Comments about the Wilcoxon Rank Sum Test

Scott S. Emerson, M.D., Ph.D.

This document presents some general comments about the Wilcoxon rank sum test. Even the most casual reader will gather that I am not too impressed with the scientific usefulness of the Wilcoxon test. However, the actual motivation is more to illustrate differences between parametric, semiparametric, and nonparametric (distribution-free) inference, and to use this example to illustrate how many misconceptions have been propagated through a focus on (semi)parametric probability models as the basis for evaluating commonly used statistical analysis models.

The Wilcoxon rank sum test was defined to compare the probability distributions for measurements taken from two independent samples. This document describes

1. A general notation that applies to all two sample problems.
2. A definition of parametric, semiparametric, and nonparametric probability models that might be used in the two sample setting.
3. A characterization of the null hypotheses commonly tested in the two sample setting.
4. The transformation of the data used in the definition of the Wilcoxon rank sum test.
5. The formulation of the Wilcoxon rank sum statistic, including its relationship to the Mann-Whitney U statistic.
6. The sampling distribution of the Wilcoxon rank sum statistic under the various choices for the null and alternative hypotheses, and the formulation of hypothesis tests and test statistics.
7. The interpretation of the statistic with respect to common summary measures of probability distributions.
8. The intransitivity of the functional $Pr(X \geq Y)$.
9. Some results about the relative efficiency of the Wilcoxon rank sum test.
10. Relevance of the above comments to other parametric and semiparametric testing / estimation settings.

1. General Notation

We consider a scientific question that is to be statistically addressed by comparing the distribution of some random variable across two populations. Without loss of generality, we adopt the nomenclature from the setting of a randomized clinical trial in which we have a "treated" population and a "control" population. For notational convenience we denote the random variable by X when measured on the "treated" group and by Y when measured on the "control" group.

We thus consider the two sample problem is which we have:

- independent, identically distributed observations $X_i \sim F(x) = Pr(X_i \leq x)$ for $i = 1, \ldots, n$, and
- independent, identically distributed observations $Y_i \sim G(y) = Pr(Y_i \leq y)$ for $i = 1, \ldots, m$.

We further assume that X_i and Y_j are independent for all $1 \leq i \leq n$ and $1 \leq j \leq m$.

2. Probability Models for the Two Sample Problem

2.1. Parametric Probability Models

By a parametric probability model, we assume that there exists a finite p dimensional parameter $\vec{\omega}$ and a <u>known</u> probability distribution function F_0 depending on $\vec{\omega}$ such that $F(x) = F_0(x; \vec{\omega} = \vec{\omega}_X)$ and $G(y) = F_0(y; \vec{\omega} = \vec{\omega}_Y)$ for two specific values $\vec{\omega}_X$ and $\vec{\omega}_Y$. Typically, a statistical problem involves the estimation and testing of some $\theta = h(\vec{\omega}_X, \vec{\omega}_Y)$.

Estimation of $\vec{\omega}_X$ amd $\vec{\omega}_Y$ will typically proceed using parametric likelihood theory or parametric methods of moments estimation. Estimation of F and G will then use the parametric estimates $\hat{F}(x) = F_0(x, \hat{\vec{\omega}}_X)$ and $\hat{G}(y) = F_0(y, \hat{\vec{\omega}}_Y)$. Estimation of other functionals of those distributions will then be based on the parametric estimates derived from $\hat{F}(x)$ and $\hat{G}(y)$.

Examples of commonly used parametric probability models include

- Normal: A continuous probability model having parameter $\vec{\omega} = (\mu, \sigma^2)$ with $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. $F_0(x, \vec{\omega} = (\mu, \sigma^2)) = \Phi\left((x - \mu)/\sigma\right)$, where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2} \, du$$

  is the standard normal cumulative distribution function. Most often our statistical question of interest relates to the difference of means $\theta = \mu_X - \mu_Y$.

- Exponential: A continuous probability model having parameter $\vec{\omega} = \lambda$ with $0 < \lambda < \infty$.

$$F_0(x, \vec{\omega} = \lambda) = (1 - e^{-\lambda x}) 1_{(0,\infty)}(x)$$

  In this setting we might consider statistical questions based on the difference of means $\theta = 1/\lambda_X - 1/\lambda_Y$ or the hazard ratio $\theta = \lambda_X / \lambda_Y$.

- Poisson: A discrete probability model having parameter $\vec{\omega} = \lambda$ with $0 < \lambda < \infty$, and, for some known measure t of time and space,

$$F_0(x; \vec{\omega} = \lambda) = \sum_{k=0}^{\lfloor x \rfloor} \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$

  In this setting we might consider statistical questions based on the difference of mean rates $\theta = \lambda_X - \lambda_Y$ or the ratio of mean rates $\theta = \lambda_X / \lambda_Y$.

2.2. Semiparametric Probability Models

The following definition of semiparametric probability models is a bit more restrictive than those used by some other authors. This definition is used because 1) it is satisfied by the most commonly used semiparametric statistical analysis models, and 2) there are some important common issues that arise in all semiparametric models that satisfy this definition.

By a semiparametric probability model, we assume that there exists a finite p dimensional parameter $\vec{\omega}$ and an unknown probability distribution function F_0 depending on $\vec{\omega}$ such that $F(x) = F_0(x; \vec{\omega} = \vec{\omega}_X)$ and $G(y) = F_0(y; \vec{\omega} = \vec{\omega}_Y)$ for some specific values $\vec{\omega}_X$ and $\vec{\omega}_Y$. The unknown, infinite dimensional $F_0(x, \vec{\omega}_0)$ for some "standard" choice of $\vec{\omega}_0$ is generally just regarded as a nuisance parameter. For identifiability it is sometimes convenient to put constraints on the moments of $F_0(x, \vec{\omega}_0)$ for the "standard" choice of $\vec{\omega}_0$. In other settings, it is convenient to choose $F_0(x, \vec{\omega}_0) = G(x)$, the distribution in some control population.

The salient feature of a semiparametric model (under this definition) is the existence of some $\theta = h(\vec{\omega}_X, \vec{\omega}_Y)$ that allows the transformation of F to G, and estimation and testing of θ can typically be performed without appealing directly to estimation of $\vec{\omega}_X$ and $\vec{\omega}_Y$. We do note that in many cases the semiparametric model also specifies a way in which θ can be used to transform the individual X_i's in such a way that the transformed variables, say $W_i = \psi(X_i, \theta)$ are distributed according to G. In such cases one could imagine an estimation approach that finds the choice $\hat{\theta}$ such that $W_i = \psi(X_I, \hat{\theta})$ would have an empirical distribution function $\hat{F}_W$ that was closest (in some sense) to the empirical distribution function $\hat{G}$ of the Y_j's.

When it is of interest to do so, the nuisance parameter $F_0(x, \vec{\omega}_0)$ is estimated by using the parameter estimate $\hat{\theta}$, and then to perform suitable transformations of the X_i's or their empirical distribution function $\hat{F}$ to an estimate $\hat{F}_0$ based on the entire sample.

Examples of commonly used semi-parametric probability models include

- Location shift: A continuous probability model having parameter $\omega = \mu$ with baseline probability distribution $F_0(x; \omega = 0)$ typically chosen to have mean 0. Then $F(x) = F_0(x; \omega = \mu_X) = F_0(x - \mu_X; \omega = 0)$ and $G(y) = F_0(y; \omega = \mu_Y) = F_0(y - \mu_Y; \omega = 0)$. This then also ensures that $G(y) = F(y - (\mu_X - \mu_Y))$, and $E(X) = E(Y) + (\mu_X - \mu_Y)$. Hence, in this setting, we typically consider statistical questions based on the difference $\theta = \mu_X - \mu_Y$, although that same number represents the difference of any quantile, as well. Note that under this model $X - \theta$ and Y are distributed according to G, so $F(x) = G(x - \theta)$ and $G(y) = F(y + \theta)$.

- Shift-scale: A continuous probability model having parameter $\vec{\omega} = (\mu, \sigma)$ and "baseline" distribution $F_0(x; \vec{\omega}_0)$ typically chosen such that $\vec{\omega} = (0, 1)$ or such that $G(y) = F_0(y, \vec{\omega}_0)$. Then

$$F(x) = F_0(x; \vec{\omega} = \vec{\omega}_X) = F_0\left(\frac{(x - \mu_X)}{\sigma_X}; \vec{\omega} = \vec{\omega}_0\right) \quad \text{and} \quad G(y) = F_0(y; \vec{\omega} = \vec{\omega}_Y) = F_0\left(\frac{(y - \mu_Y)}{\sigma_Y}; \vec{\omega} = \vec{\omega}_0\right).$$

  This then also ensures that

$$\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) + \mu_X \sim X \quad \text{and} \quad \frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \mu_Y \sim Y$$

  and

$$F(x) = G\left(\frac{\sigma_Y}{\sigma_X}x - \left(\frac{\sigma_Y}{\sigma_X}\mu_X - \mu_Y\right)\right) \quad \text{and} \quad G(y) = F\left(\frac{\sigma_X}{\sigma_Y}y - \left(\frac{\sigma_X}{\sigma_Y}\mu_Y - \mu_X\right)\right).$$

  In this setting, we could consider statistical questions based on two dimensional parameter

$$\vec{\theta} = \left(\mu_X - \frac{\sigma_X}{\sigma_Y}\mu_Y, \frac{\sigma_X}{\sigma_Y}\right),$$

  where $(X - \theta_1)/\theta_2$ and Y are distributed according to G. However, more typically inference is based on the unscaled difference in means $\mu_X - \mu_Y$, with the scale parameters treated as nuisance parameters.

- Accelerated failure time: A nonnegative random variable has a continuous probability distribution defined by parameter $\vec{\omega} = \lambda$ and some "baseline" distribution $F_0(x; \vec{\omega} = \lambda_0)$. Then

$$F(x) = F_0(x; \vec{\omega} = \lambda_X) = F_0(\lambda_X x; \vec{\omega} = \lambda_0) \qquad \text{and} \qquad G(y) = F_0(y; \vec{\omega} = \lambda_Y) = F_0(\lambda_Y y; \vec{\omega} = \lambda_0).$$

  This then ensures that $\lambda_X X/\lambda_Y \sim Y$, $\lambda_Y Y/\lambda_X \sim X$,

$$F(x) = G\left(x\frac{\lambda_X}{\lambda_Y}\right) \qquad \text{and} \qquad G(y) = F\left(y\frac{\lambda_Y}{\lambda_X}\right).$$

  In this setting we might consider statistical questions based on $\theta = \lambda_Y/\lambda_X$, which can be shown to be the ratio of any quantile of the distribution of X to the corresponding quantile of the distribution of Y. Under this model, X/θ and Y are both distributed according to G, hence the accelerated failure time model is a subset of a larger semiparametric scale family (which larger family might allow random variables that could also take on negative values).

- Proportional hazards: A nonnegative random variable has a continuous probability distribution defined by parameter $\vec{\omega} = \lambda$ and some "baseline" distribution $F_0(x; \vec{\omega} = \lambda_0)$. Then

$$F(x) = F_0(x; \vec{\omega} = \lambda_X) = 1 - [1 - F_0(x; \vec{\omega} = \lambda_0)]^{\lambda_X} \quad \text{and} \quad G(y) = F_0(y; \vec{\omega} = \lambda_Y) = 1 - [1 - F_0(y; \vec{\omega} = \lambda_0)]^{\lambda_Y}.$$

  This then ensures that

$$F(x) = 1 - [1 - G(x)]^{\frac{\lambda_X}{\lambda_Y}} \qquad \text{and} \qquad G(y) = 1 - [1 - F(y)]^{\frac{\lambda_Y}{\lambda_X}}.$$

  In this setting we might consider statistical questions based on $\theta = \lambda_X/\lambda_Y$, which can be shown to be the ratio of the hazard function for the distribution of X to that of the distribution of Y. In its general form, there is no specific transformation of X that would lead to the transformed variable having the same distribution as

Y. However, any monotonic transformation of both X and Y will lead to the same relationship between the distribution of the transformed X and the distribution of the transformed Y.

2.3. Nonparametric Probability Models

By a nonparametric probability model, we assume that the distribution functions F and G are unknown with no pre-specified relationship between them.

Typically, a statistical problem involves the estimation and testing of some $\theta = d(F(x), G(y))$, where $d(\cdot,\cdot)$ measures some difference between two distribution functions.

Common choices for θ might be contrasts (differences or ratios) of univariate functionals (e.g., means, geometric means, medians):

- difference of means: $\theta = \int x dF(x) - \int y dG(y)$
- ratio of geometric means: $\theta = \exp\left[\int \log(x) dF(x) - \int \log(y) dG(y)\right]$
- difference of medians: $\theta = F^{-1}(0.5) - G^{-1}(0.5)$
- difference of the probability of exceeding some threshold c: $\theta = G(c) - F(c)$

At times θ is defined based on a bivariate functional:

- median difference: $\theta = F^{-1}_{X-Y}(0.5)$
- maximal difference between cumulative distribution functions: $\theta = \max |F(x) - G(x)|$
- probability that a randomly chosen value of X exceeds a randomly chosen value of Y: $\theta = Pr(X > Y)$

3. Characterization of null hypotheses

In two sample tests, we are often interested in inference about general tendencies for measurements in the treatment group ($X \sim F$) to be larger than measurements in the control group ($Y \sim G$). The null hypothesis to be disproved is generally one of some tendency for measurements to be similar in the two populations. As noted above, we generally define some estimand θ that contrasts the distributions F and G.

We thus find it of interest to consider two distinct levels of null hypothesis.

- the “Strong” null hypothesis: $H_0 : F(x) = G(x) \forall x$
- the “Weak” null hypothesis: $H_0 : \theta = \theta_0$ where θ_0 is typically chosen to be the value of θ when the strong null is true.

There are two main distinctions that need to be made between these hypotheses:

First, scientifically, if we have chosen the form of θ to capture scientifically important differences in the distribution, then we might only want to detect differences between the distributions that do affect θ. It is of course possible that a treatment might modify aspects of a probability distribution in a way that θ is not affected. For instance, if θ is measuring the difference in medians, a treatment that only modifies the upper 10% of the probability distribution will have $\theta = \theta_0$. Hence, the strong null would be false, but the weak null would be true.

Second, statistically, if our true goal is to make statements about whether the weak null is true or not, calculating the variance of our test statistic under the strong null can lead to tests of the wrong statistical level (the type I error might be wrong as a test of the weak null). When this is true, we can only interpret our results as rejecting the strong null hypothesis and cannot make a statistically valid statement about the weak null unless we use a different variance estimate.

The issues that arise in common practice are that:

- Use of a parametric or semiparametric probability model might suggest (for reasons of efficiency) the testing and estimation of a particular choice of θ. Such a choice might directly address the scientific question at hand, or it might be used to derive tests and estimates of some function scientifically important $\psi(\theta)$ using a parametric or semiparametric estimator $\psi(\hat{\theta})$.

- In a distribution free setting, $\psi(\hat{\theta})$ may not be consistent for the scientifically important functional. For instance, the parametric estimator of the median in a lognormal model is not consistent for the median of an exponential model. If the estimation of the median was the scientifically important task, then the use of the wrong assumption about the shape of the distribution might make the analysis scientifically invalid.
- Even if the parametric or semiparametric estimator can be shown to be consistent in the distribution free setting, the use of the parametric or semiparametric model to estimate the variability of the estimator might lead to invalid statistical inference. For instance, in a Poisson probability model, the efficient estimator of the rate is the sample mean, and the parametric estimator of the standard error is also based on the sample mean due to the known mean-variance relationship of the Poisson distribution. However, should the count data not be Poisson distributed, that estimated standard error may be smaller than or larger than the true standard error for the sample mean.
- Even if the parametric or semiparametric estimator of the standard error can be shown to be consistent for the true standard error under the null hypothesis of the parametric or semiparametric model, that null hypothesis most often corresponds to the strong null hypothesis. Hence, under the weak null hypothesis (which may be scientifically more relevant), the statistical test is possibly of the wrong size.
- The statistical efficiency of a parametric or semiparametric estimator might be substantially affected by even small (statistically undetectable) departures from the parametric and semiparametric probability model.

There are settings in which there is no distinction between the strong null and the weak null. For instance, in the setting of independent binary data, the strong and weak null are identical: The measurements have to follow the one parameter Bernoulli family, differences in the proportion are synonymous with differences in the distribution. Similarly, in the setting of any ordered random variable when defining $\theta = \max|F(x) - G(x)|$ as the maximum difference between the cumulative distribution functions, if $\theta = 0$, the strong null has to be true and vice versa. (This is the functional tested in the Kolmogorov-Smirnov test.)

The distinctions between the strong and weak null hypothesis will be illustrated below with the Wilcoxon rank sum statistic.

4. Transformation of the data

The Wilcoxon rank sum test can be thought of as a transformation of the original data to their ranks. That is, given a sample of independent, identically distributed X_i, $i = 1, \ldots, n$, and a sample of independent, identically distributed Y_i, $i = 1, \ldots, m$, we transform all of the random variables from the scale they were originally measured on to their ranks. When there are no ties, this can be written as

$$R_i^* = \sum_{j=1}^{n} 1_{[X_j \leq X_i]} + \sum_{j=1}^{m} 1_{[Y_j \leq X_i]}$$

$$S_i^* = \sum_{j=1}^{n} 1_{[X_j \leq Y_i]} + \sum_{j=1}^{m} 1_{[Y_j \leq Y_i]}$$

where the indicator function 1_A is 1 if A is true and 0 otherwise. Under the above notation, the observation with the lowest value in the combined sample will have rank 1, the observation with the highest value will have rank $m + n$.

In the presence of ties in the sample, we modify the ranks to use the midrank among the tied observations:

$$R_i = rank(X_i) = R_i^* - (\sum_{j=1}^{n} 1_{[X_j = X_i]} + \sum_{j=1}^{m} 1_{[Y_j = X_i]} - 1)/2$$

$$S_i = rank(Y_i) = S_i^* - (\sum_{j=1}^{n} 1_{[X_j = Y_i]} + \sum_{j=1}^{m} 1_{[Y_j = Y_i]} - 1)/2$$

It should be noted that in the absence of ties, $R_i^* = R_i$ and $S_i^* = S_i$, so these latter definitions are taken to be the defining transformation.

5. Definition of the statistic

5.1. Formulation as the rank sum

The Wilcoxon rank sum test statistic is then (as its name implies) based on the sum of the ranks for each group

$$R = \sum_{i=1}^{n} R_i$$
$$S = \sum_{i=1}^{m} S_i$$

Note that considering just one of R or S is sufficient because

$$R + S = \sum_{i=1}^{m+n} i = \frac{(m+n)(m+n+1)}{2}$$

where we make use of the result $\sum_{i=1}^{N} i = N(N+1)/2$. This can be established by noting that $\sum_{i=1}^{N} i = \sum_{i=1}^{N}(N-i+1)$, thus

$$\begin{aligned}\sum_{i=1}^{N} i &= \frac{1}{2}\left(\sum_{i=1}^{N} i + \sum_{i=1}^{N}(N-i+1)\right)\\ &= \frac{1}{2}\sum_{i=1}^{N}(i+N-i+1)\\ &= \frac{1}{2}\sum_{i=1}^{N}(N+1)\\ &= \frac{1}{2}N(N+1)\end{aligned}$$

(Though this was of course known to others, Gauss derived this result on his own in about first grade when his teacher gave his class the busy work of adding the numbers 1 to 100.)

The intuitive motivation for such a statistic is obvious: If X tends toward larger values than Y, it stands to reason that the measurements of X will correspond to the larger ranks. We could look at the average rank or the sum of the ranks, it does not really matter.

5.2. Formulation as a U-statistic

There is an alternative form of the Wilcoxon rank sum test, the Mann-Whitney U statistic, that is perhaps a more useful derivation of the test, because it provides more insight into what scientific quantity is being tested and a useful structure for similar methods to other settings. In the Mann-Whitney U statistic, we are interested in the probability that a randomly chosen X will be greater than a randomly chosen Y. We estimate this by

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[1_{[X_i>Y_j]} + 0.5 \times 1_{[X_i=Y_j]}\right].$$

From our definition of R_i and R, we find that

$$R - U = \sum_{i=1}^{n}\sum_{i=1}^{n} 1_{[X_i \geq X_j]} = \frac{n(n+1)}{2},$$

which, for a given sampling scheme, is constant. Thus the sampling distribution for R and the sampling distribution for U just differ by that constant, and tests based on R are equivalent to tests based on U.

6. Null sampling distribution

The null hypothesis considered by the Wilcoxon rank sum test (and, equivalently, the Mann-Whitney U statistic) is the strong null hypothesis that the X_i's and the Y_i's have the same probability distribution. The test does not make any assumptions about that common probability distribution.

In the derivations given below, we assume that there are no ties in the data. In the presence of ties, some modifications must be made to the variance of the null sampling distribution. Interested readers can see Lehmann's *Nonparametric Statistics.*

6.1. Derivation of moments using the Wilcoxon rank sum

Under the null hypothesis, then, the sampling distribution of R is the same as that of the sum of n numbers chosen at random without replacement from the set of numbers $\{1, 2, \ldots, m+n\}$. We can find the moments of this sampling distribution of R as follows.

The expectation of R is

$$\begin{aligned} E[R] &= E\left[\sum_{i=1}^{n} R_i\right] \\ &= \sum_{i=1}^{n} E[R_i] \\ &= nE[R_1] \end{aligned}$$

where the last step follows by the fact that all the R_i's are identically distributed (but not independent due to the sampling without replacement). Because each of the $m+n$ ranks are equally likely to be chosen for R_i under the null hypothesis, it follows that

$$\begin{aligned} E[R_i] &= \frac{1}{m+n} \sum_{i=1}^{m+n} i \\ &= \frac{1}{m+n} \frac{(m+n)(m+n+1)}{2} = \frac{(m+n+1)}{2} \end{aligned}$$

which yields

$$E[R] = \frac{n(m+n+1)}{2}$$

The variance of R is

$$\begin{aligned} Var(R) &= E\left[(R - E[R])^2\right] \\ &= E\left[R^2\right] - E^2[R] \end{aligned}$$

and

$$\begin{aligned}
E\left[R^2\right] &= E\left[\left(\sum_{i=1}^{n} R_i\right)^2\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} R_i R_j\right] \\
&= E\left[\sum_{i=1}^{n} R_i^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} R_i R_j\right] \\
&= \sum_{i=1}^{n} E[R_i^2] + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} E[R_i R_j] \\
&= nE[R_1^2] + n(n-1)E[R_1 R_2]
\end{aligned}$$

where, again, the last step follows by the fact that all the R_i's are identically distributed and the joint distribution of (R_i, R_j) is the same for all values $i = 1, \ldots, n$, $j = 1, \ldots, n$, and $i \neq j$. Now

$$E[R_1^2] = \frac{1}{m+n}\sum_{i=1}^{m+n} i^2$$

and because $i^2 = \sum_{j=1}^{i}(2j-1)$ we can find

$$\begin{aligned}
\sum_{i=1}^{N} i^2 &= \sum_{i=1}^{N}\sum_{j=1}^{i}(2j-1) \\
&= \sum_{j=1}^{N}\sum_{i=j}^{N}(2j-1) && \text{(reversing order of summation)} \\
&= \sum_{j=1}^{N}(N-j+1)(2j-1) && \text{(summand does not depend on i)} \\
&= (2N+3)\sum_{j=1}^{N} j - 2\sum_{j=1}^{N} j^2 - \sum_{j=1}^{N}(N+1) \\
3\sum_{i=1}^{N} i^2 &= (2N+3)\frac{N(N+1)}{2} - N(N+1) && \text{(moving sum of } j^2 \text{ to LHS)} \\
\sum_{i=1}^{N} i^2 &= \frac{(2N+1)N(N+1)}{6} && \text{(simplifying terms)}
\end{aligned}$$

Therefore we have

$$E[R_1^2] = \frac{1}{m+n}\frac{(2(m+n)+1)(m+n)(m+n+1)}{6} = \frac{(2(m+n)+1)(m+n+1)}{6}$$

Similarly, the joint distribution for (R_1, R_2) is just the distribution of choosing two numbers without replacement from $\{1, 2, \ldots, m+n\}$, so

$$\begin{aligned}
E[R_1R_2] &= \frac{1}{(m+n)(m+n-1)} \sum_{i=1}^{m+n} \sum_{j \neq i} ij \\
&= \frac{1}{(m+n)(m+n-1)} \left(\sum_{i=1}^{m+n} \sum_{j=1}^{m+n} ij - \sum_{i=1}^{m+n} i^2 \right) \\
&= \frac{1}{(m+n)(m+n-1)} \left(\frac{(m+n)(m+n+1)}{2} \frac{(m+n)(m+n+1)}{2} - \frac{(2(m+n)+1)(m+n)(m+n+1)}{6} \right) \\
&= \left(\frac{(m+n)(m+n+1)^2}{4(m+n-1)} - \frac{(2(m+n)+1)(m+n+1)}{6(m+n-1)} \right) \\
&= \frac{(3(m+n)+2)(m+n+1)}{12}
\end{aligned}$$

and

$$\begin{aligned}
Var(R) &= nE[R_1^2] + n(n-1)E[R_1R_2] - E^2[R] \\
&= n\frac{(2(m+n)+1)(m+n+1)}{6} + n(n-1)\frac{(3(m+n)+2)(m+n+1)}{12} - \frac{n^2(m+n+1)^2}{4} \\
&= \frac{mn(m+n+1)}{12}
\end{aligned}$$

6.2. Derivation of moments using the Mann-Whitney U statistic

Note that in its Mann-Whitney form, finding the moments of U directly is fairly straightforward. The mean of the sampling distribution for U under any hypothesis is easily found to be

$$\begin{aligned}
E[U] &= E\left[\sum_{i=1}^{n} \sum_{j=1}^{m} 1_{[X_i \geq Y_j]} \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} E\left[1_{[X_i \geq Y_j]} \right] \\
&= mnE\left[1_{[X_1 \geq Y_1]} \right] \\
&= mnPr(X \geq Y)
\end{aligned}$$

where we use the fact that the X_i's are identically distributed and the Y_i's are identically distributed, as well as the fact that the expectation of a binary indicator variable is just the probability that the event measured by the indicator variable occurs. Under the strong null hypothesis that X and Y have the same distribution, $Pr(X \geq Y)$ is just the probability that the larger of two randomly sampled independent measurements from the same population would have the first measurement larger than the second. So $Pr(X \geq Y) = 0.5$ under the null hypothesis, and

$$E[U] = \frac{mn}{2}.$$

To find $Var[U]$, we would again use $Var[U] = E[U^2] - E^2[U]$. And

$$U^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{n} \sum_{\ell=1}^{m} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]}$$

This can be most easily solved by considering cases where $i = k$ and $j = \ell$, where $i = k$ but $j \neq \ell$, where $i \neq k$ but $j = \ell$, and where $i \neq k$ and $j \neq \ell$. So

$$\begin{aligned}
U^2 &= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=i}\sum_{\ell=j} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]} + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=i}\sum_{\ell\neq j} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}\sum_{\ell=j} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]} + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}\sum_{\ell\neq j} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} (1_{[X_i \geq Y_j]})^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell\neq j} 1_{[X_i \geq Y_j]} 1_{[X_i \geq Y_\ell]} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}\sum_{\ell=j} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_j]} + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}^{n}\sum_{\ell\neq j}^{m} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]}
\end{aligned}$$

Now, the square of an indicator function is just the indicator function, so the expectation of the first term is

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} (1_{[X_i \geq Y_j]})^2\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} 1_{[X_i \geq Y_j]}\right] = E[U] = \frac{mn}{2}.$$

Then, owing to the exchangeability of the X_i's and Y_j's, we find the expectation of the second term is

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell\neq j} 1_{[X_i \geq Y_j]} 1_{[X_i \geq Y_\ell]}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell\neq j} E\left[1_{[X_i \geq Y_j]} 1_{[X_i \geq Y_\ell]}\right] = mn(m-1)E\left[1_{[X_1 \geq Y_1]} 1_{[X_1 \geq Y_2]}\right],$$

with

$$E\left[1_{[X_1 \geq Y_1]} 1_{[X_1 \geq Y_2]}\right] = Pr(X_1 \geq Y_1, X_1 \geq Y_2).$$

Now under the null hypothesis that X and Y have the same distribution and the independence of X_1, Y_1, and Y_2, this is just equal to the probability that the largest of three randomly chosen measurements would be the first measurement chosen. That is easily computed by considering all permutations of three distinct numbers. Each permutation should be equally likely. There are 6 such permutations, and 2 of those permutations have the largest value first, so

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell\neq j} 1_{[X_i \geq Y_j]} 1_{[X_i \geq Y_\ell]}\right] = \frac{nm(m-1)}{3}.$$

Similarly, the expectation of the third term is

$$E\left[\sum_{i=1}^{n}\sum j = 1^m \sum_{k\neq i} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_j]}\right] = \frac{nm(n-1)}{3}.$$

Owing to the exchangeability of the X_i's and Y_j's, we find the expectation of the fourth term is

$$E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}^{n}\sum_{\ell\neq j}^{m} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}^{n}\sum_{\ell\neq j}^{m} E\left[1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]}\right] = mn(n-1)(m-1)E\left[1_{[X_1 \geq Y_1]} 1_{[X_2 \geq Y_2]}\right],$$

and the independence of X_1, Y_1, X_2, and Y_2 yields

$$\begin{aligned}
E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k\neq i}^{n}\sum_{\ell\neq j}^{m} 1_{[X_i \geq Y_j]} 1_{[X_k \geq Y_\ell]}\right] &= mn(m-1)(n-1)E\left[1_{[X_1 \geq Y_1]}\right] E\left[1_{[X_2 \geq Y_2]}\right] \\
&= mn(m-1)(n-1)\left[Pr(X \geq Y)\right]^2 = \frac{mn(m-1)(n-1)}{4}.
\end{aligned}$$

We thus have

$$\begin{aligned}E[U^2] &= \frac{mn}{2} + \frac{mn(m-1)}{3} + \frac{mn(n-1)}{3} + \frac{mn(m-1)(n-1)}{4} \\ &= \frac{6mn + 4m^2n - 4mn + 4mn^2 - 4mn + 3m^2n^2 - 3mn^2 - 3m^2n + 3mn}{12} \\ &= \frac{3m^2n^2 + mn^2 + m^2n + mn}{12},\end{aligned}$$

so

$$Var(U) = E[U^2] - E^2[U] = \frac{3m^2n^2 + mn^2 + m^2n + mn}{12} - \frac{m^2n^2}{4} = \frac{mn(m+n+1)}{12}.$$

An alternative approach could have used the results from the rank sum null distribution. Then from the relationship between $U = R - n(n+1)/2$ we can find the moments for the distribution of U under the null hypothesis as (recall for random variable X and constant c, $E[X+c] = E[X] + c$ and $Var(X+c) = Var(X)$)

$$\begin{aligned}E[U] &= E[R] - \frac{n(n+1)}{2} = \frac{n(m+n+1)}{2} - \frac{n(n+1)}{2} = \frac{mn}{2} \\ Var[U] &= Var(R) = \frac{mn(m+n+1)}{12}\end{aligned}$$

6.3. Exact distribution and permutation tests

In small samples (i.e., when either m or n is small), we can find the distribution of R exactly by brute force: We can consider all the combinations of choosing n numbers out of the integers $1, 2, \ldots, m+n$, summing the numbers for each of those combinations, and then finding the percentiles by noting that each such combination is equally likely. In the absence of ties, the number of such combinations is known to be $(m+n)!/(m!n!)$, which can get big pretty quickly. Thus an alternative approach is by Monte Carlo methods. It should be obvious that the Wilcoxon rank sum test is nothing more than a permutation test based on the ranks. Hence the following S-plus (or R) function `simWilcoxonP()` given below would estimate the quantiles of the sampling distribution for R for data vectors x and y. I also had it estimate the upper one-sided P value for the test. Note that this case handles ties, because it permutes the possibly tied ranks of the observed data.

```
simWilcoxonP <- function (x, y, Nsim=10000,
        prob=c(.01,.025,.05,.1,.25,.5,.75,.9,.95,.975,.99)) {
    x <- x[!is.na(x)]
    y <- y[!is.na(y)]
    n <- length(x)
    ranks <- rank(c(x,y))
    R <- sum(ranks[1:n])
    N <- length(ranks)
    indx <- runif(Nsim * N)
    study <- rep(1:Nsim,rep(N,Nsim))
    indx <- as.vector (
        rep(1,n) %*% matrix( rep(ranks,Nsim)[order (study, indx)], N)[1:n,] )
    list (RankSum= R, Pval = sum(indx >= R)/Nsim, Pctile = quantile (indx, prob), Nsim=Nsim)
}
```

6.4. Large sample approximation to the sampling distribution under the strong null

Now, having the first two moments of the null sampling distribution for R is sufficient knowledge to construct hypothesis tests if R has a normal distribution. While R is the sum of identically distributed random variables, it is not the sum of *independent* random variables, and thus the usual central limit theorem will not work here. Instead, we would need to use the central limit theorem for sampling without replacement from a finite population (there does exist such a thing), which says that providing the number sampled n is sufficiently larger than 0 but sufficiently small relative to the size $m+n$ of the population, the sample average is approximately normally distributed. This in turn suggests that the sum of the ranks will tend to be normally distributed providing neither m nor n are too small. In that case, we expect

$$R \dot{\sim} \mathcal{N}\left(\frac{n(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right)$$

and a test statistic

$$T = \frac{R - \frac{n(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

will tend to have the standard normal distribution under the null hypothesis that the distributions of X and Y are the same. Thus a test could be constructed by comparing T to the percentiles of the standard normal distribution.

Using the Mann-Whitney formulation, we can also provide some intuitive motivation for the asymptotic distribution of U. First, the mean of the sampling distribution for U under any hypothesis is easily found to be

$$\begin{aligned} E[U] &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} 1_{[X_i \geq Y_j]}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} E\left[1_{[X_i \geq Y_j]}\right] \\ &= mnE\left[1_{[X_1 \geq Y_1]}\right] \\ &= mnPr(X \geq Y) \end{aligned}$$

where we use the fact that the X_i's are identically distributed and the Y_i's are identically distributed, as well as the fact that the expectation of a binary indicator variable is just the probability that the event measured by the indicator variable occurs.

It is clear that we can estimate $Pr(X > Y)$ using

$$\overline{U^*} = \frac{1}{\min(n,m)} \sum_{i=1}^{\min(n,m)} 1_{[X_i > Y_i]}$$

which is clearly based on independent Bernoulli random variables $W_i = 1_{X_i > Y_i} \sim \mathcal{B}(1, Pr(X \geq Y))$. We know that the sample mean $\overline{W} = \overline{U^*}$ has an asymptotically normal distribution

$$\overline{W} \dot{\sim} \mathcal{N}\left(Pr(X \geq Y), \frac{Pr(X \geq Y)\,(1 - Pr(X \geq Y))}{\min(n,m)}\right).$$

Then, because $\overline{U} = U/(nm)$ makes more efficient use of all of the data than $\overline{W}$ <u>and</u> weights all observations equally (thereby avoiding undue influence from any single observation), it seems reasonable that the U will be approximately normally distributed some tighter variance V

$$U \dot{\sim} \mathcal{N}\left(mnPr(X \geq Y), V\right).$$

In particular, $V = mn(m+n+1)/12 \leq mn/4$ under the strong null hypothesis.

The quantiles of the null distribution for $\overline{U}$ are thus

$$\frac{1}{2} + z_p\sqrt{\frac{(m+n+1)}{12mn}}$$

Thus to perform a level α two-sided test for equality of the distributions of X and Y (the strong null hypothesis), we might choose $p = 1 - \alpha/2$ and reject the null hypothesis when

$$\overline{U} < \frac{1}{2} - z_p\sqrt{\frac{(m+n+1)}{12mn}} \text{ or } \overline{U} > \frac{1}{2} + z_p\sqrt{\frac{(m+n+1)}{12mn}}$$

(recall that for the standard normal distribution, $z_p = -z_{1-p}$. This is equivalent to using the test statistic

$$T = \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}},$$

which is equivalent to the statistic defined above for the rank sum and can be compared to the quantiles of a standard normal distribution.

6.5. Large sample approximation to the sampling distribution under alternatives

More generally, we might want to use U to compute confidence intervals for $\theta = Pr(X \geq Y)$. Our first temptation might be to use the asymptotic distribution under the strong null to compute a $100(1-\alpha)\%$ confidence interval for θ as

$$\overline{U} \pm z_{\alpha/2}\sqrt{\frac{(m+n+1)}{12mn}}$$

The above formula assumes that the variance V of $\overline{U}$ does not change markedly as θ varies and that the sampling distribution under the strong null is relevant. However, as U is a sum of binary variables used to estimate a probability, we might expect that V will be of the general form

$$V = \frac{\theta(1-\theta)}{h(n,m,\theta)}$$

for some function h that will depend upon the exact shape of F and G under each possible value of θ. In fact, it is easy to show that $V = 0$ as θ approaches 0 or 1.

It is therefore also of interest to explicitly consider the variance of the sampling distribution for U under alternatives to the strong null hypothesis, i.e., under hypotheses in which the distributions for X and Y differ.

In our derivation of the variance of the Mann-Whitney U statistic, we found that the variance could be expressed in terms of the distribution of independent $X_1, X_2 \sim F$, $Y_1, Y_2 \sim G$. This can then be used to express the variance of $\overline{U}$ as

$$\begin{aligned} Var(\overline{U}) = & \frac{1}{mn}Pr(X_1 \geq Y_1) + \frac{(m-1)}{mn}Pr(X_1 \geq Y_1, X_1 \geq Y_2) \\ & + \frac{(n-1)}{mn}Pr(X_1 \geq Y_1, X_2 \geq Y_1) - \frac{(m+n-1)}{mn}\left[Pr(X_1 \geq Y_1)\right]^2. \end{aligned}$$

We thus see that we would need to know for each such alternative $\theta = Pr(X_1 \geq Y_1)$ the probability that a randomly chosen X might exceed the maximum of two independent observations of Y and the probability that the minimum of two independent observations of X might exceed a randomly chosen Y. Without knowing more about the shapes of the distributions, this will be difficult to express in general terms.

As suggested above, it is possible to estimate the variance of U under the true distributions for X and Y (which may or may not be the same distribution) using bootstrapping within each group separately (as opposed to using a permutation distribution). We will still be faced with the problem of knowing how the variance of U might change under different alternatives. This is necessary in order to construct confidence intervals.

We can put an upper bound on $Var(\overline{U})$ by noting that

$$Pr(X_1 \geq Y_1) \geq Pr(X_1 \geq Y_1, X_1 \geq Y_2) \qquad \text{and} \qquad Pr(X_1 \geq Y_1) \geq Pr(X_1 \geq Y_1, X_2 \geq Y_1).$$

Hence

$$Var(\overline{U}) \leq \frac{1}{mn}\theta + \frac{(m-1)}{mn}\theta + \frac{(n-1)}{mn}\theta - \frac{(m+n-1)}{mn}\theta^2 = \frac{m+n-1}{mn}\theta(1-\theta).$$

One particularly interesting alternative to the strong null is the case where the weak null might be true, but the strong null is not. In this case, $\theta = 0.5$, so the upper bound on the variance is

$$Var(\overline{U}) \leq \frac{m+n-1}{4mn} = \frac{1}{4n} + \frac{1}{4m} - \frac{1}{4mn}.$$

So then the question is whether any choices of F and G will attain the upper bound.

Consider, then, the distribution in which

$$\begin{aligned} Y &\sim G(y) = y1_{[0<y<1]} + 1_{[y\geq 1]} \\ X &\sim F(x) = (x+0.5)1_{[-0.5<x<0]} + (x-0.5)1_{[1<x<1.5]} + 1_{[x\geq 1.5]}. \end{aligned}$$

(So Y is uniformly distributed between 0 and 1, and X is with probability 0.5 uniformly distributed between -0.5 and 1 and with probability 0.5 uniformly distributed between 1 and 1.5.) Under these probability distributions, the event $[X_1 > Y_1]$ is exactly the same as the event $[X_1 > 1]$. Hence

$$\begin{aligned} Pr(X_1 \geq Y_1) &= Pr(X_1 > 1) = 0.5 \\ Pr(X_1 \geq Y_1, X_1 \geq Y_2) &= Pr(X_1 > 1) = 0.5 \\ Pr(X_1 \geq Y_1, X_2 \geq Y_1) &= Pr(X_1 > 1, X_2 > 1) = 0.25 \end{aligned}$$

In this setting, then,

$$\begin{aligned} Var(\overline{U}) &= \frac{1}{mn} Pr(X_1 \geq Y_1) + \frac{(m-1)}{mn} Pr(X_1 \geq Y_1, X_1 \geq Y_2) \\ &\quad + \frac{(n-1)}{mn} Pr(X_1 \geq Y_1, X_2 \geq Y_1) - \frac{(m+n-1)}{mn} \left[Pr(X_1 \geq Y_1) \right]^2 . \\ &= \frac{1}{2mn} + \frac{(m-1)}{2mn} + \frac{(n-1)}{4mn} - \frac{(m+n-1)}{4mn} \\ &= \frac{1}{4n} \end{aligned}$$

Note that as $\min(m,n) \rightarrow \infty$ with $m/n = r$ large, the upper bound on the variance of $\overline{U}$ under the weak null can be arbitrarily close to $1/(4n)$, so the upper bound on $Var(\overline{U})$ given above is a tight upper bound in general, though it may not be tight for arbitrary values of r.

6.5. Statistical properties of tests based on the Mann-Whitney U statistic

From the above results about the moments and sampling distribution we know

- $\overline{U}$ is an unbiased distribution-free (nonparametric) estimator of $\theta = Pr(X \geq Y)$.
- A test of the strong null $H_0 : F(x) = G(x),\ \forall x$ based on

$$\text{reject } H_0 \quad \Leftrightarrow \quad T = \frac{\overline{U} - 0.5}{\sqrt{\frac{m+n+1}{12mn}}} > z_{1-\alpha},$$

  where $z_{1-\alpha}$ is the upper α quantile of the standard normal distribution, is a one-sided level α test as $\min(m,n) \rightarrow \infty$.
- The above test based on T is not an unbiased test of the strong null hypothesis. (An unbiased test would always have $Pr(\text{reject } H_0 \,|\, F, G \notin H_0) > Pr(\text{reject } H_0 \,|\, F, G \notin H_0)$.) To see this, consider the setting described in the previous section in which

$$\begin{aligned} Y &\sim G(y) = y1_{[0<y<1]} + 1_{[y\geq 1]} \\ X &\sim F(x) = (x+0.5)1_{[-0.5<x<0]} + (x-0.5)1_{[1<x<1.5]} + 1_{[x\geq 1.5]}. \end{aligned}$$

  Clearly, these distributions do not satisfy the strong null distribution, because $F(x) \neq G(x)\ \forall x \neq 0.5$. In the previous section, we found that under these distributions $\overline{U} \dot{\sim} \mathcal{N}(\theta, 1/(4n))$ as $n \rightarrow \infty$. Hence, under these distributions with

$$\begin{aligned} Pr(T > z_{1-\alpha}) &= Pr\left(\frac{\overline{U} - 0.5}{\sqrt{\frac{m+n+1}{12mn}}} > z_{1-\alpha} \right) = Pr\left(\frac{\overline{U} - 0.5}{\sqrt{\frac{1}{4n}}} > \sqrt{\frac{m+n-1}{3m}} z_{1-\alpha} \right) \\ &= 1 - \Phi\left(\sqrt{\frac{m+n-1}{3m}} z_{1-\alpha} \right). \end{aligned}$$

If $n > 2m+1$, the probability of rejecting the null hypothesis is less than α. Hence, there exist some alternatives (and settings) for which the probability of rejecting the null is less than α.

- The above test based on T is not a consistent test of the strong null hypothesis. (A consistent test would have $Pr(\text{reject } H_0 \,|\, F, G \notin H_0) \to 1$ as $\min(m,n) \to \infty$.) To see this, consider again the setting described in the previous section in which $F \neq G$ and
$$Pr(T > z_{1-\alpha}) = 1 - \Phi\left(\sqrt{\frac{m+n-1}{3m}} z_{1-\alpha}\right).$$
If $n = 2m+1$, the probability of rejecting the null hypothesis is $\alpha < 1$, regardless of how large $\min(m,n)$ becomes. Hence, there exist some alternatives for which the probability of rejecting the null does not approach 1 asymptotically.

- The above test based on T is not a level α test of the weak null hypothesis $H_0 : \theta = 0.5$. To see this, consider again the setting described in the previous section in which $\theta = 0.5$ and
$$Pr(T > z_{1-\alpha}) = 1 - \Phi\left(\sqrt{\frac{m+n-1}{3m}} z_{1-\alpha}\right).$$
If $n \neq 2m+1$, the probability of rejecting the null hypothesis is not α. Note that the test is anti-conservative (has a type I error greater than α) if $n < 2m+1$, and it is conservative (has a type I error less than α if $n > 2m+1$.

- The above test based on T is a consistent test of the weak null hypothesis $H_0 : \theta = 0.5$ versus an upper alternative $H_1 : \theta > 0.5$. To see this, note that as $\min(m,n) \to \infty$, $\overline{U} \dot{\sim} \mathcal{N}(\theta, V)$ with
$$V \leq V_{bound} = \frac{1}{mn}\theta + \frac{(m-1)}{mn}\theta + \frac{(n-1)}{mn}\theta - \frac{(m+n-1)}{mn}\theta^2 = \frac{m+n-1}{mn}\theta(1-\theta).$$
So, as $\min(m,n) \to \infty$, $V_{bound} \to 0$, and thus $V \to 0$. Hence,
$$\begin{aligned}
Pr(T > z_{1-\alpha}) &= Pr\left(\frac{\overline{U} - 0.5}{\sqrt{\frac{m+n+1}{12mn}}} > z_{1-\alpha}\right) \\
&= Pr(\left(\frac{\overline{U} - \theta}{\sqrt{V}} > \sqrt{\frac{m+n+1}{12mnV}} z_{1-\alpha} - \frac{\theta - 0.5}{\sqrt{V}}\right) \\
&\doteq 1 - \Phi\left(\sqrt{\frac{m+n+1}{12mnV}} z_{1-\alpha} - \frac{\theta - 0.5}{\sqrt{V}}\right) \\
&\leq 1 - \Phi\left(\sqrt{\frac{m+n+1}{12mnV_{bound}}} z_{1-\alpha} - \frac{\theta - 0.5}{\sqrt{V}}\right) \\
&= 1 - \Phi\left(\sqrt{\frac{m+n+1}{12(m+n-1)\theta(1-\theta)}} z_{1-\alpha} - \frac{\theta - 0.5}{\sqrt{V}}\right) \to 1 \text{ as } V \to 0,
\end{aligned}$$
where $\Phi(x)$ is the cumulative distribution function for the standard normal.

7. Interpretation of test in terms of marginal distributions of X and Y

Many people are under the erroneous impression that the Wilcoxon rank sum test is somehow a nonparametric test of the median. This is not the case. Nor is it a nonparametric test of the mean. Instead for a two sample test of random variables X and Y, it is, as the Mann-Whitney form would suggest, a test of whether the $Pr(X > Y) > .5$ for independent randomly sampled X and Y. Note:

1. This will be true if X is “stochastically larger” than Y. X is “stochastically larger” than Y if $Pr(X > c) > Pr(Y > c)$ for all c. In such a setting we will also have that $E[X] > E[Y]$ and $mdn(X) > mdn(Y)$.

2. This can be true when X is not stochastically larger than Y. For instance, suppose $Y \sim \mathcal{U}(0,1)$ a uniform random variable and $X \sim \mathcal{N}(1,1)$ a normally distributed random variable. Then

    - $Pr(X > 0) = .84$ is less than $Pr(Y > 0) = 1$,
    - but

$$\begin{aligned}
Pr(X > Y) &= \int_0^1 Pr(X > Y \,|\, Y = u)\, du \\
&= \int_0^1 \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-1)^2}{2}\right\} dx\, du \\
&= \int_0^1 \int_{u-1}^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx\, du \\
&= \int_0^1 \int_{u-1}^0 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx\, du + 0.5 \\
&= \int_{-1}^0 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \int_0^{x+1} du\, dx + 0.5 \\
&= \int_{-1}^0 (x+1) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx + 0.5 \\
&= \int_{-1}^0 x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx + 0.8413 \\
&= \frac{1}{\sqrt{2\pi}}(-1 + e^{-0.5}) + 0.8413 = 0.6844
\end{aligned}$$

3. This can be true when the median of X is less than the median of Y. For instance, suppose that for some $a < b < c < d$

$$\begin{aligned}
Pr(Y < y) &= py1_{[0 \le y \le 1]} + p1_{y>1]} + (y-2)(1-p)1_{[2 \le y \le 3]} + (1-p)1_{[y>3]} \\
Pr(X < x) &= r\frac{(x-a)}{(b-a)}1_{[a \le x \le b]} + r1_{x>b]} + (x-c)(1-r)1_{[c \le x \le d]} + (1-r)1_{[x>d]}
\end{aligned}$$

   These distributions correspond to Y being uniformly distributed between 0 and 1 with probability p and uniformly distributed between 2 and 3 with probability $1-p$, and X being uniformly distributed between a and b with probability r and uniformly distributed between c and d with probability $1-r$.

    - The mean of Y is easily found to be $E[Y] = 0.5p + 2.5(1-p) = 2.5 - 2p$, and the mean of X is $E[X] = r(a+b)/2 + (1-r)(c+d)/2$.
    - The median of Y is $mdn(Y) = 0.5/p$ if $p > 0.5$ and $mdn(Y) = 2 + (0.5-p)/(1-p)$ if $p < 0.5$. The median of X is $mdn(X) = a + (0.5/r)(b-a)$ if $r > 0.5$ and $mdn(X) = c + (0.5-r)(d-c)/(1-r)$ if $r < 0.5$.

   Now suppose that we take $1 \le a < b \le 2$ and $3 \le c$ and $p = 0.4$ and $r = 0.7$. Then

    - Neither X nor Y is stochastically larger than the other, because $Pr(X > 1) = 1 > Pr(Y > 1) = 1 - p = .6$, but $Pr(X > 2) = 1 - r = .3 < Pr(Y > 2) = .6$.
    - With a sufficiently large sample size, the Wilcoxon test would suggest that X tends to be larger than Y, because

$$Pr(X > Y) = Pr(X > 1)Pr(Y < 1) + Pr(X > 3)Pr(Y > 2) = p + (1-r)(1-p) = 1 - r + rp,$$

      which for choices $p = 0.4$ and $r = 0.7$ yields $Pr(Y > X) = 0.58$

- The median of Y is $mdn(Y) = 2.167$. Then, for appropriate choices of $1 \leq a < b \leq 2$, we can make $mdn(X)$ arbitrarily close to any number between 1 and 2. In particular, if we choose $a = 1$ and $b = 2$, $mdn(X) = 1.714$. So with a sufficiently large sample size, using the differences in medians would tend to suggest that Y tends to be larger than X, because $mdn(Y) = 2.167 > mdn(X) = 1.714$.
- The mean of Y is $E[Y] = 1.7$. For the choices $r = 0.7$, $a = 1$, $b = 2$, and $c \geq 3$, $E[X] = 1.05 + 0.3 \times (c + d)/2 > 1.95$. So with a sufficiently large sample size, using the differences in means would tend to suggest that X tends to be larger than Y.

So clearly the Wilcoxon test cannot in general be interpreted as evidence about the medians.

4. The Wilcoxon can also tend to suggest that X tends to be larger than Y when $E[X] < E[Y]$. We use the same distribution as above, with $r = 0.7$, $p = 0.4$, $a = 1$, $b = 1.01$, $c = 3$, and $d = 3.01$. Then

   - Neither X nor Y is stochastically larger than the other, because $Pr(X > 1) = 1 > Pr(Y > 1) = 1 - p = .6$, but $Pr(X > 2) = 1 - r = .3 < Pr(Y > 2) = .6$.
   - With a sufficiently large sample size, the Wilcoxon test would suggest that X tends to be larger than Y, because

   $$Pr(X > Y) = Pr(X > 1)Pr(Y < 1) + Pr(X > 3)Pr(Y > 2) = p + (1 - r)(1 - p) = 1 - r + rp,$$

   which for choices $p = 0.4$ and $r = 0.7$ yields $Pr(Y > X) = 0.58$

   - The median of Y is $mdn(Y) = 2.167$, and the median of X is $mdn(X) = 1.007$. So with a sufficiently large sample size, using the differences in medians would tend to suggest that Y tends to be larger than X, because $mdn(Y) = 2.167 > mdn(X) = 1.007$.
   - The mean of Y is $E[Y] = 1.7$, and $E[X] = 1.605$. So with a sufficiently large sample size, using the differences in means would tend to suggest that Y tends to be larger than X.

   So clearly the Wilcoxon test cannot in general be interpreted as evidence about the means.

5. It should be noted that it is possible to also find settings in which $Pr(X \geq Y) > 0.5$, $mdn(X) > mdn(Y)$, and $E(X) < E(Y)$. That is, it is possible to find distributions that match any pattern of concordance or discordance among these three functionals with respect to the implied ordering of distributions.

8. Intransitivity of $Pr(X \geq Y)$

The bivariate functional $\theta = Pr(X \geq Y)$ can be shown to be intransitive. That is, given $X \sim F$, $Y \sim G$, and $W \sim H$, defined by

$$\begin{aligned} Y &\sim G(y) = y1_{[0<y<1]} + 1_{[y\geq 1]} \\ X &\sim F(x) = (x+2)1_{[-2<x<-1.6]} + (x-0.6)1_{[1<x<1.6]} + 1_{[x\geq 1.6]} \\ W &\sim H(w) = (w+1.6)1_{[-1.6<w<-1]} + (w-1)1_{[1.6<w<2]} + 1_{[w\geq 2]}. \end{aligned}$$

Then we have that

- $Pr(X \geq Y) = 0.6 > 0.5$ (implying X tends to be larger than Y)
- $Pr(Y \geq W) = 0.6 > 0.5$ (implying Y tends to be larger than W)
- $Pr(W \geq X) = 0.64 > 0.5$ (implying W tends to be larger than X)

9. Some results on relative efficiency of the Wilcoxon rank sum and t tests

Many authors have reported on the efficiency of the Wilcoxon rank sum test relative to efficiency of the t test. Unfortunately, these results are almost always presented within the context of a parametric or semiparametric probability model. What makes it worse is that those results are almost always based on the location shift semiparametric model. Hence, those generally impressive efficiency results do not necessarily generalize to fully nonparametric settings.

In considering the relative merits of the Wilcoxon rank sum test and the t test, it should be noted that we are comparing nonparametric (distribution-free) estimators of $\theta_W = Pr(X \geq Y)$ and $\theta_T = E[X] - E[Y]$. For some families, these estimators may also be the parametric efficient estimates of those functionals.

The following subsections use particular parametric models to compare the estimated statistical power of the Wilcoxon and the t test that allows unequal variances to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. (Estimates are based on simulations.) The relative efficiency can be thought of as the proportionate decrease or increase in the sample size for a t test that would provide the same power as the Wilcoxon test that had 100 subjects in each treatment arm. Hence, a relative efficiency of 0.95 suggests that a t test with 95 subjects per group would have the same power as a Wilcoxon test that had 100 subjects per group. A relative efficiency of of 1.24 suggests that a t test with 124 subjects per group would have the same power as a Wilcoxon test that had 100 subjects per group.

9.1. Normal distribution with homoscedasticity

We consider a parametric family in which (without loss of generality) $X \sim \mathcal{N}(\theta, 1)$ and $Y \sim \mathcal{N}(0, 1)$. This parametric family is a subset of a location shift semiparametric family.

In this family, the efficient estimator of θ_T is $\hat{\theta}_T = \overline{X} - \overline{Y}$. Hence, the t test that presumes equal variances will be the optimal inferential strategy, and in this balanced setting ($m = n = 100$) the t test that allows for the possibility of unequal variances will be essentially equivalent. The efficient estimator of θ_W would be the parametric estimator based on $Pr(\mathcal{N}(\overline{X}, s_X^2) > \mathcal{N}(\overline{Y}, s_Y^2))$. The distribution-free estimator $\overline{U}$ will not therefore be efficient.

The following table provides estimates of the statistical power of the Wilcoxon and t tests to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. As can be seen from these data (which agree well with, for instance, Lehmann's *Nonparametrics: Statistical Methods Based on Ranks*), the Wilcoxon rank sum test is approximately 90-95% efficient in this parametric model.

Table 9.1
Power and Relative Efficiency of Wilcoxon and t Tests
in a Parametric Normal Location Shift Model

		Power to Detect Alternative		Relative
$\theta_W = Pr(X \geq Y)$	$\theta_T = E[X] - E[Y]$	Wilcoxon	t Test	Efficiency
0.500	0.00	0.026	0.027	NA
0.529	0.10	0.104	0.103	1.021
0.556	0.20	0.279	0.292	0.943
0.585	0.30	0.549	0.560	0.975
0.611	0.40	0.780	0.806	0.937
0.639	0.50	0.934	0.944	0.955
0.663	0.60	0.985	0.989	0.947

9.2. Exponential distribution

We consider a parametric family in which (without loss of generality) $X \sim \mathcal{E}(\theta)$ and $Y \sim \mathcal{E}(1)$, where we have parameterized the exponential distribution such that $E[X] = \theta$ and $E[Y] = 1$. This parametric family is a subset of both the accelerated failure time (scale) and the proportional hazards semiparametric families.

In this family, the efficient estimator of θ_T is $\hat{\theta}_T = \overline{X} - \overline{Y}$. The distribution of the sample means would be related to a gamma distribution, but owing to the central limit theorem, $\hat{\theta}_T$ is approximately normally distributed with a variance that depends upon θ and the distribution of Y. The t test that allows for the possibility of unequal variances would be the typical choice here, but it will not be the most efficient choice, because it does not explicitly consider the mean-variance relationship. (The most efficient test of the strong null would use s_Y^2 for both groups, as it would estimate the correct within group variance under the null hypothesis.)

The efficient estimator of θ_W would be the parametric estimator based on $Pr(\mathcal{E}(\overline{X}) > \mathcal{E}(\overline{Y}))$. The distribution-free estimator $\overline{U}$ will not therefore be efficient.

The following table provides estimates of the statistical power of the Wilcoxon and t tests to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. As can be seen from these data, the Wilcoxon is less efficient (70% to 80%) than the t test in this parametric model. (I note that if in the t statistic we use an estimated standard error of $s_Y\sqrt{1/n + 1/m}$ instead of $\sqrt{s_X^2/n + s_Y^2/m}$, the efficiency advantage of the t test is even more pronounced: The Wilcoxon is only about 50% as efficient as a test of means.)

Table 9.2
Power and Relative Efficiency of Wilcoxon and t Tests
in a Parametric Exponential Scale Model

		Power to Detect Alternative		Relative
$\theta_W = Pr(X \geq Y)$	$\theta_T = E[X] - E[Y]$	Wilcoxon	t Test	Efficiency
0.501	0.00	0.026	0.024	NA
0.526	0.10	0.091	0.111	0.722
0.556	0.30	0.274	0.355	0.733
0.588	0.40	0.580	0.691	0.773
0.625	0.70	0.873	0.932	0.809
0.667	1.00	0.989	0.994	0.894
0.714	1.50	1.000	1.000	1.032

The results presented in the above table are seemingly at odds with the 3-fold greater efficiency reported for the Wilcoxon test over the t test reported in Lehmann's *Nonparametrics: Statistical Methods Based on Ranks*). And one has to wonder at Lehmann's results, given the optimality of the sample mean in the exponential distribution and the implications of the central limit theorem.

The seeming paradox is resolved by closer examination of the setting in which Lehmann examined the exponential distribution: He considered a location shift semiparametric model in which $Y \sim \mathcal{E}(1)$ and $X - \theta \sim Y$. This setting is considered in the next section.

9.3. Shifted exponential distribution

We consider a parametric family in which (without loss of generality) $Y \sim \mathcal{E}(1)$ and $X - \theta \sim \mathcal{E}(1)$, where we have parameterized the exponential distribution such that $E[X] = \theta$ and $E[Y] = 1$. This parametric family is called a "shifted exponential" and is a subset of location shift semiparametric family. (In the most general case of shifted exponential family, there are two parameters: a shift and a scale. However, we are trying to duplicate Lehmann's results, and he used a simple location shift model.)

In this family, the maximum likelihood estimator of θ_T is the difference of the sample minima $\hat{\theta}_{MLE} = X_{(1)} - Y_{(1)} = \min\{X_1, \ldots, X_n\} - \min\{Y_1, \ldots, Y_n\}$. Owing to the changing support of the distributions as θ varies (that is, the set of possible values of X depends upon θ), the asymptotic results about efficiency of maximum likelihood estimators does not apply to this problem. Nevertheless, it is easily shown that $\hat{\theta}_{MLE}$ is an n-consistent, with $n(\hat{\theta}_{MLE} - \theta) \sim \mathcal{E}(1)$ (note that this "asymptotic" distribution is actually exact).

Hence, the difference in sample means $\hat{\theta}_T = \overline{X} - \overline{Y}$ is a highly inefficient estimate of θ. The distribution of the sample means would be related to a gamma distribution, but owing to the central limit theorem, $\hat{\theta}_T$ is approximately normally distributed with a variance that depends only upon the sample sizes and the distribution of Y. The t test that allows for the possibility of unequal variances would be the typical choice here, but it will not be the most efficient choice, because under this location shift model, the variances are equal.

The efficient estimator of $\theta_W = Pr(X > Y) = 1 - 0.5e^{-\theta}$ would be the parametric estimator $\hat{\theta}_{Wmle} = 1 - 0.5e^{-\hat{\theta}_{MLE}}$. The distribution-free estimator $\hat{\theta}_W = \overline{U}$ will not therefore be efficient.

The following table provides estimates of the statistical power of the Wilcoxon and t tests to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. As can be seen from these data, the Wilcoxon is more efficient (2 to 2.5 fold) than the t test in this parametric model. However, both of these statistics are exceedingly inefficient to detect a difference in distributions within this parametric family. Using a test based on the maximum likelihood estimator, a difference in means of 0.10 could be detected with approximately 25% power with a sample size of 28 in each group (compare the Wilcoxon test's power of 22% with the sample size of 100 in each group) and with approximately 11% power with a sample size of 26 in each group (compare the t test's power of 11% with a sample size of 100 in each group). As Lehmann acknowledges, the relevance of comparing the Wilcoxon to the t test in this parametric family in practice is highly questionable.

Table 9.3
Power and Relative Efficiency of Wilcoxon and t Tests
in a Parametric Shifted Exponential Model

		Power to Detect Alternative		Relative
$\theta_W = Pr(X \geq Y)$	$\theta_T = E[X] - E[Y]$	Wilcoxon	t Test	Efficiency
0.501	0.00	0.026	0.025	NA
0.548	0.10	0.217	0.111	2.525
0.590	0.20	0.597	0.291	2.444
0.630	0.30	0.898	0.575	2.253
0.665	0.40	0.987	0.810	2.160
0.697	0.50	0.999	0.941	2.089
0.728	0.60	1.000	0.990	1.990

9.4. Lognormal distribution

We consider a parametric family in which (without loss of generality) $X \sim \mathcal{LN}(\omega, 1)$ (so $\log(X) \sim \mathcal{N}(\omega, 1)$) and $Y \sim \mathcal{LN}(0, 1)$ (so $\log(Y) \sim \mathcal{N}(0, 1)$). This parametric family is a subset of the accelerated failure time (scale) semiparametric family.

In this family, the efficient estimator of $\theta_T = e^{\omega + 0.5} - e^{0.5}$ is $\hat{\theta}_T = \exp\left\{\overline{log(X)} + s_X^2/2\right\} - \exp\left\{\overline{log(Y)} + s_Y^2/2\right\}$. Hence, the nonparametric estimator $\hat{\theta}_T$ is not efficient. Owing to the central limit theorem, $\hat{\theta}_T$ is approximately normally distributed with a variance that depends upon θ and the distribution of Y. However, because these lognormal distributions are heavily skewed, the approximation provided by the CLT is not good in very small samples.

The efficient estimator of θ_W would be the parametric estimator based on $Pr(\mathcal{LN}(\overline{X}, s_X^2) > \mathcal{LN}(\overline{Y}, s_Y^2))$. The distribution-free estimator $\overline{U}$ will not therefore be efficient.

The following table provides estimates of the statistical power of the Wilcoxon and t tests to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. As can be seen from these data, the Wilcoxon is much more efficient than the t test in this parametric model. This is in keeping with the oft quoted statement that the Wilcoxon will out perform the t test in distributions with heavy tails.

Table 9.4
Power and Relative Efficiency of Wilcoxon and t Tests
in a Parametric Lognormal Scale Model

		Power to Detect Alternative		Relative

$\theta_W = Pr(X \geq Y)$	$\theta_T = E[X] - E[Y]$	Wilcoxon	t Test	Efficiency
0.499	0.00	0.023	0.023	NA
0.529	0.20	0.103	0.082	1.510
0.558	0.40	0.289	0.187	1.720
0.585	0.60	0.547	0.366	1.649
0.612	0.80	0.788	0.552	1.742
0.637	1.10	0.929	0.724	1.798
0.665	1.40	0.987	0.861	1.903

9.5. Shifted t distributions

We can further explore the effect of heavy-tailed distributions within the family of shifted t distributions. t distributions are parameterized by a parameter k measuring the degrees of freedom. As $k \rightarrow \infty$, the t distribution converges to a standard normal distribution, in which setting the Wilcoxon test was found to be approximately 90-95% efficient. The case $k = 1$ corresponds to a Cauchy distribution, which is of particular interest because it has no mean. Similarly, the case of a t distribution with $k = 2$ has no variance. In these two cases, the t test is not asymptotically valid, but the Wilcoxon test is. Hence, the relative efficiency of the Wilcoxon test is infinite for these two lowest values of k.

We explore a few other t distributions below. We consider a parametric family in which (without loss of generality) $Y \sim t(k)$ for a specific value of $k > 2$ and $X - \theta \sim Y$. This parametric family is a subset of the location shift semiparametric family.

The following table provides estimates of the statistical power of the Wilcoxon and t tests to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. As can be seen from these data, the Wilcoxon is much more efficient (1.8 - 2 fold) than the t test in this parametric model with heavy tails ($k = 3$), approximately 10% more efficient with moderately heavy tails ($k = 7$), and approximately equally efficient when $k = 19$.

Table 9.5
Power and Relative Efficiency of Wilcoxon and t Tests
in a Parametric Shifted t Model

		Power to Detect Alternative		Relative
$\theta_W = Pr(X \geq Y)$	$\theta_T = E[X] - E[Y]$	Wilcoxon	t Test	Efficiency
		$k = 3$		
0.501	0.00	0.026	0.025	NA
0.547	0.20	0.208	0.129	1.912
0.591	0.40	0.607	0.379	1.823
0.633	0.60	0.910	0.686	1.822
0.675	0.80	0.994	0.890	1.952
0.713	1.00	1.000	0.983	1.866
0.748	1.20	1.000	0.999	1.829
		$k = 7$		
0.500	0.00	0.025	0.025	NA
0.525	0.10	0.090	0.083	1.158
0.551	0.20	0.235	0.224	1.061
0.574	0.30	0.440	0.426	1.040
0.600	0.40	0.688	0.646	1.101
0.625	0.50	0.873	0.847	1.081
0.653	0.60	0.970	0.954	1.115

$k = 19$

0.500	0.00	0.026	0.026	NA
0.528	0.10	0.099	0.105	0.915
0.556	0.20	0.274	0.275	0.995
0.579	0.30	0.489	0.508	0.951
0.609	0.40	0.763	0.766	0.993
0.634	0.50	0.915	0.918	0.989
0.659	0.60	0.980	0.980	1.004

9.6. Mixture distributions

In the preceding sections, we have explored the relative efficiency of the Wilcoxon and t tests under several simple shift alternatives. In the normal and shifted t probability models, the support of the distribution was $(-\infty, \infty)$ for all alternatives. We found that relative to the t test, the Wilcoxon test was 90-95% efficient for the normal distribution, with increasing relative efficiency as the heaviness of the tails increased. A t distribution with $k = 19$ degrees of freedom had the t test and Wilcoxon test approximately equally efficient, and a t distribution with $k = 3$ degrees of freedom had the Wilcoxon approximately twice as efficient. With $k = 1$ or 2, the Wilcoxon is infinitely more efficient than the t test, because those heavy tailed distributions have no variance.

We also explored a shifted exponential distribution in which the support of the distribution varies with the value of $\theta_T = E[X] - E[Y]$. In that parametric family of distributions, the Wilcoxon was found to be 2 - 2.5 times more efficient than the t test. We can also consider the effect that heavier tails has on the relative efficiency of these two tests in the setting of changing support. We observe the same trend of increased relative efficiency of the Wilcoxon as the tails become increasingly heavy, but with the Wilcoxon having the advantage with lighter tails when the support changes that it does under common support:

- In the setting of a uniform distribution (which distribution has lighter tails than the normal distribution) with $Y \sim \mathcal{U}(0, 1)$ and $X - \theta \sim Y$, the Wilcoxon is 90-95% as efficient as the t test.
- In the setting of a shifted folded normal distribution (so $Y \sim |\mathcal{N}(0, 1)|$, the absolute value of a standard normal random value and $X - \theta \sim Y$, the Wilcoxon is 1.2 - 1.3 times as efficient as the t test.
- In the setting of a shifted folded t distribution with $k = 19$ degrees of freedom (so $Y \sim |t(k)|$, the absolute value of a t distributed random value and $X - \theta \sim Y$, the Wilcoxon is 1.3 - 1.4 times as efficient as the t test.
- In the setting of a shifted folded t distribution with $k = 7$ degrees of freedom (so $Y \sim |t(k)|$, the absolute value of a t distributed random value and $X - \theta \sim Y$, the Wilcoxon is 1.5 - 1.7 times as efficient as the t test.
- In the setting of a shifted folded t distribution with $k = 3$ degrees of freedom (so $Y \sim |t(k)|$, the absolute value of a t distributed random value and $X - \theta \sim Y$, the Wilcoxon is 3 - 3.5 times as efficient as the t test.

We also explored two probability models in the family of accelerated failure time models. With the exponential distribution, the Wilcoxon was approximately 70-90% as efficient as the t test, while in the more heavily skewed lognormal distribution, the Wilcoxon test was 1.5 - 1.9 times more efficient than the t test.

It is also of interest to explore some parametric models mimicking a scientific setting in which only patients in a nonidentifiable subset are susceptible to the effects of the treatment. We thus consider a model in which $Y \sim \mathcal{N}(0, 1)$ and the distribution of X depends upon what the corresponding individuals (counterfactual) value of Y would have been. That is, we consider distributional parameters (π, η, ω) and latent normal random variable $Z_i \sim \mathcal{N}(0, 1)$ and latent Bernoulli random variable $W_i \sim \mathcal{B}(1, \pi)$. We then let $X_i = Z_i + \omega 1_{[\Phi(Z_i)>\eta]} 1_{[W_1=1]}$. (A corresponding untreated patient would have $Y_i = Z_i$.)

This mimics the setting in which the patients from the population having the lowest $100\eta\%$ values of Z_i receive no benefit of treatment (so η models non-susceptibility to the treatment that is related to severity of disease), while the patients in the upper $100(1 - \eta)$ percentile of the distribution have a benefit ω of treatment with probability π (so π models susceptibility to the treatment that is unrelated to the counterfactual value of outcome in the absence of treatment). Using such a model thus constitutes a mixture of parametric distributions.

The following table provides estimates of the statistical power of the Wilcoxon and t tests to detect various alternatives, as well as an estimate of the relative efficiency of the two tests under those alternatives. As can be seen from these data, the Wilcoxon is less efficient than the t test in these particular mixture models considered here. It should be noted that for fixed sample size in these mixture models, the Wilcoxon hits an upper bound on the possible power no matter how the mean changes. This is because the mixture model places an upper bound on the magnitude of $\theta_W = Pr(X \geq Y)$.

Table 9.6
Power and Relative Efficiency of Wilcoxon and t Tests
in a Parametric Mixture Model

		Power to Detect Alternative		Relative
$\theta_W = Pr(X \geq Y)$	$\theta_T = E[X] - E[Y]$	Wilcoxon	t Test	Efficiency
Complete non-susceptibility $\eta = 0$, Probability of benefit among remainder $\pi = 0.5$				
0.500	0.00	0.024	0.024	NA
0.529	0.10	0.104	0.111	0.898
0.557	0.20	0.280	0.284	0.983
0.582	0.30	0.516	0.544	0.934
0.607	0.40	0.750	0.770	0.954
0.628	0.50	0.889	0.911	0.925
0.650	0.60	0.965	0.973	0.936
Complete non-susceptibility $\eta = 0$, Probability of benefit among remainder $\pi = 0.25$				
0.503	0.00	0.028	0.027	NA
0.534	0.12	0.129	0.137	0.914
0.565	0.25	0.349	0.389	0.878
0.589	0.38	0.585	0.673	0.814
0.607	0.50	0.748	0.860	0.746
0.614	0.62	0.803	0.937	0.649
0.624	0.76	0.869	0.977	0.605
Complete non-susceptibility $\eta = 0.5$, Probability of benefit among remainder $\pi = 0.5$				
0.502	0.01	0.027	0.028	NA
0.524	0.10	0.083	0.095	0.784
0.541	0.20	0.172	0.252	0.616
0.552	0.30	0.249	0.439	0.505
0.560	0.41	0.313	0.633	0.410
0.561	0.50	0.316	0.755	0.313
0.560	0.55	0.316	0.802	0.278
Complete non-susceptibility $\eta = 0.5$, Probability of benefit among remainder $\pi = 0.25$				
0.500	0.00	0.025	0.025	NA
0.522	0.12	0.080	0.121	0.492
0.532	0.25	0.120	0.320	0.276
0.530	0.37	0.109	0.501	0.137
0.531	0.50	0.115	0.647	0.106
0.531	0.62	0.116	0.740	0.086
0.531	0.75	0.117	0.803	0.075
0.531	0.88	0.115	0.843	0.066

It should be noted that the results presented in this section are related to results that have been explored in the setting of weighted logrank statistics and censored time to event analyses. The Wilcoxon form of the logrank statistic is well-known to have greater power than the logrank statistic under alternatives that lead to "early differences" survival distributions. Hence, those models that had varying support corresponded to such early differences, and the mixture models in this section tended to lead to "late differences", especially when the probability of complete non-susceptibility was $\eta = 0.5$. The unweighted logrank statistic is generally preferred to the Wilcoxon from in the setting of "late differences" in the survival curves.

In the setting of time to event analyses, the difference in means corresponds to the area between to survival curves.

10. Implications for Parametric and Semiparametric Analyses

In the preceding sections, I have criticized the evaluation and use of the Wilcoxon rank sum statistic on multiple grounds:

- (Science) The functional of the distribution that the Wilcoxon consistently tests, $\theta_W = Pr(X \geq Y)$, does not provide any information about the scientific or clinical importance of the magnitude of differences in outcomes.
    - Arguably, the same could be also be said about such functionals as mean ratios, median ratios, odds ratios, and hazard ratios, as the scientific importance might be more related to differences of univariate functionals.
- (Science) The functional of the distribution that the Wilcoxon consistently tests, $\theta_W = Pr(X \geq Y)$, does not provide a transitive ordering of populations.
    - This property is shared by all functionals that cannot be expressed as a contrast of univariate functionals. Hence, the median difference (e.g., sign test) or mean ratio of paired observations, the maximal distance between two cumulative distribution functions (e.g., Kolmogorov-Smirnov test), and even the usual computation of a hazard ratio (due to the weighting of estimated hazards at observed failure times in Cox proportional hazards) can be shown to also be intransitive.
- (Statistics) The null sampling distribution is computed under the strong null hypothesis. The resulting test is neither unbiased nor consistent as a test of the strong null in a distribution-free sense.
    - This drawback is shared, at least in part, by any inference about the strong null hypothesis using a statistic that is consistent for a particular functional of the probability distributions, if that null value does not uniquely indicate the strong null hypothesis in a distribution-free environment. Most commonly used parametric and semiparametric analysis models are based on some functional θ and define a null value such that when $F = G$ within the presumed distributional family, $\theta = \theta_0$. Furthermore, within the parametric or semiparametric family, there is a constraint that if $\theta = \theta_0$, then $F = G$. Yet in most such analysis models, outside the parametric or semiparametric model, $\theta = \theta_0$ does not necessarily imply $F = G$. For instance, the t test that presumes equal variances will asymptotically reject the null hypothesis with probability equal to the type I error whenever the means and variances are equal across the two groups. It is of course trivial to find $X \sim F$ and $Y \sim G$ such that $E[X] = E[Y]$, $Var(X) = Var(Y)$, but $F(x) \neq G(x)$ for some x. (The two-sample parametric binomial probability model is a notable exception to this drawback, because it is a one parameter family and the sum of independent binary variables must be binomial.)
- (Statistics) The null sampling distribution is computed under the strong null hypothesis. The resulting test is not necessarily of the right size under distributions satisfying the weak null hypothesis.
    - Again, this drawback is shared, at least in part, by most of the commonly used parametric and semiparametric analysis models. For instance, the t test that presumes equal variances is asymptotically of the nominal level under the weak null only if either the variances are equal or the sample sizes in the two groups are equal.
- (Science) In light of the two previous results, rejection of the null hypothesis using the Wilcoxon statistic can only validly be interpreted as a difference in distributions, not as a difference in location.
- (Statistics) Because the operating characteristics of the Wilcoxon statistic have generally been evaluated in restrictive parametric and semiparametric settings, the generalization of those efficiency results are not clear.

I note that it is frequently the case that the historical development of useful statistics has involved a derivation of the statistic in the confines of a parametric or semiparametric model, and then the evaluation of the robustness of the inference in a distribution-free setting. Often that further evaluation leads to relatively minor modifications of the statistic that yields valid, unbiased, and consistent testing of a weak null hypothesis. Such tests can then often be inverted to obtain robust confidence intervals for a scientifically meaningful functional of the probability distribution.