Can We Generalize Our Findings?

- So far we have dealt with limited sets of subjects.
- Any inference (p-values) concerning effects of treatments or controls was based on the random assignment of treatments and controls to just these subjects.
- Would we get similar results with another group of subjects?
- Generalization of inference to other subjects may not be reasonable.
- Such generalization may not even be relevant. (Farmer’s fields)
- However, most often the expense of any treatment study warrants a generalization of any results, positive or negative (cost benefit analysis).
- Such a step will depend very much on the relationship of the study subjects to the larger population of subjects to which we wish to generalize any findings.
- How do the study subjects relate to the population of interest?
Populations

- Sometimes our subjects are judged to be typical or representative members of a larger population.
- The simplest way to make sure that this judgment holds up is to take our subjects as a random sample from this larger population.
- Any group of $N$ subjects has same chance of being selected.
- Often populations change over time or geographically and we cannot sample subjects from future or remote populations.
- Generalizing any findings to future or remote populations is a matter of a judgment.
- How do we sample from a treatment population and from a control population? Treating all members of a population of interest leaves no controls (& vice versa).
- Finesse this by randomizing treatment/control after we have gotten our random sample of subjects from the population.
Assume that $N = m + n$ subjects were drawn randomly from a given population.

Treatment is assigned at random to $n$ of them and the other $m$ act as controls.

Denote treatment responses by $Y_1, \ldots, Y_n$ and control responses by $X_1, \ldots, X_m$.

Each $Y$ can be viewed as a random element of the $Y$-population with CDF: $G(y) = P(Y \leq y)$.

The $Y$-population is the collection of all subject responses had they all been treated.

Each $X$ can be viewed as a random element of the $X$-population with CDF $F(x) = P(X \leq x)$.

The $X$-population is the collection of all subject responses had they all been controls.
Additional Simplifying Assumptions

- We assume that the sampled population is large compared to the sample size $N$.
- It is then reasonable to treat $Y_1, \ldots, Y_n$ as independent with common CDF $G$.
- Similarly we can then treat $X_1, \ldots, X_m$ as independent with common CDF $F$.
- We say: $Y_1, \ldots, Y_n$ \textit{i.i.d.} $\sim G$ and $X_1, \ldots, X_m$ \textit{i.i.d.} $\sim F$
- The two samples of responses are independent of each other.
- The hypothesis to be tested is $H_0 : G = F$, i.e., there is no difference between treatment and control.
- This is also referred to as the \textit{two-sample problem}. 
So far the population model was motivated by the desire to generalize any findings from the experimental subjects to other possible subjects.

Another important reason is that of sample size planning to achieve a given power, i.e., probability of rejecting $H_0$ when a specific treatment effect is present.

For $m + n$ randomized subjects (no population model) a power calculation depends on

1) the sample sizes $m$ and $n$
2) treatment responses are how much better than the control responses?
3) the responses of all $m + n$ subjects if they all had been controls.
We can choose \( m \) and \( n \), but can we get the desired power?

We can specify under what degree of improvement we want to be able to reject \( H_0 \) with a given probability \( \beta \) (power).

Suppose the treatment adds an amount \( \Delta \) to the control responses \( Z \), then

\[
Z_{\pi_1}, \ldots, Z_{\pi_m}, \quad Z_{\pi_{m+1}} + \Delta, \ldots, Z_{\pi_{m+n}} + \Delta
\]

The treatment randomization decides which \( n \) of the \( Z \)'s get a \( \Delta \) added.

\[
W_{XY} = \text{number of } Z_{\pi_i} < Z_{\pi_{m+j}} + \Delta, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n
\]

\[
= \text{number of } Z_{\pi_i} - Z_{\pi_{m+j}} < \Delta, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n
\]

The power \( \beta(\Delta) = P_{\Delta}(W_{XY} \geq c_\alpha) \) depends on the \( Z \)'s, some of which will never be observed, all of which will never be known prior to experimentation/measuring.

Thus the power cannot be assessed prior to (not even after(?)) the experiment.

We cannot plan for \( m \) and \( n \) to achieve a given power.
Two Extreme Examples for $Z_1, \ldots, Z_N$

$W_{XY} = \text{number of } Z_{\pi_i} - Z_{\pi_{m+j}} < \Delta, \ i = 1, \ldots, m, \ j = 1, \ldots, n$

- $\implies \beta(\Delta) = 1$ for $\Delta > \max_{i \neq j} |Z_i - Z_j|$, because then $Z_{\pi_i} - Z_{\pi_{m+j}} < \Delta$ for all $i$ and $j$.
- Even for small $\Delta > 0$ when all $Z$’s are very nearly the same.
- If the ordered $Z_{(1)} < \ldots < Z_{(N)}$ have gaps $Z(i) - Z(i-1) \geq \Delta > 0, \ \forall i$, then asking

$$Z_{\pi_i} - Z_{\pi_{m+j}} < \Delta \text{ is equivalent to } Z_{\pi_i} - Z_{\pi_{m+j}} < 0$$

- In that case the distribution of $W_{X,Y}$ does not change for $0 \leq \Delta \leq \min\{Z(i) - Z(i-1), i = 2, \ldots, N\}$ and we have $\beta = \alpha$.
- In both the above cases it is quite clear why the power depends on $Z_1, \ldots, Z_N$. 
The previous two examples show the effect of control response variability on the power of a test.

In the first scenario the variability in control responses is very small and any $\Delta$ shift under treatment is easily recognized, provided it is larger than the control response variability range.

The treatment effect exceeds the variability of control responses. i.e., it dominates.

In the second scenario the gaps between adjacent control responses is large and any $\Delta$ less than that gap is not detected with any power $> \alpha$ by the Wilcoxon test.

Here the response variability swamps the treatment effect.

Both behaviors make intuitive sense and suggest moderated power behavior for intermediate control response variation scenarios.
Aside from the assumption of a population model and the independence of all observations (based on large sampled populations) we need to distinguish between two situations:

1. there will not be any ties among the observations
2. ties are a possibility

We postpone the treatment of ties and focus for now on the first situation: assume that the CDF’s $F$ and $G$ are continuous.

In that case we have $P(X_i = X_j) = P(X_i = Y_k) = P(Y_k = Y_\ell) = 0$ with $X_i, X_j, Y_k, Y_\ell$ independent with CDF’s $F$, $F$, $G$, and $G$, respectively, provided $i \neq j$ and $k \neq \ell$.

The probability of any ties among the $N = m + n$ independent sample values is then zero.
$P(Z = Z') = 0$

$Z, Z'$ independent $Z, Z' \sim F$ or $G$ continuous $\implies P(Z = Z') = 0$

Proof: For any integer $K$ let

$a_i = F^{-1}(i/K) = \inf\{x : F(x) \geq i/K\}, \ i = 0, 1, \ldots, K$ with the

convention that $a_K = \infty$ if $F(x) < 1$ for all $x$. Also, $a_0 = -\infty$.

For continuous $F$ we have $F(a_i) = i/K, \ i = 0, \ldots, K$. Let

$l_i = (a_{i-1}, a_i], \ i = 1, \ldots, K$


\[
P(Z = Z') \leq P(Z \in l_i \cap Z' \in l_i \text{ for some } i = 1, \ldots, K)
= \sum_{i=1}^K P(Z \in l_i \cap Z' \in l_i) = \sum_{i=1}^K P(Z \in l_i)P(Z' \in l_i)
\leq \max_i P(Z' \in l_i) \sum_{i=1}^K P(Z \in l_i) = \max_i P(Z' \in l_i)
= \max_{i=1,\ldots,K} [F(a_i) - F(a_{i-1})] = \max_{i=1,\ldots,K} \left[ \frac{i}{K} - \frac{i-1}{K} \right] = \frac{1}{K} \to_\infty 0 \ \text{q.e.d.}
\]

$\Rightarrow P(Z_1, \ldots, Z_N \text{ are all distinct}) = 1$ for independent $Z_i \sim F$ or $G$. 
**Equally Probable Rankings (1)**

**Theorem:** Let $X_1, \ldots, X_m$ i.i.d. $\sim F$ and $Y_1, \ldots, Y_n$ i.i.d. $\sim G$.

Under $H_0 : F = G$ with $F$ continuous we have

$$P_{H_0}(S_1 = s_1, \ldots, S_n = s_n) = \frac{1}{\binom{m+n}{n}} = \frac{1}{\binom{N}{n}}$$

for any ordered ranking subset $\{s_1 < \ldots < s_n\}$ of $\{1, 2, \ldots, N\}$.

**Proof:** The vector $(Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ has the same distribution as the vector $(Z_{\pi_1}, Z_{\pi_2}, \ldots, Z_{\pi_N})$ for any permutation $(\pi_1, \ldots, \pi_N)$ of $(1, \ldots, N)$. 
The region \( \{Z_1 < \ldots < Z_N\} \subset R^N \) has the same probability as the region \( \{Z_{\pi_1} < \ldots < Z_{\pi_N}\} \subset R^N \) for any other permutation \((\pi_1, \ldots, \pi_N)\) of \((1, \ldots, N)\).

There are \(N!\) such regions, each region having probability \(1/N!\).

The regions \( \{Z_{\pi_1} < \ldots < Z_{\pi_N}\} \) give rise to the same ordered rankings \(S_1 = s_1 < \ldots < S_n = s_n\) as long as the \(Z\)'s in the rank positions \(s_1 < \ldots < s_n\) of \( \{Z_{\pi_1} < \ldots < Z_{\pi_N}\} \) are occupied by \(Y\)'s and the rest are occupied by \(X\)'s.

There are \(m! \cdot n!\) such regions, all with same ranking \(S_1 = s_1 < \ldots < S_n = s_n\).
The Y’s with these rank positions can have their indices permuted in $n!$ ways.

The X’s with ranks $r_1 < \ldots < r_m$ can have their indices permuted in $m!$ ways.

Such permutations don’t change $s_1 < \ldots < s_n$ and $r_1 < \ldots < r_m$.

Thus each group of such regions with $S_1 = s_1 < \ldots < S_n = s_n$ has probability

$$\frac{m!n!}{N!} = \frac{1}{\binom{N}{n}} = P_{H_0}(S_1 = s_1, \ldots, S_n = s_n)$$
Consequences

- The null distribution of the Wilcoxon rank-sum statistic, also known as the Wilcoxon two-sample statistic, is as derived before within the randomization model.

- The null distribution of any rank statistic is based on

\[ P_{H_0}(S_1 = s_1, \ldots, S_n = s_n) = 1/N^n. \]

- This applies for example to the Siegel-Tukey and the KS statistics.

- \( H_0 : F = G \) does not specify the common CDF \( F \).

- Thus one calls rank tests distribution-free or nonparametric.

- The significance level is free of the assumption that \( F \) belongs to some specific parametrized family of distributions, such as the normal distributions.
Dropping the Continuity Assumption

- Ties among ranks will be replaced again by midranks.
- An extreme example for discontinuous $F$:
  Assume $H_0 : F = G$ with $F$ assigning respective probabilities $p$ and $q = 1 - p$ to $a$ and $b$, with $a < b$.
- For sample sizes $m = 2$ and $n = 1$ there is only one midrank $S_{1}^{*}$ for the single $Y$.
- $S_{1}^{*} = 1 \iff a = Y < X_1 = X_2 = b$ with probability $pq^2 = P(Y = a, X_1 = X_2 = b)$.
- $S_{1}^{*} = 1.5 \iff a = Y = X_1 < X_2 = b$ or $a = Y = X_2 < X_1 = b$ with probability $2p^2q$.
- $S_{1}^{*} = 2 \iff Y = X_1 = X_2 = a$ or $Y = X_1 = X_2 = b$ with probability $p^3 + q^3$, etc.

<table>
<thead>
<tr>
<th>$s$</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{H_0}(S_{1}^{*} = s)$</td>
<td>$pq^2$</td>
<td>$2p^2q$</td>
<td>$p^3 + q^3$</td>
<td>$2pq^2$</td>
<td>$p^2q$</td>
</tr>
</tbody>
</table>

- The null distribution of $S_{1}^{*}$ depends on $p$ and thus on $F$. 
The phenomenon on the previous slide may hold in general.

The distribution of midranks depends on the multiplicities $d_1, \ldots, d_e$ with which the $e$ distinct observations are observed.

These numbers are sometimes called the configuration of ties.

In our randomization model from Chapter 1 these configurations were given a priori.

The randomization assigns the labels $X$ and $Y$ to the $N$ associated midranks.

Now these configurations are random variables with distribution depending on $F$ under $H_0$ (previous example).
Two Stage Generation of $X$ and $Y$ Sample

- Under $H_0$ the $X$’s and $Y$’s all have the same distribution $F$.
- We may just view them as $Z_1, \ldots, Z_N$ i.i.d. $\sim F$,
- with an independent choice afterwards as to which $n$ to call $Y$ and which $m$ to call $X$.
- These subsequent $\binom{N}{n}$ choices are all equally likely.
- From these $Z$’s get the random configuration $\{e, d_1, \ldots, d_e\}$.
- Let $Z_{1}^{*} \leq \ldots \leq Z_{N}^{*}$ be the $Z_1, \ldots, Z_N$ in increasing order.
- Our previous designation of $X$ and $Y$ labels to the $Z$’s was completely independent of the $Z$’s.
- We might as well assign such labels now also retroactively to these $Z_{1}^{*} \leq \ldots \leq Z_{N}^{*}$.
- Any such assignment has chance $1/\binom{N}{n}$, independently of the values of $Z_{1}^{*} \leq \ldots \leq Z_{N}^{*}$. 
The ordered $Z_1^* \leq \ldots \leq Z_N^*$ also determine the corresponding $N$ midranks, denoted here by $Q_1^* \leq \ldots \leq Q_N^*$.

The random selection of midranks $S_1^* \leq \ldots \leq S_n^*$ to associate with the $Y$ labels gives us the $n$ ordered midranks of the $Y$’s, conditional on the full set of midranks $Q_1^* \leq \ldots \leq Q_N^*$, which are equivalent to $\{e, d_1, \ldots, d_e\}$.

Each such random choice of $S_1^* \leq \ldots \leq S_n^*$ from $Q_1^* \leq \ldots \leq Q_N^*$ has the same chance $1/(N\binom{N}{n})$,
Conditional $W_s^*$ Test

- Using $W_s^* = \sum_{i=1}^{n} S_i^*$ we test $H_0 : F = G$ by rejecting $H_0$ for large values of $W_s^*$. We can proceed as follows:
- Use the fact that the conditional distribution of midranks given $T = \{e, d_1, \ldots, d_e\}$ is independent of $F$, with probability $1/(\binom{N}{n})$ for each midrank selection from $Q_1^* \leq \ldots \leq Q_N^*$.
- Find $C_T$ such that the conditional significance level

$$P_{H_0}[W_s^* \geq C_T | T]$$

of the rejection rule $W_s^* \geq C_T$ is as close to (and below) the desired significance level $\alpha$.
- The overall significance level of this test procedure is

$$P_{H_0}(\text{rejection}) = \sum_T P_{H_0}(T) \ P_{H_0}[W_s^* \geq C_T | T]$$

summed over all possible configurations $T$. 
What about $P_{H_0}(\text{rejection})$?

- This overall probability of rejection or achieved significance level will be $\leq \alpha$ if we conservatively chose each conditional test with significance level $\leq \alpha$.
- It then is conservatively distribution-free.
- Otherwise the achieved significance level is somewhat close to $\alpha$, or more generally between the minimum and maximum of all possible conditional significance levels.
- $P_{H_0}(\text{rejection})$ depends on $F$ to some extent, through the probabilities $P_{H_0}(T)$ in the previous summation,
- but not through the varying $P_{H_0}[W_s^* \geq C_T | T]$.
- Thus the Wilcoxon test is no longer strictly distribution-free when the population model allows ties.
The Situation Improves for Large $m$ and $n$

- As $m$ and $n$ get larger it is usually possible to get fairly close to $\alpha$ with the conditional significance level for most $T$.
- Those $T$ with $\max(d_1/N, \ldots, d_e/N) \approx 1$ usually have small probability weight, except for extreme discrete distributions $F$.
- Using the normal approximation we have as rejection rule

$$\frac{W_s^* - E[W_s^*]}{\sqrt{\text{var}(W_s^*)}} \geq C_T' = z_{1-\alpha} = u_\alpha$$

- The unconditional distribution of the standardized $W_s^*$ becomes the standard normal distribution.
- The unconditional or overall significance level becomes approximately $\alpha$ as $m$ and $n$ get large.
- The test becomes approximately distribution-free, because the CLT holds for a wide spectrum of distributions $F$. 

Attribute Populations

- Not always able to assign treatments to sampled subjects.
- Sometimes members of a population can be distinguished by some attribute.
- These attributes are integral part of the subjects and cannot be assigned at will.
- We focus here on attributes with two levels.
- Examples: 1st and 2nd born twins, males and females, smokers and nonsmokers, voters favoring candidate A or B, low and high degree of education, and so on.
- Such attributes can be used to view the full population as two subpopulations.
- We want to compare certain measurements or responses for such subpopulations.
For small subpopulations one may measure all responses and compare them.

For large populations this is no longer practical and we will have to settle for samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from the respective subpopulations.

If these subpopulations are sufficiently large we can view these samples as independent observations with respective distribution functions $F$ and $G$, describing the distribution of such values over the full subpopulations, respectively.

Our hypothesis of interest then is $H_0 : F = G$, i.e. there is no difference in the subpopulations defined by the two attribute values.

We again are dealing with the two-sample problem.
A form of cancer has a wide variation in its type of progression.

We distinguish between

- Group I: rapidly progressing, uncontrollable form of the disease
- Group II: a slow progression, controllable form of the disease.

Question: Are any psychological factors linked to this?

Each subject was given a psychological test.

See next slide for the test scores.

If there was any difference between scores from the two groups it was expected to show in high negative scores for Group I.

Read the Text’s careful discussion of viewing these two groups as samples from subpopulations.
The Data and Analysis

Scores of Group II: -2 -11 -16 -12 -13 -6 -9 -7 -11 -9

Midranks of Group I: 22.5 22.5 22.5 22.5 32 35 36.5 39 42 43.5 45.5 45.5
Midranks of Group II: 32 32 32 36.5 38 40 41 43.5 47

> PsychEx1(Nsim=100000)

<table>
<thead>
<tr>
<th>Ws.star</th>
<th>EWs.star</th>
<th>varWs.star</th>
<th>p.val</th>
<th>p.val.sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.050e+02</td>
<td>5.280e+02</td>
<td>2.188e+03</td>
<td>4.985e-02</td>
<td>4.898e-02</td>
</tr>
</tbody>
</table>

- **p.val** is based on the normal approximation.
- For a two-sided test the p-values would be double.
Causality and Association

- Analysis for two randomly assigned treatments and for two sampled attributes proceeds along the same lines as a solution to the two sample problem.

- Treatments are assigned at random, but attributes are assigned or given a priori.

- Does smoking increase the risk of cancer?

- One could make a case for it (causality), if we could assign smoking as a treatment and nonsmoking as control.

- As it is, we can only make a strong case for association.

- Smoking causes cancer? Cancer causes smoking? A hidden factor causes both?

- Showing the same association in many stratifications of the population narrows down the possibilities for hidden factors.
The Text distinguishes the following 5 data models, all with the same analysis.

Model 1. Randomization for the comparison of two treatments.
Model 3. Comparison of two attributes or subpopulations through a sample from each.
Model 4. Comparison of two attributes through a sample from the total population
Model 5. Model for the comparison of two sets of measurements.

The discussion of the Text is instructive (read!)
The power of a test is the probability of rejecting the hypothesis when it is false. In that case ($H_0$ false) it is the probability that we make the correct decision.

We had $H_0 : F = G$ in the two-sample problem.

Any $F$ and $G$ with $F \neq G$ is an instance of $H_0$ being false. Such instances are called alternatives.

While there is an infinity of $(F, G)$ with $F = G$, there are even more alternatives.

We are typically most interested in the alternative that the treatment has beneficial effect when compared to the control.

Assume that “beneficial” means: $Y$ responses from $G$ are high more frequently than $X$ responses from $F$, and will correspondingly also be low less frequently.
This last description of a beneficial effect can be expressed in terms of $F$ and $G$ as:

$$G(x) \preceq F(x) \quad \forall x$$

where $\preceq$ indicates inequality for at least one $x$.

We say: $Y$ is stochastically larger than $X$ or $Y \geq^\text{st} X$.

$G(x) \preceq F(x)$ seems to be in contrary direction to $Y \geq^\text{st} X$.

However, $P(Y \leq x) = G(x) \preceq F(x) = P(X \leq x)$ means that small values of $Y$ are less likely than small values of $X$.

$P(Y > x) = 1 - G(x) \succeq 1 - F(x) = P(X > x)$,

$\Rightarrow$ large values of $Y$ are more likely than large values of $X$. 
Illustration for Stochastic Ordering

\[ F(x) = \Phi(x) \] standard normal

\[ G(x) = x^2, \quad x \in (0,2) \] Uniform(0,2)

\[ F(x) = x^4, \quad x \in (0,4) \] Uniform(0,4)

\[ G(x) = F(x)^{1.2} \]

not stochastically ordered

stochastically ordered
An important special case of $Y \geq X$ is the shift model:

For some $\Delta > 0$ it assumes $G(x) = F(x - \Delta)$ for all $x$.

It is a special case of $G \preceq F$ (Problem 29).

The shift model can be viewed as the treatment adding a constant amount $\Delta > 0$ to any control response.

\[
Y = X + \Delta \Rightarrow G(y) = P(Y \leq y) = P(X + \Delta \leq y) = P(X \leq y - \Delta) = F(y - \Delta)
\]

Conversely, $Y \sim G(y) = F(y - \Delta)$

\[
\Rightarrow P(Y - \Delta \leq x) = P(Y \leq x + \Delta) = F(x + \Delta - \Delta) = F(x)
\]

Thus view $Y = (Y - \Delta) + \Delta = X' + \Delta$, where $X' \sim F$. 
Illustration for the Shift Model

\[ F(x) = \Phi(x) \text{ standard normal} \]
\[ G(x) = \Phi(x - 1) \]

\[ F(x) = (x - 1)/2, x \in (1,3) \text{ Uniform}(1,3) \]
\[ G(x) = F(x - 1) \text{ Uniform}(2,4) \]
Comments on Stochastic Ordering

- $Y \geq_{st} X$ does not necessarily mean that every realization of $Y$ is $\geq$ to any realization of $X$, i.e., $Y \geq X$.

- However, this could happen when $X$ and $Y$ are highly correlated (not independent).

- As example, consider $X \sim F(x)$ and $Y = X + \Delta \sim F(x - \Delta)$. Then $X, Y$ are highly correlated. We indeed have for $\Delta > 0$

  \[ Y = X + \Delta > X \]

  for any realization of $X$ with consequent (dependent) realization of $Y = X + \Delta$.

- With independent $X$ and $Y$ we can have complete ordering $X \leq_{st} Y$ only when the support of the $X$ distribution is completely to the left of the support of the $Y$ distribution.

- Example: $X \sim U(0, 1)$ and $Y \sim U(2, 4)$, then $P(X \leq_{st} Y) = 1$. 
For the Wilcoxon rank-sum test the power function is given by

$\Pi_F(\Delta) = P_\Delta(W_s \geq \tilde{c}) = P_\Delta(W_{XY} \geq c)$ with $c = \tilde{c} - n(n + 1)/2$,

where $P_\Delta$ indicates that probabilities are calculated under the shift model $(F(x), G(x)) = (F(x), F(x - \Delta))$.

We should write $P_{F,\Delta}$. That is captured in $\Pi_F(\Delta)$.

We extend the definition of $\Pi_F(\Delta)$ to $\Delta = 0$ (no treatment effect $G = F$) and to $\Delta < 0$, negative treatment effect.

Intuitively it seems very plausible that the power function should be increasing in $\Delta$.

If the treatment adds $\Delta > 0$ ($\Delta < 0$) to the control response it should result in higher (lower) treatment ranks and a higher (lower) $W_s$, i.e., more (less) chance of rejecting $H_0 : \Delta = 0$. 
Theorem 2: $\Pi_F(\Delta)$ is an increasing function of $\Delta$.

Proof: Let $\Delta_0 < \Delta_1$. Let $X_1, \ldots, X_m$ be independent $\sim F(x)$ and $Y_1, \ldots, Y_n$ be independent $\sim G_0(y) = F(y - \Delta_0)$. Let $V_j = Y_j + (\Delta_1 - \Delta_0)(> Y_j)$, which has CDF

$$P(V_j \leq y) = P(Y_j \leq y-(\Delta_1-\Delta_0)) = F(y-(\Delta_1-\Delta_0)-\Delta_0) = F(y-\Delta_1)$$

$$\implies \Pi_F(\Delta_0) = P_{\Delta_0}(W_{XY} \geq c) \leq P_{\Delta_1}(W_{XV} \geq c) = \Pi_F(\Delta_1)$$

- To understand the $\leq$: $W_{XY}$ is the number $X_i < Y_j$.
- It is always $\leq$ to $W_{XV}$, the number of $X_i < V_j$, since $V_j > Y_j$.
- The same argument applies in the case of ties, i.e., for $W_{XY}^*$. 
Some Consequences

- If $\Pi_F(0) = \alpha_c = \alpha$ is the significance level of our Wilcoxon rank-sum test we have $\Pi_F(\Delta) \geq \alpha$ for all $\Delta > 0$.
- Tests for which the power does not fall below the significance level $\alpha$ are called unbiased against such alternatives.
- In this case unbiasedness holds against all shift alternatives, i.e., for all $F, \Delta > 0$.
- We also have $\Pi_F(\Delta) \leq \alpha$ for all $\Delta < 0$.
- Our chance of rejecting $H_0 : \Delta = 0$ in favor of a positive treatment effect ($\Delta > 0$), when in fact the treatment has no effect or even a negative effect ($\Delta < 0$), never exceeds the significance level $\alpha$.
- Thus we can view our testing problem also as testing $H'_0 : \Delta \leq 0$ against $A : \Delta > 0$.
- Our Wilcoxon test here is still an unbiased level $\alpha$ test.
Symmetric Comparison of Two Treatments Revisited

- In the symmetric comparison of A and B we chose treatment B as better when $W_B \geq n(N + 1)/2 + c$, A as better when $W_B \leq n(N + 1)/2 - c$ and suspended judgment otherwise.
- Assuming no ties, the critical value $c$ is determined so that

$$P_{H_0} (\text{choosing } A) = P_{H_0} (W_B \leq n(N + 1)/2 - c) = \alpha'$$

$$P_{H_0} (\text{choosing } B) = P_{H_0} (W_B \geq n(N + 1)/2 + c) = \alpha'$$

- $F(x)$ and $F(x - \Delta)$ be the (continuous) CDFs for A and B.
- $\Delta = 0 \implies$ none of the three decisions constitutes an error, because both treatments are equally good.
- If $\Delta < 0$ ($A$ is better than $B$), we only commit an error if $W_B \geq n(N + 1)/2 + c$ with probability

$$P_{\Delta} (W_B \geq n(N + 1)/2 + c) \leq P_0 (W_B \geq n(N + 1)/2 + c) = \alpha'$$

- Similarly for $\Delta > 0$. The error probability is always $\leq \alpha'$.
(see Slide 101, Chapter 1)
Our shift model assumed a constant shift effect $\Delta$ for all treated subjects.

A more general shift model allows the effect of the treatment to depend on $X$, the response under the control, i.e., $Y = X + \Delta(X)$.

The treatment effect will be beneficial if $\Delta(x) \geq 0$ for all $x$.

It can be shown that $Y = X + \Delta(X)$ with $\Delta(x) \geq 0$ for all $x$ implies $G(x) \leq F(x)$ for all $x$, i.e., stochastic ordering.

If $F$ and $G$ are continuous and strictly increasing then $G(x) \leq F(x)$ for all $x \Rightarrow$ the existence of a function $\Delta(x) \geq 0$ such that $G$ is the CDF of $X + \Delta(X)$ when $F$ is the CDF of $X$. 
For the generalized shift model $Y = X + \Delta(X)$ we have

(i') The Wilcoxon test is unbiased against alternatives $\Delta(x) \geq 0$ for all $x$

(ii') The Wilcoxon test has level $\alpha$ over the wider hypothesis $H'_0: \Delta(x) \leq 0$ for all $x$

(iii') For the symmetric comparison of two treatments the maximum error probability still is $\alpha'$, provided the treatment effect is $\Delta(x) \geq 0$ for all $x$ or $\Delta(x) \leq 0$ for all $x$. 
So far our statements concerning power properties were mainly qualitative.

The power $\Pi(F, G) = P_{F,G}(W_{XY} \geq c)$ depends strongly on $F$ and $G$ and is typically not easy to compute, except in some special instances.

In R the value of any power $\Pi(F, G)$ can easily be estimated via simulation, if we know how to generate independent random samples $X_1, \ldots, X_m \sim F$ and $Y_1, \ldots, Y_n \sim G$.

The accuracy of such estimates is easily controlled by the number $N_{\text{sim}}$ of simulations to be run.

The time to run such simulations is proportional to $N_{\text{sim}}$ and increases with $m$ and $n$.

Fortunately we have another option for large $m$ and $n$. 
Mean and Variance of $W_{XY}$

\[ W_{XY} = \sum_{i=1}^{m} \sum_{j=1}^{n} I\{X_i < Y_j\} \]

\[ = \text{number of pairs } (X_i, Y_j) \text{ with } X_i < Y_j \]

For $F$ and $G$ continuous the mean and variance of $W_{XY}$ are

\[ E(W_{XY}) = mnp_1 \quad \text{with} \quad p_1 = P_{F,G}(X < Y) \]

\[ \text{var}(W_{XY}) = mn p_1 (1 - p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2) \]

where

\[ p_2 = P(X < Y \cap X < Y') \]

and

\[ p_3 = P(X < Y \cap X' < Y) \]

with $X, X', Y, Y'$ independent with $X, X' \sim F$ and $Y, Y' \sim G$. 
Mean and Variance when $F = G$

- Assuming $F = G$ is continuous we have

$$p_1 = P(X < Y) = P(Y < X) = \frac{1}{2} \quad \text{since} \quad P(X = Y) = 0$$

Thus $E(W_{XY}) = \frac{mn}{2}$ and

$$p_2 = P(X < Y \cap X < Y') = \frac{1}{3}$$

since each of $X, Y, Y'$ has the same chance to be the smallest.

$$p_3 = P(X < Y \cap X' < Y) = \frac{1}{3}$$

since each of $X, X', Y$ has the same chance to be the largest.

$$\text{var}(W_{XY}) = \frac{mn}{4} + mn(n - 1) \left(\frac{1}{3} - \frac{1}{4}\right) + mn(m - 1) \left(\frac{1}{3} - \frac{1}{4}\right)$$

$$= \frac{mn}{12} (3 + n - 1 + m - 1) = \frac{mn(N + 1)}{12}$$
For large $m$ and $n$ the distribution of $W_{XY}$ is approximately normal with the previously given mean and variance so that

$$\frac{W_{XY} - E(W_{XY})}{\sqrt{\text{var}(W_{XY})}} \approx N(0, 1) \quad \text{provided} \quad 0 < p_1 < 1.$$ 

$p_1 = 0$ and $p_1 = 1$ are trivial, with separated distributions.

We approximate the CDF of $W_{XY}$ as follows

$$P(W_{XY} \leq w) = P\left(\frac{W_{XY} - E(W_{XY})}{\sqrt{\text{var}(W_{XY})}} \leq \frac{w - E(W_{XY})}{\sqrt{\text{var}(W_{XY})}}\right) \approx \Phi\left(\frac{w + .5 - E(W_{XY})}{\sqrt{\text{var}(W_{XY})}}\right)$$

We use the continuity correction $w + .5$ in place of $w$. 

Asymptotic Distribution of $W_{XY}$
Question: Does background music enhance the average page output of a typing pool?

20 consecutive working days were randomly split into 10 and 10 to receive background music or not.

For each day the average page output was recorded.

Expecting no ties we can use a full enumeration of all $\binom{20}{10} = 184756$ splits and evaluate $W_{XY}$ each time.

We find a proportion $\alpha_c = 0.052561$ of $W_{XY}$ values $\geq c = 72$ and a proportion $\alpha_c = 0.044605$ of $W_{XY}$ values $\geq c = 73$.

Rejecting $H_0$ for $W_{XY} \geq 72$ is closest to a level $\alpha = 0.05$ test.
A shift alternative is a reasonable focus, in particular a normal shift alternative, i.e.,
$$ F = \mathcal{N}(\xi, \sigma^2) \quad \text{and} \quad G = \mathcal{N}(\xi + \Delta, \sigma^2), $$
with $\Delta = 5$ of particular interest.

$\Pi_{F,G}$ does not depend on $\xi$ but depends on $\sigma^2$, i.e.,
$$ \Pi_{F,G} = \Pi_{F,\sigma^2}(\Delta) = \Pi_F(\Delta). $$

Moving $\xi$ in both distributions has no effect on rankings.

The probability of detecting a shift increases as the background variation decreases, i.e., as $\sigma^2 \downarrow 0$.

Past experience shows that $\sigma^2 = 32$ is a reasonable value.
Example 2: Normal Approximation for Computing Power

\[ p_1 = P(X < Y) = P(X - Y < 0) = \Phi \left( \frac{\Delta}{\sigma \sqrt{2}} \right) = \Phi \left( \frac{5}{\sqrt{32}\sqrt{2}} \right) = 0.734 \]

since \( X - Y \sim \mathcal{N}(-\Delta, 2\sigma^2) \).

\[
p_2 = P(X < Y, X < Y') = P \left( \frac{X - (Y - \Delta)}{\sigma \sqrt{2}} < \frac{\Delta}{\sigma \sqrt{2}}, \frac{X - (Y' - \Delta)}{\sigma \sqrt{2}} < \frac{\Delta}{\sigma \sqrt{2}} \right)
\]

\[ = P(Z < z, Z' < z) \text{ with } z = \Delta/(\sigma \sqrt{2}) = 5/8 = .625 \]

and \( Z \) and \( Z' \sim \mathcal{N}(0, 1) \) with correlation 1/2, since

\[
\text{cov}(Z, Z') = \text{cov} \left( \frac{X - Y + \Delta}{\sigma \sqrt{2}}, \frac{X - Y' + \Delta}{\sigma \sqrt{2}} \right) = \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}
\]
Example 2: Normal Approximation (continued)

- Using `pmnorm` from package `mnormt` (see next slide)
  
  ```r
  p2 <- pmnorm(c(.625,.625), c(0,0),
    varcov=matrix(c(1,.5,.5,1),ncol=2))
  results in \( p_2 = .5996 \).
  ```

- We have \( p_3 = p_2 \).
- Thus \( E(W_{XY}) = mnp_1 = 73.4, \)
- \( \text{var}(W_{XY}) = 129.03 \)

\[
\Pi_F(\Delta = 5) \approx 1 - \Phi \left( \frac{71.5 - 73.4}{\sqrt{129.03}} \right) = 0.5665.
\]
Installing the Package `mnormt`

- Install package `mnormt` in R either by:
  - menu choice `Packages → Install package(s)`
  - then choosing the nearest download mirror site, here USA (WA), and then choose `mnormt` from the packages menu list
  - or alternatively, execute the command line `install.packages("mnormt")` in an R session and choose the mirror site when prompted.
- This package install is done once for a particular R installation.
- For each new R session you need to execute `library(mnormt)` **prior to using** `pmnorm`.
- See documentation for `mnormt` under `help.start()`→ packages for all the functions that the package `mnormt` offers.
Recapitulating the Power

- We reject $H_0$ for $W_{XY} \geq c = 72 \rightarrow$ a level $\alpha_c = .0526$ test.
- Our chance of rejecting $H_0$ when in fact we have $\Delta = 5$ is about $.5665$, based on the given $\sigma^2 = 32$.
- When $\Delta = 10$ we get $p_1 = \Phi(1.25) = 0.8944$ and $p_2 = p_3 = .8237$ from
  
  $$p_2 \leftarrow \text{pmnorm}(c(1.25,1.25),c(0,0),$$  
  $$\text{varcov}=\text{matrix}(c(1,.5,.5,1),\text{ncol}=2))$$

  with resulting $E(W_{XY}) = 89.44$ and $\text{var}(W_{XY}) = 52.41$

  and thus

  $$\Pi_F(\Delta = 10) = 1 - \Phi((71.5 - 89.44)/\sqrt{52.41}) = .9934$$

  i.e., our chance of rejecting the hypothesis $H_0$ is sufficiently high in that case.
This method should work in principle for other alternatives. However, the determination of $p_1, p_2, p_3$ may not be this easy. We made use of normal distribution properties and used the bivariate normal probability function `pmnorm`. If we can generate random variables $X \sim F$, $Y \sim G$, we can obtain good estimates of $p_1, p_2, p_3$ through simulations, followed up by the normal approximation power calculation. Especially for large $m$ and $n$ this should be quicker than a simulation of the distribution of $W_{XY}$ or $W_s$ based on repeated samples from $F$ and $G$. 
Simulations of the $W_{XY}$ distribution (using $N_{\text{sim}} = 100000$) produced the following estimates

0.59123 for $P_{F,G}(\Delta = 5)$ and 0.98021 for $P_{F,G}(\Delta = 10)$

These are not too different from our normal approximations.

However, 95% confidence intervals for the true values $\Pi_F(\Delta = 5)$ and $\Pi_F(\Delta = 10)$ can be computed as $(0.588, 0.594)$ and $(0.979, 0.981)$.

Neither one of our normal approximation values falls within its respective interval.

This is not of much practical concern, since

$.5665 < (0.588, 0.594)$ and $(0.979, 0.981) < .9934$. 
An alternate power approximation within the shift model is available for small $\Delta$, namely

$$\Pi_F(\Delta) \approx \tilde{\Pi}_F(\Delta) = \Phi \left[ \sqrt{\frac{12mn}{N+1}} f^*(0) \Delta - u_\alpha \right]$$

- $u_\alpha = \Phi^{-1}(1 - \alpha)$ is the upper $\alpha$ point of $\mathcal{N}(0, 1)$.
- $f^*(z)$ is the density of $F^*(z) = P(X - Y \leq z)$, where $X$ and $Y$ are independent with common distribution function $F$.

$$P_F(X-Y \leq z) = \int_{-\infty}^{\infty} P(X \leq y+z)f(y)dy = \int_{-\infty}^{\infty} F(y+z)f(y)dy$$

$$\Rightarrow f^*(z) = \int_{-\infty}^{\infty} f(y+z)f(y)dy \quad \Rightarrow \quad f^*(0) = \int_{-\infty}^{\infty} f^2(y)dy$$
Alternate Approximation in Normal Shift Model

- When $F = \mathcal{N}(\xi, \sigma^2)$ then $X - Y \sim \mathcal{N}(0, 2\sigma^2)$ with density

$$f^*(z) = \frac{1}{\sqrt{2\pi \cdot 2\sigma^2}} \exp\left(-\frac{z^2}{2 \cdot 2\sigma^2}\right) \Rightarrow f^*(0) = \frac{1}{2\sigma \sqrt{\pi}}$$

- We get as approximation for $\Pi_\Phi(\Delta)$

$$\tilde{\Pi}_\Phi(\Delta) = \Phi \left[ \sqrt{\frac{12mn}{N+1}} \cdot \frac{\Delta}{2\sigma \sqrt{\pi}} - u_\alpha \right] = \Phi \left[ \sqrt{\frac{3mn}{(N+1)\pi}} \cdot \frac{\Delta}{\sigma} - u_\alpha \right]$$

- This captures the clear dependence of $\Pi_F(\Delta)$ on just $\Delta/\sigma$.

- This should be clear from the equivalence of

$(\mathcal{N}(\xi, \sigma^2), \mathcal{N}(\xi + \Delta, \sigma^2))$, $(\mathcal{N}(\xi/\sigma, 1), \mathcal{N}(\xi/\sigma + \Delta/\sigma, 1))$, and $(\mathcal{N}(0, 1), \mathcal{N}(\Delta/\sigma, 1))$

with same effect on the distribution of ranks.
As an illustration of the alternate approximation in the normal shift model consider again our previous example of \( \Delta = 5, \sigma^2 = 32, m = n = 10 \) and \( \alpha = .05 \).

Then \( u_\alpha = 1.645 \) gives

\[
\tilde{\Pi}_\Phi (\Delta = 5) = \Phi \left[ \sqrt{\frac{3 \cdot 10 \cdot 10}{21 \cdot \pi}} \frac{5}{\sqrt{32}} - 1.645 \right] = .5948 \cong .5665 \ (\text{.5912})_{\text{sim}}
\]

For \( \Delta = 10 \) we get

\[
\tilde{\Pi}_\Phi (\Delta = 10) = \Phi \left[ \sqrt{\frac{3 \cdot 10 \cdot 10}{21 \cdot \pi}} \frac{10}{\sqrt{32}} - 1.645 \right] = .9832 \cong .9934 \ (\text{.9802})_{\text{sim}}
\]

The alternate approximation appears to be remarkably good even though \( \Delta/\sigma = 5/\sqrt{32} = .884 \) and \( \Delta/\sigma = 10/\sqrt{32} = 1.77 \) are not exactly small.
Comparison with $\alpha = .05$ or $\alpha = .0526$?

- Recall that our original test could not attain exactly the desired significance level of $\alpha = .05$.
- We settled for $\alpha = .0526$, rejecting $H_0$ for $W_{XY} \geq c = 72$.
- Thus it might be fairer in our previous comparison to use $\alpha = .0526$ in the alternate approximation, since we used $c = 72$ both in our simulation and in the first approximation.
- Using $u_{.0526} = 1.620$ in place of $u_{.05} = 1.645$ gives
  \[ \tilde{\Pi}_\phi(\Delta = 0) = 0.0526 \]
  \[ \tilde{\Pi}_\phi(\Delta = 5) = 0.6042 \cong .5665 ( .5912 )_{sim} \]
  \[ \tilde{\Pi}_\phi(\Delta = 10) = 0.9842 \cong .9934 ( .9802 )_{sim} \]
- This raised the approximate power slightly, as it should.
- The comparison is still quite favorable compared with the simulated values.
Compareding Exact Power with Approximations

Table 2.1 Power of the Wilcoxon rank-sum test for normal shift alternatives; \( m = n = 7, \alpha = .049 \)

<table>
<thead>
<tr>
<th>( \Delta / \sigma )</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>0.049</td>
<td>0.094</td>
<td>0.165</td>
<td>0.264</td>
<td>0.386</td>
<td>0.520</td>
<td>0.815</td>
<td>0.958</td>
</tr>
<tr>
<td>Simulated (400,000)</td>
<td>0.049</td>
<td>0.095</td>
<td>0.165</td>
<td>0.264</td>
<td>0.386</td>
<td>0.521</td>
<td>0.815</td>
<td>0.959</td>
</tr>
<tr>
<td>Approximation</td>
<td>0.048</td>
<td>0.094</td>
<td>0.160</td>
<td>0.249</td>
<td>0.359</td>
<td>0.485</td>
<td>0.807</td>
<td>0.981</td>
</tr>
<tr>
<td>Alternate Approximation</td>
<td>0.049</td>
<td>0.096</td>
<td>0.171</td>
<td>0.275</td>
<td>0.403</td>
<td>0.543</td>
<td>0.839</td>
<td>0.970</td>
</tr>
</tbody>
</table>

- .049 is rounded from .04865967 which is the attainable level closest to .05, by rejecting \( H_0 \) whenever \( W_{XY} \geq 38 \).
- The exact (Milton, 1970) and simulated values agree well.
- The last row differs from the last row of Table 2.1 in the Text, which seems to use

\[
\Pi_F(\Delta) \approx \Phi \left[ \sqrt{\frac{12mn}{N}} f^*(0)\Delta - u_\alpha \right]
\]

\( N \) in place of \( N + 1 \) in denominator.
Sample Size Planning

- Having a power of less than .6 with an output increase as large as 5 pages in Example 2 may not be satisfactory.
- Higher power ⇒ plan a larger sample size.
- Our alternate approximation suggests a maximal power for fixed \( N = m + n \) when \( m = n = N/2 \) when \( N \) is even.
- Assuming \( N = 2n \) and using \( N = 2n \) in place of \( N + 1 \) in our alternate approximation we have for desired power \( \Pi \)

\[
\Pi_F(\Delta) \approx \Phi[\sqrt{6}nf^*(0)\Delta - u_\alpha] = \Pi = 1 - \Phi(u_\Pi) = \Phi(-u_\Pi)
\]

\[
n \approx \frac{1}{6} \left( \frac{u_\alpha - u_\Pi}{\Delta f^*(0)} \right)^2 \text{ for normal } F \Rightarrow n \approx \frac{2\pi}{3} \left( \frac{u_\alpha - u_\Pi}{\Delta/\sigma} \right)^2 = \frac{2\pi}{3} \left( \frac{(u_\alpha - u_\Pi)\sigma}{\Delta} \right)^2
\]

- For \( \Delta = 5, \sigma = \sqrt{32}, \Pi = .95, \alpha = .05 \) get \( n = m \approx 29.58 \).
Example 3: Cultural Influences on IQ

- A group of underprivileged children are to meet individually for 2 hours per week with college students, to be compared with a control group without this experience at the end of the year.
- With $N = 2n$ how large should $n$ be for a level $\alpha = .01$ test to have power .95 in case the IQ increases by two points?
- We will add an experimental feature that greatly improves the experiment efficiency.
- Rather than using $n$ controls and $n$ treated subjects and measuring their IQ at the end of the year we will measure their IQ at the beginning and at the end of the year.
- The difference of IQ scores on the same subject acts as the subject response.
Two Sources of Variation

- With a single measurement per subject we would experience a greater variation in either group of responses, because of the inherent subject to subject variation.

- By taking two measurements on each subject and taking the difference as response we compound two sources of repeat measurement variability, but we eliminate the variability from subject to subject, which usually is a much larger.

\[ X_i = \mu + U_i + V_i \quad \text{and} \quad X_i' = \mu + U_i + V_i' \quad \text{with} \quad U_i, V_i, V_i' \text{ independent} \]

\[ \Rightarrow \quad X_i' - X_i = V_i' - V_i \quad \Rightarrow \quad \text{var}(X_i' - X_i) = \sigma_V^2 + \sigma_V'^2 = 2\sigma_V^2 \]

\[ \text{var}(X_i) = \sigma_U^2 + \sigma_V^2 \gg 2\sigma_V^2 \text{ since typically } \sigma_U \gg \sigma_V. \]

\[ Y_i = \mu + U_i + V_i \quad \text{and} \quad Y_i' = \mu + \Delta + U_i + V_i' \quad \text{with} \quad U_i, V_i, V_i' \text{ independent} \]

\[ Y_i' - Y_i = \Delta + V_i' - V_i \quad \Rightarrow \quad \text{var}(Y_i' - Y_i) = \sigma_V^2 + \sigma_V'^2 = 2\sigma_V^2 \]
Why Greater Efficiency?

- Recall

\[ \Pi_F(\Delta) \approx \Phi \left[ \sqrt{\frac{3mn}{(N + 1)\pi}} \frac{\Delta}{\sigma} - u_\alpha \right] \]  
  when \( F \) is normal.

- For fixed \( m \) and \( n \) the power increases as \( \sigma \) gets reduced, which is exactly what we hope to accomplish by taking the difference in responses on the same subject.

- Conversely,

\[ n \approx \frac{2\pi}{3} \left( \frac{(u_\alpha - u_\Pi)\sigma}{\Delta} \right)^2 \]  
  when \( F \) is normal.

- Smaller \( \sigma \) \( \implies \) smaller \( n = N/2 \) to achieve the same power \( \Pi \).

- This makes sense, since the presence of a signal \( \Delta \) is more easily discernible against less response variation.
Sample Size Calculation

- Determine $m = n$ for which the rank-sum test at level $\alpha = .01$ will give power .95 for a $\Delta = 2$ improvement of IQ scores.
- For the calculation we need one further piece of information. Assume that the subject score difference is approximately normally distributed with variance $\sigma^2 = 2$.
- The normality assumption is helped by the fact that we take the difference of two independent test scores as our response.
- We get

$$n \approx \frac{2\pi}{3} \left( \frac{(u_\alpha - u_\Pi)\sigma}{\Delta} \right)^2$$

$$= \frac{2\pi}{3} \left( \frac{(2.326 - (-1.645))\sqrt{2}}{2} \right)^2 = 16.515$$

thus $n = 17$ should do.
Recall that for the sample size determination we replaced \( N + 1 \) by \( N \) in the alternate approximation.

We will now calculate the approximate power for \( \alpha = .01 \) at \( \Delta = 2 \) and \( \sigma = \sqrt{2} \) when using \( n = m = 17 \) and the correct \( N + 1 = 35 \).

The alternate power approximation for normal \( F \) is

\[
\Phi \left( \sqrt{\frac{12mn}{N+1}} \frac{\Delta}{2\sigma\sqrt{\pi}} - u_\alpha \right) = \Phi \left( \sqrt{\frac{12 \cdot 17 \cdot 17}{35}} \frac{2}{2\sqrt{2\pi}} - u_\alpha \right) = 0.949994
\]

Ignoring approximation error this agrees with the desired power of \( \beta = .95 \).
Student’s $t$-Test

- The classical test for comparing two treatments is Student’s $t$-test which rejects the hypothesis $H_0$ of no effect when

$$t(X, Y) = \frac{\bar{Y} - \bar{X}}{S \sqrt{1/n + 1/m}} \geq c = t_{n+m-2}(1 - \alpha)$$

with

$$S^2 = \frac{\sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{j=1}^{n} (Y_j - \bar{Y})^2}{m + n - 2}$$

$$t_f(\gamma) = \gamma\text{-quantile of } t\text{-distribution with } f \text{ degrees of freedom}.$$ 

- This test assumes the normal shift model in which $X = (X_1 \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$ are independently, normally distributed with common unknown variance $\sigma^2$ and with unknown means $E(X_i) = \xi$ and $E(Y_j) = \xi + \Delta$.

- In this normal shift model the $t$-test is uniformly most powerful among all unbiased tests of $H_0 : \Delta = 0, \sigma > 0$ (or $H'_0 : \Delta \leq 0, \sigma > 0$) against alternatives $A : \Delta > 0, \sigma > 0$. 
Under $H_0 : \Delta = 0, \sigma > 0$ the $t$-statistic $t(X, Y)$ has a (central) Student $t$-distribution with $m + n - 2$ degrees of freedom.

When $\Delta \neq 0$ the test statistic $t(X, Y)$ has a noncentral Student $t$-distribution with $m + n - 2$ degrees of freedom and with noncentrality parameter $\delta = \Delta / [\sigma \sqrt{1/m + 1/n}]$.

It becomes the central Student $t$ distribution when $\Delta = 0$.

$R$ gives us the density, CDF, quantile, and random samples from the Student $t$ distribution through the functions: $dt(x, f, ncp)$, $pt(q, f, ncp)$, $qt(p, f, ncp)$, and $rt(n, f, ncp)$, respectively.

$ncp = \text{noncentrality parameter} = \delta = \Delta / [\sigma \sqrt{1/m + 1/n}]$ and $f$ denotes the degrees of freedom. Here $f = m + n - 2$.

The critical point $c$ from the previous slide is obtained as $c = t_{m+n-2}(1 - \alpha) = qt(1-\alpha, m + n - 2)$. 
Often the assumption of normality is not satisfied or it is quite tenuous. We will examine the behavior of both tests with respect to significance level and power in such situations. We know that the significance level of the Wilcoxon test is independent of $F$ when testing $H_0 : F = G$. This is not true for the $t$-test when testing $H_0 : F = G$ and $F$ is not normal.
Large Sample Properties of $\bar{X}, \bar{Y}, S^2$

When $\Delta = 0$ we have from the CLT that

$$Z_1 = \frac{\bar{X} - \xi}{\sigma/\sqrt{m}} = \sqrt{m} \frac{\bar{X} - \xi}{\sigma} \approx \mathcal{N}(0, 1) \quad \text{and} \quad Z_2 = \frac{\bar{Y} - \xi - 0}{\sigma/\sqrt{n}} = \sqrt{n} \frac{\bar{Y} - \xi}{\sigma} \approx \mathcal{N}(0, 1)$$

$$\implies \frac{\bar{Y} - \bar{X}}{\sigma \sqrt{1/n + 1/m}} = \frac{(\bar{Y} - \xi)/\sigma - (\bar{X} - \xi)/\sigma}{\sqrt{1/n + 1/m}} = \frac{Z_2/\sqrt{n} - Z_1/\sqrt{m}}{\sqrt{1/n + 1/m}} \approx \mathcal{N}(0, 1)$$

By the law of large numbers (LLN) we have as $m, n \to \infty$

$$\sum_{i=1}^{m} (X_i - \bar{X})^2 \quad \sum_{i=1}^{m} (X_i - \bar{X})^2 - m(\bar{X} - \xi)^2 \quad \sum_{i=1}^{m} (X_i - \xi)^2 \quad (\bar{X} - \xi)^2 \quad P \to \sigma^2$$

$$and \text{ similarly } \sum_{i=1}^{n} (Y_j - \bar{Y})^2 \quad P \to \sigma^2 \quad and \text{ thus }$$

$$S^2 = \frac{m \sum_{i=1}^{m} (X_i - \bar{X})^2 / m + n \sum_{i=1}^{m} (Y_j - \bar{Y})^2 / n}{m + n - 2} \quad P \to \sigma^2 \quad and \text{ thus } \frac{S}{\sigma} \quad P \to 1$$
Combining the results from the previous slide we have

\[
t(X, Y) = \frac{\bar{Y} - \bar{X}}{\sigma \sqrt{1/n + 1/m} \times S/\sigma} \to N(0, 1) \quad \text{as} \quad m, n \to \infty.
\]

the above result shows that \( t_{m+n-2}(1 - \alpha) \to u_\alpha \), the upper \( \alpha \) quantile \( N(0, 1) \).

In large samples we can as well use \( u_\alpha \) in place of \( t_{m+n-2}(1 - \alpha) \) with the added advantage that the level is approximately \( \alpha \) no matter what the nature of \( F \) is, aside from it having a finite variance \( \sigma^2 \).

The approximation quality may still depend on the nature of \( F \) and on \( m \) and \( n \).

In that sense the \( t \)-test is asymptotically distribution-free.
Comparing the Power of $t(\mathbf{X}, \mathbf{Y})$ and $W_{XY}$

- The power function of the $t$-test is

$$\Pi'_F(\Delta) = P \left( \frac{\bar{Y} - \bar{X}}{S \sqrt{1/m + 1/n}} \geq c \right)$$

$$= P \left( \frac{\bar{Y} - \bar{X} - \Delta}{\sigma \sqrt{1/m + 1/n}} \geq \frac{cS - \Delta}{\sigma \sqrt{1/m + 1/n}} \right)$$

$$\approx P \left( Z \geq u_\alpha - \frac{\Delta}{\sigma \sqrt{1/m + 1/n}} \right) = \Phi \left( \frac{\Delta \sqrt{mn}}{N} - u_\alpha \right)$$

- We used $(\bar{Y} - \bar{X} - \Delta)/[\sigma \sqrt{1/m + 1/n}] \approx Z \sim \mathcal{N}(0, 1)$, $S/\sigma \approx 1$ and $c \approx u_\alpha$ for large samples.

- Compare this with our 2nd power approximation for $W_{XY}$:

$$\Pi_F(\Delta) \approx \Phi \left[ \sqrt{\frac{12mn}{N + 1}} \right] f^*(0) \Delta - u_\alpha$$

- Note the similarities in the approximations to the two power functions. Both involve $m$ and $n$ through roughly equal factors.

- The distribution $F$ shows in $\sigma = \sigma_F$ and $f^*(0)$, respectively.
Using $\alpha = .0526$ our 2nd approximation gave us a power of $\Pi_\Phi(\Delta = 5) = .604$ for the Wilcoxon test when $\sigma^2 = 32$.

Using the same level for the $t$-test we get approximate power

$$\Pi'(\Delta = 5) = \text{pnorm}(\sqrt{10/20} \times 5/\sqrt{32} - \text{qnorm}(1 - .0526)) = 0.639$$

This is slightly higher than the value .63 given in the Text. Due to the level of .0526 used here, instead of .05.

Based on the normality assumption the power of the $t$-test is

$$1 - \text{pt}(\text{qt}(1 - .0526, 10 + 10 - 2, 0), 10 + 10 - 2, 5/\sqrt{32 * (1/10 + 1/10)})$$

which computes to 0.612.

The Wilcoxon test achieves 94.5% (.604/0.639) or 98.7% (.604/.612) of the power of the $t$-test when $F$ is normal.

This is remarkable, since $t(X, Y)$ uses the actual sample values and $W_{XY}$ uses only the ranks.
Comparing Sample Size Requirements

- Suppose we determine the sample size $m = n$ to achieve a power $\beta$ with the Wilcoxon test for a given $\Delta > 0$.
- What sample sizes $m' = n'$ would it require for the $t$-test to achieve the same power? Equating

$$
\beta = \Phi \left[ \frac{\Delta}{\sigma} \sqrt{\frac{m'n'}{N'}} - u_\alpha \right] = \Phi \left[ \Delta f^*(0) \sqrt{\frac{12mn}{N + 1}} - u_\alpha \right]
$$

we need to equate

$$
\frac{1}{\sigma} \sqrt{\frac{m'}{2}} = \frac{1}{\sigma} \sqrt{\frac{m'n'}{N'}} = f^*(0) \sqrt{\frac{12mn}{N + 1}} = f^*(0) \sqrt[4]{\frac{6m}{1 + \frac{1}{2m}}}
$$

$$\implies m' = (\sigma f^*(0))^2 \frac{12m}{1 + \frac{1}{2m}} \approx 12m\sigma^2 f^{*2}(0)$$
The ratio $m'/m$ is called the efficiency of the Wilcoxon test relative to the $t$-test.

For example, an efficiency $m'/m = 1/2$ means that the $t$-test requires half as many observations as are needed by the Wilcoxon test in order to get the same power.

The limiting value of $m'/m$, denoted by $e_{W,t}(F)$, is called the Pitman efficiency or asymptotic relative efficiency (ARE) of the Wilcoxon test w.r.t. the $t$-test. We have

$$e_{W,t}(F) = \lim_{m \to \infty} (m'/m) = 12\sigma^2 f^*(0)$$

When $F$ is normal we have $f^*(0) = 1/(2\sigma\sqrt{\pi})$ so that $e_{W,t}(\Phi) = 3/\pi = .955$. 
The previous efficiency formulas are based on approximations that require large sample sizes.

One benefit: the dependence on $\Delta/\sigma$ and $\alpha$ dropped out.

Dixon (1954) made exact power calculations for the normal shift model with $m = n = 5$ and $\alpha = 4/126$.

Note that $\binom{10}{5} = 252$. Thus Dixon (1954) had to calculate the probability of the four most extreme rank sums on either end.

He considered two-sided tests: $\alpha = 2 \times 4/252 = .0317$.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pi$</td>
<td>.072</td>
<td>.210</td>
<td>.431</td>
<td>.674</td>
<td>.858</td>
<td>.953</td>
<td>.988</td>
<td>.998</td>
</tr>
<tr>
<td>$e$</td>
<td>.968</td>
<td>.978</td>
<td>.961</td>
<td>.956</td>
<td>.960</td>
<td>.960</td>
<td>.964</td>
<td>.976</td>
</tr>
</tbody>
</table>

The row $e$ is a modification by Hodges and Lehmann (1956).

The small sample efficiency is well represented by the ARE.
\( e_{W, t}(F) \) for Other Distributions \( F \)

<table>
<thead>
<tr>
<th>( F )</th>
<th>Logistic</th>
<th>Double Exponential</th>
<th>Rectangular</th>
<th>Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e_{W, t}(F) )</td>
<td>( \pi^2 / 9 = 1.097 )</td>
<td>1.5</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

- Hodges and Lehmann (1956) showed
  
  \[
  e_{W, t}(F) \geq \frac{108}{125} = 0.864
  \]

  for all distributions \( F \) with finite variance \( \sigma^2 \).

- See the Appendix of the Text for the proof and the minimizing \( F \).

- This makes a very strong case for using the Wilcoxon test in such situations.
Sir David Cox in his 2008 UW Norm Breslow Lecture said something like this:

If you don’t have a nonparametric procedure don’t rely on a parametric one, but when you have a nonparametric solution, go ahead and use a parametric one.

My interpretation: Do not rely on procedures that impose a parametric model to understand the data, unless you have a nonparametric way to check what you are doing.
Dealt with testing $H_0 : \Delta = 0$ in the shift effect model.

Easily extended to testing $H_{\Delta_0} : \Delta = \Delta_0$ for any specified $\Delta_0$.

The solution is to subtract $\Delta_0$ from the $Y_1, \ldots, Y_n$, i.e., form $Y_i' = Y_i - \Delta_0$, and then test $H_0 : \Delta = 0$ in terms of $X_1, \ldots, X_m$ and $Y_1', \ldots, Y_n'$.

To see the equivalence note that for $Y' = Y - \Delta_0$ and $Y \sim F(y - \Delta)$ we have

$$P(Y' \leq y) = P(Y - \Delta_0 \leq y) = P(Y \leq y + \Delta_0)$$
$$= F(y + \Delta_0 - \Delta) = F(y - (\Delta - \Delta_0))$$

$Y' \sim F(y - \Delta')$ with $\Delta' = \Delta - \Delta_0$.

$H_0 : \Delta' = 0 \iff H_0 : \Delta - \Delta_0 = 0 \iff H_{\Delta_0} : \Delta = \Delta_0$. 

\[ \text{Testing } H_{\Delta_0} : \Delta = \Delta_0 \]
The two-sided Wilcoxon test rejects $H_0 : F = G$ whenever

$$\left| W_{XY} - \frac{mn}{2} \right| = \left| W_s - \frac{n(N + 1)}{2} \right| \geq k$$

This symmetry of the two-tailed rejection region is justified by the symmetry of the null distribution.

The question arises whether the power function $\Pi(F, G) = \Pi_F(\Delta)$ of this test is symmetric under the shift model $G(x) = F(x - \Delta)$.

Do we have $\Pi_F(\Delta) = \Pi_F(-\Delta)$?

This symmetry indeed holds in two situations, namely when

i) $m = n$ or

ii) $F$ is symmetric around some point $a$. 
Proof of $\Pi_F(\Delta) = \Pi_F(-\Delta)$

Using the notation: $X_i, Y'_j \sim F(x)$ and $Y_j = Y'_j + \Delta \sim F(x - \Delta)$

$$\Pi_F(\Delta) = P \left( \left| \sum_{ij} I[X_i < Y_j] - \frac{mn}{2} \right| \geq k \right) = P \left( \left| \sum_{ij} I[X_i - Y'_j < \Delta] - \frac{mn}{2} \right| \geq k \right)$$

$$\Pi_F(-\Delta) = P \left( \left| \sum_{ij} I[X_i - Y'_j < -\Delta] - \frac{mn}{2} \right| \geq k \right) = P \left( \left| \sum_{ij} I[Y'_j - X_i > \Delta] - \frac{mn}{2} \right| \geq k \right)$$

$$= P \left( \left| \frac{mn}{2} - \sum_{ij} I[Y'_j - X_i < \Delta] \right| \geq k \right) \ast = P \left( \left| \sum_{ij} I[X_i - Y'_j < \Delta] - \frac{mn}{2} \right| \geq k \right)$$

In $\ast$ we made the following legitimate interchanges:

when $m = n$, we interchanged $X_i$ with $Y'_j$ or
when $F$ is symmetric around $a$, we have

$$X_i - Y'_j = X_i - a - (Y'_j - a) \overset{D}{=} a - X_i - (a - Y'_j) = Y'_j - X_i$$

since $X_i - a$ and $Y'_j - a$ are distributed symmetrically around zero.
Often we are interested in estimating the shift effect $\Delta$ itself, without saying anything about the CDF $F$, except maybe that it should be continuous.

Even more ambitiously we may want to obtain a confidence interval for $\Delta$.

There is a strong connection between confidence intervals and testing hypotheses.
The hypothesis $H_0 : \Delta = 0$ is least likely rejected whenever the $X_i$’s and $Y_j$’s are in close agreement with each other.

Close agreement is indicated by a high $p$-value for the two-sided Wilcoxon test.

We get the highest $p$-value ($= 1$) and thus the least reason for rejecting $H_0 : \Delta = 0$ when half of the $Y_j - X_i > 0$ and half of the $Y_j - X_i < 0$, i.e., when $W_{XY} = mn/2$.

When the alignment is not so good, shift the $Y$-sample by an amount $\hat{\Delta}$ that leaves the shifted $Y_j$, i.e., $Y_j' = Y_j - \hat{\Delta}$, in “closest” agreement with the $X_i$,

i.e., half of $Y_j' - X_i = Y_j - \hat{\Delta} - X_i > 0$ or half of $Y_j - X_i > \hat{\Delta}$ and half of $Y_j - X_i < \hat{\Delta}$. 
Shift Estimate Illustration

X-sample

Y-sample

aligned Y-sample

\[ \Delta \]

\[ \hat{\Delta} \]
The Hodges-Lehmann Estimator \((mn \text{ even})\)

- Denote by \(D(1) < D(2) < \ldots < D(mn)\) the ordered values of all \(mn\) differences \(Y_j - X_i\).
- The \(D(i)\) are all distinct with probability 1 for continuous \(F\).
- When \(mn = 2k\) is even (\(m\) or \(n\) even), we saw that \(W_{X,Y-a}\) is closest to its central value \(mn/2 = k\) when \(D(k) < a < D(k+1)\).
- Thus we get a whole interval of values \(a\) that give us a two-sided \(p\)-value of one.
- A natural choice as estimate is the midpoint of that interval

\[
\hat{\Delta} = \frac{D(k) + D(k+1)}{2} = \text{median}(Y_j - X_i, \ i = 1, \ldots, m, \ j = 1, \ldots, n)
\]

\[
= \text{med}(Y_j - X_i)
\]

- This is known as the Hodges-Lehmann estimator of \(\Delta\).
The case $mn = 2k + 1$ odd requires some special attention.

Under $H_0$: $W_{XY}$ is symmetric around $\frac{mn}{2} = k + \frac{1}{2}$ (fractional).

\[
W_{X,Y-a} > k + \frac{1}{2} \iff \#\{Y_j - a - X_i > 0\} > k + \frac{1}{2} \\
\iff \#\{Y_j - X_i > a\} > k + \frac{1}{2} \\
\iff \#\{D(\ell) > a\} > k + \frac{1}{2} \iff a < D_{(k+1)}
\]

and similarly

\[
W_{X,Y-a} < k + \frac{1}{2} \iff a > D_{(k+1)}
\]

- $a = D_{(k+1)} = Y_{j_0} - X_{i_0}$: observations $Y_{j_0} - a$ and $X_{i_0}$ are tied.
- $W_{X,Y}^*$ counted such comparisons with weight $\frac{1}{2}$.
- As $a$ passes over $D_{(k+1)}$ the statistic $W_{XY}$ in its modified form $W_{X,Y}^*$ achieves the central value $k + \frac{1}{2}$. It makes sense to take

\[
\hat{\Delta} = D_{(k+1)} = \text{median}(Y_j - X_i, \ i = 1, \ldots, m, \ j = 1, \ldots, n) = \text{med}(Y_j - X_i)
\]
The following drive times to work on routes A & B were collected by Gus Haggstrom

Route A: 6.0 5.8 6.5 5.8 6.3 6.0 6.3 6.4 5.9 6.5 6.0
Route B: 7.3 7.1 6.5 10.2 6.8

Finding \( \text{med}(Y_j - X_i) \) manually can be daunting, because \( mn \) can be large.

Here we have \( mn = 11 \times 5 = 55 \).

The Text discusses a relatively efficient manual computation.

We will use \( R \) instead.
> outer(RouteB, RouteA, "-")

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>[2,]</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>[3,]</td>
<td>1.3</td>
<td>1.3</td>
<td>1.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>[4,]</td>
<td>1.5</td>
<td>1.5</td>
<td>1.4</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>[5,]</td>
<td>4.4</td>
<td>4.4</td>
<td>4.3</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>3.9</td>
<td>3.9</td>
<td>3.8</td>
<td>3.7</td>
</tr>
</tbody>
</table>

The `outer(y, x, "-")` function call forms all pairs \((y_j, x_i)\) and outputs their differences in an \(n \times m\) array.

> median(outer(RouteB, RouteA, "-"))

[1] 0.9

computes the median \(\text{med}(\text{RouteB}[j] - \text{RouteA}[i])\) of the previous array, i.e., the Hodges-Lehmann estimator.
Using `wilcox.test`

```r
> wilcox.test(RouteB, RouteA, conf.int=T, exact=F)

Wilcoxon rank sum test with continuity correction

data:  RouteB and RouteA
W = 54, p-value = 0.003003
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  0.4999546 3.7000985
sample estimates:
difference in location
  0.9000572
```

- **The estimate** 0.9000572 does not match ours.
- Seems linked to ties in the data. 10 unique values.
- A normal approximation is used with iteration.
Using `wilcox.test` without Ties

```r
> x
[1] -1.44 -1.20 -1.15 -1.08 -0.90
[6] -0.75 -0.51 -0.07 0.31 0.37 1.14
> y
[1] -0.91 -0.50 0.05 0.25 0.87

> median(outer(y, x, "-"))
[1] 0.53

> wilcox.test(y, x, conf.int=T)
# exact=F not needed without ties

Wilcoxon rank sum test

data:  y and x
W = 37, p-value = 0.3196
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.43  1.40
sample estimates:
difference in location
    0.53
```
Using `wilcox_test` in the Package `coin`

Prepare data for required structure in `wilcox_test`.

```r
> Route <- factor(c(rep("A",11),rep("B",5)),
                  levels=c("B","A"))
> Route.Time <- c(RouteA,RouteB)
> Route
> Route.Time
[1]  5.8  5.8  5.9  6.0  6.0  6.0  6.3  6.3
[9]  6.4  6.5  6.5  6.5  6.8  7.1  7.3 10.2
> library(coin)
```
> wilcox_test(Route.Time~Route, conf.int=TRUE,
  distr = exact())

Exact Wilcoxon Mann-Whitney Rank Sum Test

data:  Route.Time by Route (B, A)
Z = 3.0245, p-value = 0.001374
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
  0.5 3.7
sample estimates:
difference in location
  0.9
Robustness of $\hat{\Delta} = \text{med}(Y_j - X_i)$

- How sensitive is the Hodges-Lehmann estimator to outliers?
- If $k_1$ of the $X_i \to -\infty$ and $k_2$ of the $Y_j \to \infty$ then exactly $(m - k_1)(n - k_2)$ of the differences $Y_j - X_i$ will stay put.
- The other $mn - (m - k_1)(n - k_2)$ will move to $\infty$.
- Thus the median of all these differences will stay put as long as

$$mn - (m - k_1)(n - k_2) < (m - k_1)(n - k_2) \quad \text{or} \quad \frac{1}{2} < \left(1 - \frac{k_1}{m}\right) \left(1 - \frac{k_2}{n}\right)$$

To simplify assume $k_1 = k_2 = k$ and $m = n$ and get

$$\left(1 - \frac{k}{m}\right)^2 > \frac{1}{2} \quad \text{or} \quad \frac{k}{m} < 1 - \frac{1}{\sqrt{2}} = .293$$

- We could tolerate about 29% outliers in either sample when $m = n$ without affecting the estimate $\hat{\Delta}$.
- Compare this with no outlier tolerance in $\bar{Y} - \bar{X}$ and almost 50% outlier tolerance in either sample for $\text{med}(Y_j) - \text{med}(X_i)$. 
The Effect of Ties Due to Rounding

- Often ties occur because measurements are not given to the full possible accuracy and that rounding took place.
- Assume that the original observations are \( X'_i \) and \( Y'_j \), distributed according to continuous distributions (without ties).
- The respective rounded values are \( X_i \) and \( Y_j \), which take values among a grid of rounded values given by 0, \( \pm \epsilon \), \( \pm 2\epsilon \), \ldots.
- From rounding to the nearest multiple \( k\epsilon \) we have \( |X_i - X'_i| \leq \frac{\epsilon}{2} \) and \( |Y_j - Y'_j| \leq \frac{\epsilon}{2} \)

\[
\Rightarrow |(Y_j - X_i) - (Y'_j - X'_i)| = |(Y_j - Y'_j) - (X_i - X'_i)| \leq |Y_j - Y'_j| + |X_i - X'_i| \leq \epsilon
\]

\[
|\hat{\Delta}' - \hat{\Delta}| = |\text{med}(Y'_j - X'_i) - \text{med}(Y_j - X_i)| \leq \epsilon
\]

- The rounding errors don’t build up.
\[ |a_i - a'_i| \leq \epsilon \ \forall i \implies |a(k) - a'(k)| \leq \epsilon \ \forall k \]

- Here \( a(k) \) \((a'_{(k)})\) is the \( k^{\text{th}} \) ordered value of the \( a_i \) \((a'_i)\).
- Proof of the above implication:
  Assume \( a'(k) < a(k) - \epsilon \), i.e., at least \( k \) of the \( a'_i \) are \(< a(k) - \epsilon \).
  Thus we also have that the corresponding \( a_i \) are \(< a(k) \).
  Since there are at least \( k \) such \( a_i \) we get a contradiction to the definition of \( a(k) \). It only exceeds at most \( k - 1 \) of the \( a_i \).
  The assumption \( a'(k) > a(k) + \epsilon \) leads to a contradiction in a similar fashion.
- This result \( \implies |\text{med}(Y'_j - X'_i) - \text{med}(Y_j - X_i)| \leq \epsilon \) using \( a_\nu = Y_j - X_i \) and \( a'_{\mu} = Y'_j - X'_i \).
When \( X_i \sim F(x) \) and \( Y_j \sim G(y) = F(y - \Delta) \) with \( F \) continuous, then \( \hat{\Delta} \) has a continuous distribution, i.e.,

\[
P(\hat{\Delta} = d) = 0 \quad \text{for any } d \in R.
\]

For \( mn = 2k + 1 \) (odd) this follows from

\[
P(\hat{\Delta} = d) \leq \sum_i \sum_j P(Y_j - X_i = d) = 0
\]

For \( mn = 2k \) (even) this follows similarly from

\[
P(\hat{\Delta} = d) \leq \sum_i \sum_j \sum_{i'} \sum_{j'} P(Y_j - X_i + Y_{j'} - X_{i'} = 2d) = 0
\]
Lemma 1: The distribution of $\hat{\Delta} - \Delta$ is independent of $\Delta$.

Proof:

$\hat{\Delta} - \Delta = \text{med}(Y_j - X_i) - \Delta = \text{med}(Y_j - \Delta - X_i)$

has a distribution independent of $\Delta$ because the distribution of $X_i \sim F$ does not involve $\Delta$ and

$$P(Y_j - \Delta \leq y) = P(Y_j \leq y + \Delta) = F(y + \Delta - \Delta) = F(y)$$

no longer involves $\Delta$. 
Theorem 3: The estimator $\hat{\Delta}$ of the shift parameter $\Delta$ is distributed symmetrically around $\Delta$ if either of the following two conditions hold:

(i) The distribution $F$ is symmetric around some point $\mu$, i.e,
$$X - \mu \overset{D}{=} \mu - X.$$ 

(ii) The two sample sizes are equal, i.e., $m = n$. 

Proof

(i) symmetry of $F$ around $\mu$

\[ \Rightarrow X_i - \mu \overset{\mathcal{D}}{=} \mu - X_i \text{ and } Y_j - \Delta - \mu \overset{\mathcal{D}}{=} \mu + \Delta - Y_j \]

\[
\hat{\Delta} - \Delta = \text{med}(Y_j - X_i) - \Delta = \text{med}(Y_j - \Delta - \mu - (X_i - \mu)) \\
\overset{\mathcal{D}}{=} \text{med}(\mu + \Delta - Y_j - (\mu - X_i)) \\
= \Delta - \text{med}(Y_j - X_i) = \Delta - \hat{\Delta}
\]

(ii) $(X_i, Y_j - \Delta)$ and $(Y_i - \Delta, X_j)$ have the same distribution

\[
\hat{\Delta} - \Delta = \text{med}(Y_j - \Delta - X_i) \overset{\mathcal{D}}{=} \text{med}(X_j - (Y_i - \Delta)) \\
= \Delta - \text{med}(Y_i - X_j) = \Delta - \hat{\Delta}
\]

The $X$ and $Y$ interchange is possible since $m = n$. 
The symmetry property on the previous slide implies two unbiasedness properties.

\( \hat{\Delta} \) is unbiased, i.e., \( E(\hat{\Delta}) = \Delta \), provided the expectation exists.

The distribution of \( \hat{\Delta} \) has median \( \Delta \).

\( \hat{\Delta} \) is median unbiased, i.e., \( P(\hat{\Delta} < \Delta) = P(\hat{\Delta} > \Delta)(= \frac{1}{2}) \).

where the \( \frac{1}{2} \) applies when \( F \) is also continuous.

When the conditions of Theorem 3 are no longer satisfied, then \( \hat{\Delta} \) is usually no longer distributed symmetrically around \( \Delta \).

The median unbiasedness property holds when \( mn \) is odd, and approximately so when \( mn \) is even. See the next four slides.
Theorem 4: Suppose the differences $Y_j - X_i$ are distinct. If $D(1) < \ldots < D(mn)$ denote the ordered differences $Y_j - X_i$, then for any $\ell$ with $1 \leq \ell \leq mn$ and any $a \in R$ we have

$$D(\ell) \leq a \iff W_{X,Y-a} \leq mn - \ell$$

$$D(\ell) > a \iff W_{X,Y-a} \geq mn - \ell + 1$$

Proof:

$$D(\ell) \leq a \iff \text{at least } \ell \text{ of the differences } (Y_j - a) - X_i \text{ are } \leq 0,$$

$$\iff \text{at most } mn - \ell \text{ of these differences are } > 0$$

$$\iff W_{X,Y-a} \leq mn - \ell.$$
Assume $mn = 2k + 1$, we then have $\hat{\Delta} = D_{(k+1)}$ and let $F$ be continuous $\Rightarrow D_{(1)} < \ldots < D_{(mn)}$ with probability 1.

\[
P_{\Delta}(\hat{\Delta} < \Delta) = \frac{1}{2} \quad P_{\Delta}(\hat{\Delta} \leq \Delta) = 2 \quad P_{0}(\hat{\Delta} \leq 0) = 3 \quad P_{0}(D_{(k+1)} \leq 0) = 4 \quad P_{0}(W_{X,Y} \leq 2k + 1 - (k + 1)) = 5 \quad P_{0}(W_{X,Y} \leq k) = \frac{1}{2}
\]

$=^1$ comes from the continuity of $F$,

$=^2$ since $\hat{\Delta} - \Delta$ has distribution independent of $\Delta$ ($\Delta = 0$)

$=^3$ from $\hat{\Delta} = D_{(k+1)}$,

$=^4$ from Theorem 4 with $\ell = k + 1$

$=^5$ symmetry of $W_{X,Y}$ null distribution around $\frac{mn}{2} = k + \frac{1}{2}$.

The symmetry property from Theorem 3 was not invoked.

Similarly $P_{\Delta}(\hat{\Delta} > \Delta) = P_{0}(W_{X,Y} \geq k + 1) = \frac{1}{2}$

$\Rightarrow \hat{\Delta}$ is median unbiased.
In that case \( \hat{\Delta} = \frac{D(k) + D(k+1)}{2} \) and we have

\[
P_0(D(k) \leq 0) = P_0(W_X, Y \leq 2k - k) = P_0(W_X, Y \leq k)
\]
\[
P_0(D(k+1) \leq 0) = P_0(W_X, Y \leq 2k - (k + 1)) = P_0(W_X, Y \leq k - 1) = P_0(W_X, Y < k)
\]

\[
\Rightarrow P_0(W_X, Y < k) = P_0(D(k+1) \leq 0) \leq P_\Delta(\hat{\Delta} \leq \Delta)
\]
\[
\leq P_0(D(k) \leq 0) = P_0(W_X, Y \leq k)
\]

The null distribution of \( W_X, Y \) is symmetric around \( k \), thus

\[
P_0(W_X, Y < k) \leq \frac{1}{2} \quad \text{and} \quad P_0(W_X, Y \leq k) \geq \frac{1}{2}
\]

\( P_0(W_X, Y \leq k) - P_0(W_X, Y < k) \) tends to zero as \( mn \to \infty \).

\( P_\Delta(\hat{\Delta} < \Delta) = P_\Delta(\hat{\Delta} \leq \Delta) \approx \frac{1}{2} \), i.e., close to \( P_\Delta(\hat{\Delta} > \Delta) \)

\( \Rightarrow \hat{\Delta} \approx \text{median unbiased.} \)
As an example consider \( m = 8 \) and \( n = 9 \) so that \( mn = 72 \) is even.

\[
p_{\text{wilcox}}(35, 8, 9) = 0.4813 \\
p_{\text{wilcox}}(36, 8, 9) = 0.5187
\]

\[
0.4813 \leq P_{\Delta}(\hat{\Delta} < \Delta) \leq 0.5187
\]

As another example consider \( m = 20 \) and \( n = 21 \) so that \( mn = 420 \)

\[
p_{\text{wilcox}}(209, 20, 21) = 0.4949 \\
p_{\text{wilcox}}(210, 20, 21) = 0.5051
\]

\[
0.4949 \leq P_{\Delta}(\hat{\Delta} < \Delta) \leq 0.5051
\]

And finally \( m = 100 \) and \( n = 101 \) with \( mn = 10100 \)

\[
p_{\text{wilcox}}(5049, 100, 101) = 0.4995 \\
p_{\text{wilcox}}(5050, 100, 101) = 0.5005
\]

\[
0.4995 \leq P_{\Delta}(\hat{\Delta} < \Delta) \leq 0.5005
\]
The traditional $\text{var}(\hat{\Delta})$ is not easy to calculate.

Examine dispersion in terms of closeness probability, i.e.,

$$P_\Delta(|\hat{\Delta} - \Delta| \leq a) = P_0(|\hat{\Delta}| \leq a) = P_0(-a \leq \hat{\Delta} \leq a)$$

$$= P_0(\hat{\Delta} \leq a) - P_0(\hat{\Delta} < -a)$$

Proof later

$$\approx \Phi \left( \frac{mn(\frac{1}{2} - p_1)}{\text{var}(W_{X,Y-a})} \right) + \Phi \left( \frac{mn(\frac{1}{2} - p_1)}{\text{var}(W_{X,Y+a})} \right) - 1$$

The distribution of $\hat{\Delta} - \Delta$ does not depend on $\Delta \Rightarrow \Delta = 0$.

Here \( p_1 = P(X < Y - a) = 1 - P(X < Y + a) = 1 - \tilde{p}_1 \) for $X, Y \sim F$ continuous.

\[
p_1 = P(X - Y < -a) =^1 P(X - Y > a)
= P(X > Y + a) =^2 1 - P(X < Y + a) = 1 - \tilde{p}_1
\]

\(^1\) uses $X - Y \overset{D}{=} Y - X$ and \(^2\) uses the continuity of $F$. 
\[\text{var}(W_{X,Y-a}) \quad \text{and} \quad \text{var}(W_{X,Y+a})\]

\[\text{var}(W_{X,Y-a}) = mn \left[ p_1 (1 - p_1) + (n - 1) (p_2 - p_1^2) + (m - 1) (p_3 - p_1^2) \right] \]

\[p_2 = P(X < Y-a \cap X < Y'-a) \quad \text{and} \quad p_3 = P(X < Y-a \cap X' < Y-a)\]

\[\text{var}(W_{X,Y+a}) = mn \left[ \tilde{p}_1 (1 - \tilde{p}_1) + (n - 1) (\tilde{p}_2 - \tilde{p}_1^2) + (m - 1) (\tilde{p}_3 - \tilde{p}_1^2) \right] \]

\[\tilde{p}_2 = P(X < Y+a \cap X < Y'+a) \quad \text{and} \quad \tilde{p}_3 = P(X < Y+a \cap X' < Y+a)\]

with \(X, Y, X', Y'\) independent \(\sim F\).

\(F\) continuous \(\implies \tilde{p}_2 - \tilde{p}_1^2 = p_3 - p_1^2 \quad \text{and} \quad \tilde{p}_3 - \tilde{p}_1^2 = p_2 - p_1^2\)
Proof of \( \tilde{p}_2 - \tilde{p}_1^2 = p_3 - p_1^2 \)

\[
\tilde{p}_2 - \tilde{p}_1^2 = P(X < Y + a \cap X < Y' + a) - (1 - p_1)^2
\]
\[
= 2p_1 - p_1^2 - [1 - P(X < Y + a \cap X < Y' + a)]
\]
\[
= 2p_1 - p_1^2 - P(X > Y + a \cup X > Y' + a)
\]
\[
= 2p_1 - p_1^2 - P(X > Y + a) - P(X > Y' + a) + P(X > Y + a \cap X > Y' + a)
\]
\[
= 2p_1 - p_1^2 - P(Y > X + a) - P(Y > X' + a) + P(Y > X + a \cap Y > X' + a)
\]
\[
= p_3 - p_1^2
\]

Here we used the continuity of \( F \), the interchangeability of the independent random variables \( X, X', Y, Y' \sim F \),
\[
1 - P(A \cap B) = P([A \cap B]^c) = P(A^c \cup B^c) \quad \text{and} \quad P(A \cup B) = P(A) + P(B) - P(A \cap B).
\]
\[ m = n \implies \text{var}(W_{X,Y - a}) = \text{var}(W_{X,Y + a}) \]

\[
\begin{align*}
\text{var}(W_{X,Y - a}) &= m^2 \left[ p_1 (1 - p_1) + (m - 1) (p_2 - p_1^2) \right] \\
&\quad + (m - 1) (p_3 - p_1^2) \\
&= m^2 \left[ (1 - \tilde{p}_1)\tilde{p}_1 + (m - 1) (\tilde{p}_3 - \tilde{p}_1^2) \right] \\
&\quad + (m - 1) (\tilde{p}_2 - \tilde{p}_1^2) \\
&= \text{var}(W_{X,Y + a})
\end{align*}
\]
$F$ symmetric around $\mu \iff P(X > \mu + b) = P(X < \mu - b) \ \forall b$.

$X - \mu - (Y - \mu) = X - Y$: may assume $\mu = 0$ in $p_2$ and $p_3$.

\[
p_2 = P(X < Y - a \cap X < Y' - a)
= P(-X < -Y - a \cap -X < -Y' - a)
= P(Y < X - a \cap Y' < X - a)
= P(X < Y - a \cap X' < Y - a) = p_3
\]

\[
\text{var}(W_{X,Y-a}) = mn \left[ p_1(1 - p_1) + (m + n - 2) (p_2 - p_1^2) \right]
= mn \left[ (1 - \tilde{p}_1)\tilde{p}_1 + (m + n - 2) (\tilde{p}_3 - \tilde{p}_1^2) \right]
= \text{var}(W_{X,Y+a})
\]
Proof of Approximation for $P_\Delta(|\hat{\Delta} - \Delta| \leq a)$

Case 1: $mn = 2k + 1$ so that $\hat{\Delta} = D(\ell)$ with $\ell = k + 1$.

\[
P_0(\hat{\Delta} \leq a) = P_0(D(\ell) \leq a) = P_0(W_X, Y - a \leq mn - \ell)
\]

\[
= P_0(W_X, Y - a \leq k)
\]

\[
= P_0 \left( \frac{W_X, Y - a - E(W_X, Y - a)}{\sqrt{\text{var}(W_X, Y - a)}} \leq \frac{k + \frac{1}{2} - E(W_X, Y - a)}{\sqrt{\text{var}(W_X, Y - a)}} \right)
\]

\[
\approx \Phi \left( \frac{mn \left( \frac{1}{2} - p_1 \right)}{\sqrt{\text{var}(W_X, Y - a)}} \right)
\]

\[
P_0(\hat{\Delta} < -a) = P_0(D(\ell) \leq -a) = P_0(W_X, Y + a \leq mn - \ell) = P_0(W_X, Y + a \leq k)
\]

\[
= \Phi \left( \frac{mn \left( \frac{1}{2} - (1 - p_1) \right)}{\sqrt{\text{var}(W_X, Y + a)}} \right) = 1 - \Phi \left( \frac{mn \left( \frac{1}{2} - p_1 \right)}{\sqrt{\text{var}(W_X, Y + a)}} \right)
\]
Proof of Approximation for $P_\Delta(|\hat{\Delta} - \Delta| \leq a)$

Case 2: $mn = 2k$ (part 1) then $\hat{\Delta} = (D(k) + D(k+1))/2$ and we have

$W_{X,Y_a} \leq mn - (k+1) = k - 1 \iff D(k+1) \leq a$

$\implies \hat{\Delta} \leq a \implies D(k) \leq a \iff W_{X,Y_a} \leq mn - k = k$

$$
\Phi \left( \frac{k - \frac{1}{2} - mnp_1}{\sqrt{\text{var}(W_{X,Y_a})}} \right) \approx P_0(W_{X,Y_a} \leq k - 1) \leq P_0(\hat{\Delta} \leq a)
$$

$$
P_0(\hat{\Delta} \leq a) \leq P_0(W_{X,Y_a} \leq k) \approx \Phi \left( \frac{k + \frac{1}{2} - mnp_1}{\sqrt{\text{var}(W_{X,Y_a})}} \right)
$$

which suggests

$$
P_0(\hat{\Delta} \leq a) \approx \Phi \left( \frac{k - mnp_1}{\sqrt{\text{var}(W_{X,Y_a})}} \right) = \Phi \left( \frac{mn \left( \frac{1}{2} - p_1 \right)}{\sqrt{\text{var}(W_{X,Y_a})}} \right)
$$
Proof of Approximation for $P_\Delta(|\hat{\Delta} - \Delta| \leq a)$

Case 2: $mn = 2k$ (part 2)

$W_{X,Y+a} \leq mn-(k+1) = k-1 \iff D_{(k+1)} \leq -a$

$\implies \hat{\Delta} < -a \iff D_{(k)} \leq -a \iff W_{X,Y+a} \leq mn-k = k$

$$\Phi\left(\frac{k - \frac{1}{2} - mn\tilde{p}_1}{\sqrt{\text{var}(W_{X,Y+a})}}\right) \approx P_0(W_{X,Y+a} \leq k - 1) \leq P_0(\hat{\Delta} < -a)$$

$$P_0(\hat{\Delta} < -a) \leq P_0(W_{X,Y+a} \leq k) \approx \Phi\left(\frac{k + \frac{1}{2} - mn\tilde{p}_1}{\sqrt{\text{var}(W_{X,Y+a})}}\right)$$

which suggests

$$P_0(\hat{\Delta} < -a) \approx \Phi\left(\frac{k - mn\tilde{p}_1}{\sqrt{\text{var}(W_{X,Y+a})}}\right) = 1 - \Phi\left(\frac{mn \left(\frac{1}{2} - p_1\right)}{\sqrt{\text{var}(W_{X,Y+a})}}\right)$$
The approximation is further simplified by approximating for small $a$

$$p_1 - \frac{1}{2} \approx -af^*(0)$$

and

$$\text{var}(W_{X,Y-a}) \approx \text{var}(W_{X,Y+a}) \approx \text{var}(W_X,Y) = \frac{mn(N+1)}{12}$$

where $f^*(0) =$ density of $Y - X$ at zero, $X, Y$ independent $\sim F$.

$$\implies P_\Delta(|\hat{\Delta} - \Delta| \leq a) \approx 2\Phi \left[ \sqrt{\frac{12mn}{N+1}} f^*(0)a \right] - 1$$

$F = \mathcal{N}(\mu, \sigma^2) \Rightarrow Y - X \sim \mathcal{N}(0, 2\sigma^2)$ with $f^*(0) = 1/(2\sigma \sqrt{\pi})$ and

$$P_\Delta(|\hat{\Delta} - \Delta| \leq a) \approx 2\Phi \left[ \sqrt{\frac{3mn}{\pi(N+1)}} \frac{a}{\sigma} \right] - 1$$
Numerical Comparisons for $P_\Delta(|\hat{\Delta} - \Delta| \leq a)$

For $m = n = 15$, $F = \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = 2$ the previous approximations give the results in the following table (Table 2.5)

<table>
<thead>
<tr>
<th>a</th>
<th>.2</th>
<th>.4</th>
<th>.6</th>
<th>.8</th>
<th>1.0</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>0.4602</td>
<td>0.4207</td>
<td>0.3821</td>
<td>0.3446</td>
<td>0.3085</td>
<td>0.2743</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.2944</td>
<td>0.2577</td>
<td>0.2235</td>
<td>0.1920</td>
<td>0.1633</td>
<td>0.1376</td>
</tr>
<tr>
<td>$E(W_{X,Y-a})$</td>
<td>103.54</td>
<td>94.67</td>
<td>85.97</td>
<td>77.53</td>
<td>69.42</td>
<td>61.71</td>
</tr>
<tr>
<td>$\text{var}(W_{X,Y-a})$</td>
<td>576.67</td>
<td>563.16</td>
<td>541.35</td>
<td>512.26</td>
<td>477.18</td>
<td>437.62</td>
</tr>
<tr>
<td>$P_\Delta(</td>
<td>\hat{\Delta} - \Delta</td>
<td>\leq a)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>approximation</td>
<td>0.2910</td>
<td>0.5476</td>
<td>0.7458</td>
<td>0.8777</td>
<td>0.9514</td>
<td>0.9848</td>
</tr>
<tr>
<td>alt. approximation</td>
<td>0.2903</td>
<td>0.5435</td>
<td>0.7360</td>
<td>0.8636</td>
<td>0.9373</td>
<td>0.9745</td>
</tr>
<tr>
<td>$10^6$ simulations</td>
<td>0.2925</td>
<td>0.5464</td>
<td>0.7389</td>
<td>0.8662</td>
<td>0.9390</td>
<td>0.9754</td>
</tr>
</tbody>
</table>
HLsim <- function(Nsim=100,m=15,n=15,
aVEC=c(.2,.4,.6,.8,1,1.2),SIG2=2){
    k <- length(aVEC)
    out <- matrix(rep(0,k*Nsim),ncol=k)
    SIG <- sqrt(SIG2)
    for(i in 1:Nsim){
        x <- rnorm(m,0,SIG)
        y <- rnorm(n,0,SIG)
        HL <- median(outer(y,x,"-"))
        pHL <- abs(HL)<=aVEC
        out[i,] <- pHL
    }
    x <- apply(out,2,mean)
    x}

Simulation advantage: unbiased estimation of \( P_\Delta(|\hat{\Delta} - \Delta| \leq a) \),
with any desired accuracy depending on \( N_{\text{sim}} \).
We can easily handle other distributions for \( X \) and \( Y \).
The error in the previous approximations $\to 0$ as $m, n \to \infty$.
The approximation is already quite good for $m = n = 15$.
It can also be shown that for any $a > 0$ we have

$$P_\Delta(|\hat{\Delta} - \Delta| \leq a) = P_\Delta(|\hat{\Delta}_{m,n} - \Delta| \leq a) \to 1 \text{ as } m, n \to \infty$$

Such estimators are called consistent estimators of $\Delta$.
$\hat{\Delta}_{m,n} \xrightarrow{P} \Delta$, convergence in probability.
We assume the shift model:
\( X_i, Y_j - \Delta \) independent \( \sim F, \ i = 1, \ldots, m, \ j = 1, \ldots, n. \)
\( \bar{\Delta} = \bar{Y} - \bar{X} \) is a consistent estimator of \( \Delta \) (\( \mu_F \) finite).
If \( m = n \) or \( F \) is symmetric around some point \( \mu \), then the distribution of \( \bar{\Delta} \) is symmetric around \( \Delta \).
In that case \( \bar{\Delta} \) is unbiased and median unbiased (\( \mu_F \) finite).
\( \hat{\Delta} \) had same unbiasedness properties under same conditions.
Distribution of $\bar{\Delta}$ is typically no longer symmetric around $\Delta$.

$\bar{\Delta}$ is typically also no longer median unbiased.

But $\bar{\Delta}$ continues to be unbiased, i.e., $E(\bar{\Delta}) = \Delta$ ($\mu_F$ finite).

In contrast to this $\hat{\Delta}$ remained median unbiased for $mn$ odd (nearly so for $mn$ even).

Neither unbiasedness property is uniformly better than the other.
The previous efficiency results of the Wilcoxon test relative to the $t$-test carry over.

If we define the efficiency of $\hat{\Delta}$ relative to $\bar{\Delta}$ as the ratio of sample sizes $n'/n$, where $m = n$ and $m' = n'$ observations are needed to match

$$P_\Delta(|\hat{\Delta}_{n,n} - \Delta| \leq a) = P_\Delta(|\bar{\Delta}_{n',n'} - \Delta| \leq a)$$

We have

$$\lim_{n \to \infty} \frac{n'}{n} = e_{\hat{\Delta}, \bar{\Delta}}(F) = e_{W, t} \quad \text{while} \quad a = a_n = d / \sqrt{n} \to 0.$$

Thus the same efficiency results apply to these estimators, i.e., $e_{\hat{\Delta}, \bar{\Delta}}(\Phi) = 3/\pi = .955$ and $e_{\hat{\Delta}, \bar{\Delta}}(F) \geq .864 \ \forall F \ (\sigma^2_F < \infty)$. 
Rather than estimating $\Delta$ using a point estimate $\hat{\Delta}$ or $\bar{\Delta}$, we will estimate $\Delta$ by using an interval.

It will capture the true unknown $\Delta$ with a prescribed confidence probability $\gamma \in (0, 1)$.

Such intervals are known as $100\gamma\%$ confidence intervals for $\Delta$.

Do we already have such intervals with approximate confidence level $\gamma$ through

$$P_{\Delta} \left( |\hat{\Delta} - \Delta| \leq a \right) \approx \Phi \left( \frac{mn(\frac{1}{2} - p_1)}{\text{var}(W_{X,Y-a})} \right) + \Phi \left( \frac{mn(\frac{1}{2} - p_1)}{\text{var}(W_{X,Y+a})} \right) - 1 = \gamma$$

This would give us such an interval via

$$P_{\Delta} \left( \hat{\Delta} - a \leq \Delta \leq \hat{\Delta} + a \right) \approx \gamma \quad (???)$$

To determine $a$ to get $\gamma$ we need to know $p_1$, $p_2$ and thus need to know $F$, i.e., the interval would not be distribution-free.
Building Blocks for a $\Delta$ Confidence Interval

- From Theorem 4 we have
  \[ D(\ell) \leq \Delta \iff W_{X,Y-\Delta} \leq mn - \ell \]
  \[ D(\ell) > \Delta \iff W_{X,Y-\Delta} \geq mn - \ell + 1 \]
  \[ \text{or} \quad D(\ell+1) > \Delta \iff W_{X,Y-\Delta} \geq mn - (\ell + 1) + 1 = mn - \ell \]
  \[ D(\ell) \leq \Delta < D(\ell+1) \iff W_{X,Y-\Delta} = mn - \ell \quad \text{for} \quad \ell = 0, 1, \ldots, mn, \]
  where $D(0) = -\infty$ and $D(mn+1) = \infty$.
- In the shift model: $X \sim F(x)$ and $Y = X' + \Delta \sim F(x - \Delta)$ or $X' = Y - \Delta \sim F(x)$.
- Thus the distribution of $W_{X,Y-\Delta}$ is independent of $F$ (the null distribution of $W$).

\[ P_{\Delta} \left( D(\ell) \leq \Delta < D(\ell+1) \right) = P_0(W_{X,Y} = mn - \ell) \]
\[ = P_0(W_{X,Y} = \ell) \quad \ell = 0, 1, \ldots, mn \]

- The order statistics $D_{(1)}, \ldots, D_{(mn)}$ divide the real line into $mn + 1$ intervals which capture the unknown $\Delta$ with known probabilities, independent of $F$ and $\Delta$. 
Distribution-free Confidence Intervals for $\Delta$

- The previous confidence interval building blocks give us

$$P_\Delta(D_k \leq \Delta \leq D_\ell) = P_\Delta(D_k \leq \Delta < D_\ell)$$

$$= P_0(k \leq WX, Y \leq \ell - 1)$$

$$= 1 - P_0(WX, Y \leq k - 1) - P_0(WX, Y \geq \ell)$$

$$= * \ 1 - 2P_0(WX, Y \leq k - 1) \geq \gamma$$

- In * we took $\ell = mn - (k - 1) = mn - k + 1$ and exploited the symmetry of the $WX, Y$ null distribution around $mn/2$.

- To get confidence level barely $\geq \gamma$, find the largest $k$ such that

$$P_0(WX, Y \leq k - 1) \leq \frac{1 - \gamma}{2}$$
To Find Largest $k$ with $P_0(W_X, Y \leq k - 1) \leq \frac{1-\gamma}{2}$

By symmetry this is equivalent to finding largest $k$ such that

$$P_0(W_X, Y \leq k - 1) = P_0(W_X, Y \geq mn - (k - 1))$$

$$= P_0(W_X, Y \geq mn - k + 1) \leq \frac{1-\gamma}{2}$$

$$\iff 1 - P_0(W_X, Y \leq mn - k) \leq \frac{1-\gamma}{2}$$

$$\iff P_0(W_X, Y \leq mn - k) \geq 1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$$

R has the quantile function $\text{qwilcox}$ which gives the smallest $mn - k$ such that $P_0(W_X, Y \leq mn - k) \geq \frac{1+\gamma}{2}$, using

$$mn - k = \text{qwilcox}((1 + \gamma)/2, m, n)$$

$$\implies k = mn - \text{qwilcox}((1 + \gamma)/2, m, n)$$
The Achieved Confidence Level $\gamma_k$ of $[D(k), D(mn-k+1)]$

Based on $k = mn - q\text{wilcoxon}((1 + \gamma)/2, m, n)$ the achieved confidence level $\gamma_k$ is

$$\gamma_k = P_\Delta(D(k) \leq \Delta \leq D(mn-k+1)) = 1 - 2P_0(W_{X,\gamma} \leq k - 1)$$

$$= 1 - 2 * p\text{wilcoxon}(k-1,m,n)$$

$[D(k), D(mn-k+1)]$ is a distribution-free confidence interval for $\Delta$, with achieved level $\gamma_k \geq \gamma$.

$$P_\Delta(D(k+1) \leq \Delta \leq D(mn-k)) = 1 - 2 * p\text{wilcoxon}(k,m,n) = \tilde{\gamma}_k < \gamma$$

may get you closer to the desired $\gamma$, not conservatively, but the interval is also tighter.
Example 5: Augmenters and Reducers (Data: Petrie, 1967).
It concerns a pain sensory classification.
The Text used 10 and 7 observations for use of Table B.
We work with the full data set, reproduced below.
Augmenters (Reducers) were tested for pain reaction without
and with an analgesic.
For Augmenters the difference in pain scores is $Y_j$,
for Reducers it is $X_i$.

Augmenters: 17.94 13.32 11.31 10.62 7.56 5.73 5.61
5.40 3.30 3.09 .93

Reducers: 7.74 5.04 1.68 0.00 −3.03 −3.09 −10.53

Based on Petrie’s theory it is suspected that Augmenters will
show higher response differences, in particular, shifted
differences compared to those of Reducers.
It is desired to obtain a confidence interval for this shift $\Delta$. 
Sorted $D_{ij} = Y_j - X_i$

```r
> Dmatfun()
[1] -6.81 -4.65 -4.44 -4.11 -2.34 -2.13 -2.01 -1.95 -1.74 -0.75
[11] -0.18  0.36  0.57  0.69  0.93  1.41  1.62  2.52  2.88  3.09
[21]  3.30  3.57  3.72  3.93  3.96  4.02  4.05  5.40  5.58  5.58
[31]  5.61  5.73  5.88  6.12  6.18  6.27  6.33  6.39  7.56  8.28
[41]  8.43  8.49  8.64  8.70  8.76  8.82  8.94  9.63 10.20 10.59

Dmatfun <- function(){
  Augmenters <- c(17.94, 13.32, 11.31, 10.62, 7.56,
                  5.73, 5.61, 5.4, 3.3, 3.09, .93)
  Reducers <- c(7.74, 5.04, 1.68, 0.0,
               -3.03, -3.09, -10.53)
  Dmat <- t(outer(Augmenters, Reducers, "-"))
  Dmat
}
```
Example 5: A 90% Confidence Interval

- We want a 90% confidence interval for \( \Delta \).

\[
\Rightarrow \quad k = m \times n - q\text{wilcoxon}((1 + .9)/2, 7, 11) = 20
\]

- Thus \( \gamma_k = 1 - 2 \times p\text{wilcoxon}(19, 7, 11) = 0.91466 \).

- Desired confidence interval with confidence level \( \gamma_k = .912 \) is

\[
[D(20), D(77-20+1)] = [D(20), D(58)] = [3.09, 13.62]
\]

- With closer confidence, but not conservatively, we get \( \tilde{\gamma}_k = 1 - 2 \times p\text{wilcoxon}(20, 7, 11) = 0.8958 \) and

\[
[D(21), D(77-21+1)] = [D(21), D(57)] = [3.30, 13.32]
\]
Sometimes upper or lower confidence bounds for $\Delta$ are more appropriate, i.e.,

$$P_\Delta(\hat{\Delta}_L \leq \Delta) = \gamma \quad \text{or} \quad P_\Delta(\hat{\Delta}_U \geq \Delta) = \gamma.$$ 

We split the miss-probability \(2 \times \frac{1-\gamma}{2} = 1 - \gamma\) of our confidence intervals equally.

View the left interval endpoint $D_{(k)}$ as lower bound with miss-probability $\frac{1-\gamma}{2}$ and thus with confidence $1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$.

View $D_{(mn-k+1)}$ as an upper bound with confidence level $\frac{1+\gamma}{2}$.

For the actually achieved confidence level replace $\gamma$ by $\gamma_k$.

In our example we can thus view $D_{(58)} = 13.62$ as 95% upper bound for $\Delta$, with achieved confidence level $\gamma_k = 0.957$.

$D_{(20)} = 3.09$ is a 95% lower bound, with achieved $\gamma_k = .957$. 
So far it was assumed that there are no ties.

This happens with probability 1 when $F$ is continuous.

The continuity of $F$ is reasonable for many measurement phenomena.

It is typically realized with sufficient measurement precision.

However, there is rounding.

If rounded measurements are on a grid, say $0, \pm \epsilon, \pm 2\epsilon, \ldots$, then exact observations $X'_i(Y'_j)$ deviate from their rounded versions $X_i(Y_j)$ by at most $\epsilon/2$.

\[ |Y'_j - Y_j - (X'_i - X_i)| = |(Y'_j - X'_i) - (Y_j - X_i)| = |D'_{i,j} - D_{i,j}| \leq \epsilon \]

\[ \implies |D'_{(i)} - D_{(i)}| \leq \epsilon \quad \text{(see Slide 92)} \]
Assume $X'_i \sim F$ and $Y'_j \sim F(x - \Delta)$ are independent, with $F$ continuous, and $D'_\ell$ is the $\ell^{th}$ ordered value of $D_{i,j} = Y'_j - X'_i$. $D(\ell)$ be the counterpart using the rounded values $X_i$, $Y_j$.

If $\hat{\Delta}'_L = D'_\ell$ is a (valid) lower bound for $\Delta$ with confidence level $\gamma$, then $\hat{\Delta}_L = D(\ell) - \epsilon$ is a conservative lower confidence bound for $\Delta$ with confidence level $\geq \gamma$.

If $\hat{\Delta}'_U = D'_u$ is an upper bound for $\Delta$ with confidence level $\gamma$, then $\hat{\Delta}_U = D(u) + \epsilon$ is a conservative upper confidence bound for $\Delta$ with confidence level $\geq \gamma$.

If $[D'_\ell, D'_u]$ is a confidence interval for $\Delta$ with confidence coefficient $\gamma$, then $[D(\ell) - \epsilon, D(u) + \epsilon]$ is a conservative confidence interval for $\Delta$ with confidence level $\geq \gamma$. 
DConfInt <- function(x=Augmenters,y=Reducers,gam=.95){
  n <- length(x); m <- length(y)
  score.factor <- factor(c(rep("A",n),rep("B",m)))
  out1 <- wilcox.test(x,y,conf.int=T,
    conf.level=gam,exact=T)
  dat.fr <- data.frame(scores=c(x,y),type=score.factor)
  out2 <- wilcox_test(scores~type,data=dat.fr,
    conf.int=T,conf.level=gam,dist=exact())
  list(out1=out1,out2=out2)}

- **wilcox.test** and **wilcox_test** give intervals with confidence $\geq \gamma = .95$, where $\text{gam} = \gamma$ is the specified level.
- For example, if you specify $\gamma = .95$ you may get an achieved confidence level of .976 in the confidence interval returned by **wilcox.test**, although it still refers to it as a 95% interval.
- However, it might be possible to get a tighter confidence interval with achieved level .948, by specifying $\gamma = .948$. 
Trying \texttt{DConfInt (gam=\gamma)} for Various $\gamma$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$D(\ell)$</th>
<th>$D(u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.9</td>
<td>3.09</td>
<td>13.62</td>
</tr>
<tr>
<td>.87</td>
<td>3.57</td>
<td>12.90</td>
</tr>
<tr>
<td>.84</td>
<td>3.72</td>
<td>11.64</td>
</tr>
<tr>
<td>.92</td>
<td>2.88</td>
<td>13.65</td>
</tr>
<tr>
<td>.94</td>
<td>2.52</td>
<td>13.71</td>
</tr>
</tbody>
</table>
An alternative to the Wilcoxon test is the normal scores test. Rather than summing the ranks in the treatment group we sum transformed ranks.

The ordinary ranks $1, 2, \ldots, N$ are equally spaced.

Normal data are crowded in the center and sparser in the tails.

The normal scores try to emulate this via transformed ranks.

Make the ranks $1 < 2 < \ldots < N$ correspond to a standard normal random sample $Z_{(1)} < \ldots < Z_{(N)}$,

or more practically, to their expected values, the normal scores:

$$a_N(s) = E_\Phi \left( Z_{(s)} \right), \quad \text{for } s = 1, 2, \ldots, N$$
As test statistic take

\[ T_s = a_N(S_1) + a_N(S_2) + \ldots + a_N(S_n) \]

where the \( S_1, \ldots, S_n \) are the treatment sample ranks.

Depending on the anticipated direction of the treatment effect under the alternative to \( H_0 \), reject \( H_0 \) when \( T_s \) is too large, or when it is too small, or, for two-sided alternatives, when either situation occurs.

\( (a_N(1), \ldots, a_N(N)) \) is obtained in R via \texttt{normOrder(N)}.

\texttt{normOrder} is a function in the package \texttt{SuppDist}.

Before using this function you need to install that package.
Once the scores are known it is straightforward to obtain the null distribution of $T_S$.

Simply perform all splits of $a_N(1) < \ldots < a_N(N)$ into $n$ and $m$ such scores, using again `combn`, and obtain the transformed rank sum $T_S$ for each such split.

The vector of these $\binom{N}{n}$ sums represents the full null distribution of $T_S$ and can be used to compute any $p$-values.

$p$-value = the proportion of these $\binom{N}{n}$ sums $T_S$ that are $\geq$ (or $\leq$) to the observed value $T_{S,\text{obs.}}$ of $T_S$.

For a two-sided test compute the proportion of $|T_S| \geq |T_{S,\text{obs.}}|$.

The null distribution of $T_S$ is symmetric around zero.
The Wilcoxon test compared quite favorably with the two-sample $t$-test.

The same holds with respect to the normal scores test ($N$) and the two-sample $t$-test.

The ARE $e_N, t(F) \geq 1$ for all $F$, with $= \iff F = \Phi$.

A remarkable result (Chernoff and Savage 1958).
The van der Waerden Test

- The van der Waerden test is essentially a close relative to the normal scores test.
- In large samples the two tests are equivalent.
- The normal scores in $T_s$ are simply replaced by the scores

$$a_N(s) = \Phi^{-1} \left( \frac{s}{N + 1} \right) \quad \text{for } s = 1, \ldots, N,$$

denoting the transformed rank sum again by $T_s$.
- These scores are much easier to compute in R via

$$a_N(s) = \text{qnorm}(s/(N + 1)).$$
- The exact null distribution is easily obtained via \texttt{combn} when $m$ and $n$ are not too large.
- The above efficiency result holds for this test as well.
The coin package has a function normal_test that carries out the normal scores test using van der Waerden scores.

```r
x0 <- c(10, 12, 15, 21, 32, 40, 41, 45)
y0 <- c(6, 20, 27, 38, 46, 51, 54, 57)
m <- length(x0); n <- length(y0); z <- c(x0, y0)
fac <- factor(c(rep("C", m), rep("T", n)),
              levels = c("T", "C"))
dat <- data.frame(z, fac)
normal_test(z ~ fac, data = dat,
            ties.method = "mid-ranks",
            alternative = "two.sided",
            dist = exact())
```

**Exact Normal Quantile (van der Waerden) Test**

data:  z by fac (T, C)
Z = 1.2278, p-value = 0.2275
alternative hypothesis: true mu is not equal to 0
• `ties.method` can take the values "mid-ranks" or "average-scores".

• The argument `ties.method="mid-ranks"` has no impact since there are no ties in this case.

• `dist` can take values as in the case of `wilcox_test`.

• `alternative` can also take values "greater", "less".

• For "greater", scores under the first factor level are expected to be greater than those for the second level.

• The function `normal_test` still handles $m = n = 16$ with ease, impressive since $\binom{32}{16} = 601080390$.

• Simulation with `dist=approximate(B=1e7)` and $m = n = 16$ is still possible in 7.5 sec.

• `B=1e8` ran into memory allocation problems on this laptop.

• In Linux that did not happen, but took time.
Assume the 2-sample problem, samples $X$ and $Y$ with shift $\Delta$.

Test the hypothesis $H_{\Delta_0} : \Delta = \Delta_0$.

Suppose we have a test that rejects $H_{\Delta_0}$ with maximum type I error probability $\alpha$, i.e.,

$$\max_{\Delta_0} P_{\Delta_0}(\text{reject } H_{\Delta_0}) = \alpha$$

Let $C(X, Y)$ be a set of $\Delta_0$ values for which $(X, Y)$ leads to acceptance of $H_{\Delta_0}$.

Then $C(X, Y)$ is a $(1 - \alpha)$-level confidence set for $\Delta$, i.e.,

$$\min_{\Delta_0} P_{\Delta_0}(\Delta_0 \in C(X, Y)) = 1 - \alpha$$

Given a $(1 - \alpha)$-level confidence set $C(X, Y)$, we can reject at level $\alpha$ any $\Delta_0$ that is not in $C(X, Y)$.

$\implies$ Equivalence of testing and confidence sets.
The Duality for the Wilcoxon Test

- We saw that

\[ D(i) \leq \Delta_0 \leq D(mn-i+1) \iff i \leq W_{X,Y-\Delta_0} \leq mn - i \]

\[ \iff \left| W_{X,Y-\Delta_0} - \frac{mn}{2} \right| \leq \frac{mn}{2} - i \]

- The latter can be viewed as \( H_{\Delta_0} \) acceptance criterion for the 2-sided Wilcoxon test with significance level

\[ \alpha_i = P_{\Delta_0} \left( \left| W_{X,Y-\Delta_0} - \frac{mn}{2} \right| \geq \frac{mn}{2} - i + 1 \right) = 2P_0(W_{X,Y} \leq i - 1) \]

- As corresponding confidence set coverage probability we have

\[ P_{\Delta_0} \left( D(i) \leq \Delta_0 \leq D(mn-i+1) \right) = P_{\Delta_0} \left( i \leq W_{X,Y-\Delta_0} \leq mn - i \right) = 1 - \alpha_i \]
This duality holds quite generally.
Not just for the Wilcoxon test.
Not just for the 2-sample shift problem.
It also applies for vector parameters.
When we deal with one-sided tests the confidence sets typically usually become open ended, i.e., we get lower or upper confidence bounds.
The latter makes sense for 1-dimensional parameters.
In the 2-sample problem the relative ranks of $X_1, \ldots, X_m$ within and the relative ranks of $Y_1 - \Delta_0, \ldots, Y_n - \Delta_0$ within don’t change when we move $\Delta_0$.

The overall ranks only change when one of the $Y_j - \Delta_0$ crosses one of the $X_i$, or when $\Delta_0$ crosses $D_{ij} = Y_j - X_i$.

For testing $H_{\Delta_0} : \Delta = \Delta_0$ any rank statistic $T_s(\Delta_0)$, i.e., based on ranking $X_1, \ldots, X_m, Y_1 - \Delta_0, \ldots, Y_n - \Delta_0$, will change as function of $\Delta_0$ only when $\Delta_0$ crosses any of the values $D_{ij} = Y_j - X_i$, $i = 1, \ldots, m, j = 1, \ldots, n$.

When the scores $a_N(s)$ of the linear rank statistic $T_s(\Delta_0) = \sum_{i=1}^{n} a_N(S_i(\Delta_0))$ are monotone increasing, then $T(\Delta_0)$ is a decreasing step function in $\Delta_0$, with steps at the ordered values $D_{(1)} < \ldots < D_{(mn)}$ of the $D_{ij}$.

The above observations are due to Bauer (1972).
We illustrate the previous with two samples

```r
x <- c(-0.212, -1.042, -1.153, 0.322)
y <- c(0.500, 1.554, 3.734, 2.511)
```

Slide 142 shows the step function plot for $T_s(\Delta_0)$.  
41 dashed lines = null distribution values of $T_s = T_s(0)$.  
$T_s$ has $17 = mn + 1$ distinct values.  
Adjacent (null distr.) frequency counts of these $T_s$ values.  
Note that these frequencies add to $70 = \binom{8}{4}$.  
A step change at each of the $mn = 16$ ordered values $D(i)$.  
The null distribution of $T_s$ is symmetric around zero.
van der Waerden rank–sum $T_s(\Delta)$ for the ranks of $y_1 - \Delta, \ldots, y_n - \Delta$
when ranking $x_1, \ldots, x_m, y_1 - \Delta, \ldots, y_n - \Delta$
(\binom{4}{2} = 6) ways to get $T_s = 0$ for 2 symmetric rank pairs, e.g.,

$$T_s = \Phi^{-1}\left(\frac{1}{9}\right) + \Phi^{-1}\left(\frac{4}{9}\right) + \Phi^{-1}\left(\frac{5}{9}\right) + \Phi^{-1}\left(\frac{8}{9}\right) = 0$$

because of symmetry of score function around 0.

For 1 symmetric rank pair we get same $T_s$ for 2 rankings, e.g.,

$(s_1, s_2, s_3, s_4) = (1, 2, 4, 5)$ and $(1, 2, 3, 6)$.

$\frac{8 \cdot 6}{2} \cdot 2 = 48$ asymmetric pairs (above 1,2) together with one symmetric pair (above 4,5 or 3,6), for 24 distinct $T_s$ values.

No symmetric pairs, $2^4 = 16$ ways to take one each from $(1, 8), (2, 7), (3, 6), (4, 5)$.

$16 + 24 + 1 = 41$ distinct values for $T_s$

with respective multiplicities 1, 2, 6.
Similar to the Hodges-Lehmann point estimate based on the Wilcoxon test we can motivate an estimate based on $T_s(\Delta)$.

As estimate $\hat{\Delta}_{\Phi^{-1}}$ take either

- the midpoint of the interval $(D(\ell), D(\ell+1))$ where $T_s(\Delta) = 0$.
- or the point $D(\ell)$ where $T_s(\Delta)$ crosses 0.

In case of ties due to rounding, one could break the ties by adding tiny random amounts to the $X_i, Y_j$.

Then remove this small perturbation from the estimate resulting from the $D(\ell)$.

Report $\hat{\Delta}_{\Phi^{-1}}$, with $\pm \epsilon$, indicating the rounding error.
normal.scores.estimate(y, x) computes this estimate.

```r
x0 <- c(10, 12, 15, 21, 32, 40, 41, 45)
y0 <- c(6, 20, 27, 38, 46, 51, 54, 57)
normal.scores.estimate(y0, x0)
[1] 11
```

```r
z0 <- c(x0, y0)
fac <- factor(c(rep("C", 8), rep("T", 8)),
              levels=c("T","C"))
dat <- data.frame(z0, fac)
library(coin)
normal_test(z0~fac, data=dat, alt="two.sided",
            dist=exact(), conf.int=T, conf.level=.95)
```

```
Exact Normal Quantile (van der Waerden) Test

data:  z0 by fac (T, C)
Z = 1.2278, p-value = 0.2275
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -6 30
sample estimates:
difference in location
 11
```
Lower Confidence Bounds Derived from the $T_s$-Test

- Reject $H_{\Delta_0} : \Delta = \Delta_0$ whenever $T_s(\Delta_0) \geq c_\alpha$, where
  \[ P_{\Delta_0}(T_s(\Delta_0) \geq c_\alpha) = P_0(T_s \geq c_\alpha) = \alpha \]

  \( \forall \) achievable significance levels $\alpha$ under the $T_s$ null distribution.

- By the duality principle we get as confidence set
  \[ C(X, Y) = \{\Delta_0 : T_s(\Delta_0) < c_\alpha\} = \{\Delta : T_s(\Delta) < c_\alpha\} \]

  with confidence coefficient
  \[ P_{\Delta_0}(\Delta_0 \in C(X, Y)) = P_{\Delta_0}(T_s(\Delta_0) < c_\alpha) = P_0(T_s < c_\alpha) = 1 - \alpha \]

- Since $T_s(\Delta)$ is a non-increasing step function in $\Delta$
  \[ \implies C_L(X, Y) = \{\Delta : T_s(\Delta) < c_\alpha\} = [D(\ell_0), \infty) \]

  where $D(\ell_0)$ is the largest $D(\ell)$ with $T_s(D(\ell) -) \geq c_\alpha$.

  \[ P_{\Delta}(\Delta \in C_L(X, Y)) = P_{\Delta}(D(\ell_0) \leq \Delta) = P_0(T_s < c_\alpha) = 1 - \alpha \]
The 5 smallest achievable significance levels correspond to the 5 highest dashed lines which $T_s(\Delta)$ may assume as values.

They correspond to significance probabilities:
$$\alpha_1 = 1/70, \quad \alpha_2 = 2/70, \quad \alpha_3 = 4/70, \quad \alpha_4 = 5/70, \quad \alpha_5 = 7/70$$

The corresponding $(1 - \alpha_i)$-level lower confidence bounds are $D_{(1)}, D_{(2)}, D_{(3)}, \ldots$ at appropriate $\alpha_i$ jump locations.

Which of these dashed line values $T_s(\Delta)$ takes on depends on the $(X, Y)$ configurations.

Note the difference of 6th $T_s(\Delta)$-level from top in Slides 142 & 148.

For the Wilcoxon test we have $mn + 1 = 4 \cdot 4 + 1 = 17$ levels, with 16 jump locations. All levels are taken on by $W_s(\Delta)$.

For the $T_s$ test we have 41 levels, with same 16 jump locations, $D_{(i)}$.

This gives us more to play with in terms of $\alpha_i = 1 - \gamma_i$.

For $\alpha_i < \alpha < \alpha_{i+1}$ take $\alpha_i$ (conservative) or the closer one.
van der Waerden rank–sum $T_s(\Delta)$ for the ranks of $y_1-\Delta$, ..., $y_n-\Delta$
when ranking $x_1$, ..., $x_m$, $y_1-\Delta$, ..., $y_n-\Delta$
Slide 145: `normal_test` produces confidence intervals.

Achieved confidence level not given, but it is ≥ nominal level.

To get a better idea about the achieved level, vary the nominal level and observe at which point the reported interval changes.

For the example on Slide 145 the achieved level appears to be .951, since the nominal level .952 gives interval $[-7, 30]$.

For `alt = "greater"` the upper end point is $\text{Inf}$, i.e., we get a lower bound for $\Delta$.

For `alt = "less"` the lower end point is $-\text{Inf}$, i.e., we get an upper bound for $\Delta$. 