WILCOXON RANK-SUM TEST

The Wilcoxon rank sum test is the quintessential classic nonparametric test. These notes study it in detail; other rank-based methods develop similarly.

SETUP:

Suppose we collect random samples from "control" and "treatment" populations:

$$X_1, \dots, X_m \overset{iid}{\sim} F \qquad \text{"control"}$$

$$Y_1, \dots, Y_n \overset{iid}{\sim} G \qquad \text{"treatment"},$$

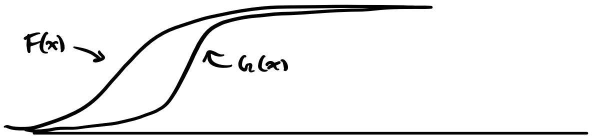
with $N = m+n$ the total number of observations.

It may be that we draw N subjects from a single population and randomly assign n to the treatment group and m to the control group.

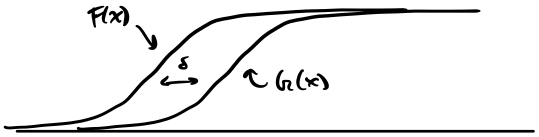
The Wilcoxon rank sum test was contrived as a way to test the effectiveness of a treatment. For our development, suppose the treatment is effective if it tends to increase the measured outcomes — that is if it tends to make the Y_i's bigger than the X_i's.

There are various senses in which a treatment could "tend to make the Y_i's bigger than the X_i's". For example, in terms of the cdfs F and G, we could have:

(i) $G(x) \le F(x)$ $\forall x$, i.e. Y_i is <u>stochastically greater</u> than X_i.



(ii) $G(x) = F(x-\delta)$ $\forall x$, i.e. F and G differ by a <u>location shift</u>.



The Wilcoxon rank sum test tests the null hypothesis

$$H_0: \; F = G.$$

The types of differences between F and G in (i) and (ii) represent different alternate states — different ways in which the null hypothesis could be false. We will consider them later when we study the power of the Wilcoxon rank sum test.

THE TEST:

To test whether the treatment tends to increase the measured outcome, the Wilcoxon rank sum test prescribes rejecting $H_0: F = G$ when the test statistic

$$W_{XY} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(X_i < Y_j)$$

is large, i.e. when $W_{XY} \geq c$ for some c.

The critical value c can be chosen to bound the Type I error rate.

Why is this considered a <u>nonparametric</u> test?

In classical nonparametric literature, the terms "nonparametric" and "distribution-free" were used more or less interchangeably. The term "distribution-free" meant "free of distributional assumptions." The term applies to the Wilcoxon rank-sum test because we can exactly find the distribution of the test statistic W_{XY} <u>without</u> making any assumptions whatsoever about the distribution F (which is equal to G under H_0). Therefore, we do not need to assume anything (like Normality, existence of moments, etc.) about the population distributions in order to trust the test. For this reason it belongs to the classic nonparametric battery of tests.

The test statistic W_{XY} counts the number of (X_i, Y_j) pairs such that $X_i < Y_j$.

The more effective the treatment at increasing the Y_i's, the greater we expect this number to be, so we reject the null hypothesis of ineffectiveness when W_{XY} exceeds a certain threshold.

We can also write W_{XY} in "rank-sum" form as follows:

(i) Sort the set of all the data $(X_1, \dots, X_m, Y_1, \dots, Y_n)$, (assume for now no ties).

(ii) obtain the ranks.

(iii) Keep the ranks corresponding to $Y_1, \dots, Y_n$; denote these by $S_1, \dots, S_n$.

<u>Example:</u> Suppose $(X_1, X_2, X_3) = (0.5, 2.0, 0.75)$, $(Y_1, Y_2) = (0.9, 3.0)$.

Sorting all the data and assigning ranks gives

data point	0.5	0.75	0.9	2.0	3.0
rank	1	2	3	4	5

So $(S_1, S_2) = (3, 5)$.

We find that $W_{XY} = S_1 + \dots + S_n - \frac{1}{2}n(n+1)$, by

$$
\begin{aligned}
W_{XY} &= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(X_i < Y_j) \\
&= \sum_{j=1}^{n} \# \{ X_i < Y_j \} \\
&= \sum_{j=1}^{n} \# \{ X_i < Y_{(j)} \} \\
&= \sum_{j=1}^{n} \Big[\underbrace{\# \{ X_i < Y_{(j)} \} + \# \{ Y_i < Y_{(j)} \}}_{S_j - 1} - \underbrace{\# \{ Y_i < Y_{(j)} \}}_{j-1} \Big] \\
&= \sum_{j=1}^{n} \Big[(S_j - 1) + (j - 1) \Big] \\
&= S_1 + \dots + S_n - \frac{1}{2} n(n+1).
\end{aligned}
$$

It will be convenient to define $W_S = S_1 + \dots + S_n$.

Note that the smallest possible value of $W_S = S_1 + \dots + S_n$ occurs when $(S_1, \dots, S_n) = (1, \dots, n)$, in which case $S_1 + \dots + S_n = \frac{1}{2}n(n+1)$.

A larger sum of ranks $S_1 + \dots + S_n$ casts greater doubt on $H_0: F = G$, as it indicates a tendency for the Y_i's to be higher than the X_i's.

NULL DISTRIBUTION OF W_{XY}:

so that ties occur w/ prob 0

Result: Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ be iid from the same continuous distribution. Then

$$P(\{S_1, \dots, S_n\} = \{s_1, \dots, s_n\}) = \frac{1}{\binom{N}{n}} \text{ for all sets of } n \text{ ranks } \{s_1, \dots, s_n\} \subset \{1, \dots, N\}.$$

Proof: Let $Z_1, \dots, Z_N$ denote the combined $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$.

Sort $Z_1, \dots, Z_N$ to obtain the order statistics $Z_{(1)} < \dots < Z_{(N)}$ (Assume no ties).

Now $\{S_1 = s_1, \dots, S_n = s_n\} \iff \{Y_1, \dots, Y_n$ occupy positions $s_1, \dots, s_n$ in $Z_{(1)}, \dots, Z_{(N)}\}$.

There are a total of $\binom{N}{n}$ sets of n positions in $Z_{(1)}, \dots, Z_{(N)}$, and each is occupied by $Y_1, \dots, Y_n$ with equal probability, since $X_1, \dots, X_m, Y_1, \dots, Y_n$ are iid.

The result follows.

The above result allows us to find the exact distribution of W_{XY}.

<u>Example</u>: For $N=5$, $n=2$, we have

$W_S = S_1 + \dots + S_n$

$\left(W_{XY} = W_S - \frac{1}{2}n(n+1)\right)$

$\binom{N}{n} = \binom{5}{2} = 10$ possible 2-tuples (s_1, s_2)

Each occurs with probability 1/10

s_1	s_2	W_S	W_{XY}
1	2	3	0
1	3	4	1
1	4	5	2
1	5	6	3
2	3	5	2
2	4	6	3
2	5	7	4
3	4	7	4
3	5	8	5
4	5	9	6

From the above we can tabulate the null distribution of W_{XY} as

w	0	1	2	3	4	5	6
$P(W_{XY}=w)$	1/10	1/10	2/10	2/10	2/10	1/10	1/10

The rule

$$\text{Reject } H_0: F = G \quad \text{if} \quad W_{XY} \geq 6$$

has a Type I error rate of 1/10.

[Since W_{XY} has a discrete dist. there may not exist a value c such that $P(W_{XY} \geq c) = \alpha$.]

One can easily imagine that for large N and n, finding the exact distribution of W_{XY} becomes tedious.

- "Nonparametrics" by Lehmann has several pages of tables in the back giving values of $P(W_{XY} \le a)$ for different (small) values of n, $m = N-n$, and a.
- pwilcox() function in R evaluates the cdf of W_{XY}. It is slow (can crash) when n, N are large.

For large n, N, we can use the asymptotic null distribution of W_{XY}.

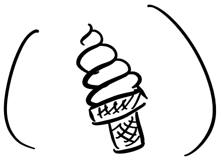
Actually, since $W_{XY} = W_S - \frac{1}{2}n(n+1)$, we can equivalently base tests on W_S.

Next we obtain a Normal approximation to $P(W_S \le a)$.

ASYMPTOTIC ANALYSIS OF W_S :

Result: If $F = G$ then

$$\frac{W_S - EW_S}{\sqrt{Var\, W_S}} \xrightarrow{D} N(0,1)$$



as $N \to \infty$, provided $n \to \infty$ and $N-n \to \infty$.

We present expressions for EW_S and $Var\, W_S$ before jumping into the proof:

We have

(Each $S_1, \ldots, S_n$ has the same marginal dist., but they are not independent)

$$P(S_1 = s_1) = P(Y_1 \text{ in position } s_1) = \frac{1}{N}$$

$$\begin{aligned} EW_S &= E \sum_{j=1}^{n} S_j \\ &= \sum_{j=1}^{n} ES_j \\ &= n\, ES_1 \\ &= n \sum_{s_1=1}^{N} s_1 \cdot \frac{1}{N} \\ &= \frac{n}{N} \frac{N(N+1)}{2} \\ &= \frac{1}{2} n (N+1) \end{aligned}$$

In addition

$$\operatorname{Var} W_s = \operatorname{Var}\left(\sum_{j=1}^{n} S_j\right)$$

Every pair has the same covariance.

$$= \sum_{j=1}^{n} \operatorname{Var} S_j + \sum_{j \neq i}\sum \operatorname{Cov}(S_i, S_j)$$

$$= n \operatorname{Var} S_1 + n(n-1) \operatorname{Cov}(S_1, S_2), \qquad (A)$$

where

$$\operatorname{Var} S_1 = \mathbb{E} S_1^2 - (\mathbb{E} S_1)^2$$

$$= \sum_{s_1=1}^{N} s_1^2 \cdot \frac{1}{N} - \left(\frac{N+1}{2}\right)^2$$

$$= \frac{1}{N} \frac{N(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4}$$

$$= \frac{(N+1)(2(2N+1) - 3(N+1))}{12}$$

$$= \frac{(N+1)(N-1)}{12}$$

$$= \frac{N^2-1}{12}.$$

We can simplify finding $\operatorname{Cov}(S_1, S_2)$ by a wily trick:

If $n = N$, then $W_s = \frac{N(N+1)}{2}$, so $\operatorname{Var} W_s = 0$, and we may write

$$0 = N \operatorname{Var} S_1 + N(N-1) \operatorname{Cov}(S_1, S_2)$$

$$\iff \operatorname{Cov}(S_1, S_2) = -\frac{\operatorname{Var} S_1}{N-1} = -\frac{\overset{(N+1)(N-1)}{(N^2-1)}}{12}\,\frac{1}{N-1} = -\frac{N+1}{12}.$$

Plugging this back into $(\star)$, we obtain

$$\operatorname{Var} W_s = n \operatorname{Var} S_1 + n(n-1)\operatorname{Cov}(S_1, S_2),$$

$$= n\,\frac{(N-1)(N+1)}{12} - \frac{n(n-1)(N+1)}{12}$$

$$= \frac{n}{12}(N-n)(N+1)$$

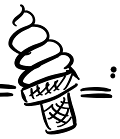
So the result above tells us that

$$P(W_s \le a) \approx \Phi\left(\frac{a - \frac{1}{2}n(N+1)}{\sqrt{\frac{n}{12}(N-n)(N+1)}}\right)$$

provided n and $N-n$ are large.

Since W_s is discrete, a "continuity correction" is generally employed:

$$P(W_s \le a) \approx \Phi\left(\frac{a - \frac{1}{2}n(N+1) + \frac{1}{2}}{\sqrt{\frac{n}{12}(N-n)(N+1)}}\right).$$

Application of  :

An asymptotic size-α test of $H_0 : F = G$ vs the "right-sided" alternative — that the treatment tends to increase the Y_i's over the X_i's is

$$\text{Reject } H_0 \text{ if } \underbrace{1 - \Phi\left(\frac{W_s - \frac{1}{2}n(N+1) + \frac{1}{2}}{\sqrt{\frac{n}{12}(N-n)(N+1)}}\right)}_{\text{asymptotic p-value}} < \alpha.$$

PROOF OF (Adapted from Appendix of Lehmann's "Nonparametrics"):

The main complication is the fact that W_s is a sum of dependent rvs.

Our strategy is to

I. Prove $\xrightarrow{D} N(0,1)$ of a sum of independent rvs which approximates W_s.

II. Show that the difference between W_s and the approximation vanishes.

I. Convergence to $N(0,1)$ of an approximation to $(W_s - \mathbb{E}W_s)/\sqrt{\text{Var}W_s}$:

Let $U_1, \ldots, U_N \overset{iid}{\sim} \text{Unif}(0,1)$.

A way to draw n from among N items with replacement is to assign $U_1, \ldots, U_N$ to the items $1, \ldots, N$; then take only those items whose uniform realization is among the smallest n values of $U_1, \ldots, U_N$.

With this fun fact in mind, we see that under H_0 we can write

$$W_s = \sum_{i=1}^{N} i \cdot J_i \quad , \quad J_i = \begin{cases} 1 & \text{if} \quad U_i \leq U_{(n)} \\ 0 & \text{o.w.,} \end{cases}$$

which is a sum of dependent rvs.

Now introduce $\tilde{W}_s$ as an approximation of W_s which is a sum of independent rvs:

$$\tilde{W}_s = \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right) K_i + \frac{n(N+1)}{2} \quad , \quad K_i = \begin{cases} 1 & \text{if} \quad U_i \leq \frac{n}{N} \\ 0 & \text{o.w.} \end{cases}$$

We have

$$\begin{aligned} \mathbb{E}\tilde{W}_s &= \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right) \frac{n}{N} + \frac{n(N+1)}{2} \\ &= \underbrace{\left(\frac{N(N+1)}{2} - \frac{N(N+1)}{2}\right)}_{=0} \frac{n}{N} + \frac{n(N+1)}{2} \\ &= \frac{1}{2} n (N+1), \end{aligned}$$

as well as

$$\operatorname{Var} \tilde{W}_S = \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2 \operatorname{Var} K_i$$

$$= \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2 \frac{n}{N}\left(1 - \frac{n}{N}\right) .$$

$$= \frac{n}{N^2}(N-n) \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2$$

$$= \frac{n}{N^2}(N-n)\left[\sum_{i=1}^{N} i^2 - 2\left(\frac{N+1}{2}\right)\sum_{i=1}^{N} i + N\left(\frac{N+1}{2}\right)^2\right]$$

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

$$= \frac{n}{N^2}(N-n)\frac{1}{12}\left[2N(N+1)(2N+1) - 3N(N+1)^2\right]$$

$$= \frac{n}{N^2}(N-n)\frac{1}{12}(N+1)\left[2N(2N+1) - 3N(N+1)\right]$$

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

$$= \frac{n}{N^2}(N-n)\frac{1}{12}N(N+1)(N-1)$$

$$= \frac{1}{12}n(N-n)(N+1)\cdot\left(\frac{N-1}{N}\right)$$

$$= \left(\operatorname{Var} W_S\right)\cdot\frac{N-1}{N} .$$

Now we see that

$$\frac{\tilde{W}_s - \mathbb{E}\tilde{W}_s}{\sqrt{\mathrm{Var}\,\tilde{W}_s}} = \frac{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)K_i + \frac{n(N+1)}{2} - \frac{n(N+1)}{2}}{\sqrt{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 \frac{n}{N}\left(1 - \frac{n}{N}\right)}}$$

$$\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right) = 0$$

$$= \frac{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)\left[\left(K_i - \frac{n}{N}\right)\Big/\sqrt{\frac{n}{N}\left(1 - \frac{n}{N}\right)}\right]}{\sqrt{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2}}$$

$$= \frac{\sum_{i=1}^{N} a_i \xi_i}{\sqrt{\sum_{i=1}^{N} a_i^2}} \qquad \left(\begin{array}{l}\xi_1,\ldots,\xi_N \text{ ind. w/ zero mean and unit variance.} \\ a_i = i - \frac{N+1}{2}, \quad i = 1,\ldots,N.\end{array}\right)$$

$$\xrightarrow{D} N(0,1)$$

by the Lindeberg C.L.T. provided $\frac{\max_{1\le i\le N}|a_i|}{\sqrt{\sum_{i=1}^{N} a_i^2}} \to 0$ as $N \to \infty$.

We have

$$\frac{\max_{1\le i\le N}|a_i|}{\sqrt{\sum_{i=1}^{N} a_i^2}} = \frac{\frac{N-1}{2}}{\sqrt{N(N^2-1)/12}} \to 0 \quad \text{as} \quad N \to \infty.$$

We used

$$\max_{1\le i\le N}|a_i| = \left(\left|1 - \frac{N+1}{2}\right| \vee \left|N - \frac{N+1}{2}\right|\right) = \frac{N-1}{2}$$

$$\sum_{i=1}^{N} a_i^2 = \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 \overset{\text{(earlier work)}}{=} \frac{1}{12}N(N+1)(N-1) = \frac{N(N^2-1)}{12}.$$

II. Showing the goodness of the approximation to $(W_S - EW_S)/\sqrt{\text{Var}\, W_S}$:

Having established

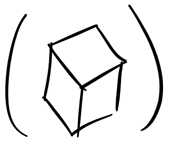
$$\frac{\tilde{W}_S - \mathbb{E}\tilde{W}_S}{\sqrt{\text{Var}\,\tilde{W}_S}} \xrightarrow{D} N(0,1) \quad \text{as} \quad N \to \infty,$$

Hajek's Theorem (Corollary 2 on pg. 349 of "Nonparametrics" by Lehmann) says

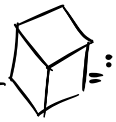
$$\frac{W_S - \mathbb{E}W_S}{\sqrt{\text{Var}\, W_S}} \xrightarrow{D} N(0,1)$$

provided

$$\frac{\mathbb{E}\left(\tilde{W}_S - W_S\right)^2}{\text{Var}\,\tilde{W}_S} \longrightarrow 0 \quad \text{as} \quad N \to \infty.$$



In essence, if $\tilde{W}_S$ gets very close to W_S, then the limiting distribution of W_S will also be the limiting distribution of $\tilde{W}_S$.

Showing  :

First write

$$\tilde{W}_S - W_S = \sum_{i=1}^{N}\left(i - \tfrac{N+1}{2}\right)K_i + \frac{n(N+1)}{2} - \sum_{i=1}^{N} i \cdot J_i$$

$$= \sum_{i=1}^{N}\left(i - \tfrac{N+1}{2}\right)\left(K_i - J_i\right).$$

↙ $\sum_{i=1}^{N} J_i = n$

Then, using iterated expectation, conditioning on $U_{(1)}, \ldots, U_{(N)}$, we have

$$\mathbb{E}\left(\tilde{W}_S - W_S\right)^2 = \mathbb{E}\left(\mathbb{E}\left[\left(\tilde{W}_S - W_S\right)^2 \middle| U_{(1)}, \ldots, U_{(N)}\right]\right)$$

$$= \mathbb{E}\left(\mathrm{Var}\left[\tilde{W}_S - W_S \,\middle|\, U_{(1)},\dots,U_{(N)} \right]\right)$$

$$+ \,\mathbb{E}\left(\left(\mathbb{E}\left[\tilde{W}_S - W_S \,\middle|\, U_{(1)},\dots,U_{(N)}\right]\right)^2\right),$$

where

$$\mathbb{E}\left[\tilde{W}_S - W_S \,\middle|\, U_{(1)},\dots,U_{(N)}\right] = \mathbb{E}\left[\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)\left(K_i - J_i\right) \,\middle|\, U_{(1)},\dots,U_{(N)}\right]$$

$$= \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)\underbrace{\mathbb{E}\left[K_i - J_i \,\middle|\, U_{(1)},\dots,U_{(N)}\right]}_{\text{Does not change with } i}$$

$$= \underbrace{\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)}_{=0}\mathbb{E}\left[K_1 - J_1 \,\middle|\, U_{(1)},\dots,U_{(N)}\right]$$

$$= 0$$

and

$$\mathrm{Var}\left[\tilde{W}_S - W_S \,\middle|\, U_{(1)},\dots,U_{(N)}\right] = \mathrm{Var}\left[\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2\left(K_i - J_i\right) \,\middle|\, U_{(1)},\dots,U_{(N)}\right].$$

To find the variance, it helps to note that after conditioning on $U_{(1)},\dots,U_{(N)}$, the values $U_1,\dots,U_N$ on which $K_1,\dots,K_N$ and $J_1,\dots,J_N$ are a random permutation of $U_{(1)},\dots,U_{(N)}$.

Let $c_1,\dots,c_N$ and $a(1),\dots,a(N)$ be constants and let $T_1,\dots,T_N$ be a random permutation of $\{1,\dots,N\}$. Then

$$\mathbb{E}\left(\sum_{i=1}^{N} c_i\, a(T_i)\right) = \bar{a}\sum_{i=1}^{N} c_i$$

$$\mathrm{Var}\left(\sum_{i=1}^{N} c_i\, a(T_i)\right) = \frac{\sum_{i=1}^{N}(c_i - \bar{c})^2\sum_{i=1}^{N}(a(i) - \bar{a})^2}{N-1}$$

Result on pg 334 of Lehmann.

$$\mathrm{Var}\left[\tilde{W}_S - W_S \,\middle|\, U_{(1)},\dots,U_{(N)}\right] = \frac{1}{N-1}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2\sum_{i=1}^{N}\left(K_i - J_i - \underbrace{\bar{K} - \bar{J}}\right)^2$$

$$\le \frac{1}{N-1}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2\sum_{i=1}^{N}\left(K_i - J_i\right)^2$$

mean minimizes the least-squares criterion.

Now

$$\mathbb{E}\left(\tilde{W}_S - W_S\right)^2 = \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 \sum_{i=1}^{N}\left(K_i - J_i\right)^2\right]$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 \mathbb{E}\left[\sum_{i=1}^{N}\left(K_i - J_i\right)^2\right].$$

Let K_i^* and J_i^* be K_i and J_i when associated with the corresponding $U_{(i)}$. In addition, let $D = \#\{U_i \le \frac{n}{N}\}$. Then $\sum_{i=1}^{N}\left(K_i - J_i\right)^2 = \sum_{i=1}^{N}\left(K_i^* - J_i^*\right)^2$, and

$$K_i^* = \begin{cases} 1 & i=1,\dots,D \\ 0 & i=D+1,\dots,N \end{cases} \qquad J_i^* = \begin{cases} 1 & i=1,\dots,n \\ 0 & i=n+1,\dots,N. \end{cases}$$

Now

$D \le n$

$$\Rightarrow \sum_{i=1}^{N}\left(K_i^*-J_i^*\right)^2 = \underbrace{\sum_{i=1}^{D}\left(K_i^*-J_i^*\right)^2}_{=0} + \underbrace{\sum_{i=D+1}^{n}\underbrace{\left(K_i^*-J_i^*\right)^2}_{1}}_{(n-D)} + \underbrace{\sum_{i=n+1}^{N}\left(K_i^*-J_i^*\right)^2}_{0} = n-D$$

$D > n$

$$\Rightarrow \sum_{i=1}^{N}\left(K_i^*-J_i^*\right)^2 = \underbrace{\sum_{i=1}^{n}\left(K_i^*-J_i^*\right)^2}_{=0} + \underbrace{\sum_{i=n+1}^{D}\underbrace{\left(K_i^*-J_i^*\right)^2}}_{D-n} + \underbrace{\sum_{i=n+1}^{N}\left(K_i^*-J_i^*\right)^2}_{0} = D-n,$$

so

$$\sum_{i=1}^{N}\left(K_i-J_i\right)^2 = |D-n|.$$

From here, noting that $D \sim \text{Binomial}\left(N, \frac{n}{N}\right)$, we have

$$\mathbb{E}\left[\sum_{i=1}^{N}\left(K_i-J_i\right)^2\right] = \mathbb{E}\left|D-n\right| \le \sqrt{\mathbb{E}(D-n)^2} = \sqrt{N\frac{n}{N}\left(1-\frac{n}{N}\right)}.$$

Putting everything together gives

$$\frac{\mathbb{E}\left(\tilde{W}_S - W_S\right)^2}{\operatorname{Var} \tilde{W}_S} \le \frac{\frac{1}{N-1} \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2 \sqrt{N \frac{n}{N}\left(1-\frac{n}{N}\right)}}{\sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2 \frac{n}{N}\left(1-\frac{n}{N}\right)}$$

$$= \frac{1}{N-1} \frac{\sqrt{n(N-n)/N}}{n(N-n)/N^2}$$

$$= \frac{N}{N-1} \sqrt{\frac{N}{n(N-n)}}$$

$$\le \begin{cases} \frac{N}{N-1} \sqrt{\frac{N}{\frac{1}{2}N(\underline{N}-n)}} = \frac{N}{N-1} \sqrt{\frac{2}{N-n}} & n \ge N-n \quad \left(\Rightarrow \ n \ge \frac{1}{2}N\right) \\ \frac{N}{N-1} \sqrt{\frac{N}{n \ \frac{1}{2}N}} = \frac{N}{N-1} \sqrt{\frac{2}{n}} & n < N-n \quad \left(\Rightarrow \ N-n > \frac{1}{2}N\right) \end{cases}$$

$$N-n > N-(N-n)$$
$$2(N-n) > N$$
$$(\underline{N}-n) > \tfrac{1}{2}N$$

$$\to 0$$

since $n \to \infty$ and $N-n \to \infty$.

This completes the proof. □

Good exercises would be:

(i) Prove Hajeki result.

(ii) Prove the result ( ).

(iii) Show that $\operatorname{Cov}(J_i, J_{i'}) = -\frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1}$ and hence

$$\operatorname{Var}\left(\sum_{i=1}^{N} i J_i\right) = \frac{n}{12}(N-n)(N+1) = \operatorname{Var}(W_s).$$

POWER OF THE WILCOXON RANK-SUM TEST:

It is difficult to analyze the power of the Wilcoxon rank-sum test unless we assume a specific form for the alternative.

For our discussions of power, we will assume the "location shift" scenario:

Suppose

$$G(x) = F(x - \Delta)$$

for some Δ and consider testing

$$H_0: \Delta \le 0 \quad \text{vs} \quad H_1: \Delta > 0.$$

We study the power of the test which rejects when

$$W_{XY} \ge c .$$

Proposition: The power function $\gamma(\Delta)$ is non-decreasing in Δ.

Proof: The power function is given by

$$\begin{aligned} \gamma(\Delta) &= P_\Delta\left(W_{XY} \ge c\right) \\ &= P_\Delta\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{1}(X_i < Y_j) \ge c\right) \\ &= P_\Delta\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{1}(X_i < X_j' + \Delta) \ge c\right), \end{aligned}$$

Since $Y_j \sim F(\cdot - \Delta)$, we have $Y_j \overset{D}{=} X_j' + \Delta$, $X_j' \sim F$.

where $X_1', \dots, X_n' \overset{iid}{\sim} F$ are independent of $X_1, \dots, X_n$ but have the same dist.

$\square$

Note that $\gamma(\Delta_1) \le \gamma(\Delta_2)$ for all $\Delta_1 \le \Delta_2$.

Asymptotic power under the location shift model:

Assuming $G(x) = F(x-\Delta)$, we have

$$\begin{aligned}
\beta(\Delta) &= P_\Delta\left(W_{XY} \geq c \right) \\
&= P_\Delta\left(\frac{W_{XY} - \mathbb{E}W_{XY}}{\sqrt{\mathrm{Var}\, W_{XY}}} \geq \frac{c - \mathbb{E}W_{XY}}{\sqrt{\mathrm{Var}\, W_{XY}}} \right) \\
&\approx 1 - \Phi\left(\frac{c - \mathbb{E}W_{XY}}{\sqrt{\mathrm{Var}\, W_{XY}}} \right)
\end{aligned}$$

$$\frac{W_{XY} - \mathbb{E}W_{XY}}{\sqrt{\mathrm{Var}\, W_{XY}}} = \frac{W_S - \mathbb{E}W_S}{\sqrt{\mathrm{Var}\, W_S}} \xrightarrow{D} N(0,1)$$

for large N, $N-n$, n. Now we have

$$\begin{aligned}
\mathbb{E}W_{XY} &= \mathbb{E} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(X_i < Y_j) \\
&= mn\, \mathbb{E}\, \mathbb{1}(X_1 < Y_1) \\
&= mn\, P(X_1 < Y_1) \\
&= mn\, P(X_1 - \Delta < Y_1 - \Delta) \\
&= mn\, P(X_1 - (Y_1 - \Delta) < \Delta) \\
&= mn\, P(X_1 - X_1' < \Delta), \qquad X_1' \sim F, \text{ independent of } X_1. \\
&= mn\, p_1(\Delta),
\end{aligned}$$

where $p_1(\Delta) = P(X_1 - X_1' < \Delta)$ can be computed if F is known.

Moreover

$$\mathrm{Var}\, W_{XY} = \mathrm{Var}\left(\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(X_i < Y_j) \right)$$

$$= \sum_{i=1}^{m} \mathrm{Var}\left(\sum_{j=1}^{n} \mathbb{1}(X_i < Y_j) \right) + \sum_{i \neq i'} \mathrm{Cov}\left(\sum_{j=1}^{n} \mathbb{1}(X_i < Y_j), \sum_{j'=1}^{n} \mathbb{1}(X_{i'} < Y_{j'}) \right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{Var}\left(\mathbb{1}(X_i < Y_j) \right) + \sum_{i=1}^{m} \sum_{j \neq j'} \mathrm{Cov}\left(\mathbb{1}(X_i < Y_j), \mathbb{1}(X_i < Y_{j'}) \right)$$

$$+ \sum_{i \neq i'} \sum_{j=1}^{n} \sum_{j'=1}^{n} \mathrm{Cov}\left(\mathbb{1}(X_i < Y_j), \mathbb{1}(X_{i'} < Y_{j'}) \right).$$

The three terms simplify to

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{Var}\left(\mathbb{1}(X_i < Y_j) \right) = mn\, p_1(\Delta) [1 - p_1(\Delta)]$$

$$\sum_{i=1}^{m} \sum_{j \neq j'} \mathrm{Cov}\left(\mathbb{1}(X_i < Y_j), \mathbb{1}(X_i < Y_{j'}) \right) = \sum_{i=1}^{m} \sum_{j \neq j'} \left[\mathbb{E}\, \mathbb{1}(X_i < Y_j) \mathbb{1}(X_i < Y_{j'}) - \mathbb{E}\, \mathbb{1}(X_i < Y_j) \mathbb{E}\, \mathbb{1}(X_i < Y_{j'}) \right]$$

$$= m\, n(n-1) \left[\underbrace{P(X_1 < Y_1 \cap X_1 < Y_2)}_{p_2(\Delta)} - p_1^2(\Delta) \right]$$

$$= m\, n(n-1) \left[p_2(\Delta) - p_1^2(\Delta) \right]$$

$$\sum_{i \neq i'} \sum_{j=1}^{n} \sum_{j'=1}^{n} \mathrm{Cov}\left(\mathbb{1}(X_i < Y_j), \mathbb{1}(X_{i'} < Y_{j'}) \right)$$

$$= m(m-1) \sum_{j=1}^{n} \sum_{j'=1}^{n} \mathrm{Cov}\left(\mathbb{1}(X_1 < Y_j), \mathbb{1}(X_2 < Y_{j'}) \right)$$

$$= n\, m(m-1)\, \mathrm{Cov}\left(\mathbb{1}(X_1 < Y_1), \mathbb{1}(X_2 < Y_1) \right)$$

$$= n\, m(m-1) \left[\mathbb{E}\, \mathbb{1}(X_1 < Y_1)\, \mathbb{1}(X_2 < Y_1) - \mathbb{E}\, \mathbb{1}(X_1 < Y_1)\, \mathbb{E}\, \mathbb{1}(X_2 < Y_1) \right]$$

$$= n\, m(m-1) \left[\underbrace{P(X_1 < Y_1 \cap X_2 < Y_1)}_{p_3(\Delta)} - P(X_1 < Y_1) \cdot P(X_2 < Y_1) \right]$$

$$= nm\, (m-1) \left[p_3(\Delta) - p_1^2(\Delta) \right]$$

So we can write

$$\begin{aligned}\text{Var } W_{XY} &= mn\, p_1(\Delta)[1 - p_1(\Delta)] + mn(n-1)\left[p_2(\Delta) - p_1^2(\Delta)\right] \\ &\qquad + nm(m-1)\left[p_3(\Delta) - p_1^2(\Delta)\right] \\ &=: \vartheta(\Delta),\end{aligned}$$

where $p_1(\Delta)$, $p_2(\Delta)$, and $p_3(\Delta)$ can be computed if F is known.

So our Normal approximation to the power is

$$\gamma(\Delta) \approx 1 - \Phi\left(\frac{c - nm\, p_1(\Delta)}{\sqrt{\vartheta(\Delta)}}\right).$$

Asymptotic power of size-α test in the location shift model:

In the location shift model the Wilcoxon rank sum test will have size tending to α as $n \to \infty$ under the rule $W_{XY} \geq c_\alpha$, with c_α the solution to

$$\alpha = 1 - \Phi\left(\frac{c_\alpha - nm\, p_1(0)}{\sqrt{\vartheta(0)}}\right) \quad \left(\approx \gamma(0)\right),$$

$$\iff$$

$$z_\alpha = \frac{c_\alpha - nm\, p_1(0)}{\sqrt{\vartheta(0)}},$$

that is with c_α given by

$$c_\alpha = z_\alpha \sqrt{\vartheta(0)} + nm\, p_1(0).$$

Under $\Delta = 0$ we have

$$p_1(0) = P(X_1 < Y_1) = 1/2$$

$$p_2(0) = P(X_1 < Y_1 \cap X_1 < Y_2) = 1/3 \quad \left(\tfrac{1}{3} \text{ prob } X_1 \text{ is smallest among } X_1, Y_1, Y_2\right)$$

$$p_3(0) = P(X_1 < Y_1 \cap X_2 < Y_1) = 1/3, \quad \left(\tfrac{1}{3} \text{ prob } Y_1 \text{ is greatest among } X_1, X_2, Y_1\right)$$

so that

$$\vartheta(0) = mn \, \tfrac{1}{2}\left[1 - \tfrac{1}{2}\right] + mn(n-1)\left[1/3 - \left(\tfrac{1}{2}\right)^2\right] + nm(m-1)\left[\tfrac{1}{3} - \left(\tfrac{1}{2}\right)^2\right]$$

$$= \frac{mn}{4} + \Big(mn(n-1) + nm(m-1)\Big)\frac{1}{12}$$

$$= \frac{1}{12}\cdot\left[3mn + mn^2 - mn + nm^2 - nm\right]$$

$$= \frac{mn}{12}\left(n + m + 1\right)$$

$$= \frac{1}{12}\, n\,(N-n)(N+1),$$

which matches Var W_s that we computed earlier. So

$$c_\alpha = z_\alpha \sqrt{\tfrac{1}{12}\, nm\,(N+1)} + \frac{nm}{2}.$$

Now the power of the size-α test is approximately

$$\gamma^\alpha(\Delta) \approx 1 - \Phi\left(\frac{z_\alpha\sqrt{\tfrac{1}{12}\,nm\,(N+1)} + \frac{nm}{2} - nm\,p_1(\Delta)}{\sqrt{\vartheta(\Delta)}}\right)$$

$$= 1 - \Phi\left(\frac{nm\left(\tfrac{1}{2} - p_1(\Delta)\right) + z_\alpha\sqrt{\tfrac{1}{12}\,nm\,(N+1)}}{\sqrt{\vartheta(\Delta)}}\right).$$

It is convenient to employ a further approximation obtained by setting

(i) $p_1(\Delta) = P(X_1 - X_1' < \Delta) = F^*(\Delta) \approx F^*(0) + \Delta\cdot f^*(0) = \underset{\uparrow \text{ b/c } X_1 - X_1' \text{ symm around } 0}{\tfrac{1}{2}} + \Delta\cdot f^*(0)$

where F^* is the cdf of $X_1 - X_1'$ and f^* the corresponding density, and

(ii) $\vartheta(\Delta) \approx \vartheta(0) = \sqrt{\tfrac{1}{12}\,nm\,(N+1)}$.

Making these substitutions yields the approximation to the power given by

$$\tilde{\gamma}^{\alpha}(\Delta) = 1 - \Phi\left(z_{\alpha} - \sqrt{\frac{12\, m n}{N+1}} \cdot \Delta \cdot f^{*}(0) \right).$$

Approximate power under Normal location shift model:

If F is Normal then

$$f^{*}(0) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}\sigma} \phi\left(0/\sqrt{2}\sigma \right) = \frac{1}{2\sqrt{\pi}} \frac{1}{\sigma} .$$

This gives

$$\tilde{\gamma}^{\alpha}(\Delta) = 1 - \Phi\left(z_{\alpha} - \sqrt{\frac{6\, m n}{N+1}} \, \frac{\Delta}{2\sigma\sqrt{\pi}} \right).$$

For quick sample size calculations, we can set $n = m$ and treat $N+1 \approx 2n$.

Then set

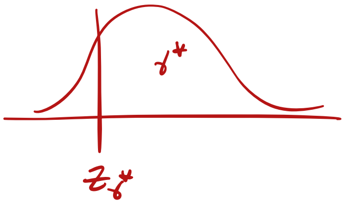
$$\tilde{\gamma}_{n}^{\alpha}(\Delta) = 1 - \Phi\left(z_{\alpha} - \frac{\sqrt{6n}}{2\sigma\sqrt{\pi}} \Delta \right)$$

Now, if we wish to reject $H_0: \Delta \leq 0$ vs $H_1: \Delta > 0$ for some Δ^{*} with power at least γ^{*}, choose the smallest n such that

$$\tilde{\gamma}_{n}^{\alpha}(\Delta^{*}) \geq \gamma^{*}$$

$$\iff \quad 1 - \Phi\left(z_{\alpha} - \frac{\sqrt{6n}}{2\sigma\sqrt{\pi}} \Delta^{*} \right) \geq \gamma^{*}$$

$$\iff \quad z_{\alpha} - \frac{\sqrt{6n}}{2\sigma\sqrt{\pi}} \Delta^{*} \leq z_{\gamma^{*}}$$



$\iff$

$$z_\alpha - z_{\gamma^*} \le \frac{\Delta^* \sqrt{6n}}{\sigma\sqrt{\pi}}$$

$\iff$

$$\frac{2\sigma\sqrt{\pi}\left(z_\alpha + z_{1-\gamma^*}\right)}{\sqrt{6}\;\Delta^*} \le \sqrt{n}$$

$\iff$

$$n \ge \frac{\pi}{3} \cdot \frac{2\sigma^2\left(z_\alpha + z_{1-\gamma^*}\right)^2}{(\Delta^*)^2} .$$

Note: The sample size required by a Z test for the same hypotheses in the Normal location-shift model would be

$$n \ge \frac{2\sigma^2\left(z_\alpha + z_{1-\gamma^*}\right)^2}{(\Delta^*)^2} .$$

So the Wilcoxon rank sum test (according to the approximations) needs sample sizes larger by a factor of $\pi/3 = 1.05$, so not very much larger!

More on $p_1(\Delta)$, $p_2(\Delta)$, and $p_3(\Delta)$:

The values $p_1(\Delta)$, $p_2(\Delta)$, and $p_3(\Delta)$ depend on F.

If F is the Normal(μ, σ^2) distribution then under the location shift alternative, G is the Normal$(\mu+\Delta, \sigma^2)$ distribution. So we have

$$
\begin{aligned}
p_1(\Delta) &= P(X_1 < Y_1) \\
&= P(X_1 - \mu < Y_1 - (\mu + \Delta) + \Delta) \\
&= P\Big(\underbrace{(X_1 - \mu) - (Y_1 - (\mu + \Delta))}_{\sim \text{Normal}(0,\, 2\sigma^2)} < \Delta \Big) \\
&= P\Big(Z < \frac{\Delta}{\sqrt{2}\sigma} \Big), \qquad Z \sim \text{Normal}(0,1).
\end{aligned}
$$

Moreover

$$
\begin{aligned}
p_2(\Delta) &= P\big(\{X_1 < Y_1\} \cap \{X_1 < Y_2\} \big) \\
&= P\big(\{X_1 - \mu < Y_1 - (\mu+\Delta) + \Delta\} \cap \{(X_1 - \mu) < Y_2 - (\mu + \Delta) + \Delta\} \big) \\
&= P\left(\left\{ \frac{(X_1 - \mu) - (Y_1 - (\mu+\Delta))}{\sqrt{2}\sigma} < \frac{\Delta}{\sqrt{2}\sigma} \right\} \cap \left\{ \frac{(X_1 - \mu) - (Y_2 - (\mu+\Delta))}{\sqrt{2}\sigma} < \frac{\Delta}{\sqrt{2}\sigma} \right\} \right) \\
&= P\Big(Z_1 < \frac{\Delta}{\sqrt{2}\sigma} \cap Z_2 < \frac{\Delta}{\sqrt{2}\sigma} \Big), \quad \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \right).
\end{aligned}
$$

We obtain the same expression for $p_3(\Delta)$.

This can be evaluated from the bivariate Normal joint cdf.

If we obtain these values we can get a closer approximation to the power (by not replacing $\vartheta(\Delta)$ with $\vartheta(0)$).