University of Washington

STATISTICS

Stat 425

Introduction to Nonparametric Statistics

Comparing Two Treatments or Attributes in a Population Model

Fritz Scholz

Spring Quarter 2009*

*May 4, 2009

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/controls to just these subjects.

Would we have gotten similar results with another group of subjects?

Often the generalization of inference results to other subjects may not be reasonable. Sometimes such generalization is not even 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 for generalization?

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 study subjects as a random sample from this larger population.
Any group of N subjects has the 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).

We finesse this by randomizing treatment and control after we have gotten our random sample of subjects from the population of interest.

Population Model

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 cumulative distribution function (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,

and 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 two-sample problem.

Further Reason for Population Model

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.

Discussion of the Three Power Drivers

We can choose m and n, but can we achieve the desired power?

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

Suppose the treatment adds an amount Δ 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 Δ added.

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

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 appropriate m and n to achieve a given power.

Two Extreme Examples for $Z_1, \ldots, Z_N$

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

$\Longrightarrow \;\; \beta(\Delta) = 1 \quad$ for $\quad \Delta > \max_{i,j} |Z_i - Z_j| \;\; (\Longrightarrow \;\; Z_{\pi_i} - Z_{\pi_{m+j}} < \Delta$ for all i and j).

This is the case for even 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

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

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 Effect of Variability in Control Responses

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 Δ shift under treament is easily recognized, provided it is large compared to 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 Δ 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.

The Continuity Assumption

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 by assuming 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_ℓ 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.

$P(Z = Z') = 0$

$P(Z = Z') = 0$ where Z and Z' are independent with

$Z \sim F$ or $\sim G$ and $Z' \sim F$ or $\sim G$. This in turn implies

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

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

$$
\begin{aligned}
P(Z = Z') &\leq P(Z \in I_i \cap Z' \in I_i \quad \text{for some} \quad i = 1, \ldots, K) \\
&= \sum_{i=1}^{K} P(Z \in I_i \cap Z' \in I_i) = \sum_{i=1}^{K} P(Z \in I_i) P(Z' \in I_i) \\
&\leq \max_i P(Z' \in I_i) \sum_{i=1}^{K} P(Z \in I_i) = \max_i P(Z' \in I_i) \\
&= \max_{i=1,\ldots,K} [F(a_i) - F(a_{i-1})] = \max_{i=1,\ldots,K} \left[\frac{i}{K} - \frac{i-1}{K} \right] = \frac{1}{K} \overset{K \to \infty}{\longrightarrow} 0 \quad \text{q.e.d.}
\end{aligned}
$$

Equally Probable Rankings

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, length n ranking subset $\{s_1 < \ldots < s_n\}$ of $\{1, 2, \ldots, N\}$.

This follows immediately from the fact that the vector

$$(Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$$

has the same distribution as the vector

$$(Z_{\pi_1}, Z_{\pi_2}, \ldots, Z_{\pi_N}) \qquad \text{for any permutation } (\pi_1, \ldots, \pi_N) \text{ of } (1, \ldots, N).$$

Thus 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 equal probability $1/N!$

Equally Probable Rankings

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

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

$$P_{H_0}(S_1 = s_1, \ldots, S_n = s_n) = 1/\binom{N}{n}.$$

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

Here the null hypothesis is $H_0 : F = G$ without specifying the common CDF F.

For that reason one calls rank tests also 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

Then we may get ties among our ranks and we will replace them 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.

s	1	1.5	2	2.5	3
$P_{H_0}(S_1^* = s)$	pq^2	$2p^2q$	p^3+q^3	$2pq^2$	p^2q

The null distribution of S_1^* depends on p and thus on F.

Disontinuity of F: General Case

The phenomenon observed on the previous slide holds in general.

It relates to the fact that in the randomization model the distribution of midranks depends on the numbers $d_1, \ldots, d_e$ which denote the multiplicies with which the e distinct observations are observed.

These numbers are sometimes called the configuration of ties.

In our randomization model from Chapter 1, under H_0 the number e, these configurations, and the e corresponding unique observations were given a priori. The randomization assigns the labels X and Y to the N associated midranks.

Now e, these configurations and the corresponding distinct observations are random variables with distribution depending on F under H_0 (previous example).

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 $Z_1, \ldots, Z_N$ independent, identically distributed $\sim F$, with an independent choice afterwards as to which n to call Y and which m to call X. The $\binom{N}{n}$ such choices are all equally likely.

From these Z's we will get a random number e of distinct, ordered Z values and their corresponding configuration $d_1, \ldots, d_e$.

These can be used to construct an ordered sequence $Z_1^* \leq \ldots \leq Z_N^*$ which are nothing but the original $Z_1, \ldots, Z_N$ in increasing order.

Since 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 $Z_1^* \leq \ldots \leq Z_N^*$ and any such assignment with same chance $1/\binom{N}{n}$, independently of the values of $Z_1^* \leq \ldots \leq Z_N^*$.

The Conditional Distribution of $S_1^*, \ldots, S_n^*$

The ordered $Z_1^* \leq \ldots \leq Z_N^*$ also determine the corresponding N midranks, denoted here by $Q_1^* \leq \ldots \leq Q_N^*$.

Thus 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 is equivalent to e and $d_1, \ldots, d_e$.

Each such random choice of $S_1^* \leq \ldots \leq S_n^*$ from $Q_1^* \leq \ldots \leq Q_N^*$ has the same chance $1/\binom{N}{n}$, just as in our previous randomization model (Chapter 1).

Conditional W_s^* Test

When testing $H_0 : F = G$ by using the midrank statistics W_s^*, and considering large values of W_s^* a good reason for rejecting H_0, we can proceed as follows:

Using the fact that the conditional distribution of midranks given e and $d_1, \ldots, d_e$ is independent of F, with probability $1/\binom{N}{n}$ for each midrank selection from $Q_1^* \leq \ldots \leq Q_N^*$, we can find critical points $C(e, d_1, \ldots, d_e)$ such that the conditional significance level

$$P_{H_0}[W_s^* \geq C(e, d_1, \ldots, d_e) | e, d_1, \ldots, d_e]$$

of the rejection rule $W_s^* \geq C(e, d_1, \ldots, d_e)$ is as close to (and below) the desired significance level α.

The overall significance level of this test procedure is

$$P_{H_0}(\text{rejection}) = \sum P_{H_0}(e, d_1, \ldots, d_e)\, P_{H_0}[W_s^* \geq C(e, d_1, \ldots, d_e) | e, d_1, \ldots, d_e]$$

summed over all configurations $(e, d_1, \ldots, d_e)$.

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 $\leq \alpha$.
Hence it is conservatively distribution-free.

Otherwise the achieved significance level is somewhat close to α, or more generally between the minimum and maximum of the conditional significance levels.

$P_{H_0}(\text{rejection})$ will depend on F to some extent, through the probability weights $P_{H_0}(e, d_1, \ldots, d_e)$ in the previous summation and also through the configurations which induce variations in $P_{H_0}[W_s^* \geq C(e, d_1, \ldots, d_e) | e, d_1, \ldots, d_e]$.

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 α with the conditional significance level for most configurations.

Those configurations with small e will usually have small probability weight, unless we deal with rather extreme discrete distributions F.

Using the normal approximation we have as rejection rule

$$\frac{W_s^* - E[W_s^*]}{\sqrt{\mathrm{var}(W_s^*)}} \geq C'(e, d_1, \ldots, d_e) = z_{1-\alpha} = u_\alpha$$

The unconditional distribution of the standardized W_s^* becomes the standard normal distribution and the unconditional or overall significance level becomes approximately α 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

We are 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: 1^{st} and 2^{nd} 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.

Attribute Sampling

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 $H_0 : F = G$, namely there is no difference in the subpopulations defined by the two attribute values.

We again are dealing with the two-sample problem.

Psychological Factors and Cancer

A form of cancer is known to have wide variation in its type of progression.

We distinguish between

Group I: which has a rapidly progressing, uncontrollable form of the disease

Group II: which shows a slow progression and controllable form of the disease.

The question was whether any psychological factors are 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	-14	3	1	-16	-21	7	-7	-13	-22	-17	-14	-8	7
of Group I:	-18	-13	-18	-9	-22	-25	-24	-18	-13	-13	-18	-5	
Scores	-18	-16	-9	-14	3	-9	-16	10	-11	-3	-13	-21	
of Group II:	-2	-11	-16	-12	-13	-6	-9	-7	-11	-9			

Midranks	1	2	3.5	3.5	5.5	9	9	9	9	12	14.5	18	18
of Group I:	22.5	22.5	22.5	22.5	32	35	36.5	39	42	43.5	45.5	45.5	
Midranks	5.5	9	14.5	14.5	14.5	18	22.5	22.5	26	28	28	28	32
of Group II:	32	32	32	36.5	38	40	41	43.5	47				

```
> PsychEx1(Nsim=100000)

    Ws.star   EWs.star varWs.star      p.val  p.val.sim
 6.050e+02  5.280e+02  2.188e+03  4.985e-02  4.898e-02
```

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.

Models 1-5

The Text distinguishes the following 5 data models, all with the same analysis.

Model 1. Randomization for the comparison of two treatments.

Model 2. Population model 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!)

Power of the Wilcoxon Rank-Sum Test

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 $F \neq G$ represents 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.

However, we are typically most interested in the alternative that the treatment has a 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.

Stochastic Ordering

This last description of a beneficial effect can be expressed in terms of F and G as:

$$G(x) \lneq F(x) \quad \forall x \quad \text{where } \lneq \text{ indicates inequality for at least one } x$$

and we say that Y is stochastically larger than X, also expressed as $Y \overset{\text{st}}{\geq} X$.

The inequality $G(x) \lneq F(x)$ seems to be in contrary direction in relation to $Y \overset{\text{st}}{\geq} X$.

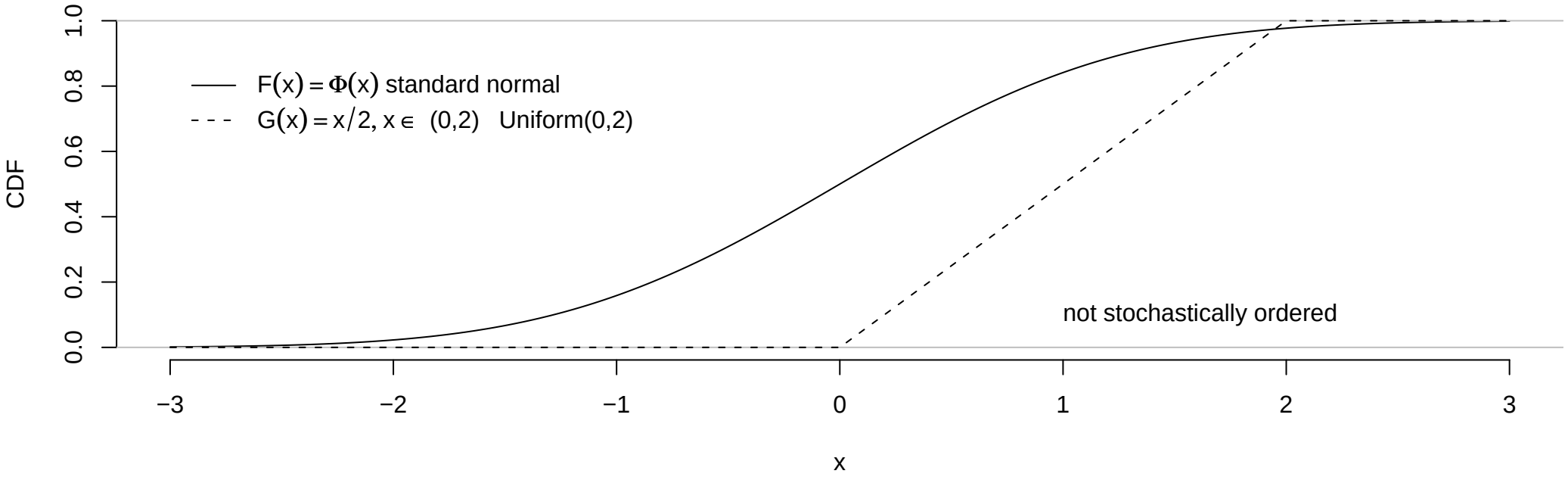
However, $\quad P(Y \leq x) = G(x) \lneq F(x) = P(X \leq x) \quad$ means that

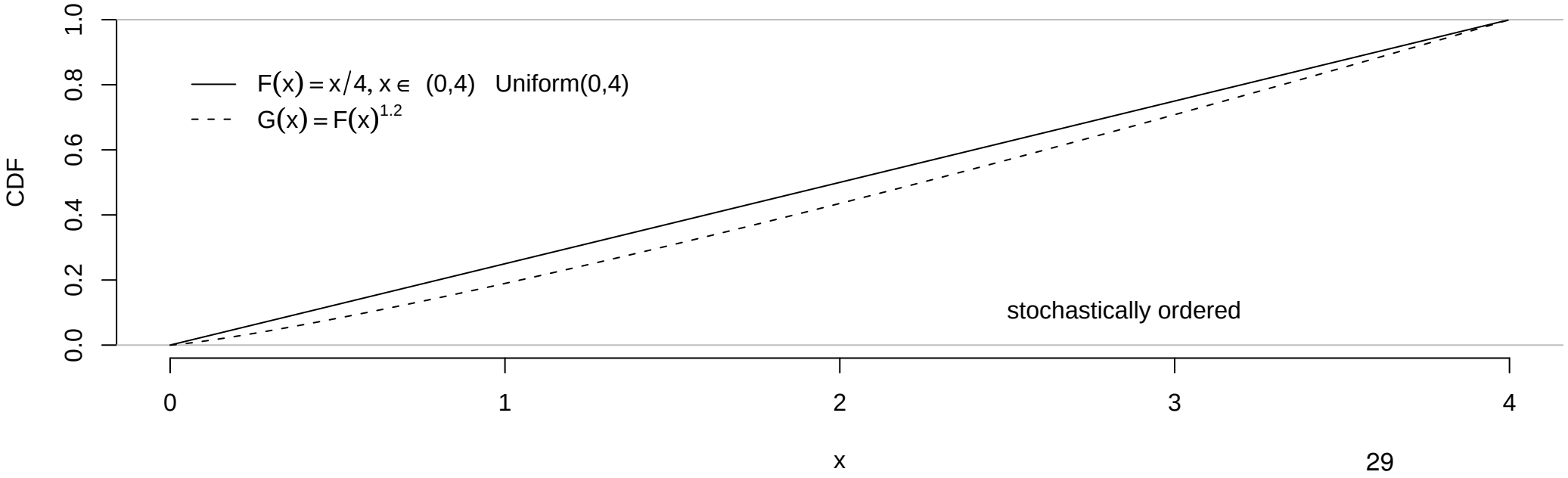
small values of Y are less likely than small values of X.

Also, $\quad P(Y > x) = 1 - G(x) \gneq 1 - F(x) = P(X > x), \quad$ means that

large values of Y are more likely than large values of X, as we had stipulated.

Illustration for Stochastic Ordering





Shift Alternatives

An important special case of $Y \overset{\text{st}}{\geq} X$ is the shift model: For some $\Delta > 0$ it assumes

$$G(x) = F(x - \Delta) \quad \text{for all } x.$$

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

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

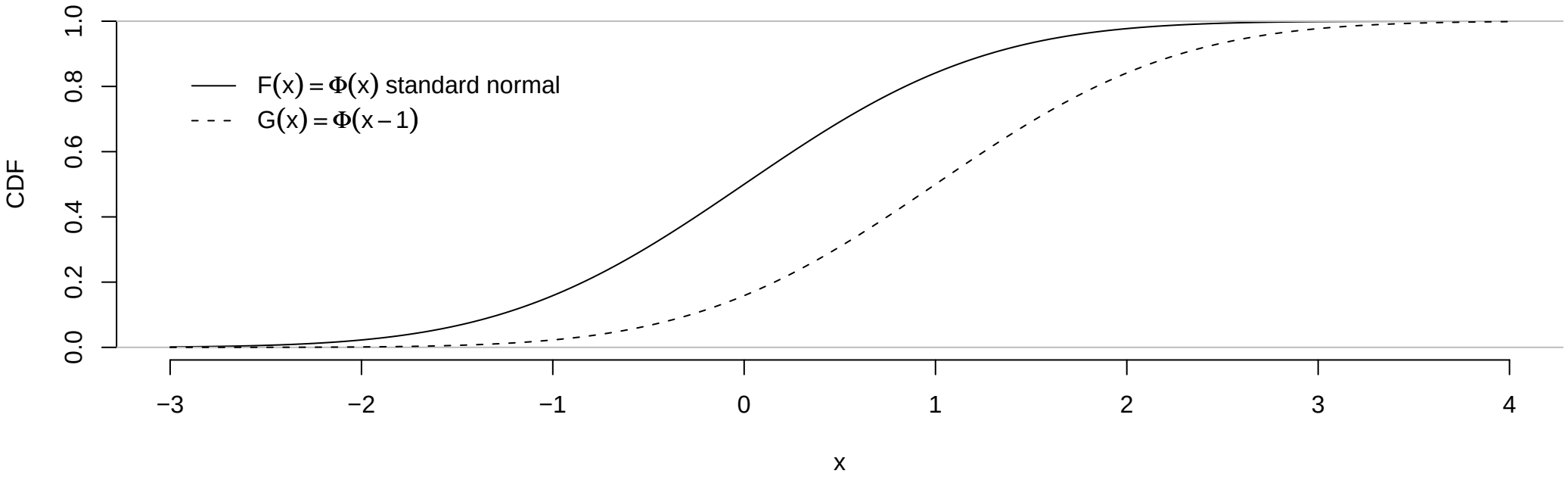
$$Y = X + \Delta \quad \Longrightarrow \quad G(y) = P(Y \leq y) = P(X + \Delta \leq y) = P(X \leq y - \Delta) = F(y - \Delta)$$

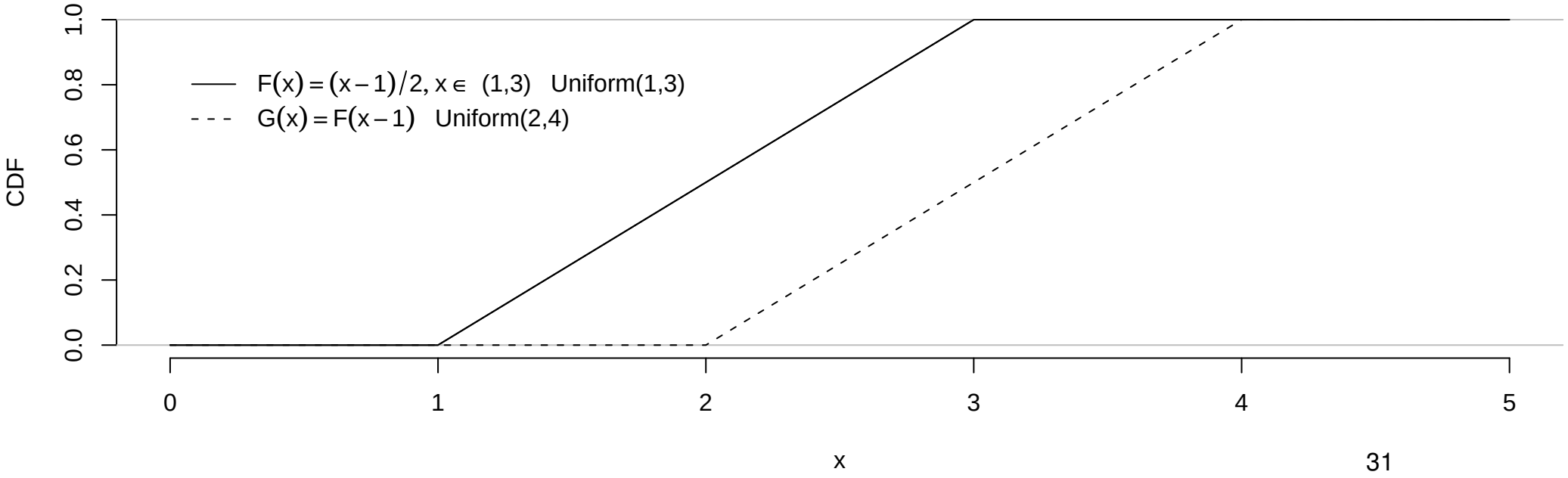
Conversely,

$$Y \sim G(y) = F(y - \Delta) \quad \Longrightarrow \quad P(Y - \Delta \leq x) = P(Y \leq x + \Delta) = F(x + \Delta - \Delta) = F(x)$$

Thus we can view $Y = (Y - \Delta) + \Delta = X' + \Delta$, where $X' \sim F$.

Illustration for the Shift Model





Comment on Stochastic Ordering

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

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 and Y are highly correlated and 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 $X \lneq Y$ only when the support of the X distribution is completey to the left of the support of the Y distribution. Example: $X \sim U(0,1)$ and $Y \sim U(2,4)$, then $P(X \lneq Y) = 1$.

Power Function under Shift Model

For the Wilcoxon rank-sum test the power function is given by

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

where P_Δ indicates that probabilities are calculated under the shift model

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

Strictly speaking we should write $P_{F,\Delta}$ and that is captured in $\Pi_F(\Delta)$.

We extend the definition of the power function 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 Δ.

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 $H_0 : \Delta = 0$.

Increasing Power Function under Shift Model

Theorem 2: The power function $\Pi_F(\Delta)$ is an increasing function of Δ.

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

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

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

To understand the inequality, recall that W_{XY} is the number of pairs with $X_i < Y_j$
and that is always $\leq$ to W_{XV}, the number of pairs with $X_i < V_j$, 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

$$\Pi_F(\Delta) \geq \alpha \quad \text{for all } \Delta > 0.$$

Tests for which the power does not fall below the significance level α 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 α.

Thus we can view our testing problem also as testing $H_0' : \Delta \leq 0$ against $A : \Delta > 0$. Our Wilcoxon rank-sum test is still a level α test and unbiased.

Symmetric Comparison of Two Treatments Revisited

In the symmetric comparison of two treatments A and B we chose treatment B as better when $W_B \geq n(N+1)/2+c$, treatment A as better if $W_B \leq n(N+1)/2-c$ and suspend judgment otherwise. The critical value c is determined so that

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

$$P_{H_0}(\text{choosing } B) = P_{H_0}(W_B \geq n(N+1)/2+c) = \alpha' \quad \text{no ties}$$

$F(x)$ and $F(x-\Delta)$ be the (continuous) response CDFs for A and B, respectively.

$\Delta = 0 \implies$ none of the three decisions constitutes an error, because both treatments are equally good. Interpret "better" as $\geq$.

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 77, Chapter 1)

More Flexible Shift Alternatives

Our shift model assumed a constant shift effect Δ 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 $\Delta(x) \gneq 0$ for all x.

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

If F and G are continuous and strictly increasing then $G(x) \lneq F(x)$ for all x implies the existence of a function $\Delta(x) \gneq 0$ such that G is the CDF of $X + \Delta(X)$ when F is the CDF of X.

More General Unbiasedness Property

For the generalized shift model $Y = X + \Delta(X)$ we have

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

(ii') The Wilcoxon test has level α 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 α', provided the treatment effect is $\Delta(x) \gneq 0$ for all x or $\Delta(x) \lneq 0$ for all x.

Quantitative Power Calculations

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.

However, in R the value of any power $\Pi(F, G)$ can easily be estimated via simulation, provided 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_{sim} of simulations to be run.

The time to run such simulations is proprtional to N_{sim} and increases with m and n.

Fortunately we have another option for large m and n.

Mean and Variance of W_{XY}

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

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

and

$$\text{var}(W_{XY}) = mnp_1(1-p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2)$$

where

$$p_2 = P(X < Y \ \cap \ X < Y') \qquad \text{and} \qquad p_3 = P(X < Y \ \cap \ X' < Y)$$

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

$$p_1 = P(X < Y) = P(Y < X) = \frac{1}{2} \quad \text{since} \quad P(X = Y) = 0\,. \quad \text{Thus} \quad E(W_{XY}) = \frac{mn}{2}.$$

and

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

since all three random variables X, Y,Y' have the same chance to be the smallest.

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

since all three random variables X,X', Y have the same chance to be the largest.

$$\begin{aligned} \operatorname{var}(W_{XY}) &= \frac{mn}{4} + mn(n-1)\left(\frac{1}{3} - \frac{1}{4}\right) + mn(m-1)\left(\frac{1}{3} - \frac{1}{4}\right) \\ &= \frac{mn}{12}(3 + n - 1 + m - 1) = \frac{mn(N+1)}{12} \end{aligned}$$

Asymptotic Distribution of W_{XY}

For large m and n the distribution of W_{XY} is approximately normal with the previously given mean and variance so that

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

$p_1 = 0$ and $p_1 = 1$ are trivial situations with the distributions completely separated.

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 again employ the continuity correction via $w + .5$ in place of the integer w.

Example 2: The Effect of Background Music.

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 full a enumeration of all $\binom{20}{10} = 184756$ splits and evaluate W_{XY} each time.
We find a proportion $\alpha_c = 0.05256122$ of W_{XY} values $\geq c = 72$ and
a proportion $\alpha_c = 0.04460478$ of W_{XY} values $\geq c = 73$.
Rejecting H_0 for $W_{XY} \geq 72$ comes closest to a level $\alpha = .05$ test.

A shift alternative is a reasonable focus, in particular a normal shift alternative,
i.e., $F = \mathcal{N}(\xi, \sigma^2)$ and $G = \mathcal{N}(\xi + \Delta, \sigma^2)$, with $\Delta = 5$ of particular interest.

$\Pi_{F,G}$ does not depend on ξ but depends on σ^2, i.e., $\Pi_{F,G} = \Pi_{F,\sigma^2}(\Delta) = \Pi_F(\Delta)$.
The probability of detecting a shift increases as the background variation decreases,
i.e., as $\sigma^2 \searrow 0$. Past experience shows that $\sigma^2 = 32$ is a reasonable value.

Example 2: Normal Approximation for Computing Power

$$p_1 = P(X < Y) = P(X - Y < 0) = \Phi\left(\frac{\Delta}{\sigma\sqrt{2}}\right) = \Phi\left(\frac{5}{\sqrt{32}\sqrt{2}}\right) = \Phi(.625) = 0.734$$

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

$$\begin{aligned} p_2 &= P(X < Y,\, X < Y') = P\left(\frac{X-(Y-\Delta)}{\sigma\sqrt{2}} < \frac{\Delta}{\sigma\sqrt{2}}, \frac{X-(Y'-\Delta)}{\sigma\sqrt{2}} < \frac{\Delta}{\sigma\sqrt{2}}\right) \\ &= P(Z < z,\, Z' < z) \quad \text{with } z = \Delta/(\sigma\sqrt{2}) = 5/8 = .625 \end{aligned}$$

and Z and Z' are standard normal random variables with correlation $1/2$ since

$$\operatorname{cov}(Z, Z') = \operatorname{cov}\left(\frac{X - Y + \Delta}{\sigma\sqrt{2}}, \frac{X - Y' + \Delta}{\sigma\sqrt{2}}\right) = \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}$$

```
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 $E(W_{XY}) = mnp_1 = 73.4$, $\operatorname{var}(W_{XY}) = 129.03$ and thus

$$\Pi_F(\Delta = 5) \approx 1 - \Phi((71.5 - 73.4)/\sqrt{129.03}) = 0.5665.$$

Installing the Package mnormt

Install package mnormt in R either by using the menu choice

Packages → Installpackage(s)

then choosing the nearest download mirror site, i.e., here USA (WA), and then choose mnormt from the menu list of packages,

or alternatively, execute the command line install.packages("mnormt") in an R session and choose the mirror site when prompted.

This package install is just 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 when $W_{XY} \geq c = 72$ and obtain 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

$$\texttt{pmnorm}(\texttt{c}(1.25, 1.25), \texttt{c}(0,0), \texttt{varcov} = \texttt{matrix}(\texttt{c}(1,.5,.5,1), \texttt{ncol} = 2))$$

with resulting

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

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.

Other Distribution Alternatives

The same method should work in principle for other alternatives, except that the determination of p_1, p_2 and p_3 may not be this easy.

We made use of normal distribution properties and used the bivariate normal probability function `pmnorm` (from package `mnormt`) in R.

However, as long as we are able to generate random variables $X \sim F$ and $Y \sim G$ we can obtain good estimates of $p_1, p_2,$ and 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.

Some Checking of Approximation

Simulations of the W_{XY} distribution (using $N_{\text{sim}} = 100000$) produced the following estimates

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

These are not too different from our normal approximation calculations.

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 previous approximate values (from the normal approximation) falls within its respective interval. This should not be of much practical concern.

Note $.5665 < (0.588, 0.594)$ and $(0.979, 0.981) < .9934$.

Alternate Power Approximation

An alternate power approximation within the shift model is available for small Δ, namely

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

where $u_\alpha = \Phi^{-1}(1-\alpha)$ is the upper α point of the standard normal distribution.

$f^*(z)$ is the density of $F^*(z) = P(X - Y \leq z)$, where X and Y are independent random variables with common distribution function F.

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

$$\Longrightarrow \quad f^*(z) = \int_{-\infty}^{\infty} f(y+z) f(y) dy \quad \Longrightarrow \quad f^*(0) = \int_{-\infty}^{\infty} f^2(y) dy$$

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

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

and we get

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

The latter expression captures the clear dependence of $\Pi_F(\Delta)$ on just Δ/σ.

This should be clear from the equivalence of (also true for other shift models)

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

as far as the distribution of ranks is concerned.

Illustration of Alternate Approximation

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

$$\Pi_\Phi(\Delta = 5) \approx \widetilde{\Pi}_\Phi(\Delta = 5) = \Phi\left[\sqrt{\frac{3 \cdot 10 \cdot 10}{21 \cdot \pi}}\,\frac{5}{\sqrt{32}} - 1.645\right] = .5948 \,\widehat{=}\, .5665\ (.5912)_{\text{sim}}$$

while for $\Delta = 10$ we get

$$\Pi_\Phi(\Delta = 10) \approx \widetilde{\Pi}_\Phi(\Delta = 10) = \Phi\left[\sqrt{\frac{3 \cdot 10 \cdot 10}{21 \cdot \pi}}\,\frac{10}{\sqrt{32}} - 1.645\right] = .9832 \,\widehat{=}\, .9934\ (.9802)_{\text{sim}}$$

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 $\alpha = .05$. Instead we settled for $\alpha = .0526$, while 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

$$\begin{array}{lcllcl} \Pi_\Phi(\Delta = 0) & \approx & \widetilde{\Pi}_\Phi(\Delta = 0) & = & 0.0526 & \\ \Pi_\Phi(\Delta = 5) & \approx & \widetilde{\Pi}_\Phi(\Delta = 5) & = & 0.6042 & \widehat{=} \quad .5665\ (.5912)_{\text{sim}} \\ \Pi_\Phi(\Delta = 10) & \approx & \widetilde{\Pi}_\Phi(\Delta = 10) & = & 0.9842 & \widehat{=} \quad .9934\ (.9802)_{\text{sim}} \end{array}$$

This raised the approximate power slightly, as it should.

The comparison is still quite favorable compared with the simulated values.

Comparing Exact Power with Approximations

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

Δ/σ	0	0.2	0.4	0.6	0.8	1.0	1.5	2.0
Exact	0.049	0.094	0.165	0.264	0.386	0.520	0.815	0.958
Simulated (400,000)	0.049	0.095	0.165	0.264	0.386	0.521	0.815	0.959
Approximation	0.048	0.094	0.160	0.249	0.359	0.485	0.807	0.981
Alternate Approximation	0.049	0.096	0.171	0.275	0.403	0.543	0.839	0.970

$.049$ is rounded from $.04865967$ which is the attainable level closest to $.05$ and obtains by rejecting H_0 whenever $W_{XY} \geq 38$.

The exact (obtained from Milton (1970)) and simulated values agree fairly 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 $\Rightarrow$ plan a larger sample size.

Our alternate approximation suggests that the power is maximal for fixed $N = m+n$ when $m = n = N/2$ when N is even. When $N = 2k+1$ the optimal choices are to allocate sample sizes as $(m,n) = (k,k+1)$ or $(m,n) = (k+1,k)$.

Assuming $N = 2n$ and using $N = 2n$ in place of $N+1$ in our alternate approximation we have

$$\Pi_F(\Delta) \approx \Phi[\sqrt{6n} f^*(0)\Delta - u_\alpha] = \Pi = 1 - \Phi(u_\Pi) = \Phi(-u_\Pi) \quad (\Pi = \text{ desired power})$$

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

For $\Delta = 5$, $\sigma = \sqrt{32}$, $\Pi = .95$, $\alpha = .05$ we get $n = m \approx 29$.

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

If we had 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.

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

$$\Rightarrow \quad X_i' - X_i = V_i' - V_i \quad \text{with variance } \sigma_V^2 + \sigma_V^2 = 2\sigma_V^2$$

while the variance of X_i is $\sigma_U^2 + \sigma_V^2 \gg 2\sigma_V^2$ since typically $\sigma_U \gg \sigma_V$.

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

$$Y_i' - Y_i = \Delta + V_i' - V_i \quad \text{again with variance } \sigma_V^2 + \sigma_V^2 = 2\sigma_V^2$$

Why Greater Efficiency?

Recall

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

We see that for fixed m and n the power increases as σ gets reduced, which is exactly what we hope to accomplish by taking the difference in responses on the same subject.

Conversely,

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

For smaller σ this leads to a smaller $n = N/2$ to achieve the same power Π.

This makes sense since the presence of a signal Δ is more easily discernable against less response variation.

Sample Size Calculation

We want to 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.

Checking $N+1 \longrightarrow N$ Effect

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 treatment difference when

$$t(\mathbf{X},\mathbf{Y}) = \frac{\bar{Y}-\bar{X}}{S\sqrt{1/n+1/m}} \geq c \quad \text{with} \quad S^2 = \frac{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_j-\bar{Y})^2}{m+n-2}$$

and c is the $(1-\alpha)$-quantile of the t-distribution with $m+n-2$ degrees of freedom.

This test assumes the normal shift model in which $\mathbf{X} = (X_1 \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independently, normally distributed with common unknown variance σ^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 $A : \Delta > 0, \sigma > 0$.

Distribution of Student's t-Test Statistic

Under $H_0 : \Delta = 0, \sigma > 0$ the Student t-statistic $t(\mathbf{X}, \mathbf{Y})$ has a (central) Student t-distribution with $m+n-2$ degrees of freedom.

When $\Delta \neq 0$ the test statistic $t(\mathbf{X}, \mathbf{Y})$ 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.

Here `ncp` is the 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 $c =$`qt(1-`α`,`$m+n-2$`)`.

Comparing Student's t-Test with the Wilcoxon Test

Often the assumption of normality is not satisfied or 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

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

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

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

$$\frac{\sum_{i=1}^m (X_i - \bar{X})^2}{m} = \frac{\sum_{i=1}^m (X_i - \xi)^2 - m(\bar{X} - \xi)^2}{m} = \frac{\sum_{i=1}^m (X_i - \xi)^2}{m} - (\bar{X} - \xi)^2 \longrightarrow \sigma^2$$

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

$$S^2 = \frac{m\sum_{i=1}^m (X_i - \bar{X})^2/m + n\sum_{i=1}^m (Y_j - \bar{Y})^2/n}{m + n - 2} \longrightarrow \sigma^2 \quad \text{and thus} \quad S/\sigma \longrightarrow 1$$

Asymptotic Normality of $t(\mathbf{X}, \mathbf{Y})$

Combining the results from the previous slide we have

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

If we denote the critical point of the t-test by $c_{m,n}(\alpha)$ (based on normality) then the above result shows that $c_{m,n}(\alpha) \longrightarrow u_\alpha$, the upper α quantile of the standard normal distribution.

In large samples we can thus as well use u_α in place of $c_{m,n}(\alpha)$ with the added advantage that the level is approximately α no matter what the nature of F is, aside from it having a finite variance σ^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

$$\begin{aligned}\Pi'_F(\Delta) &= P\left(\frac{\bar{Y}-\bar{X}}{S\sqrt{1/m+1/n}} \geq c\right) = P\left(\frac{\bar{Y}-\bar{X}-\Delta}{\sigma\sqrt{1/m+1/n}} \geq \frac{cS}{\sigma} - \frac{\Delta}{\sigma\sqrt{1/m+1/n}}\right) \\ &\approx P\left(Z \geq u_\alpha - \frac{\Delta}{\sigma\sqrt{1/m+1/n}}\right) = \Phi\left(\frac{\Delta}{\sigma}\sqrt{\frac{mn}{N}} - u_\alpha\right)\end{aligned}$$

where 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. We can compare this with our second approximation to the power of W_{XY}, namely

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

Note the similarities in the approximations to the two power functions. Both involve m and n through roughly equal factors since $\sqrt{mn/N}/\sqrt{mn/(N+1)} \approx 1$.

The distribution nature F is reflected in $\sigma = \sigma_F$ and $f^*(0)$, respectively.

Calculation for Background Music Example

Using $\alpha = .0526$ our alternate approximation gave us a power of $\Pi_\Phi(\Delta = 5) = .604$ for the Wilcoxon test when we assumed that $\sigma^2 = 32$.

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

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

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

Given that we assumed normality we can compute the true power of the t-test as

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

which computes to 0.612.

Note that 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 $t(\mathbf{X}, \mathbf{Y})$ 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 β 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

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

$$\Longrightarrow \quad m' = (\sigma f^*(0))^2 \frac{12m}{1+\frac{1}{2m}} \approx 12m\sigma^2 f^{*2}(0)$$

Efficiency of W_{XY} Relative to $t(\mathbf{X}, \mathbf{Y})$

The ratio

$$\frac{m'}{m} = (\sigma f^*(0))^2 \frac{12}{1 + \frac{1}{2m}} \approx 12\sigma^2 f^{*2}(0)$$

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.

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

is called the Pitman efficiency or asymptotic relative efficiency (ARE) of the Wilcoxon test w.r.t. the t-test.

When F is normal we have $f^*(0) = 1/(2\sigma\sqrt{\pi})$ so that $e_{W,t}(\Phi) = 3/\pi = .955$.

Exact Efficiency of W_{XY} Relative to $t(\mathbf{X}, \mathbf{Y})$

The previous efficiency formulas are based on approximations that require large sample sizes. One benefit was that the dependence on Δ/σ and α 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, thus $\alpha = 2 \times 4/252 = 4/126 = .0317$.

Δ	.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Π	.072	.210	.431	.674	.858	.953	.988	.998
e	.968	.978	.961	.956	.960	.960	.964	.976

The last row for *e* is a modification by Hodges and Lehmann (1956).

Thus the small sample efficiency is well represented by the ARE.

$e_{W,\,t}(F)$ for Other Distibutions F

F	Logistic	Double Exponential	Rectangular	Exponential
$e_{W,\,t}(F)$	$\pi^2/9 = 1.097$	1.5	1	3

Hodges and Lehmann (1956) showed

$$e_{W,\,t}(F) \geq \frac{108}{125} = .864$$

for all distributions F with finite variance, with the minimum attained for some F.

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

Sir David Cox in his recent Norm Breslow Lecture here 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.

Testing $H_{\Delta_0} : \Delta = \Delta_0$

We have dealt with testing the hypothesis $H_0 : \Delta = 0$ in the shift effect model.

This is easily extended to testing $H_{\Delta_0} : \Delta = \Delta_0$ for any specified Δ_0.

The solution is to subtract Δ_0 from the $Y_1, \ldots, Y_n$, i.e., form $Y_i' = Y_i - \Delta_0$, and then test the previous hypothesis $H_0 : \Delta = 0$ 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

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

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

The hypothesis $H_0 : \Delta' = 0$ is equivalent to $H_0 : \Delta - \Delta_0 = 0 \iff H_{\Delta_0} : \Delta = \Delta_0$.

Power Symmetry of the Two-sided Wilcoxon Test

Recall: 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 $G(x) = F(x-\Delta)$, i.e., $\Pi_F(\Delta) = \Pi_F(-\Delta)$.

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 thus $Y_j = Y_j' + \Delta \sim F(x-\Delta)$

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

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

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

where in $\stackrel{*}{=}$ we made the following legitimate interchanges:

when $m = n$ we interchanged X_i with Y_j' or

when F is symmetric around $a = 0$ we interchanged $-X_i$ with X_i and Y_j' with $-Y_j'$.

Note that $X_i - Y_j' = X_i - a - (Y_j' - a)$ and $X_i - a$ and $Y_j' - a$ are distributed symmetrically around zero. □

Estimating the Treatment Shift Effect

Often we are interested in estimating the shift effect Δ 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 Δ.

There is a strong connection between confidence intervals and testing hypotheses.

Rationale for Shift Effect Estimate

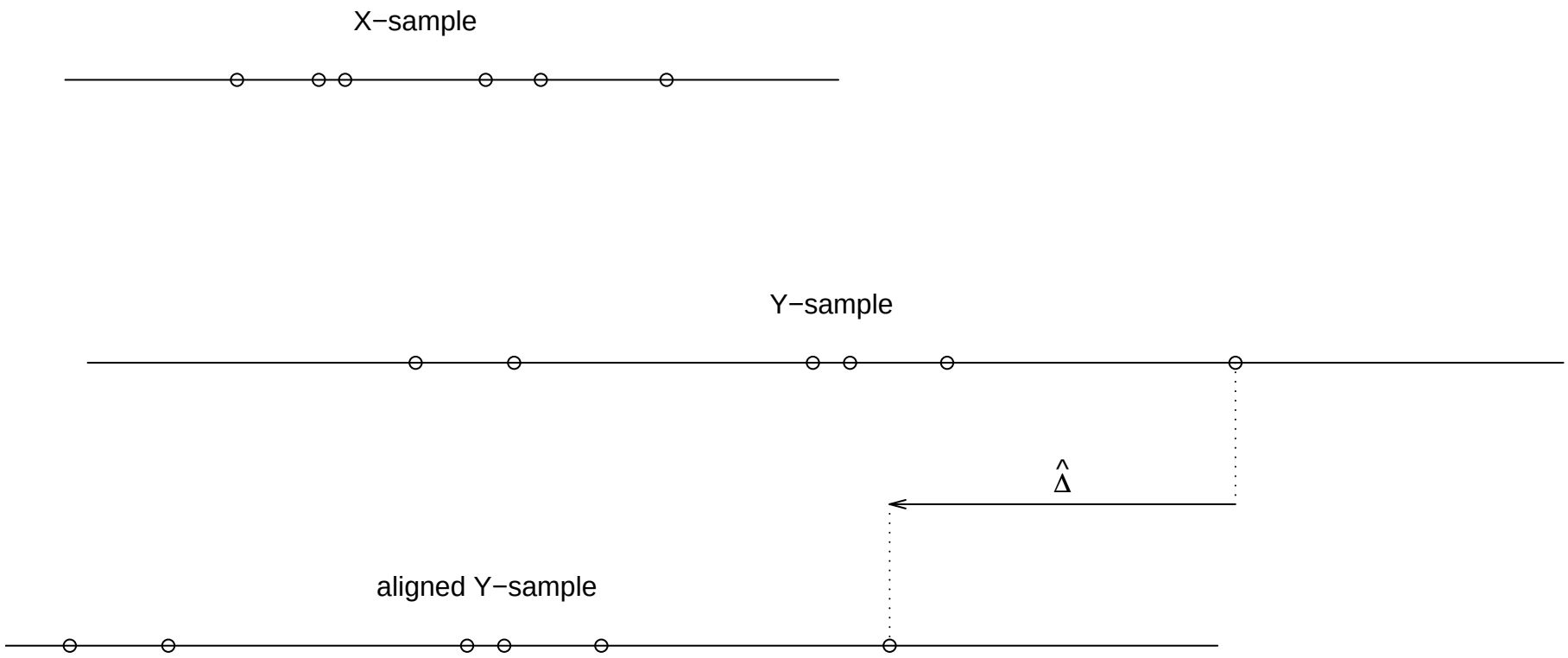
The hypothesis $H_0 : \Delta = 0$ is least likely to be rejected whenever the X_i's and Y_j's are in fairly close agreement with each other.

Close agreement can be measured by the p-value of the 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



The Hodges-Lehmann Estimator (mn even)

Denote by $D_{(1)} < D_{(2)} < \ldots < D_{(mn)}$ the ordered values of all mn differences $Y_j - X_i$, which are all distinct with probability one when F is continuous.

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 preference is to take the midpoint of that interval as our estimate

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

It is known as the Hodges-Lehmann estimator of Δ.

The Hodges-Lehmann Estimator (mn odd)

The case $mn = 2k+1$ odd requires some special attention. Then the center of symmetry for the W_{XY} distribution under H_0 is $mn/2 = k+1/2$ (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 $$W_{X,Y-a} < k+\frac{1}{2} \iff a > D_{(k+1)}$$

$D_{(k+1)} = Y_{j_0} - X_{i_0}$, we see that $a = D_{(k+1)} \iff a = Y_{j_0} - X_{i_0}$ or $Y_{j_0} - a = X_{i_0}$, i.e., the observations $Y_{j_0} - a$ and X_{i_0} are tied.

The tie version W^*_{XY} of the statistic W_{XY} counted such comparisons with weight $\frac{1}{2}$.

As a passes over $D_{(k+1)}$ the statistic W_{XY} in its modified form W^*_{XY} achieves the central, non-integer value $k+\frac{1}{2}$. Thus it makes sense to take

$$\hat{\Delta} = D_{(k+1)} = \text{median}(Y_j - X_i,\ i = 1,\ldots,m,\ j = 1,\ldots,n) = \text{med}(Y_j - X_i)$$

Commuting Example

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 $mn = 55$.

The Text discusses a manual procedure for relatively efficient computation.

We will use R instead.

Using R

```
> outer(RouteB,RouteA,"-")
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
[1,]  0.7  0.7  0.6  0.5  0.5  0.5  0.2  0.2  0.1   0.0   0.0
[2,]  1.0  1.0  0.9  0.8  0.8  0.8  0.5  0.5  0.4   0.3   0.3
[3,]  1.3  1.3  1.2  1.1  1.1  1.1  0.8  0.8  0.7   0.6   0.6
[4,]  1.5  1.5  1.4  1.3  1.3  1.3  1.0  1.0  0.9   0.8   0.8
[5,]  4.4  4.4  4.3  4.2  4.2  4.2  3.9  3.9  3.8   3.7   3.7
```

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 $\mathrm{med}(\mathtt{RouteB}_j - \mathtt{RouteA}_i)$ of the previous array, i.e., the Hodges-Lehmann estimator.

Using `wilcox.test`

```
> 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
             0.9553555
```

For some reason the estimate `0.9553555` does not match up here.

I suspect it is linked to the presence of ties in the data.

We get a warning when we omit `exact=F` or use `exact=T`.

Using `wilcox.test` without Ties

```
> x
 [1] -1.44 -1.20 -1.15 -1.08 -0.90 -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`

Install the package `coin` (for installation see previous example `mnormt`).

For each new R session invoke `library(coin)` prior to using `wilcox_test`.

Prepare data for required structure in `wilcox_test`.

```
> Route=as.factor(c(rep("A",11),rep("B",5)))
> Route.Time = c(RouteA,RouteB)
> Route
 [1] A A A A A A A A A A A B B B B B
Levels: A B
> Route.Time
 [1]  5.8  5.8  5.9  6.0  6.0  6.0  6.3  6.3  6.4  6.5  6.5  6.5  6.8
[14]  7.1  7.3 10.2
```

calling wilcox_test

```
> wilcox_test(Route.Time~Route,conf.int=TRUE, distr = exact())

Exact Wilcoxon Mann-Whitney Rank Sum Test

data:  Route.Time by Route (A, B)
Z = -3.0245, p-value = 0.001374
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -3.7 -0.5
sample estimates:
difference in location
                 -0.9
```

The answer is the exact $.9$, except that it has a negative sign.

That is caused by the alphabetical order of the factor label values A and B.

Getting the Correct Sign

```
> Route2 <- factor(as.character(Route), levels = c("B", "A"))

> Route2
 [1] A A A A A A A A A A A B B B B B
Levels: B A
> Route
 [1] A A A A A A A A A A A B B B B B
Levels: A B
> as.numeric(Route)
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2
> as.numeric(Route2)
 [1] 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1
```

Note that the order of levels has changed in `Route2`

The numerical values 1 and 2 have changed to 2 and 1.

Analysis Using Route2

```
> wilcox_test(Route.Time~Route2,conf.int=TRUE, distr = exact())

Exact Wilcoxon Mann-Whitney Rank Sum Test

data:  Route.Time by Route2 (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 $\widehat{\Delta} = \text{med}(Y_j - X_i)$

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 ∞.

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 matters we assume $k_1 = k_2 = k$ and $m = n$ and the criterion becomes

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

i.e., we could tolerate about 29% outliers in either sample when $m = n$ without affecting the estimate $\widehat{\Delta}$. 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, e.g., rounding to the nearest decimal.

Assume that the original observations are X_i' and Y_j', assumed to be 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 $0, \pm\varepsilon, \pm 2\varepsilon, \ldots$.

From rounding to the nearest multiple $k\varepsilon$ we have $|X_i - X_i'| \leq \frac{\varepsilon}{2}$ and $|Y_j - Y_j'| \leq \frac{\varepsilon}{2}$

$$\implies \quad |(Y_j - X_i) - (Y_j' - X_i')| = |(Y_j - Y_j') - (X_i - X_i')| \leq |Y_j - Y_j'| + |X_i - X_i'| \leq \varepsilon$$

$$|\widehat{\Delta}' - \widehat{\Delta}| = |\mathrm{med}(Y_j' - X_i') - \mathrm{med}(Y_j - X_i)| \leq \varepsilon$$

Thus the rounding errors don't build up.

$$|a_i - a'_i| \leq \varepsilon \; \forall i \implies |a_{(k)} - a'_{(k)}| \leq \varepsilon \; \forall k$$

Here $a_{(k)}$ ($a'_{(k)}$) is the k^{th} ordered value of the a_i (a'_i).

Proof of the above implication:

Assume $a'_{(k)} < a_{(k)} - \varepsilon$, i.e., we have at least k of the a'_i which are $< a_{(k)} - \varepsilon$.

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 $a_{(k)}$, which can only exceed at most $k-1$ of the a_i.

The assumption $a'_{(k)} > a_{(k)} + \varepsilon$ leads to a contradiction in a similar fashion. □

This result $\implies$ $|\text{med}(Y'_j - X'_i) - \text{med}(Y_j - X_i)| \leq \varepsilon$ on the previous slide.

Shift Model with Continuous F

When $X_i \sim F(x)$ and $Y_j \sim G(y) = F(y-\Delta)$ with F continuous, then $\widehat{\Delta}$ has a continuous distribution, i.e.,

$$P(\widehat{\Delta} = d) = 0 \quad \text{for any } d \in R.$$

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

$$P(\widehat{\Delta} = d) \leq \sum_i \sum_j P(Y_j - X_i = d) = 0$$

and for $mn = 2k$ (even) this follows similarly from

$$P(\widehat{\Delta} = d) \leq \sum_i \sum_j \sum_{i'} \sum_{j'} P(Y_j - X_i + Y_{j'} - X_{i'} = 2d) = 0$$

Distribution of $\widehat{\Delta} - \Delta$

Lemma 1: The distribution of $\widehat{\Delta} - \Delta$ is independent of Δ.

Proof: The difference $\widehat{\Delta} - \Delta = \text{med}(Y_j - X_i) - \Delta = \text{med}(Y_j - \Delta - X_i)$ has a distribution independent of Δ because the distribution of $X_i \sim F$ does not involve Δ and

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

no longer involves Δ.

Symmetry of the $\widehat{\Delta}$ Distribution around Δ

Theorem 3: The estimator $\widehat{\Delta}$ of the shift parameter Δ is distributed symmetrically around Δ if either of the following two conditions hold:

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

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

Proof:

(i) symmetry of F around $\mu \Longrightarrow Y_j - \Delta - \mu \stackrel{\mathcal{D}}{=} \mu + \Delta - Y_j$ and $X_i - \mu \stackrel{\mathcal{D}}{=} \mu - X_i$

$$\begin{aligned} \widehat{\Delta} - \Delta &= \operatorname{med}(Y_j - X_i) - \Delta = \operatorname{med}(Y_j - \Delta - \mu - (X_i - \mu)) \\ &\stackrel{\mathcal{D}}{=} \operatorname{med}(\mu + \Delta - Y_j - (\mu - X_i)) = \Delta - \operatorname{med}(Y_j - X_i) = \Delta - \widehat{\Delta} \end{aligned}$$ □

(ii) $(X, Y-\Delta)$ and $(Y-\Delta, X)$ have the same distribution

$$\widehat{\Delta} - \Delta = \operatorname{med}(Y_j - \Delta - X_i) \stackrel{\mathcal{D}}{=} \operatorname{med}(X_j - (Y_i - \Delta)) = \Delta - \operatorname{med}(Y_i - X_j) = \Delta - \widehat{\Delta}$$

The interchange $(i, j) \to (j, i)$ in $(Y_j, X_i) \to (Y_i, X_j)$ is possible since $m = n$. □

Symmetry of the $\widehat{\Delta}$ Distribution and Unbiasedness

The symmetry property on the previous slide (provided conditions are satisfied) implies two unbiasedness properties.

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

The distribution of $\widehat{\Delta}$ has median Δ. $\widehat{\Delta}$ is thus called is median unbiased. We have $P(\widehat{\Delta} < \Delta) = P(\widehat{\Delta} > \Delta)$, i.e., $\widehat{\Delta}$ underestimates Δ as likely as it overestimates Δ.

If in addition F is continuous, we have $P(\widehat{\Delta} = \Delta) = 0$ and thus

$$P(\widehat{\Delta} < \Delta) = P(\widehat{\Delta} > \Delta) = \frac{1}{2}.$$

When the conditions of Theorem 3 are no longer satisfied, then $\widehat{\Delta}$ will usually no longer be distributed symmetrically around Δ. The median unbiasedness property continues to hold when mn is odd, and approximately so when mn is even.

See the next four slides.

Relation between Ordered Differences $D_{(\ell)}$ and W_{XY}

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 ℓ with $1 \leq \ell \leq mn$ and any $a \in R$ we have

$$D_{(\ell)} \leq a \qquad \Longleftrightarrow \qquad W_{X,Y-a} \leq mn - \ell$$

and hence

$$D_{(\ell)} > a \qquad \Longleftrightarrow \qquad W_{X,Y-a} \geq mn - \ell + 1$$

Proof:

$$\begin{aligned} D_{(\ell)} \leq a \quad &\Longleftrightarrow \quad \text{at least } \ell \text{ of the differences } (Y_j - a) - X_i \text{ are } \leq 0, \\ &\Longleftrightarrow \quad \text{at most } mn - \ell \text{ of these differences are } > 0 \\ &\Longleftrightarrow \quad W_{X,Y-a} \leq mn - \ell. \end{aligned}$$

□

Median Unbiasedness Revisited

Assume $mn = 2k+1$ is odd and F is continuous, thus $D_{(1)} < \ldots < D_{(mn)}$ with probability one. We then have $\widehat{\Delta} = D_{(k+1)}$ and

$$\begin{aligned} P_\Delta(\widehat{\Delta} < \Delta) &=^1 P_\Delta(\widehat{\Delta} \leq \Delta) =^2 P_0(\widehat{\Delta} \leq 0) =^3 P_0(D_{(k+1)} \leq 0) \\ &=^4 P_0(W_{X,Y} \leq 2k+1-(k+1)) =^5 P_0(W_{X,Y} \leq k) = \frac{1}{2} \end{aligned}$$

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

$=^2$ since $\widehat{\Delta} - \Delta$ has distribution independent of Δ (hence w.l.o.g. $\Delta = 0$)

$=^3$ from $\widehat{\Delta} = D_{(k+1)}$, $\quad =^4$ from Theorem 4

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

Note that the symmetry property from Theorem 3 was not invoked.

Similarly $P_\Delta(\widehat{\Delta} > \Delta) = P_0(W_{X,Y} \geq k+1) = \frac{1}{2}$ $\quad \Longrightarrow \quad$ $\widehat{\Delta}$ is median unbiased.

Median Unbiasedness ($mn = 2k$ Even)

In that case $\widehat{\Delta} = [D_{(k)} + D_{(k+1)}]/2$ and we have

$$P_0(D_{(k)} \leq 0) \quad = \quad P_0(W_{X,Y} \leq 2k-k) = P_0(W_{X,Y} \leq k)$$

$$P_0(D_{(k+1)} \leq 0) \quad = \quad P_0(W_{X,Y} \leq 2k-(k+1)) = P_0(W_{X,Y} \leq k-1) = P_0(W_{X,Y} < k)$$

and thus

$$P_0(W_{X,Y} < k) = P_0(D_{(k+1)} \leq 0) \leq P_\Delta(\widehat{\Delta} \leq \Delta) \leq P_0(D_{(k)} \leq 0) = P_0(W_{X,Y} \leq k)$$

The null distribution of $W_{X,Y}$ is symmetric around k, thus

$$P_0(W_{X,Y} < k) \leq \frac{1}{2} \qquad \text{and} \qquad P_0(W_{X,Y} \leq k) \geq \frac{1}{2}$$

$P_0(W_{X,Y} \leq k) - P_0(W_{X,Y} < k)$ is usually small and tends to zero as $mn \to \infty$.

$P_\Delta(\widehat{\Delta} < \Delta) = P_\Delta(\widehat{\Delta} \leq \Delta) \approx \frac{1}{2}$, i.e., close to $P_\Delta(\widehat{\Delta} > \Delta) \Rightarrow \quad \approx$ median unbiased.

Numerical Assessment of Median Unbiasedness

As an example consider $m = 8$ and $n = 9$ so that $mn = 72$ is even.

$$\texttt{pwilcox}(35, 8, 9) = 0.4813 \leq P_\Delta(\widehat{\Delta} < \Delta) \leq 0.5187 = \texttt{pwilcox}(36, 8, 9)$$

As another example consider $m = 20$ and $n = 21$ so that $mn = 420$

$$\texttt{pwilcox}(209, 20, 21) = 0.4949 \leq P_\Delta(\widehat{\Delta} < \Delta) \leq 0.5051 = \texttt{pwilcox}(210, 20, 21)$$

And finally $m = 100$ and $n = 101$ with $mn = 10100$

$$\texttt{pwilcox}(5049, 100, 101) = 0.4995 \leq P_\Delta(\widehat{\Delta} < \Delta) \leq 0.5005 = \texttt{pwilcox}(5050, 100, 101)$$

Dispersion of $\widehat{\Delta}$

The traditional measure of dispersion, $\mathrm{var}(\widehat{\Delta})$, is not easy to calculate.

We will examine the dispersion in terms of probability of closeness to Δ, i.e.,

$$P_\Delta(|\widehat{\Delta}-\Delta|\le a) = P_0(|\widehat{\Delta}|\le a) = P_0(-a\le\widehat{\Delta}\le a) = P_0(\widehat{\Delta}\le a) - P_0(\widehat{\Delta}<-a)$$

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

The switch from P_Δ to P_0 uses the fact that the distribution of $\widehat{\Delta}-\Delta$ under P_Δ does not depend on Δ. Thus we may as well just use $\Delta = 0$ for the remainder.

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

$p_1 = P(X-Y<-a) = P(X-Y>a) = P(X>Y+a) = 1-P(X<Y+a) = 1-\tilde{p}_1$

where the second $=$ uses the symmetry of the $X-Y$ distribution around 0 and the fourth $=$ uses the continuity of F.

$\mathrm{var}(W_{X,Y-a})$ and $\mathrm{var}(W_{X,Y+a})$

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

with $p_2 = P(X < Y-a \cap X < Y'-a)$ and $p_3 = P(X < Y-a \cap X' < Y-a)$

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

with $\tilde{p}_2 = P(X < Y+a \cap X < Y'+a)$ and $\tilde{p}_3 = P(X < Y+a \cap X' < Y+a)$

with X, Y, X', Y' independent $\sim F$.

F continuous $\Longrightarrow$ $\tilde{p}_2 - \tilde{p}_1^2 = p_3 - p_1^2$ and $\tilde{p}_3 - \tilde{p}_1^2 = p_2 - p_1^2$

Proof of $\tilde{p}_2 - \tilde{p}_1^2 = p_3 - p_1^2$

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

Here we used the continuity of F, the interchangeability of the independent random variables $X, X', Y, Y' \sim F$, $\quad 1 - P(A \cap B) = P([A \cap B]^c) = P(A^c \cup B^c) \quad$ and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$$m = n \implies \mathrm{var}(W_{X,Y-a}) = \mathrm{var}(W_{X,Y+a})$$

$$\begin{aligned}
\mathrm{var}(W_{X,Y-a}) &= m^2 \left[p_1(1-p_1) + (m-1)\left(p_2 - p_1^2\right) + (m-1)\left(p_3 - p_1^2\right) \right] \\
&= m^2 \left[(1-\tilde{p}_1)\tilde{p}_1 + (m-1)\left(\tilde{p}_3 - \tilde{p}_1^2\right) + (m-1)\left(\tilde{p}_2 - \tilde{p}_1^2\right) \right] \\
&= \mathrm{var}(W_{X,Y+a}) \qquad \square
\end{aligned}$$

F Symmetric around $\mu \implies p_2 = p_3$

F symmetric around μ means $P(X > \mu + b) = P(X < \mu - b)$ for all b.

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

$$\begin{aligned} p_2 &= P(X < Y - a \cap X < Y' - a) \\ &= P(-X < -Y - a \cap -X < -Y' - a) \\ &= P(Y < X - a \cap Y' < X - a) \\ &= P(X < Y - a \cap X' < Y - a) = p_3 \end{aligned} \qquad \square$$

$$\begin{aligned} \operatorname{var}(W_{X,Y-a}) &= mn\left[p_1(1-p_1) + (m+n-2)\left(p_2 - p_1^2\right)\right] \\ &= mn\left[(1-\tilde{p}_1)\tilde{p}_1 + (m+n-2)\left(\tilde{p}_3 - \tilde{p}_1^2\right)\right] \\ &= \operatorname{var}(W_{X,Y+a}) \end{aligned}$$

Proof of Approximation for $P_\Delta(|\widehat{\Delta}-\Delta| \le a)$

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

$$
\begin{aligned}
P_0(\widehat{\Delta} \le a) &= P_0(D_{(\ell)} \le a) = P_0(W_{X,Y-a} \le mn-\ell) = P_0(W_{X,Y-a} \le k) \\
&= P_0\left(\frac{W_{X,Y-a} - E(W_{X,Y-a})}{\sqrt{\mathrm{var}(W_{X,Y-a})}} \le \frac{k+\frac{1}{2} - E(W_{X,Y-a})}{\sqrt{\mathrm{var}(W_{X,Y-a})}}\right) \\
&\approx \Phi\left(\frac{mn\left(\frac{1}{2}-p_1\right)}{\sqrt{\mathrm{var}(W_{X,Y-a})}}\right)
\end{aligned}
$$

$$
\begin{aligned}
P_0(\widehat{\Delta} < -a) &= P_0(D_{(\ell)} \overset{<}{\le} -a) = P_0(W_{X,Y+a} \le mn-\ell) = P_0(W_{X,Y+a} \le k) \\
&= \Phi\left(\frac{mn\left(\frac{1}{2} - \overset{\tilde{p}_1}{(1-p_1)}\right)}{\sqrt{\mathrm{var}(W_{X,Y+a})}}\right) = 1-\Phi\left(\frac{mn\left(\frac{1}{2}-p_1\right)}{\sqrt{\mathrm{var}(W_{X,Y+a})}}\right) \qquad \square
\end{aligned}
$$

Proof of Approximation for $P_\Delta(|\widehat{\Delta}-\Delta| \le a)$

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

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

$$\implies \quad \widehat{\Delta} \le a \quad \implies \quad D_{(k)} \le a \iff W_{X,Y-a} \le mn-k = k$$

$$\Phi\left(\frac{k-\frac{1}{2}-mnp_1}{\sqrt{\mathrm{var}(W_{X,Y-a})}}\right) \approx P_0(W_{X,Y-a} \le k-1) \le P_0(\widehat{\Delta} \le a)$$

$$P_0(\widehat{\Delta} \le a) \le P_0(W_{X,Y-a} \le k) \approx \Phi\left(\frac{k+\frac{1}{2}-mnp_1}{\sqrt{\mathrm{var}(W_{X,Y-a})}}\right)$$

which suggests

$$P_0(\widehat{\Delta} \le a) \approx \Phi\left(\frac{k-mnp_1}{\sqrt{\mathrm{var}(W_{X,Y-a})}}\right) = \Phi\left(\frac{mn\left(\frac{1}{2}-p_1\right)}{\sqrt{\mathrm{var}(W_{X,Y-a})}}\right)$$

Proof of Approximation for $P_\Delta(|\widehat{\Delta}-\Delta| \le a)$

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

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

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

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

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

which suggests

$$P_0(\widehat{\Delta} < -a) \approx \Phi\left(\frac{k-mn\tilde{p}_1}{\sqrt{\text{var}(W_{X,Y+a})}}\right) = 1-\Phi\left(\frac{mn\left(\frac{1}{2}-p_1\right)}{\sqrt{\text{var}(W_{X,Y+a})}}\right) \qquad \square$$

Alternate Approximation for $P_\Delta(|\widehat{\Delta} - \Delta| \le a)$

As before, the approximation can be further simplified by approximating for small a

$$p_1 - \frac{1}{2} \approx -a f^*(0) \quad \text{and} \quad \operatorname{var}(W_{X,Y-a}) \approx \operatorname{var}(W_{X,Y+a}) \approx \operatorname{var}(W_{X,Y}) = \frac{mn(N+1)}{12}$$

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

$$\Longrightarrow \qquad P_\Delta(|\widehat{\Delta} - \Delta| \le a) \approx 2\Phi\left[\sqrt{\frac{12mn}{N+1}}\, f^*(0) a\right] - 1$$

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

$$P_\Delta(|\widehat{\Delta} - \Delta| \le a) \approx 2\Phi\left[\sqrt{\frac{3mn}{\pi(N+1)}}\, \frac{a}{\sigma}\right] - 1$$

Numerical Comparisons for $P_\Delta(|\widehat{\Delta} - \Delta| \leq a)$

For $m = n = 15$, $F = \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = 2$ the previous approximations give the results in the following table (from `Table2.5()`)

a	.2	.4	.6	.8	1.0	1.2
p_1	0.4602	0.4207	0.3821	0.3446	0.3085	0.2743
p_2	0.2944	0.2577	0.2235	0.1920	0.1633	0.1376
$E(W_{X,Y-a})$	103.54	94.67	85.97	77.53	69.42	61.71
$\text{var}(W_{X,Y-a})$	576.67	563.16	541.35	512.26	477.18	437.62
$P_\Delta(\lvert\widehat{\Delta} - \Delta\rvert \leq a)$						
approximation	0.2910	0.5476	0.7458	0.8777	0.9514	0.9848
alt. approximation	0.2903	0.5435	0.7360	0.8636	0.9373	0.9745
10^6 simulations	0.2925	0.5464	0.7389	0.8662	0.9390	0.9754

HLsim

```
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}
```

The advantage of simulation is unbiased estimation of $P_\Delta(|\widehat{\Delta}-\Delta| \leq a)$,

with any desired accuracy depending on N_{sim}.

Furthermore, we can easily handle other distributions for X and Y.

Consistency of $\widehat{\Delta}$

The error in the previous approximations tends to zero as $m, n \longrightarrow \infty$.

We saw that the approximation is already quite good for $m = n = 15$.

It can also be shown that for any $a > 0$ we have

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

Such estimators are called consistent estimators of Δ.

The sequence of estimators tends to the true value Δ in probability.

Comparison of $\widehat{\Delta}$ with $\bar{Y} - \bar{X}$

We assume the shift model: $X_i, Y_j - \Delta$ independent $\sim F$, $i = 1, \ldots, m,\ j = 1, \ldots, n$.

It can be shown that $\bar{\Delta} = \bar{Y} - \bar{X}$ also is a consistent estimator of Δ, provided F has a finite mean.

If $m = n$ or F is symmetric around some point μ, then the distribution of $\bar{\Delta}$ is symmetric around Δ.

In that case $\bar{\Delta}$ is unbiased and median unbiased, provided F has a finite mean.

Recall that $\widehat{\Delta}$ had the same unbiasedness properties under the same conditions.

$m \neq n$ & F Not Symmetric

In that case the distribution of $\bar{\Delta}$ is typically no longer symmetric around Δ.

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

But $\bar{\Delta}$ continues to be unbiased, i.e., $E(\bar{\Delta}) = \Delta$, provided the expectation exists.

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

Neither unbiasedness property is uniformly better than the other.

Efficiency of $\widehat{\Delta}$ Relative to $\bar{\Delta}$

The previous efficiency results of the Wilcoxon test relative to the t-test carry over.

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

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

We have

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

Thus the same efficiency results apply to these estimators,i.e.,

$e_{\widehat{\Delta},\bar{\Delta}}(\Phi) = 3/\pi = .955$ and $e_{\widehat{\Delta},\bar{\Delta}}(F) \ge .864$ for all F with finite variance.

Confidence Intervals

Rather than estimating Δ using a point estimate $\widehat{\Delta}$ or $\bar{\Delta}$ we will estimate Δ by using an interval, which will capture the true unknown Δ with a prescribed confidence probability $\gamma \in (0,1)$.

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

Do we already have such intervals with approximate confidence level γ through

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

which would give us such an interval via

$$P_\Delta\left(\widehat{\Delta}-a \le \Delta \le \widehat{\Delta}+a\right) \approx \gamma \qquad (?)$$

To determine a to get γ we need to know p_1, p_2 and thus need to know F, i.e., the interval would not be distribution-free.

Confidence Interval Building Blocks for Δ

From Theorem 4 we have

$$D_{(\ell)} \le \Delta \iff W_{X,Y-\Delta} \le mn-\ell \quad \text{and} \quad D_{(\ell)} > \Delta \iff W_{X,Y-\Delta} \ge mn-\ell+1$$

$$\text{or} \quad D_{(\ell+1)} > \Delta \iff W_{X,Y-\Delta} \ge mn-(\ell+1)+1 = mn-\ell$$

$$\Longrightarrow \qquad D_{(\ell)} \le \Delta < D_{(\ell+1)} \iff W_{X,Y-\Delta} = mn-\ell \qquad \text{for} \quad \ell = 0,1,\ldots,mn,$$

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).

$$P_\Delta\left(D_{(\ell)} \le \Delta < D_{(\ell+1)}\right) = P_0(W_{X,Y} = mn-\ell) = P_0(W_{X,Y} = \ell) \quad \ell = 0,1,\ldots,mn$$

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

Distribution-free Confidence Intervals for Δ

From the previous confidence interval building blocks we easily get

$$\begin{aligned}
P_\Delta(D_{(k)} \le \Delta \le D_{(\ell)}) &= P_\Delta(D_{(k)} < \Delta \le D_{(\ell)}) = P_\Delta(\Delta \le D_{(\ell)}) - P_\Delta(\Delta \le D_{(k)}) \\
&= P_0(W_{X,Y} \le \ell - 1) - P_0(W_{X,Y} \le k - 1) \\
&= P_0(k \le W_{X,Y} \le \ell - 1) \\
&= 1 - P_0(W_{X,Y} \le k - 1) - P_0(W_{X,Y} \ge \ell) \\
&=^* 1 - 2P_0(W_{X,Y} \le k - 1) \ge \gamma
\end{aligned}$$

where in $=^*$ we took $\ell = mn - (k-1) = mn - k + 1$ and exploited the symmetry of the $W_{X,Y}$ null distribution around $mn/2$.

To achieve a confidence level barely $\ge \gamma$ we want to find the largest k such that

$$P_0(W_{X,Y} \le k - 1) \le \frac{1-\gamma}{2}$$

Finding k for Target Confidence Level γ

Let

$$\texttt{k}_0 \;=\; \texttt{qwilcox}(\frac{1-\gamma}{2},\texttt{m},\texttt{n})$$

$$\texttt{k}_0 \;=\; \text{smallest } i \text{ such that } P_0(W_{X,Y} \le i) \ge \tfrac{1-\gamma}{2}$$

$$\iff \quad \texttt{pwilcox}(\texttt{k}_0-1,\texttt{m},\texttt{n}) < \frac{1-\gamma}{2} \quad \text{and} \quad \texttt{pwilcox}(\texttt{k}_0,\texttt{m},\texttt{n}) \ge \frac{1-\gamma}{2}$$

$$\text{If} \quad \texttt{pwilcox}(\texttt{k}_0,\texttt{m},\texttt{n}) \;>\; \frac{1-\gamma}{2} \quad \text{then} \quad \texttt{k}=\texttt{k}_0 \quad \text{is it.}$$

$$\text{If} \quad \texttt{pwilcox}(\texttt{k}_0,\texttt{m},\texttt{n}) \;=\; \frac{1-\gamma}{2} \quad \text{then} \quad \texttt{k}=\texttt{k}_0+1 \quad \text{is it.}$$

With that choice of k we have that

$$P_\Delta(D_{(k)} \le \Delta \le D_{(mn-k+1)}) = 1-2*\texttt{pwilcox}(\texttt{k}-1,\texttt{m},\texttt{n}) = \gamma_k \ge \gamma$$

a distribution-free confidence interval for Δ with achieved confidence level γ_c.

Example 5: Augmenters and Reducers

The Text describes an interesting pain sensory classification developed by Petrie.

Augmenters (Reducers) were tested for pain reaction without & with an analgesic.

For Augmenters the difference in pain scores is Y_j, for Reducers it is X_i.

	Differences in Pain Scores: Without − With Analgesic									
Augmenters	17.9	13.3	10.6	7.6	5.7	5.6	5.4	3.3	3.1	.9
Reducers	7.7	5.0	1.7	0.0	-3.0	-3.1	-10.5			

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 Δ.

Matrix `Dmat` of $D_{ij} = Y_j - X_i$

	Y_j									
	10.2	5.6	2.9	-0.1	-2.0	-2.1	-2.3	-4.4	-4.6	-6.8
	12.9	8.3	5.6	2.6	0.7	0.6	0.4	-1.7	-1.9	-4.1
	16.2	11.6	8.9	5.9	4.0	3.9	3.7	1.6	1.4	-0.8
X_i	17.9	13.3	10.6	7.6	5.7	5.6	5.4	3.3	3.1	0.9
	20.9	16.3	13.6	10.6	8.7	8.6	8.4	6.3	6.1	3.9
	21.0	16.4	13.7	10.7	8.8	8.7	8.5	6.4	6.2	4.0
	28.4	23.8	21.1	18.1	16.2	16.1	15.9	13.8	13.6	11.4

```
function(){
   Augmenters=c(17.9, 13.3, 10.6, 7.6, 5.7, 5.6, 5.4, 3.3, 3.1, .9)
   Reducers=c(7.7, 5.0, 1.7, 0.0, -3.0, -3.1, -10.5)
   Dmat=t(outer(Augmenters,Reducers,"-"))
   Dmat
}
```

Sorted Matrix `Dmat.sort` of $D_{(\ell)}$

```
Dmat.sort=matrix(sort(Dmat),byrow=T,ncol=10)
```

gives the sorted values $D_{(\ell)}$ in a 7×10 array

Sorted $D_{(\ell)}$

-6.8	-4.6	-4.4	-4.1	-2.3	-2.1	-2.0	-1.9	-1.7	-0.8
-0.1	0.4	0.6	0.7	0.9	1.4	1.6	2.6	2.9	3.1
3.3	3.7	3.9	3.9	4.0	4.0	5.4	5.6	5.6	5.6
5.7	5.9	6.1	6.2	6.3	6.4	7.6	8.3	8.4	8.5
8.6	8.7	8.7	8.8	8.9	10.2	10.6	10.6	10.7	11.4
11.6	12.9	13.3	13.6	13.6	13.7	13.8	15.9	16.1	16.2
16.2	16.3	16.4	17.9	18.1	20.9	21.0	21.1	23.8	28.4

Example 5: A 90% Confidence Interval

We want a 90% confidence interval for Δ.

$\texttt{qwilcox}(.05,7,10) = 18$ and $\texttt{pwilcox}(18,7,10) = 0.05440148 > .05$

and `pwilcox(17,7,10)=0.04391197`

and thus $\gamma_k = 1 - 2 * \texttt{pwilcox}(17,7,10) = 0.912176.$

Hence

$$[D_{(18)}, D_{(70-18+1)}] = [D_{(18)}, D_{(53)}] = [2.6,\ 13.3]$$

is our desired confidence interval with achieved confidence level $\gamma_k = .912$.

Upper or Lower Confidence Bounds

Sometimes upper or lower confidence bounds for Δ are more appropriate, i.e.,

$$P_\Delta(\widehat{\Delta}_L \leq \Delta) = \gamma \qquad \text{or} \qquad P_\Delta(\widehat{\Delta}_U \geq \Delta) = \gamma\,.$$

Because we split the miss-probability ($2 \times \frac{1-\gamma}{2} = 1-\gamma$) of our confidence intervals equally, we can view the left interval endpoint $D_{(k)}$ as lower bound with miss-probability $\frac{1-\gamma}{2}$ and thus with confidence $1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$.

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

For the actually achieved confidence level replace γ by γ_k.

In our example we can thus view $D_{(53)} = 13.3$ as 95% upper bound for Δ, with achieved confidence level $\gamma_k = .956$.

$D_{(18)} = 2.6$ is a 95% lower bound for Δ, with achieved confidence level $\gamma_k = .956$.

Ties and Rounding

So far it was assumed that there are no ties.

This will happen with probability one when we assume that F is continuous.

The continuity of F is reasonable for many measurement phenomena and will typically be realized under sufficient measurement precision.

However, there is rounding.

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

$$|Y_j' - Y_j - (X_i' - X_i)| = |(Y_j' - X_i') - (Y_j - X_i)| = |D_{i,j}' - D_{i,j}| \leq \varepsilon \implies |D_{(i)}' - D_{(i)}| \leq \varepsilon$$

Confidence Bounds with Rounding

Assume that the $X'_i \sim F$ and $Y'_j \sim F(x-\Delta)$ are independent, with F continuous, and that $D'_{(\ell)}$ is the ℓ^{th} ordered value of $D'_{i,j} = Y'_j - X'_i$.

$D_{(\ell)}$ be the counterpart using the rounded values X_i, Y_j, $i = 1,\ldots,m,\ j = 1,\ldots,n$.

If $\widehat{\Delta}'_L = D'_{(\ell)}$ is a (valid) lower bound for Δ with confidene level γ, then $\widehat{\Delta}_L = D_{(\ell)} - \varepsilon$ is a conservative lower confidence bound for Δ with confidence level $\geq \gamma$.

If $\widehat{\Delta}'_U = D'_{(u)}$ is an upper bound for Δ with confidene level γ, then $\widehat{\Delta}_U = D_{(u)} + \varepsilon$ is a conservative upper confidence bound for Δ with confidence level $\geq \gamma$.

If $[D'_{(\ell)}, D'_{(u)}]$ is a confidence interval for Δ with confidence coefficient γ, then $[D_{(\ell)} - \varepsilon, D_{(u)} + \varepsilon]$ is a conservative confidence interval for Δ with confidence level $\geq \gamma$.

Using `wilcox.test` or `wilcox_test`

```
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)}
```

It seems that `wilcox.test` and `wilcox_test` give intervals with confidence $\geq \gamma = .95$, where $\texttt{gam} = \gamma$ 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 `DConfInt(gam=`γ`)` for Various γ

γ	$D_{(\ell)}$	$D_{(u)}$
.944	1.4	13.6
.945	.9	13.7
.950	.9	13.7
.956	.9	13.7
.957	.7	13.8

Normal Scores

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 will look more crowded in the center and sparser out in the tails.

The normal scores try to make the transformed ranks look like normal observations, i.e., 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 $$a_N(s) = E_\Phi\left(Z_{(s)}\right) , \qquad \text{for } s = 1, 2, \ldots, N$$

The Normal Scores Test

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 ranks of the treatment sample of Y's, while $R_1, \ldots, R_m$ are again the ranks of the control sample of X's.

Depending on the anticipated direction of the treatment effect under the alternative to the null hypothesis H_0 of no treatment effect, we would reject H_0 when T_s is too large, or when it is too small, or, for two-sided alternatives, when either situation occurs.

The main problem with this test is the evaluation of the scores, which require N numerical integrations.

Null Distribution of the Normal Scores Test

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 $\binom{N}{n}$ rank-sums again represents the full null distribution of T_s and can be used to compute any p-values, simply by computing the proportion of these $\binom{N}{n}$ rank-sums T_s that are $\geq$ (or $\leq$) to the observed value $T_{s,\text{obs.}}$ of T_s.

For a two-sided test we would compute the proportion of $|T_s| \geq |T_{s,\text{obs.}}|$.

The null distribution of T_s is symmetric around zero.

Efficiency of the Normal Scores Test

Just as we compared the Wilcoxon test quite favorably with the two-sample t-test, we can do the same with respect to the normal scores test (N) and the t-test.

Here it can be shown that the ARE

$$e_{N,\,t}(F) \geq 1 \qquad \text{for all } F \text{ and with equality when } F = \Phi.$$

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

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

denoting the transformed rank sum again by T_S.

These scores are much easier to compute in R via $a_N(s) = \texttt{qnorm(s/(N+1))}$.

The exact null distribution can now be easily obtained via `combn` when m and n are not too large.

The above efficiciency result holds for this test as well.