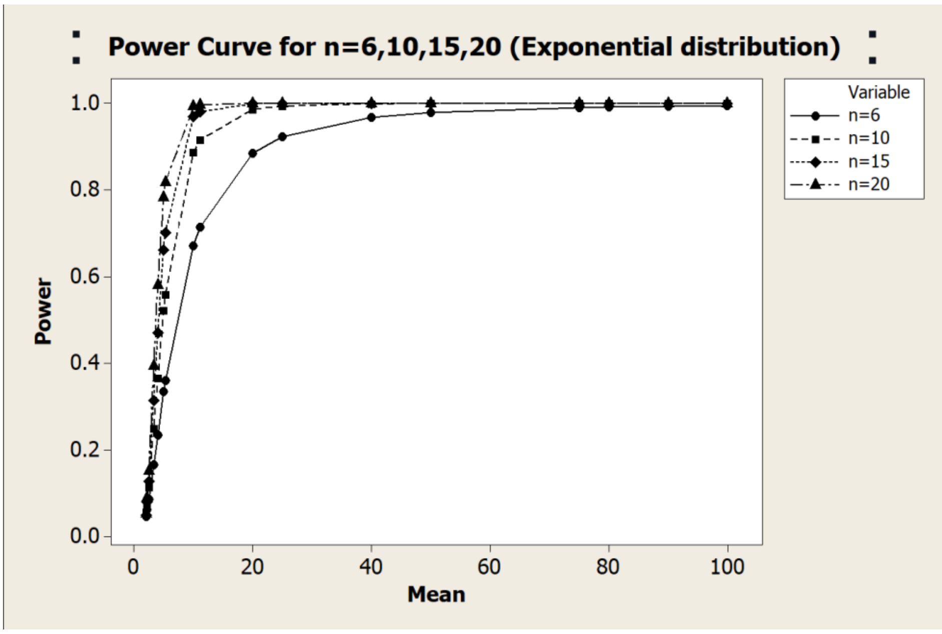
- Note that, a consistent test is the one for which its power increases to 1 as the number of observations considered increases. Now, it is also observed that when the two samples are drawn from two different continuous populations with the difference in their location parameters more than zero, the value of p (the desired estimated probability, which is also the power of the test) increases as n increases and tends to 1. So, it implies that the test becomes more and more consistent.

- Then the following power curves[15] for the four sample sizes corresponding to each continuous distribution considered, are obtained:

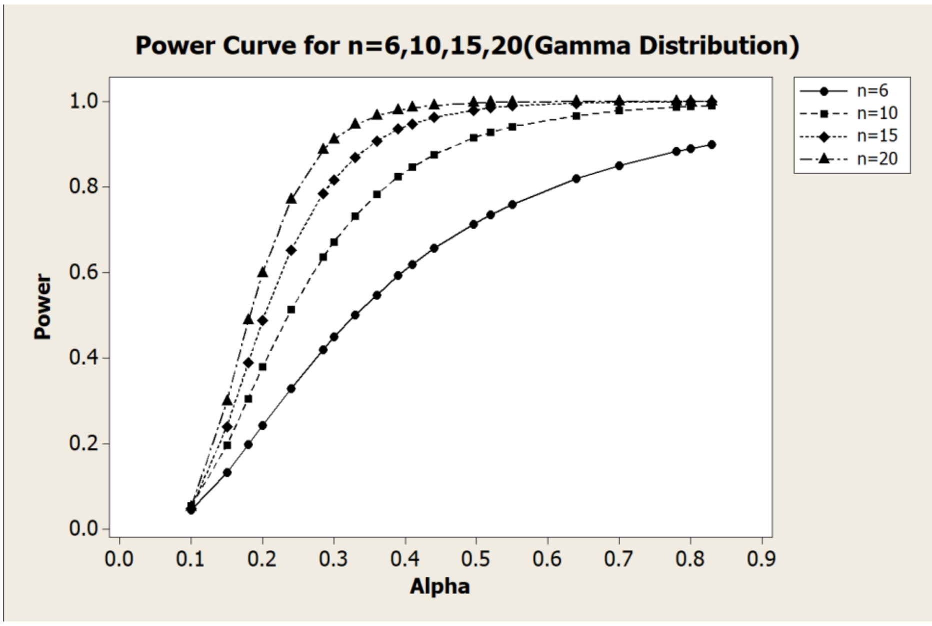
  - <u>For Normal Distribution:</u>



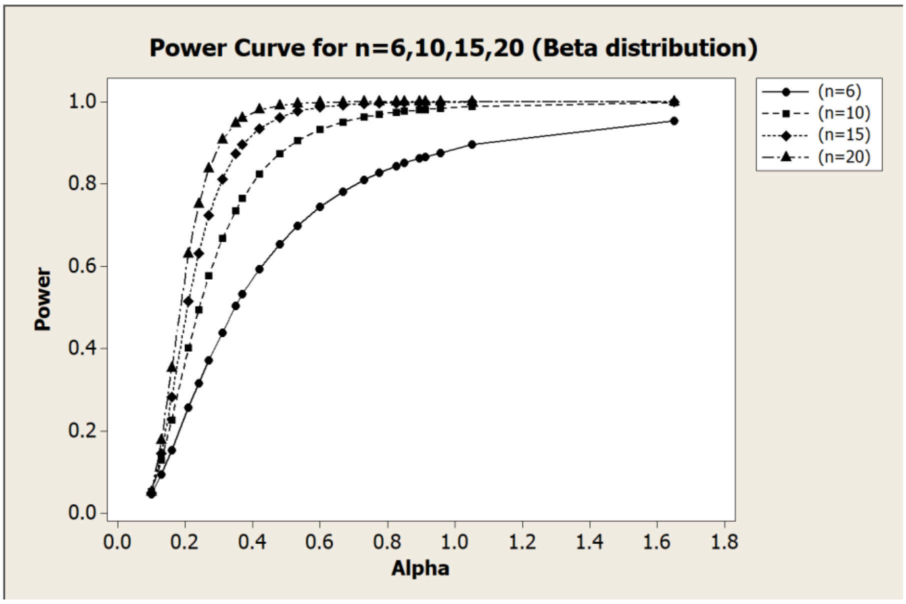
- <u>For Exponential Distribution:</u>



- <u>For Gamma Distribution:</u>



▪ <u>For Beta Distribution:</u>



- **<u>Conclusion:</u>**

It is known that for a number of tests the one with a steeper power curve is more consistent. Thus, it is evident from the above four graphs that as the sample sizes are increased, for each of the continuous distributions the power curves become more and more steeper which implies that the <u>test becomes more and more consistent</u>.

Appendix

Key Words:

[1] **Null Hypothesis:** A hypothesis of no difference where we assume that there is actually no difference between the true value of the unknown parameter and the hypothesised value is known as a Null Hypothesis. A null hypothesis may simple or composite.

[2] **Alternative Hypothesis:** Any hypothesis that contradicts the null hypothesis is known as an Alternative Hypothesis. An alternative hypothesis may either be simple or composite.

[Example: Suppose, $X \sim N(\theta, \sigma^2)$. Here we want to test the null hypothesis H_0: $\theta = 2$ against the alternative hypothesis H_1: $\theta \neq 2$.]

[3] **Type 1 error:** It is the error committed when the null hypothesis is rejected despite being true. Consequently, the error in accepting a false null hypothesis is Type 2 error.

[4] **Size:** Let us consider a test such that, P (Type 1 error) $\leq \alpha$, α is a preselected nominal upper bound to the probability of type 1 error or simply Level of the test. In case of equality, i.e., the least upper bound value α is said to be the Size of the test. Level is the upper bound to the size of the test, which may not be attained if the relevant probability function is discrete.

[5] **Power:** By power of a test, we mean the probability with which it rejects a false null hypothesis, i.e., Power= P (Rejection of H_0 | H_0 is false) = 1-P(Type 2 error). The statistical power ranges from 0 to 1, and as statistical power increases, the probability of making a type II error decreases. In addition, the concept of power is used to make comparisons between different statistical testing procedures: for example, between a parametric test and a nonparametric test of the same hypothesis.

[6] **Consistency (also called stochastic convergence and convergence in probability):**

A consistent estimator is one for which, when the estimate is considered as a random variable indexed by the number *n* of items in the data set, as *n* increases the estimates converge in probability to the value that the estimator is designed to estimate. An estimator that has Fisher consistency is one for which, if the estimator were applied to the entire population rather than a sample, the true value of the estimated parameter would be obtained. A test is consistent for a specified alternative if the power of the test, when that alternative is true, approaches 1 as the sample size approaches infinity. A test is consistent for a class (or subclass) of alternatives if the power of the test when any member of the class (subclass) of alternatives is true approaches 1 as the sample size increases. Consistency of a test can often be shown by investigating whether or not the test statistic converges in probability to the parameter of interest.

[7] **Wilcoxon Signed Rank Test:**

The Wilcoxon signed- rank test is another example of a non-parametric or distribution free test. The Wilcoxon signed- rank test is used to test the null hypothesis that the median of a distribution is equal to some value. It can be used a) in place of a one-sample t-test, b) in place of a paired t-test, or c) for ordered categorial data where a numerical scale is inappropriate but where it is possible to rank the observations. The Wilcoxon signed-rank statistics can also be considered to test for symmetry if the only assumption made is that the random sample is drawn from a continuous distribution.

[8] **Hodges–Lehman estimator**:

In statistics, the Hodges–Lehmann estimator is a robust nonparametric estimator of a population's location parameter. For populations that are symmetric about one median, such as the (Gaussian) normal distribution or the Student *t*-distribution, the Hodges–Lehmann estimator is a consistent and median-unbiased estimate of the population median. For non-symmetric populations, the Hodges–Lehmann estimator estimates the "pseudo–median", which is closely related to the population median. In the simplest case, the "Hodges–Lehmann" statistic estimates the location parameter for a univariate population. The Hodges–Lehmann statistic also estimates the difference between two populations.

[9] **Efficiency:**

In the theory of point estimation, the efficiency of two unbiased estimators for a parameter is defined as the ratio of their variances. Let A and B be two consistent tests of a null hypothesis H_0 and alternative hypothesis H_1, at significance level α. The asymptotic relative efficiency (ARE) of test A relative to test B is the limiting value of the ratio n_a / n_b, where n_a is the number of observations required by test A for the power of test A to equal the power of test B based on n_b observations. Let T_n be a best asymptotic normal estimator of a parametric function γ(θ). Then asymptotic efficiency of any other consistent asymptotically

normal (CAN) estimator T_n' of γ(θ) is defined as $\mathbf{ARE(T' : T) = \sigma_T^{\,2} / \sigma_{T'}^{\,2}}$, where $\sigma_T^{\,2} = \mathrm{Var}[\sqrt{n}\{T_n - \gamma(\theta)\}]$, $\sigma_{T'}^{\,2} = \mathrm{Var}[\sqrt{n}\{T_{n'} - \gamma(\theta)\}]$.

[10] **t-test:** The t-test assesses whether the means of two groups are *statistically* different from each other. A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t* distribution.

[11] **Heteroscedasticity:**

In statistics, a collection of random variables is heteroscedastic if there are sub-populations that have different variabilities from others. The existence of heteroscedasticity is a major concern in the application of regression analysis, including the analysis of variance, as it can invalidate statistical tests of significance that assume that the modelling errors are uncorrelated and uniform, hence that their variances do not vary with the effects being modelled. Similarly, in testing for differences between sub-populations using a location test, some standard tests assume that variances within groups are equal.

[12] **Kendall's tau correlation coefficient:**

It is a measure of ordinal association between two measured quantities. Let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be a set of observations of the joint random variables X and Y respectively, such that all the values of (x_i) and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where i<j, are said to be *concordant* if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

The Kendall τ coefficient is defined as:

$$\boldsymbol{\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{n(n-1)/2}}$$

[13] **SRSWOR:** SRSWOR is a method of selection of n units out of the N units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected, i.e. 1/N . The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population.

[14] **Frequency Curves:** A frequency-curve is a smooth curve for which the total area is taken to be unity. It is a limiting form of a histogram or frequency polygon. The frequency-curve for a distribution can be obtained by drawing a smooth and free hand curve through the mid-points of the upper sides of the rectangles forming the histogram. There exist four types of frequency-curves namely: (a) Bell-shaped curve, (b) U-shaped curve, (c) J-shaped curve, (d)Mixed curve.

[15] **Power Curves:** It represents the power of the test on one axis and deviation of the mean from the target on another axis. A power curve is a plot of the power function ϕ. It helps to assess the suitable size of the sample. At a particular sample size and power value, the values on the graph are examined to find out the deviation of the mean from the target. The power curve is represented by $(1-\beta)$ wherein β stands to be the probability pertaining to which type II error is committed.

[16] **p-value:** An alternative approach to hypothesis testing is provided by computing a quantity called the p-value, sometimes called a Probability value or the associated probability or the significance probability. A p-value is defined as the probability, when the null hypothesis H_0 is true, of obtaining a sample result as extreme as, or more extreme than (in the direction of the alternative), the observed sample result. This probability can be computed for the observed value of the test statistic or some function of it like the sample estimate of the parameter in the null hypothesis. A p-value, in testing of hypothesis, provides us with the means to reject or accept H_0 on the basis of the observed value of the test statistic. Basically, the p-value gives us the probability that the test statistic exceeds the observed value when the null hypothesis is assumed to be true (this is when we are considering a right-tail test). If this probability is less than the size of the test, then it is easy to visualize that the observed value of the test statistic lies very much in the critical region. So, the observed value is highly improbable value of the test statistic. Hence, it leads to rejection of H_0 .Similarly p-value for a left-tail test is defined.

[17] **Non-parametric test:** The term non-parametric test implies a test for a hypothesis which is not a statement about parameter values. The type of statement permissible then depends on the definition accepted for the term parameter. If parameter is interpreted in the broader sense, the hypothesis can be concerned only with the form of the population, as in goodness of fit tests, or with some characteristic of the probability distribution of the sample data, as in tests of randomness and trend.

}

[18] **R codes:**

- **R code to obtain the estimated probabilities (p):**

  - For Normal Distribution [e.g., X~N(0,1),Y~N(0.2,1), n=6]:

```
set.seed(1234)
R=100000
L=function(n)
{
 X=rnorm(n,0,1)
Y=rnorm(n,0.2,1)
 S=c(X,Y)
 r=rank(S)
W=sum(r[(n+1):(n+n)])
U=W-(n*(n+1)/2)
 U
}
n=6
U=replicate(R,L(n))
t=length(U[U>=u_alpha])
p=t/R
p
```

  - For Exponential Distribution[e.g., X~Exp(0.5),Y~Exp(0.45) , n=20]:

```
set.seed(1234)
R=100000
L=function(n)
{
 X=rexp(n,rate=0.5)
```

```
 Y=rexp(n,rate=0.45)
 S=c(X,Y)
 r=rank(S)
 W=sum(r[(n+1):(n+n)])
 U=W-(n*(n+1)/2)
 U
}
n=20
U=replicate(R,L(n))
t=length(U[U>=U_alpha])
p=t/R
p
```

- <u>For Gamma Distribution[X~Gamma(0.1,1), Y~Gamma(0.15,1), n=15]:</u>

```
set.seed(1234)
R=100000
L=function(n)
{
 X=rgamma(n,0.1,rate=1)
 Y=rgamma(n,0.15,rate=1)
 S=c(X,Y)
 r=rank(S)
 W=sum(r[(n+1):(n+n)])
 U=W-(n*(n+1)/2)
 U
}
n=15
```

```
U=replicate(R,L(n))
t=length(U[U>=U_alpha])
p=t/R
p
```

- For Beta distribution[X~Beta(0.1,1),Y~Beta(1.65,1), n=10]:

```
set.seed(1234)
R=100000
L=function(n)
{
X=rbeta(n,0.1,1)
 Y=rbeta(n,1.65,1)
 S=c(X,Y)
r=rank(S)
 W=sum(r[(n+1):(n+n)])
 U=W-(n*(n+1)/2)
 U
}
n=10
U=replicate(R,L(n))
t=length(U[U>=U_alpha])
p=t/R
p
```

- **R Code to obtain the values of U_α :**

(e.g., for $n=n_1=n_2=15$)

```
n1=15
n2=15
```

```
mu=(n1*n2)/2
mu
Var=(n1*n2*(n1+n2+1)/12)
Var
s=sqrt(Var)
s
U_alpha=qnorm(0.95,mu,s)
U_alpha
```

- **R Code for the Normality test:**

(e.g., Exponential distribution, n=10)

```
set.seed(1234)
R=100
L=function(n)
{
 X=rexp(n,rate=0.5)
Y=rexp(n,rate=0.5)
S=c(X,Y)
 r=rank(S)
W=sum(r[(n+1):(n+n)])
U=W-(n*(n+1)/2)
U
}
n=10
U2=replicate(R,L(n))
shapiro.test(U2)
```

ACKNOWLEDGMENT:

I would like to express my deepest gratitude to my dissertation guide, Prof. Pallabi Ghosh for her able guidance and continuous support in completing my project. I cannot thank her enough for her patience and immense knowledge as my mentor and professor.

I am grateful to all those who have supported me during my dissertation course.

I am also grateful to St. Xavier's College (Autonomous), Kolkata for giving me this golden opportunity to enhance my knowledge and test my skills accumulated over the past years in my undergraduate course.

Bibliography:

1. Nonparametric Statistical Inference, Jean Dickinson Gibbons and Subhabrata Chakraborti.
2. Fundamentals of Statistics Volume One, Gun Gupta Dasgupta.
3. Class notes of semester 5 Non- parametric Inference.
4. https://www.wikipedia.org
5. https://www.idpublications.org
6. https://www.onlinemath4all.com/frequency-curve.html.