

Stat 425  
Introduction to Nonparametric Statistics  
Rank Tests for Comparing Two Treatments

Fritz Scholz

Department of Statistics, University of Washington

May 20, 2013

- Erich L. Lehmann,  
Nonparametrics, Statistical Methods Based on Ranks,  
Springer Verlag 2008 (required)
- W. John Braun and Duncan J. Murdoch,  
A First Course in STATISTICAL PROGRAMMING WITH R,  
Cambridge University Press 2007 (optional).
- Joris Meys and Andrie de Vries,  
R for Dummies,  
John Wiley & Sons 2012, Chichester, West Sussex, England  
(optional).
- For other introductory material on R see the class web page.  
[http://www.stat.washington.edu/fritz/Stat425\\_Spring2013R.html](http://www.stat.washington.edu/fritz/Stat425_Spring2013R.html)

# Comparing Two Treatments

- Often it is desired to understand whether an innovation constitutes an improvement over current methods/treatments.
- Application Areas: New medical drug treatment, surgical procedures, industrial process change, teaching method, ...
- Example 1 New Drug: A mental hospital wants to investigate the (beneficial?) effect of some new drug on a particular type of mental or emotional disorder.
- We discuss this in the context of five patients (suffering from the disorder to roughly the same degree), to simplify the logistics of getting the basic ideas out in the open.
- Three patients are **randomly** assigned the new treatment while the other two get a placebo pill (also called the “control treatment”).

# Precautions: Triple Blind!

- The assignments of new drug and placebo should be blind to the patient to avoid “placebo effects”.
- It should be blind to the staff of the mental hospital to avoid conscious or subconscious staff treatment alignment with either one of these treatments.
- It should also be blind to the physician who evaluates the patients after a few weeks on this treatment.
- A placebo effect occurs when a patient thinks he/she is on the beneficial drug, shows some beneficial effect, but in reality is on a placebo pill.
- The power of positive thinking is worth something.

# Evaluation by Ranking

- The physician (an outside observer) evaluates the condition of each patient, giving rank 1 to the patient with the most severe level of the studied disorder, rank 2 to the patient with next most severe level, . . . , and rank 5 to the patient with mildest disorder level.
- Ranking is usually a lot easier than objectively measuring such disorder levels.
- However, it may become problematic for a large number of patients, possibly requiring several ranking iterations  
coarse ranking  $\Rightarrow$  refined ranking.

# How to Judge the Ranking Results?

- If the patients on the new drug all receive high ranks we might consider that to be evidence in favor of the new drug.
- How do we judge the strength of that evidence?
- As benchmark we will consider the null hypothesis  $H_0$  that the new drug is acting just like a placebo, i.e., has no effect or just a placebo effect w.r.t. the disorder condition. In that case there is no difference between placebo and new treatment.
- The rankings would have been the same, no matter which way the new drug and the placebo were assigned, i.e., at study begin or after ranking.

## Possible Ranking Results?

Treated	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
Controls	(1,2)	(1,3)	(2,3)	(1,4)	(2,4)
Treated	(1,2,5)	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)
Controls	(3,4)	(1,5)	(2,5)	(3,5)	(4,5)

- 10 splits of ranks between treatment and control group.
- Under  $H_0$  the rankings would have been the same regardless of treatment/control.
- Under  $H_0$  our randomization makes all 10 splits of ranks equally likely, probability 1/10.
- This can be used to assess the statistical significance (rarity) of any observed result, when trying to judge the effectivity of the new treatment as opposed to just the luck of the draw.

## More Study Subjects

- For ease of exposition we dealt with  $N = 5$  subjects, split into  $n = 3$  with treatment and  $m = N - n = 2$  with placebo.
- What do we get for larger  $N$  and different  $m, n$  with  $N = m + n$ ?
- The number of ways of selecting a group of  $n$  (without regard to order) from  $N$  is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N(N-1) \cdot \dots \cdot (N-n+1)}{1 \cdot 2 \cdot \dots \cdot n} \quad \Rightarrow \quad \binom{5}{3} = \frac{5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3} = 10$$

- This *binomial coefficient* is also referred to as *the number of combinations of  $N$  things, taken  $n$  at a time*.
- The text gives Table A for a range of  $m$  and  $n$ . In R we get

$$\binom{N}{n} = \text{choose}(N, n)$$

for a much wider range of  $n$  and  $m = N - n$ , e.g.,  
 $\text{choose}(50, 25) = 1.264106e+14$ .

# The Distribution of Ranks

Under  $H_0$  each set of ordered ranks  $S_1 < S_2 < \dots < S_n$  associated with the treatment group has the same chance  $1/\binom{N}{n}$ , i.e.,

$$P(S_1 = s_1, \dots, S_n = s_n) = \frac{1}{\binom{N}{n}}$$

for each of the possible ordered  $n$ -tuples  $(s_1, \dots, s_n)$ .

## Example 2: *Effect of Discouragement*

- 10 subjects were split up **randomly** into two groups of 5 each.
- All 10 subjects were given form L of the revised Stanford-Binet (IQ) test, under the conditions prescribed by this test.
- Two weeks later they were given form F of this test,
  - the controls under the usual conditions
  - the 'treatment' group receiving in addition some prior discouraging comments concerning their previous performance.
- The prior test scores were not disclosed to the subjects.
- The following differences (second test score – first test score) were observed

Controls : 5 0 16 2 9      Treated : 6 -5 -6 1 4

# Ranks of Test Score Difference

- Giving rank 1 to the subject with smallest difference in scores, rank 2 to the second smallest difference, etc., the ranks in the treatment group were 1 2 4 6 8 while the control group had ranks 3 5 7 9 10.
- The ranks as well as the original test score differences suggest: the discouragement treatment has some detrimental effect.
- Its statistical significance will be assessed later.
- Again we use as benchmark the null hypothesis  $H_0$  under which the discouragement has no effect at all, i.e., the test score differences would be the same, with or without discouragement.
- In that case all  $\binom{10}{5} = 252$  splits of ranks into treatment and control groups would be equally likely, with chance  $1/252$  each.

## Comparing Examples 1 and 2

- Both examples are similar, resulting in ranks for treatment and control groups.
- Under the hypothesis  $H_0$  of no treatment effect all possible splits of the ranks into treatment and control group are equally likely.
- Yet, some may argue that we should be using the original test score differences, since by ranking them we lose some detail/information.
- We will return to this issue later.

# The Role of Randomness

- In both discussed examples we dealt with fixed subjects, they were not randomly chosen from some population.
- Randomness entered through our deliberate action of randomizing treatment assignments.
- Under  $H_0$  this randomness is very well understood and does not involve unknown parameters (nonparametric).
- The distribution of the ranks is known.
- In the case of test score differences it is conceivable to view them as a random sample from some (hypothetical) population.
- In the case of ranking emotional disorder it does not appear feasible to view it in the context of a population model, without absolute objective scores (not just rankings within the 5 subjects).

# The Randomization Model

- The imposed or induced randomization in the previous two examples gives us a known statistical model under  $H_0$ .
- We call it the **randomization model**.
- This is in contrast to **population models**, where subjects are drawn at random from a population of such subjects.
- In that case any conclusions drawn reflect back on that **assumed** sampled distribution.
- In the case of the randomization model any such conclusion can logically only pertain to the much smaller “universe” of randomized subjects.
- Generalizing from that to general populations needs to be weighed carefully.

## Some Useful R Function

```
choose(8,4) # gives us the number of combinations
            # of 8 taking 4 at a time
[1] 70 # famous from Fisher's tea tasting example
```

```
sample(1:8,4,replace=F) # gives us a random sample
                        # of 4 indices from 1:8
                        # without replacement
[1] 4 7 8 5
```

```
sort(c(10,2,3,2,5,8,7,9,10)) # sorts input vector
[1] 2 2 3 5 7 8 9 10 10
```

```
rank(c(10,2,3,2,5,8,7,9,10))
  # ranks observations in the input vector
  # using midranks
[1] 8.5 1.5 3.0 1.5 4.0 6.0 5.0 7.0 8.5
```

# Difficulties in Comparing Sets of Ranks

- Having obtained a set of ranks for the treatment group and assuming that a treatment effect would result in higher subject rankings, we need a criterion that allows us to judge whether a set of treatment ranks  $S_1 < \dots < S_n$  is generally higher than the set of control group ranks  $R_1 < \dots < R_m$ .
- Comparing such vectors  $(S_1, \dots, S_n)$  and  $(R_1, \dots, R_m)$  faces two difficulties:
  - 1) the vectors may have different lengths, i.e.,  $n \neq m$ , and
  - 2) vectors have several components, which may be compared on a component by component basis (if  $n = m$ ), but different order relations may result for the various rank coordinates.
- There are many ways of dealing with these issues, but the simplest one seems to sort such rank vectors by their rank-sum

$$W_s = S_1 + \dots + S_n$$

rejecting  $H_0$  whenever  $W_s$  is sufficiently large, say,  $W_s \geq c$ .

# The Wilcoxon Rank-Sum Test

- The test defined by

$$W_s = S_1 + \dots + S_n \quad \text{and rejecting } H_0 \text{ whenever } W_s \geq c$$

is called the *Wilcoxon rank-sum test*.

- This term distinguishes it from another Wilcoxon test to be discussed later.
- When there is no confusion we omit the qualifier *rank-sum*.
- The constant  $c$ , the **critical value**, is chosen such that  $P_{H_0}(W_s \geq c) = \alpha$ .
- The **significance level**  $\alpha$  is some specified small number.
- Equivalently one could reject  $H_0$  when the control rank sum  $W_r = R_1 + \dots + R_m$  is too small.  $W_r + W_s = N(N + 1)/2$ .

# Determining the Critical Values

Treatment Ranks	(3,4,5)	(2,4,5)	(1,4,5)	(2,3,5)	(1,3,5)
$w$	12	11	10	10	9
Treatment Ranks	(1,2,5)	(2,3,4)	(1,3,4)	(1,2,4)	(1,2,3)
$w$	8	9	8	7	6

$w$	6	7	8	9	10	11	12
$P_{H_0}(W_s = w)$	.1	.1	.2	.2	.2	.1	.1

$c$	6	7	8	9	10	11	12	13
$P_{H_0}(W_s \geq c)$	1.0	.9	.8	.6	.4	.2	.1	0

# Limited Set of Significance Levels

- As we saw, the number of possible significance levels is limited.
- This persists to a lesser degree even for larger  $N$ .
- We can compromise and choose one of the possible significance levels near the desired  $\alpha$
- or we can report the **p-value** or **significance probability**, i.e.,  $p(w_{\text{obs}}) = P_{H_0}(W_s \geq w_{\text{obs}})$ , where  $w_{\text{obs}}$  is the actually observed value of  $W_s$ .
- The latter is preferable.
- It expresses more clearly the strength of evidence against  $H_0$  offered by observed value of  $W_s$ .
- We reject  $H_0$  at level  $\alpha \iff p\text{-value} = p(w_{\text{obs}}) \leq \alpha$ .

Reject  $H_0$  at level  $\alpha \iff p(w_{\text{obs}}) \leq \alpha$

- For a level  $\alpha$  test we choose the smallest critical point  $c$  such that  $P_{H_0}(W_s \geq c) \leq \alpha$ .
- Denote this  $c$  by  $c_\alpha$  with corresponding type I error probability

$$P_{H_0}(W_s \geq c_\alpha) = \alpha_{c_\alpha} \leq \alpha .$$

- Note that by definition we have  $P_{H_0}(W_s \geq c_\alpha - 1) > \alpha$  (\*)
- For any observed value  $w_{\text{obs}} \geq c_\alpha$  we reject  $H_0$  at level  $\alpha$  and

$$p(w_{\text{obs}}) = P_{H_0}(W_s \geq w_{\text{obs}}) \leq P_{H_0}(W_s \geq c_\alpha) = \alpha_{c_\alpha} \leq \alpha .$$

- Conversely,  $p(w_{\text{obs}}) \leq \alpha \implies w_{\text{obs}} \geq c_\alpha$ , i.e., reject  $H_0$
- If  $w_{\text{obs}} < c_\alpha$ , i.e.,  $w_{\text{obs}} \leq c_\alpha - 1$  ( $w_{\text{obs}}$  and  $c_\alpha$  are integers), we would then have  $p(w_{\text{obs}}) = P_{H_0}(W_s \geq w_{\text{obs}}) > \alpha$  (see (\*))  $\implies$  a contradiction.

# Significance Levels for $N = 13$ & $n = 8$

The maximum  $W_s$  is  $13 + 12 + 11 + 10 + 9 + 8 + 7 + 6 = 76$

the minimum is  $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 8 \times 9/2 = 36$

$c$	$P_{H_0}(W_s \geq c)$	$c$	$P_{H_0}(W_s \geq c)$	$c$	$P_{H_0}(W_s \geq c)$
77	0	63	0.17716	49	0.85781
76	0.00078	62	0.21756	48	0.88889
75	0.00155	61	0.26185	47	0.91453
74	0.00311	60	0.31080	46	0.93629
73	0.00544	59	0.36208	45	0.95338
72	0.00932	58	0.41647	44	0.96737
71	0.01476	57	0.47164	43	0.97747
70	0.02253	56	0.52836	42	0.98524
69	0.03263	55	0.58353	41	0.99068
68	0.04662	54	0.63792	40	0.99456
67	0.06371	53	0.68920	39	0.99689
66	0.08547	52	0.73815	38	0.99845
65	0.11111	51	0.78244	37	0.99922
64	0.14219	50	0.82284	36	1

# R Code for Previous Table

```
Wilcoxonsig.levels <- function (N = 5, n = 3)
{
  out <- combn(1:N, n, FUN = sum)
  Com <- choose(N, n)
  out <- sort(out)
  c.unique <- rev(unique(out))
  cx <- c(max(c.unique) + 1, c.unique)
  k <- length(cx)
  alpha <- numeric(k)
  i <- 0
  for (ci in cx) {
    i <- i + 1
    alpha[i] <- sum(out >= ci)/Com
  }
  cbind(cx, round(alpha,5))
}
```

- Note the use of the function `combn`.
- It evaluates the sum of all combinations of  $N$  taken  $n$  at a time, i.e., it evaluates the rank-sum for all possible splits.

# `combn(5, 3) = combn(5, 3, FUN=NULL)`

```
> combn(5, 3)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]     1     1     1     1     1     1     2     2     2     3
[2,]     2     2     2     3     3     4     3     3     4     4
[3,]     3     4     5     4     5     5     4     5     5     5
```

The columns are all combinations of  $N = 5$  taken  $n = 3$  at a time.

```
> combn(5, 3, FUN=sum)
[1]  6  7  8  8  9 10  9 10 11 12
```

This gives the corresponding sums for each column.

# Tabulation of Rank-Sum Distribution

- The Text gives Table B for  $P_{H_0}(W_{XY} \leq a)$
- $W_{XY} = W_s - n(n+1)/2$ .
- Table B covers  $\{m = 3, 4, m \leq n \leq 12\}$  and  $\{5 \leq m \leq 10, m \leq n \leq 10\}$ .
- The preference for tabulating the null distribution of  $W_{XY}$  instead of  $W_s$  is based on the property that the null distribution of  $W_{XY}$  is the same for  $(n, m) = (k_1, k_2)$  as for  $(n, m) = (k_2, k_1)$ .  $W_{XY} = 0, 1, \dots, mn$ .
- The choice of the symbol  $W_{XY}$  for  $W_s - n(n+1)/2$  will become clear when we discuss the Mann-Whitney test.
- It also makes Table B more compact, since the smallest value of  $W_s$  is  $1 + 2 + \dots + n = n(n+1)/2$ , which can be quite large.

- R gives the  $W_{XY}$  distribution for a wider range of  $n, m$  values.
- `dwilcox(x, m, n, log = FALSE)` =  $P_{H_0}(W_{XY} = x)$
- `pwilcox(q, m, n)` =  $P_{H_0}(W_{XY} \leq q)$
- `qwilcox(p, m, n)` =  $p$ -quantile of  $W_{XY}$   
=  $\min \{x : P_{H_0}(W_{XY} \leq x) \geq p\}$ .
- `rwilcox(nn, m, n)` random sample of size `nn` from the  $W_{XY}$  null distribution.

## Warning

These functions can use large amounts of memory and stack (and even crash R if the stack limit is exceeded and stack-checking is not in place) if one sample is large (several thousands or more).

## Symmetry of Null Distribution of $W_s$

- The null distribution of  $W_s$  is symmetric around  $n(N+1)/2$ , i.e., the null dist. of  $W_s - n(N+1)/2$  is symmetric around 0

$$P(W_s = n(N+1)/2 + a) = P(W_s = n(N+1)/2 - a)$$

$$\text{or } P(W_s - n(N+1)/2 = a) = P(W_s - n(N+1)/2 = -a)$$

- To see this consider ranking all subjects in reverse order, rank 1 becomes rank  $N$ , rank 2 becomes rank  $N-1$ , etc.
- Denote these reverse ranks by  $S'_1, \dots, S'_n$  with  $S'_j = N+1 - S_j$ .
- Since  $P(S'_1 = s_1, \dots, S'_n = s_n) = 1/\binom{N}{n}$ , the rank-sum

$$\begin{aligned} W_{s'} &= S'_1 + \dots + S'_n = [(N+1) - S_1] + \dots + [(N+1) - S_n] \\ &= n(N+1) - W_s \end{aligned}$$

has the same null distribution as  $W_s$ .

- Subtracting  $n(N+1)/2$  on both sides we get

$$W_s - n(N+1)/2 \stackrel{D}{=} W_{s'} - n(N+1)/2 = n(N+1)/2 - W_s$$

# The Mann-Whitney Test Statistic

- Suppose we have scores  $X_1, \dots, X_m$  for the control group and scores  $Y_1, \dots, Y_n$  for the treatment group.
- Assume that all scores are different.
- Define the Mann-Whitney statistics

$$\widetilde{W}_{XY} = \text{number of pairs } (X_i, Y_j) \text{ with } X_i < Y_j$$

$$\widetilde{W}_{YX} = \text{number of pairs } (X_i, Y_j) \text{ with } Y_j < X_i.$$

$$W_s - n(n+1)/2 = W_{XY}$$

- Let  $Y_{(1)} < \dots < Y_{(n)}$  be the order statistics of  $Y_1, \dots, Y_n$  with corresponding ranks  $S_1 < \dots < S_n$ .
- Since  $Y_{(1)}$  has rank  $S_1$ , there are  $(S_1 - 1)$   $X$ 's  $< Y_{(1)}$
- We have  $(S_2 - 2)$   $X$ 's  $< Y_{(2)}$ , since  $Y_{(1)} < Y_{(2)}$
- $(S_n - n)$   $X$ 's  $< Y_{(n)}$ , since  $Y_{(1)} < \dots < Y_{(n-1)} < Y_{(n)}$

$$\begin{aligned}\widetilde{W}_{XY} &= (S_1 - 1) + (S_2 - 2) + \dots + (S_n - n) \\ &= W_s - (1 + 2 + \dots + n) = W_s - n(n+1)/2 = W_{XY}\end{aligned}$$

The Wilcoxon and the Mann-Whitney statistics are equivalent (differ by  $n(n+1)/2$ ). Use notation  $W_{XY}$  instead of  $\widetilde{W}_{XY}$ .

## $W_{XY}$ and $W_{YX}$ Have the Same Null Distribution

$$W_s + W_r = 1 + 2 + \dots + N = \frac{N(N+1)}{2} = \frac{(N+1)(m+n)}{2}$$

$$\implies W_r - \frac{m(N+1)}{2} = \frac{n(N+1)}{2} - W_s \stackrel{\mathcal{D}}{=} W_s - \frac{n(N+1)}{2}$$

$$\begin{aligned} \implies W_r - \frac{m(m+1)}{2} &= W_r - \frac{m(N+1)}{2} + \frac{mn}{2} \\ &\stackrel{\mathcal{D}}{=} W_s - \frac{n(N+1)}{2} + \frac{mn}{2} = W_s - \frac{n(n+1)}{2} \\ W_{YX} &\stackrel{\mathcal{D}}{=} W_{XY} \end{aligned}$$

# R Function `wilcox.test`

```
>xw
```

```
9.56 10.07 10.10 10.14 10.33 10.81 10.84 11.14 11.31 11.42 11.61 12.41
```

```
>yw
```

```
11.30 12.02 12.25 12.71 12.81 12.84 13.31 13.76 13.85 14.12
```

```
> wilcox.test(yw,xw,alternative="greater")
```

Wilcoxon rank sum test

data: yw and xw

W = 114, p-value = 4.639e-05

alternative hypothesis: true location shift is greater than 0

- Here we test the hypothesis,  $H_0$  : no effect, against the alternative that the first sample  $y_w$  tends to have somewhat higher scores than the second sample  $x_w$ .
- This function works for observations without ties and gives exact p-values when sample sizes are less than 50.

# Asymptotic Null Distribution of $W_s$

- According to the *central limit theorem (CLT)* a sum  $T$  of a large number of independent random variables (subject to mild regularity conditions) is approximately normally distributed, i.e.,  $T \approx \mathcal{N}(E(T), \text{var}(T))$  or

$$P\left(\frac{T - E(T)}{\sqrt{\text{var}(T)}} \leq b\right) \approx \Phi(b) = \text{pnorm}(b),$$

where  $\Phi(b)$  denotes the area to the left of  $b$  under a standard normal density.

- For large  $n$  and  $m$  the rank-sum  $W_s = S_1 + \dots + S_n$  can be viewed as the sum of many (only weakly dependent) random variables.
- The reason for insisting on a large  $n$  and  $m$  is that  $W_s + W_r = N(N + 1)/2$ , i.e., if one is normal, so is the other.

# CLT for Sampling without Replacement

- When taking  $Y_1, \dots, Y_n$  **randomly** and **without replacement** from a finite population  $\{Z_1, \dots, Z_N\}$ , then sampling theory  $\Rightarrow$

$$E(Y_i) = \mu_Z = \bar{Z} \quad \text{and} \quad \text{var}(Y_i) = \sigma_Z^2 = \frac{1}{N} \sum_{j=1}^N (Z_j - \bar{Z})^2$$

$$T_n = \sum_{i=1}^n Y_i \Rightarrow E(T_n) = n \mu_Z \quad \text{and} \quad \text{var}(T_n) = n \sigma_Z^2 \frac{N-n}{N-1}$$

$\kappa = (N-n)/(N-1) =$  **finite population correction factor**.

- $n = 1$  reduces these expressions to the previous ones.
- $\kappa_N \rightarrow 1$  as  $N \rightarrow \infty$  and  $\text{var}(T_n) \rightarrow$  variance under sampling with replacement.
- Further CLT theory  $\Rightarrow T_n \approx \mathcal{N}(E(T_n), \text{var}(T_n))$  as  $n \rightarrow \infty$ .

## Null Distribution Mean and Variance of $W_s$

In view of the previous slide we can view  $W_s$  as the sum of  $n$  integers taken randomly and without replacement from  $\{Z_1, \dots, Z_N\} = \{1, 2, \dots, N\}$ .

Thus  $\mu_Z = \bar{Z} = (N + 1)/2$  and

$$\begin{aligned}\sigma_Z^2 &= \mu_{Z^2} - (\mu_Z)^2 = \frac{1}{N} \sum_{j=1}^N j^2 - \bar{Z}^2 \\ &= \frac{1}{N} \frac{N(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{(N-1)(N+1)}{12} \\ &\implies E(W_s) = \frac{n(N+1)}{2}\end{aligned}$$

and

$$\text{var}(W_s) = \frac{n(N-n)(N-1)(N+1)}{12(N-1)} = \frac{mn(N+1)}{12}$$

Since  $W_s + W_r = N(N + 1)/2$  and  $W_{XY} = W_s - n(n + 1)/2$  we get immediately

$$E(W_r) = \frac{m(N + 1)}{2} \quad \text{and} \quad E(W_{XY}) = E(W_{YX}) = \frac{mn}{2}$$

while

$$\text{var}(W_r) = \text{var}(W_{XY}) = \text{var}(W_{YX}) = \frac{mn(N + 1)}{12}$$

- These variances all coincide with  $\text{var}(W_s)$ , because the respective random sums differ by constants from each other.
- $W_s$ ,  $W_r$ ,  $W_{XY}$  and  $W_{YX}$  are all approximately normally distributed with respective means and variances.

# Normal Approximation

- The distribution of  $W_{XY}$ :  $P(W_{XY} = a)$ ,  $a = 0, 1, \dots, nm$ .
- These probabilities can be represented by the **heights**  $P(W_{XY} = a)$  of vertical rods at  $a = 0, 1, \dots, nm$  (see dashed lines with dots at top, on Slide ??)
- Or they can be represented by **areas** of the rectangles, centered at  $a$ , with width one and height  $P(W_{XY} = a)$  for  $a = 0, 1, 2, \dots, nm$ .
- The straight normal approximation uses

$$\begin{aligned} P(W_{XY} \leq a) &= P\left(\frac{W_{XY} - mn/2}{\sqrt{mn(N+1)/12}} \leq \frac{a - mn/2}{\sqrt{mn(N+1)/12}}\right) \\ &\approx \Phi\left(\frac{a - mn/2}{\sqrt{mn(N+1)/12}}\right) \end{aligned}$$

# Normal Approximation with Continuity Correction

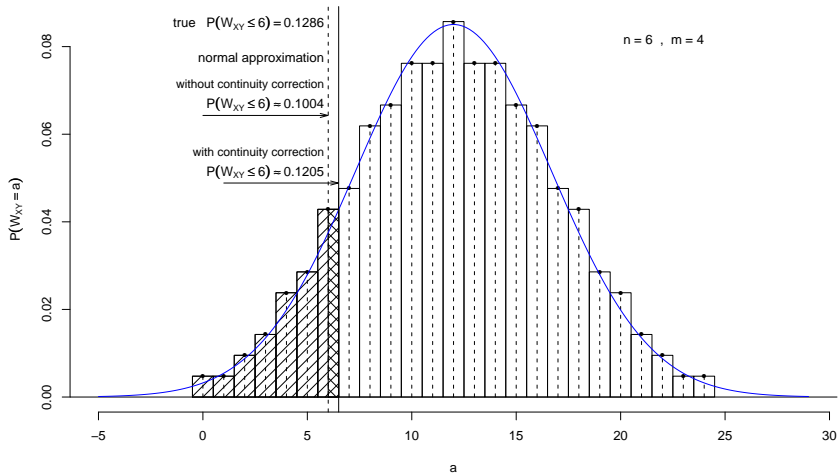
The **continuity correction** matches box areas with corresponding normal curve areas

$$P(W_{XY} \leq a) = P(W_{XY} \leq a + 1/2) \approx \Phi \left( \frac{a + 1/2 - mn/2}{\sqrt{mn(N+1)/12}} \right)$$

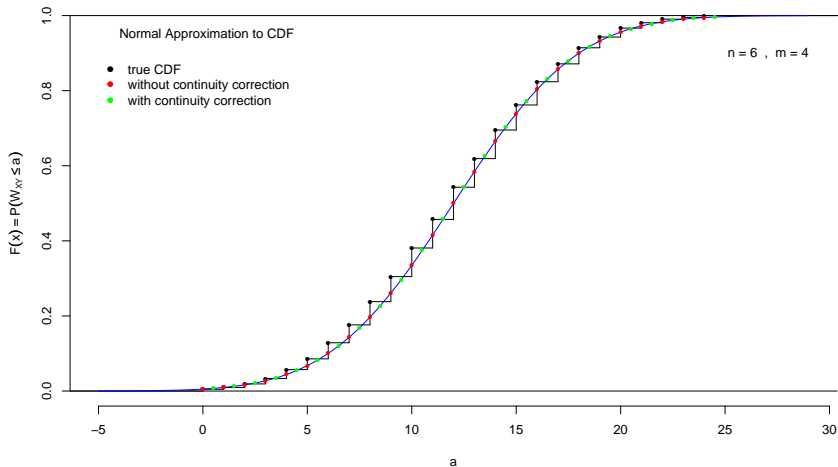
$$P(W_{XY} \geq a) = P(W_{XY} \geq a - 1/2) \approx 1 - \Phi \left( \frac{a - 1/2 - mn/2}{\sqrt{mn(N+1)/12}} \right)$$

$$\begin{aligned} P(a \leq W_{XY} \leq b) &= P(a - 1/2 \leq W_{XY} \leq b + 1/2) \\ &\approx \Phi \left( \frac{b + 1/2 - mn/2}{\sqrt{mn(N+1)/12}} \right) - \Phi \left( \frac{a - 1/2 - mn/2}{\sqrt{mn(N+1)/12}} \right) \end{aligned}$$

# Illustration of Continuity Correction (Density)



# Illustration of Continuity Correction (CDF)



# What About Tied Ranks?

- So far we have only dealt with clear rankings  $1, 2, \dots, N$ .
- Sometimes the ranking of two or more subjects is not so clear.
- The ranker would prefer to give the same rank to “tied” cases.
- When we have measurements or scores it is also possible to have observations that are the same.
- This may be due to rounding or natural discreteness in the data (e.g., counts).

## Treatment of Ties (Midranks, Special Case)

- Suppose we have  $n = 2, m = 2$  with observations 1.3, 1.7, 1.7, 2.5.
- Then the ranks of 1.3 and 2.5 should clearly be 1 and 4.
- It is most natural to assign the same average rank or **midrank** 2.5 to 1.7 and 1.7, which otherwise would have received ranks 2 and 3 in some order, had there been no ties.
- For example, 1.7, 1.7 could have been measured more accurately at 1.73 and 1.69.
- One may have second thoughts about breaking such ties by forcing more accurate measurements.
- Should such minute differences matter?

# Treatment of Ties (Midranks, in General)

- With  $\ell$  tied observations or subjects among  $N$ , with  $k$  lower rankings and  $N - k - \ell$  with higher ranks, the midrank of these  $\ell$  tied observations is the average of the corresponding ranks  $k + 1, k + 2, \dots, k + \ell$ , had there been no ties, i.e.,

$$\begin{aligned}\text{midrank} &= \frac{(k + 1) + (k + 2) + \dots + (k + \ell)}{\ell} \\ &= k + \frac{\ell(\ell + 1)}{2\ell} = k + \frac{\ell + 1}{2}\end{aligned}$$

- The midrank depends only on the number  $\ell$  of tied rankings and the number  $k$  of subjects with lower rank.
- We denote the midranks of the  $n$  treatment subjects and  $m$  controls by  $S_1^*, \dots, S_n^*$  and  $R_1^*, \dots, R_m^*$ , respectively.  
 $W_s^* = S_1^* + \dots + S_n^*$  and  $W_r^* = R_1^* + \dots + R_m^*$

# Null Distribution of Midranks

- In the presence of ties the null distribution of  $S_1^*, \dots, S_n^*$  is not the same as that of  $S_1, \dots, S_n$ , a random combination of  $N$ , taken  $n$  at a time.
- The possible set of values for  $S_1^*, \dots, S_n^*$  is different.
- However, the idea of derivation still holds, namely  $S_1^*, \dots, S_n^*$  is a random selection (without replacement) of  $n$  items from the  $N$  values  $(S_1^*, \dots, S_n^*, R_1^*, \dots, R_m^*)$ .
- Under the null hypothesis the rankings have nothing to do with the randomized assignments of treatment and control.
- Any such split of these  $N$  rankings (including midranks) into two groups of  $n$  and  $m$  is as likely as any other, namely has chance  $1/\binom{N}{n}$ .

# Null Distribution of $W_s^*$ (Special Case)

In our previous special case  $n = m = 2$  with midranks 1, 2.5, 2.5, 4 we have the following distribution of the midranks  $S_1^*, S_2^*$ :  $\binom{4}{2} = 6$

$$P_{H_0}(S_1^* = 1, S_2^* = 2.5) = \frac{2}{6} \qquad P_{H_0}(S_1^* = 1, S_2^* = 4) = \frac{1}{6}$$

$$P_{H_0}(S_1^* = S_2^* = 2.5) = \frac{1}{6} \qquad P_{H_0}(S_1^* = 2.5, S_2^* = 4) = \frac{2}{6}$$

$W_s^*$  takes on the following 3 values 3.5, 5, 6.5 with probabilities

$$P_{H_0}(W_s^* = 3.5) = P_{H_0}(S_1^* = 1, S_2^* = 2.5) = \frac{2}{6} = \frac{1}{3}$$

$$P_{H_0}(W_s^* = 5) = P_{H_0}(\{S_1^* = 1, S_2^* = 4\} \cup \{S_1^* = S_2^* = 2.5\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$P_{H_0}(W_s^* = 6.5) = P_{H_0}(S_1^* = 2.5, S_2^* = 4) = \frac{2}{6} = \frac{1}{3}$$

# Null Distribution of $W_s^*$ (Another Example)

- $m = n = 3$  with control observations 2, 2, 9 and treatment observations 4, 9, 9.
- The 6 midranks are: 1.5, 1.5, 3, 5, 5, 5 with  $W_s^* = 3 + 5 + 5 = 13$ .  $P_{H_0}(W_s^* \geq 13) = k/\binom{6}{3} = k/20$

$(s_1^*, s_2^*, s_3^*)$	1.5, 1.5, 3	1.5, 1.5, 5	1.5, 3, 5	1.5, 5, 5	3, 5, 5	5, 5, 5
$P_{H_0}(s_1^*, s_2^*, s_3^*)$	1/20	3/20	6/20	6/20	3/20	1/20
$W_s^*$	6	8	9.5	11.5	13	15

$$P_{H_0}(W_s^* \geq 13) = 4/20 = 1/5$$

# An Ambiguous Situation

- Suppose the treatment observations are 6, 6, 6, 9 with control observations 1, 3, 4, 10, large values of  $W_s$  being significant.
- Since all ties occur within the treatment group, it does not matter whether we rank them 4, 5, 6 or 5, 5, 5.
- In either case we get  
$$W_s = 4 + 5 + 6 + 7 = W_s^* = 5 + 5 + 5 + 7 = 22.$$

Should we use  $P_{H_0}(W_s \geq 22) = 12/70 = .1714$  as our  $p$ -value

or is  $P_{H_0}(W_s^* \geq 22) = 11/70 = .1571$  the correct value?

```
> 1-pwilcox(21-4*5/2, 4, 4)
0.1714286
> mean(combn(rank(c(6, 6, 6, 9, 1, 3, 4, 10)),
               4, FUN=sum) >= 22)
0.1571429
```

# Discussion of Ambiguity

- It does not matter in the computation of  $W_s$  whether we use ranks 4, 5, 6 or 5, 5, 5, they all belong to the treatment group.
- For the null distribution of  $W_s$  we need to be able to evaluate  $W_s$  for all possible splits of the 8 observations into two groups of 4 and 4.
- When the 3 observations 6,6,6 are split among the treatment and control groups, it is not clear which ranks of 4,5,6 they should get. In that case  $W_s$  is undefined.
- Just changing 6,6,6 to 5.9, 6.1, 6.4 to get distinct ranks turns the problem into a different one. It does not solve our problem.
- Would we have taken that step, when the 6,6,6 had been split among treatment and control, and how would we have assigned 5.9, 6.1, 6.4? Each results in a different rank sum.

- $P_{H_0}(W_s^* \geq 22) = 11/70 = .1571$  is the correct approach.
- Midranks and  $W_s^*$  are clearly defined for all splits of the data.
- $\Rightarrow$  proper null distribution for  $W_s^*$  and thus the above  $p$ -value.
- Our treatment should not depend on the way the rankings came out in this particular case.
- The  $p$ -value makes sense only in the context of all other splits that could have arisen.

# Problems in Tabulating $P_{H_0}(W_s^* \geq a)$

- The null distribution of  $W_s^*$  not only depends on  $n$  and  $m$ , it also depends on the pattern of tied observations or ranks.
- The variety of such tie patterns grows rather rapidly and it becomes impractical to tabulate these null distributions.
- We need other ways to assess  $p$ -values.
- Four options:
  - normal approximation, good for large  $n$  and  $m$  and without too many ties. It is always fast.
  - exact distribution via `combn`, perfect and feasible only for moderate  $m$  and  $n$ ,
  - clever algorithms for exact distribution by Streitberg and Röhmel, van de Wiel  $\Rightarrow$  `coin` package.
  - simulating the distribution of  $W_s^*$ , OK for any  $m$  and  $n$ , accuracy controlled by  $N_{\text{sim}}$ .

- The functions `dWs` and `pWs` compute the probability mass function and lower/upper tail probabilities of the Wilcoxon rank sum statistic (no ties).
- `Ws.pval.exact` gives the test statistic and p-value for the one-sided Wilcoxon rank sum test (ties or no ties).
- These calculations make use of full enumeration using `combn`, thus their usage is limited.
- `Ws.pval.sim` gives the test statistic and p-value (estimated via simulation) for the one-sided Wilcoxon rank sum test (ties or no ties).
- Study the code of these functions. It may not be optimal, but it should give you a feel for the involved methodology.

```
> dWs(5,4)
      Ws probability
[1,] 10 0.007936508
[2,] 11 0.007936508
[3,] 12 0.015873016
[4,] 13 0.023809524
[5,] 14 0.039682540
[6,] 15 0.047619048
[7,] 16 0.063492063
[8,] 17 0.071428571
[9,] 18 0.087301587
[10,] 19 0.087301587
[11,] 20 0.095238095
[12,] 21 0.087301587
[13,] 22 0.087301587
[14,] 23 0.071428571
[15,] 24 0.063492063
[16,] 25 0.047619048
[17,] 26 0.039682540
[18,] 27 0.023809524
[19,] 28 0.015873016
[20,] 29 0.007936508
[21,] 30 0.007936508
```

```
> pWs(16, 5, 4, lower.tail=T)
[1] 0.2063492
```

```
> Ws.pval.exact(c(6, -5, -6, 1, 4), c(5, 0, 16, 2, 9),
                alt="less")
p.value Ws.observed
0.1111      21
```

```
> Ws.pval.exact(c(5, 0, 6, 2, 9), c(6, -5, -6, 1, 4),
                alt="greater")
p.value Ws.observed
0.1230159      33.5
```

```
# note the tie at 6 in the last case
```

```
> set.seed(28)
> Ws.pval.sim(c(6,-5,-6,1,4),c(5,0,16,2,9),
             alt="less",Nsim = 10000)
```

```
p.value Ws.observed
0.1108          21
```

```
> set.seed(2013)
> Ws.pval.sim(c(5,0,6,2,9),c(6,-5,-6,1,4),
             alt="greater", Nsim=10000)
```

```
p.value Ws.observed
0.1208          33.5
```

```
# examine closeness to p-values on previous slide
```

# Normal Approximation

- The normal approximation for  $W_s^*$  is again a direct application of the CLT for sampling  $n$  items from the finite population  $\{Z_1, \dots, Z_N\} = \{S_1^*, \dots, S_n^*, R_1^*, \dots, R_m^*\}$ .
- The finite population depends strongly on the pattern of ties.
- It can be shown that mean and variance of  $W_s^*$  are

$$E(W_s^*) = \frac{n(N+1)}{2} \quad \text{and} \quad \text{var}(W_s^*) = \frac{mn(N+1)}{12} - \frac{mn \sum_{i=1}^{\ell} (d_i^3 - d_i)}{12N(N-1)}$$

- $\ell$  is the number of distinct midranks and  $d_i$  denotes the multiplicity of the  $i^{\text{th}}$  distinct midrank.
- For example, if  $x = \{2, 2, 9\}$  and  $y = \{4, 9, 9\}$ , then the joint set of midranks is

$$\{S_1^*, \dots, S_n^*, R_1^*, \dots, R_m^*\} = \{1.5, 1.5, 5, 3, 5, 5\} = \{1.5, 1.5, 3, 5, 5, 5\}$$

with  $\ell = 3$  and  $d_1 = 2$ ,  $d_2 = 1$  and  $d_3 = 3$ .

## $\text{var}(W_s^*)$ Calculation in R

The previous formula for  $\text{var}(W_s^*)$  involving the multiplicities  $d_i$  was useful when calculating it manually.

In R we can just resort to the previous formula (Slide ??) for the variance of a sum  $T_n = W_s^*$  of  $n$  items drawn at random and without replacement from the full set of midranks

$$Z = \{Z_1, \dots, Z_N\} = \{S_1^*, \dots, S_n^*, R_1^*, \dots, R_m^*\}.$$

$$\text{var}(W_s^*) = n \frac{N - n}{N - 1} \sigma_Z^2 = (n * (N - n) / N) * \text{var}(Z)$$

Note that R calculates  $\text{var}(Z) = \sum_{i=1}^N (Z_i - \bar{Z})^2 / (N - 1)$ , whereas  $\sigma_Z^2 = \sum_{i=1}^N (Z_i - \bar{Z})^2 / N$ .

Here  $Z \leftarrow \text{rank}(c(x, y))$  with  $x$  and  $y$  denoting the two sample vectors, with possible ties within  $c(x, y)$ .

- The previous normal approximation should only be used when

$$\max \left( \frac{d_1}{N}, \dots, \frac{d_\ell}{N} \right) \text{ stays bounded away from 1 as } N \rightarrow \infty.$$

- Furthermore, the Text recommends to use the normal approximation **without continuity correction**.
- Gaps between consecutive values of  $W_s^*$  can be quite irregular. See later examples.

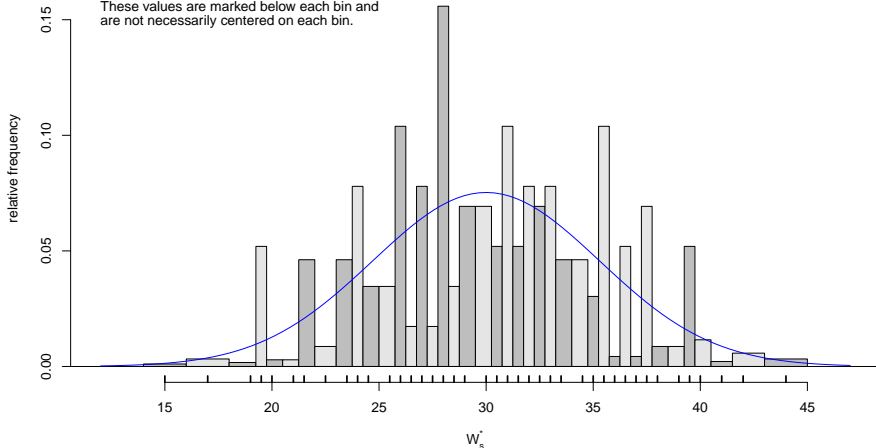
Using the full vector  $Z$  of  $N = n + m$  midranks ( $n$  treatments and  $m$  controls), we get the exact distribution of  $W_s^*$

```
out <- combn(Z, n, FUN=sum)
out.u <- unique(out)
k <- length(out.u)
freq <- numeric(k)
for(i in 1:k) {
  freq[i] = mean(out == out.u[i])
}
```

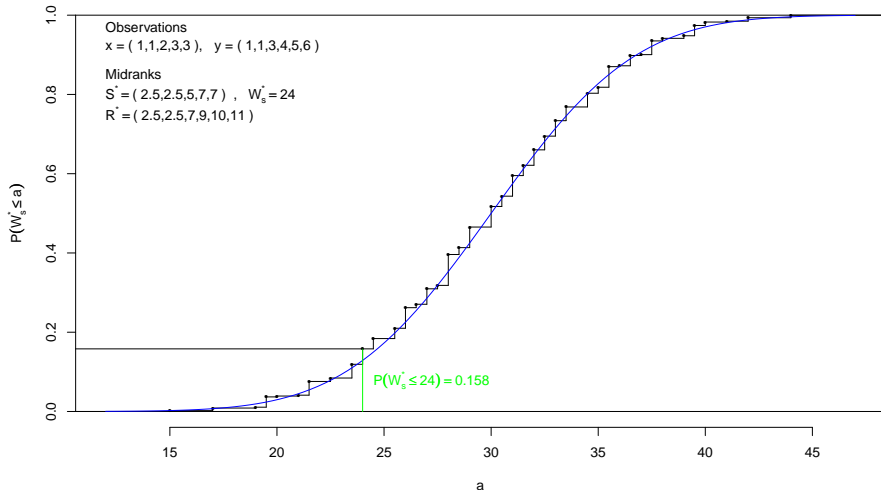
`out.u` contains the unique values of  $W_s^*$  and `freq` contains the corresponding relative frequencies (among all splits), i.e., the probabilities  $P_{H_0}(W_s^* = a)$ .

# Normal Approximation $n = 5, m = 6$ (with Ties)

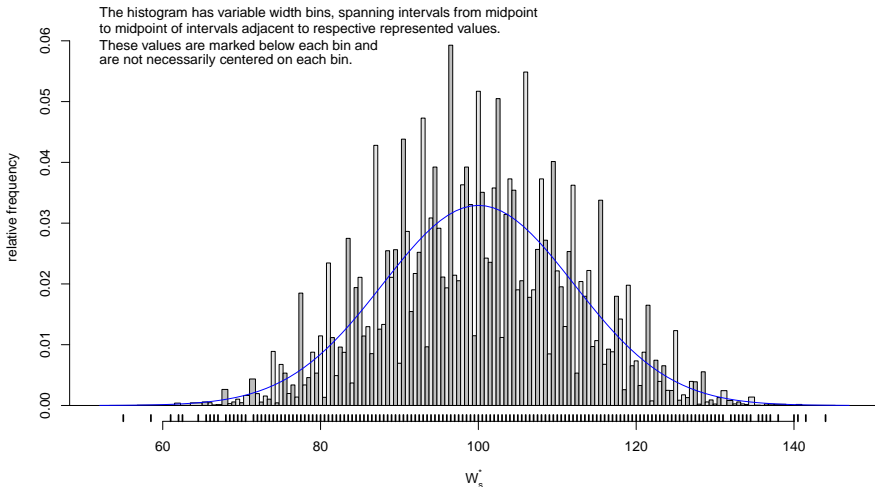
The histogram has variable width bins, spanning intervals from midpoint to midpoint of intervals adjacent to respective represented values. These values are marked below each bin and are not necessarily centered on each bin.



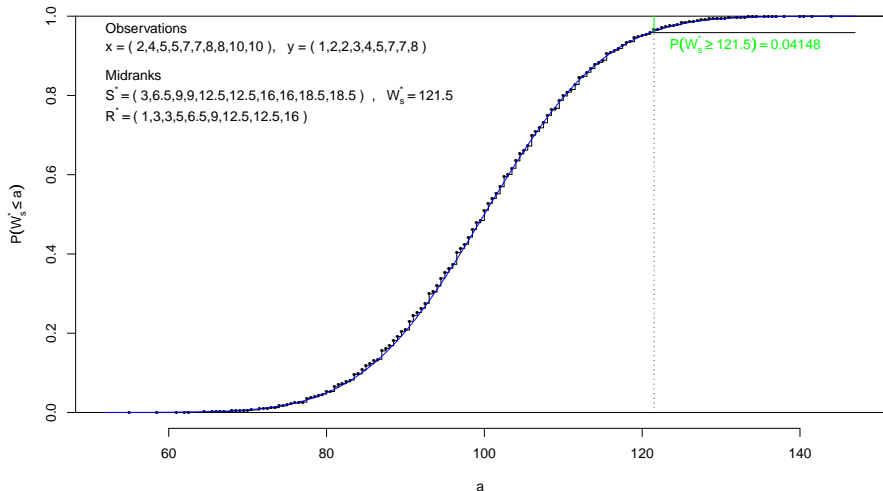
# CDF Normal Approximation $n = 5, m = 6$ (with Ties)



# Normal Approximation $n = 10, m = 9$ (with Ties)



# CDF Normal Approximation $n = 10, m = 9$ (with Ties)



- The previous histogram slides of the exact  $W_s^*$  distributions show that the distributions are no longer symmetric.
- If the combined set of midranks is symmetric around some value  $\xi$ , then the distribution of  $W_s^*$  is symmetric around  $n\xi$ .

Proof: For any sample  $X_1, \dots, X_n$  from the midranks  $Q_1, \dots, Q_N$  there is a corresponding sample  $\tilde{X}_1, \dots, \tilde{X}_n$  such that

$$(\tilde{X}_1 - \xi, \dots, \tilde{X}_n - \xi) = -(X_1 - \xi, \dots, X_n - \xi)$$

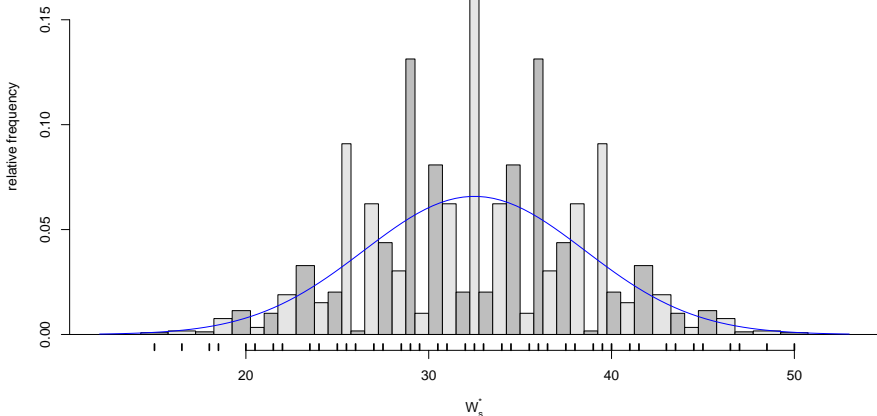
since all samples have same probability  $1/\binom{N}{n}$  it follows that

$$(X_1 - \xi, \dots, X_n - \xi) \stackrel{D}{=} -(X_1 - \xi, \dots, X_n - \xi)$$

$$\sum_{i=1}^n X_i - n\xi = \sum_{i=1}^n (X_i - \xi) \stackrel{D}{=} - \sum_{i=1}^n (X_i - \xi) = n\xi - \sum_{i=1}^n X_i$$

# Normal Approximation $n = 5, m = 7$ (with Ties)

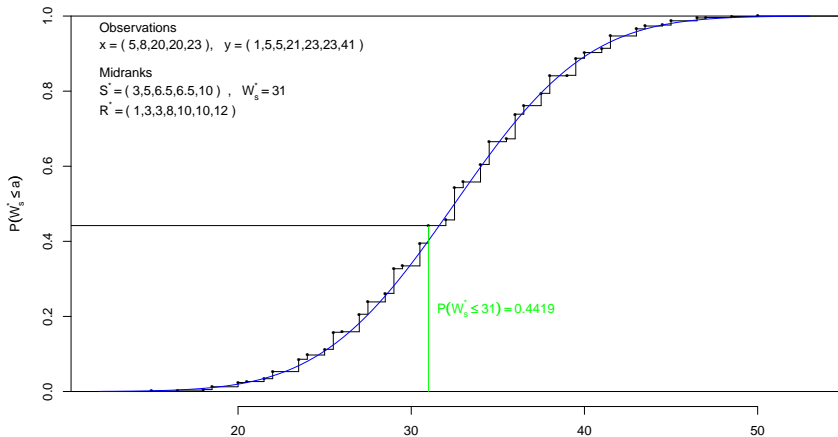
The histogram has variable width bins, spanning intervals from midpoint to midpoint of intervals adjacent to respective represented values.  
These values are marked below each bin and are not necessarily centered on each bin.



# Illustration of Symmetry Example

1, 5, 5, 5, 8, 20, 20, 21, 23, 23, 23, 41 with midranks

1, 3, 3, 3, 5, 6.5, 6.5, 8, 10, 10, 10, 12



## No Scores but Ordered Categories

Sometimes subjects don't provide numerical responses but can be classified into a set of ordered categories, such as presented in the table below.

	Poor	Fairly Poor	Fairly Good	Good	Total
Treatment	5	7	16	12	40
Control	7	9	15	9	40

The above data represent the effect of psychological counseling on 80 boys, randomly divided into two groups of 40 each.

These classifications were done by an impartial observer who did not know the treatment or control assignments.

## Midranks for Ordered Categories

	Poor	Fairly Poor	Fairly Good	Good	Total
Treatment	5	7	16	12	40
Control	7	9	15	9	40

The data can be treated as  $5 + 7 = 12$  observations tied at the lowest rank with midrank  $= (1 + \dots + 12)/12 = (1 + 12)/2 = 6.5$ ;

$7 + 9 = 16$  observations tied at the next lowest rank with midrank  $= (13 + \dots + 28)/16 = (13 + 28)/2 = 20.5$ ;

$16 + 15 = 31$  observations tied at the third lowest rank with midrank  $= (29 + \dots + 59)/31 = (29 + 59)/2 = 44$ ;

and  $12 + 9 = 21$  observations at the highest rank with midrank  $(60 + \dots + 80)/21 = (60 + 80)/2 = 70$ .

The resulting rank-sum is

$$W_s^* = 5 \times 6.5 + 7 \times 20.5 + 16 \times 44 + 12 \times 70 = 1720 .$$

For the  $\ell = 4$  distinct midranks 6.5, 20.5, 44 and 70 we have the following multiplicities  $d_1 = 12$ ,  $d_2 = 16$ ,  $d_3 = 31$  and  $d_4 = 21$ . The previous normal approximation for  $W_s^*$  is implemented in the function `psycholcounsel` (code on next slide)

```
> psycholcounsel()  
      Ws.star mean.Ws.star sdev.Ws.star pval.Ws.star  
1720.000000 1620.000000    99.2720339    0.1568874
```

# Code for psycholcounsel ()

```
psycholcounsel <- function(x=c(5,7,16,12),
                           y=c(7,9,15,9))
{
  d <- x+y
  n <- sum(x)
  m <- sum(y)
  N <- m+n
  midrank <- cumsum(d)-(d-1)/2 # see next slide
  Ws <- sum(x*midrank)
  meanWs <- n*(N+1)/2
  varWs <- m*n*(N+1)/12-m*n*sum(d^3-d)/(12*N*(N-1))
  pval <- 1-pnorm((Ws-meanWs)/sqrt(varWs))
  out <- c(Ws,meanWs,sqrt(varWs),pval)
  names(out) <- c("Ws.star","mean.Ws.star",
                 "sdev.Ws.star","pval.Ws.star")
  out
}
```

On Slide ?? we saw that the  $i^{\text{th}}$  midrank can be expressed as

$$d_1 + \dots + d_{i-1} + \frac{d_i + 1}{2} = \sum_{j=1}^i d_j - \frac{d_i - 1}{2}$$

For a vector  $d$  of midrank multiplicities

$$\text{midrank} = \text{cumsum}(d) - (d - 1)/2$$

gives us the vector of midranks.

- How to get an exact  $p$ -value when  $m$  and  $n$  become large?
- We have an  $N$ -vector  $Z$  of midranks

$$Z = (\overbrace{6.5, \dots, 6.5}^{12}, \overbrace{20.5, \dots, 20.5}^{16}, \overbrace{44, \dots, 44}^{31}, \overbrace{70, \dots, 70}^{21})$$

- Under  $H_0$  these classifications of subjects or ranks would have resulted no matter which 40 became treatment cases and which 40 became control cases.
- Need  $\binom{80}{40} = 1.075072e+23$  splits to get the exact  $p$ -value.
- This would seem insurmountable, certainly when using `combn`.
- However, `coin` still lets you do it.

- We can estimate the exact  $p$ -value.
- Sample without replacement 40 from these 80 elements in  $Z$ .
- Form their midrank sum  $W_s^*$ .
- Do this over and over  $N_{\text{sim}} = 10000$  or more times.
- Find the proportion of  $W_s^*$  values  $\geq 1720 = W_{s, \text{obs}}^*$ , the originally observed value of  $W_s^*$ .
- This proportion is our estimated  $p$ -value.
- This approach is always an option.

- The simulation approach is valid for (un)tied data (rankings).
- The simulated proportion of  $W_s^*$  values that are  $\geq W_{s, \text{obs}}^*$  is an unbiased estimate of the true exact  $p$ -value.
- It gets arbitrarily close to it by making  $N_{\text{sim}}$  sufficiently large.
- We have control over that accuracy through  $N_{\text{sim}}$ .
- We pay with simulation time (waiting for result).
- This is different from controlling the accuracy of the CLT. There we need to increase the number of subjects in the study, often not easy.
- See Lucien LeCam's principle 7 on page 19 in <http://www.stat.berkeley.edu/users/rice/LeCam/papers/tech168.pdf>

## Code for psycholcounsel.sim()

```
psycholcounsel.sim <- function (x = c(5, 7, 16, 12),  
                                y = c(7, 9, 15, 9), Nsim=1000){  
  d <- x + y; n <- sum(x); m <- sum(y);  
  ell <- length(x); N <- m + n  
  midrank <- cumsum(d) - (d - 1)/2  
  midrank.vec <- NULL  
  for(j in 1:ell){  
    midrank.vec <- c(midrank.vec, rep(midrank[j], d[j]))  
  }  
  Ws <- sum(x * midrank)  
  Ws.vec <- numeric(Nsim)  
  for(i in 1:Nsim){  
    Ws.vec[i] <- sum(sample(midrank.vec, n, replace=F))  
  }  
  pval <- mean(Ws.vec >= Ws)  
  out <- c(Ws, pval)  
  names(out) <- c("Ws.star", "pval.Ws.star")  
  out  
}
```

# Simulation Results

```
> system.time(out<-psycholcounsel.sim(Nsim=100000))
  user  system elapsed
  0.85    0.01    0.86
> out
      Ws.star pval.Ws.star
1720.00000      0.16134

> system.time(out<-psycholcounsel.sim(Nsim=10000))
  user  system elapsed
  0.09    0.00    0.10
> out
      Ws.star pval.Ws.star
1720.00000      0.1671
```

Compare this with normal approximation based  $p$ -value of .1569.

```
> Ws.pval.sim(rep(1:4, c(5, 7, 16, 12)),  
              rep(1:4, c(7, 9, 15, 9)),  
              "greater", 1000000)
```

p.value	Ws.observed
0.162588	1720

# The coin Package

- The `coin` package contains powerful nonparametric functions.
- `coin` stands short for `conditional inference`.
- The methodology is based on randomization of treatments among given subjects.
- Results are `conditional` on the tie pattern that results.
- Prior to using any function in `coin` you need to install that package via `install.packages("coin")` and choose a nearby download repository when prompted.
- This installation is done just once.
- Thereafter you need to invoke `library(coin)` for each new R session that wants to use functions in `coin`.
- Documentation on these functions can be found via `help.start()` under `packages`.

# The coin Package and wilcox\_test

- We will illustrate `coin`'s power using `wilcox_test`.
- It works with or without ties.
- We will do so using the ordered categorical data on slide ??.
- Note that after ranking there are  $\text{choose}(80, 40) = 1.075072e+23$  possible splits of the combined midranks into 2 samples of size 40 each.
- This is not conducive to full enumeration.
- See the documentation of `wilcox_test` for the method due to Streitberg and Röhmel and van de Wiel.
- Its performance is amazing (see our previous direct attempts).
- I don't know at what problem size it breaks down.

## wilcox\_test: Exact Method

```
xt <- rep(1:4, c(5, 7, 16, 12))
xc <- rep(1:4, c(7, 9, 15, 9))
z <- c(xc, xt)
m <- length(xc); n <- length(xt)
fac <- factor(c(rep("C", m), rep("T", n)),
              levels=c("T", "C"))
dat <- data.frame(z, fac) # end of setup

wilcox_test(z~fac, data=dat,
            distribution=exact(), alt="greater")
```

Exact Wilcoxon Mann-Whitney Rank Sum Test

```
data: z by fac (T, C)
Z = 1.0073, p-value = 0.1622
alternative hypothesis: true mu is greater than 0
```

## Usage of alternative in wilcox\_test

- The argument `alternative` (or short `alt`) can take three values: `"greater"`, `"less"`, and `"two.sided"`.
- The choice `"greater"` indicates that under the alternative the first factor level (here `"T"`) is expected to give greater scores than the second factor level.
- The order of factor levels is fixed by `levels=c("T", "C")` in the `factor` command.

In the previous setup let

```
wilcox.coin <- wilcox_test(z~fac, data=dat,  
                           dist=exact(), alt="greater")
```

# Getting the Full Distribution of $W_s^* = Ws.star$ (1)

```
mu <- expectation(wilcox.coin)
sig <- sqrt(variance(wilcox.coin))
# mu and sig are the mean and standard deviation
# of Ws.star and are used to standardize Ws.star

support.w <- support(wilcox.coin)
# gives the values of the standardized Ws.star

support.wstar <- sig*support.w+mu
# converts the standardized support support.w
# to that of Ws.star

p.x <- dperm(wilcox.coin,x)
# gives the probabilities at the vector x of
# standardized Ws.star values.
```

## Getting the Full Distribution of $W_s^* = \text{Ws.star}$ (2)

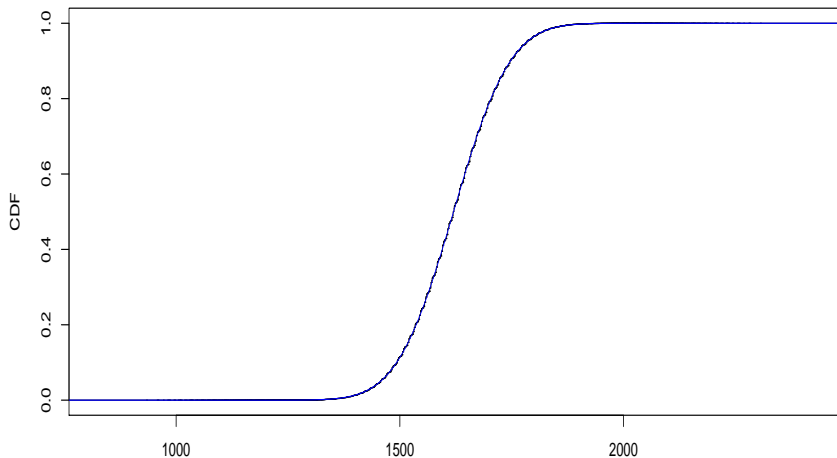
```
cdf.x <- pperm(wilcox.coin,x)
# gives the cumulative distribution function
# of the standardized Ws.star at any vector x.

q.p <- qperm(wilcox.coin,p)
# gives the quantiles of the standardized Ws.star
# for a given vector p of probabilities in (0,1).
# For a single p q.p is the smallest standardized
# q such that P( Ws.star <= q) >= p.
```

# Plotting the CDF

```
mu <- expectation(wilcox.coin)
sig <- sqrt(variance(wilcox.coin))
support.w <- support(wilcox.coin)
support.wstar <- sig*support.w+mu
cdf.w <- pperm(wilcox.coin,support.w)
plot(stepfun(support.wstar,c(0,cdf.w)),
      ylab="CDF",main="",pch=".")
readline("hit return\n")
# Note that the blue asymptotic curve
# almost completely tracks the exact cdf.
x1 <- seq(500,2500,.5)
y1 <- pnorm(x1,mu,sig)
lines(x1,y1,col="blue")
```

# Plot of CDF from Previous Slide



```
# using the same data setup as on previous slide
```

```
wilcox_test(z~fac, data=dat,  
            distribution=approximate(B=1e6),  
            alt="greater")
```

Approximative Wilcoxon Mann-Whitney Rank Sum Test

```
data: z by fac (T, C)
```

```
Z = 1.0073, p-value = 0.1626
```

```
alternative hypothesis: true mu is greater than 0
```

## wilcox\_test: Asymptotic Method

```
# using the same data setup as on previous slide
```

```
wilcox_test(z~fac, data=dat,  
            distribution=asymptotic(),  
            alt="greater")
```

Asymptotic Wilcoxon Mann-Whitney Rank Sum Test

```
data: z by fac (T, C)
```

```
Z = 1.0073, p-value = 0.1569
```

```
alternative hypothesis: true mu is greater than 0
```

Just as we had the relationship

$$W_{XY} = W_s - n(n+1)/2 \quad \text{we also have} \quad W_{XY}^* = W_s^* - n(n+1)/2$$

provided we score comparisons with  $X_i$  and  $Y_j$  as 1/2 whenever  $X_i = Y_j$ , i.e., we define

$$W_{XY}^* = [\text{number of pairs with } X_i < Y_j] \\ + \frac{1}{2} [\text{number of pairs with } X_i = Y_j]$$

For the proof we refer to the Text, pp. 22-23.

# One-Sided Treatment of Testing $H_0$

- So far our tests compared a new treatment with a standard one or a control.
- We were strongly biased in favor of the latter, requiring strong evidence to reject the hypothesis of no difference, e.g., a  $p$ -value  $\leq .05$ .
- For significance level  $\alpha = .05$  we decide for the standard treatment 95% of the time, when in fact there is no difference.
- By continuity this rate is about the same when the new treatment is only slightly better, i.e.,  $H_0$  is false.
- Under  $H_0$  we only take a  $\approx 5\%$  chance of rejecting  $H_0$  in favor of the new treatment.
- He have a strong inertia in favor of  $H_0$ .
- This uneven treatment of the two possibilities is not always appropriate.

# Alternate Scenarios

- We may be faced with two new treatments.
- We would like to decide which of them is better or whether there is no essential difference.
- We have no reason to treat either one as preferred, i.e., be biased toward either one when there is no strong a priori evidence for a difference.
- Sometimes the issue is not to decide which of two treatments is better, but to determine whether there is a difference at all.
- Sometimes it is useful to know that both treatments can be used interchangeably.
- One of the treatments may be cheaper or more practical.

## Two-Sided Alternatives and Tests

- We will be dealing with two treatments  $A$  and  $B$  and want to test whether there is a significant difference between  $A$  and  $B$ .
- The null hypothesis is  $H_0$ : there is no difference between  $A$  and  $B$ . (same as before)
- The subjects are again randomly assigned,  $m$  with treatment  $A$  and  $n$  with  $B$ .
- The  $N = m + n$  subject responses are ranked:  $\implies W_A$  and  $W_B$ , the sums of  $A$ -ranks and  $B$ -ranks, respectively.
- We deal with ties later.
- Reject  $H_0$  when  $W_B \leq c_1$  or  $W_B \geq c_2$  ( $c_1 < c_2$ ), with  $c_1$  and  $c_2$  chosen such that

$$P_{H_0}(\text{rejecting } H_0) = P_{H_0}(W_B \leq c_1) + P_{H_0}(W_B \geq c_2) = \alpha .$$

# The Choice of Critical Points $c_1$ & $c_2$

- The choice of  $c_1$  and  $c_2$  was left unspecified.
- Without a priori preference for  $A$  or  $B$  it is natural to choose  $c_1$  and  $c_2$  such that

$$P_{H_0}(W_B \leq c_1) = \alpha/2 \quad \text{and} \quad P_{H_0}(W_B \geq c_2) = \alpha/2$$

- Since the distribution of  $W_B$  is symmetric around  $n(N+1)/2$

$$\implies c_1 = \frac{n(N+1)}{2} - c \quad \text{and} \quad c_2 = \frac{n(N+1)}{2} + c$$

- This **two-sided Wilcoxon rank-sum test** rejects  $H_0$  whenever

$$W_B - \frac{n(N+1)}{2} \leq -c \quad \text{or} \quad W_B - \frac{n(N+1)}{2} \geq c, \quad \text{i.e., when}$$

$$\left| W_B - \frac{n(N+1)}{2} \right| \geq c$$

- As in the one-sided test case we cannot achieve

$$P_{H_0} \left( \left| W_B - \frac{n(N+1)}{2} \right| \geq c \right) \stackrel{\text{def.}}{=} \alpha_c = \alpha \quad \text{for all } \alpha \in (0, 1).$$

- Reason: The discrete nature of the  $W_B$  null distribution.
- Instead we choose  $c$  such that the corresponding  $\alpha_c$  is either as close as possible to the desired  $\alpha$  or yields the largest  $\alpha_c \leq \alpha$ .
- Better yet, compute the observed significance level or  $p$ -value  $p(w)$  for the observed  $W_B = w$

$$p(w) = P_{H_0} \left( \left| W_B - \frac{n(N+1)}{2} \right| \geq \left| w - \frac{n(N+1)}{2} \right| \right)$$

- $p(w)$  is more informative. How strongly to reject  $H_0$  or not.

# Normal Approximation

- For large  $m$  and  $n$  we again can use the normal approximation for  $(W_B - n(N + 1)/2)/\sqrt{mn(N + 1)/12}$ .
- Using the continuity correction, the  $p$ -value for  $w$  is obtained as the area under the standard normal curve to the left/right of

$$-\frac{\left|w - \frac{n(N+1)}{2}\right| + \frac{1}{2}}{\sqrt{mn(N+1)/12}} \quad \text{and} \quad \frac{\left|w - \frac{n(N+1)}{2}\right| - \frac{1}{2}}{\sqrt{mn(N+1)/12}},$$

respectively.  $n(N + 1)/2 \pm |w - n(N + 1)/2|$  is integer.

- Since both areas are the same (the limits being negatives of each other) we have

$$p(w) = 2 \left\{ 1 - \Phi \left( \frac{\left|w - \frac{n(N+1)}{2}\right| - \frac{1}{2}}{\sqrt{mn(N+1)/12}} \right) \right\}$$

# The Treatment of Ties

- As before, we deal with ties by assigning appropriate midranks to the tied observations.
- Compute the sum  $W_B^*$  of these midranks under treatment  $B$ .
- We reject  $H_0$  whenever  $W_B^* \leq c_1$  or  $W_B^* \geq c_2$ .
- Choose  $c_1$  and  $c_2$  so that  $P_{H_0}(W_B^* \leq c_1)$  and  $P_{H_0}(W_B^* \geq c_2)$  are close to (or just below)  $\alpha/2$  in each case.
- Complication: The null distribution of  $W_B^*$  is no longer symmetric around its mean  $n(N+1)/2$ .
- We could view  $|W_B^* - n(N+1)/2|$  as our test statistic and compute as the  $p$ -value for the observed  $W_B^* = w$

$$p(w) = P_{H_0} \left( \left| W_B^* - \frac{n(N+1)}{2} \right| \geq \left| w - \frac{n(N+1)}{2} \right| \right)$$

- $p(w)$  can be computed as before in three different ways.
- By full enumeration of the  $|W_B^* - n(N + 1)/2|$  null distribution, if feasible.
- By simulation of the  $|W_B^* - n(N + 1)/2|$  distribution, accuracy controlled by  $N_{\text{sim}}$ .
- By normal approximation (without continuity correction)

$$p(w) = 2 \left\{ 1 - \Phi \left( \frac{|w - n(N + 1)/2|}{\sqrt{\text{var}_{H_0}(W_B^*)}} \right) \right\}$$

- Of course, the normal approximation is symmetric and thus we just double the one-sided  $p$ -value.

# One-Sided Versus Two-Sided Test??

- Which test to use depends on the question that should be posed *a priori*.
- This should not be influenced by the data.
- If the question is: Is  $B$  better than  $A$  as opposed to no difference (or even  $B$  being worse than  $A$ )?  
Then we should use a one-sided test.
- Is either treatment better than the other? Then we should use a two-sided test.
- In the first case “ $B$  being worse than  $A$ ” is subsumed as part of the hypothesis, then denoted by  $H'_0$ .
- Within  $H'_0$  the null hypothesis  $H_0$  represents the least favorable case as far as type I error probabilities are concerned:  
$$P_{H'_0}(W_B \geq c) \leq P_{H_0}(W_B \geq c) = \alpha.$$

# How not to Proceed.

- We have no a priori idea which treatment is better, if they are different at all.
- After the data are taken it becomes apparent that  $B$  is better.
- After some reflection “obvious” reasons for that can be advanced.
- It is then tempting to use a one-sided test to decide rejection of  $H_0 : A = B$  in favor of  $B$  being better than  $A$ .
- This is not appropriate under the a priori circumstances.
- In the same fashion we could also have observed data that seem to favor  $A$ .
- This again may suggest “obvious” reasons why  $A$  should be better than  $B$ .

# Values of One- and Two-Sided Tests

- The critical values for one- and two-sided tests are determined respectively by

$$P_{H_0}(W_B - n(N + 1)/2 \geq c^{(1)}) = \alpha$$

$$P_{H_0}(|W_B - n(N + 1)/2| \geq c^{(2)}) = \alpha$$

where the second equation (by symmetry) is equivalent to

$$P_{H_0}(W_B - n(N + 1)/2 \geq c^{(2)}) = \alpha/2 \quad \implies \quad c^{(2)} > c^{(1)}$$

- To be judged significant at the same level  $\alpha$ , the rank-sum  $W_B$  must thus be more extreme when testing  $H_0$  against both directions  $A \succ B$  or  $B \succ A$ , than when it is tested against a specific direction, say  $B \succ A$ .
- Note that an extreme value of  $W_B$  in either direction is obtained under  $H_0$  (when only randomization is in effect) twice as likely as the same extreme value in a specified direction.

# Anticipation of Hypnosis

- 16 subjects were randomly split into two groups of 8 each, one group treated with hypnosis, the other acting as controls.
- After the experiment it was noticed that a measure of ventilation taken on each subject at the beginning of the experiment seemed higher for experimental subjects than for the controls.
- A plausible explanation was an anticipation effect. We will test for this effect.

measure of ventilation

Controls:	3.99	4.19	4.21	4.54	4.64	4.69	4.84	5.48
Treated:	4.36	4.67	4.78	5.08	5.16	5.20	5.52	5.74

- Although a one-sided test is tempting, it is not appropriate.
- A corresponding effect in the other direction would also have been noticed and tested.

# Anticipation of Hypnosis (Analysis)

- We have  $W_s = 87$ ,  $n(N + 1)/2 = 68$ ,  $\binom{16}{8} = 12870$ , which makes a full and exact evaluation of the  $p$ -value feasible.
- For the exact  $p$ -value we get

$$P_{H_0}(|W_s - 68| \geq |87 - 68|) = P_{H_0}(|W_s - 68| \geq 19) = \frac{642}{12870} = 0.04988$$

- The normal approximation gives us

$$2 \left\{ 1 - \Phi \left( \frac{|87 - 68| - .5}{\sqrt{8 \cdot 8 \cdot 17/12}} \right) \right\} = 2 \{1 - \Phi(1.943)\} = 0.0520$$

The Text incorrectly has 2.001 in place of 1.943.

# A Three Decision Approach

- The previous two-sided procedure accepts the null hypothesis  $H_0$  (no difference between  $A$  and  $B$ ) when  $|W_B - n(N + 1)/2| < c$  and rejects  $H_0$  otherwise.
- When rejecting  $H_0 \implies$  which treatment is better?
- $\implies$  choose between three decisions:  $D_0$ : accepting  $H_0$ ,  
 $D_1$ : declaring  $B$  better than  $A$ , or  
 $D_2$ : declaring  $A$  better than  $B$ .
- The natural and obvious procedure for this is to decide:

$D_1$  whenever  $W_B \geq n(N + 1)/2 + c$ ,

$D_2$  whenever  $W_B \leq n(N + 1)/2 - c$ ,

$D_0$  whenever  $|W_B - n(N + 1)/2| < c$ .

- $\alpha$  retains its interpretation as probability of falsely rejecting  $H_0$  when in fact there is no difference between  $A$  and  $B$ .

# An Alternate Three Decision Approach

- Here we focus more on deciding which treatment is better.
- We replace the somewhat artificial decision of no difference between  $A$  and  $B$  by the option of suspending judgment, if the data remain inconclusive concerning  $A \succ B$  or  $B \succ A$ .

$$\text{Choose } B \text{ if} \quad W_B \geq n(N+1)/2 + c$$

$$\text{Choose } A \text{ if} \quad W_B \leq n(N+1)/2 - c$$

$$\text{Suspend judgment if} \quad |W_B - n(N+1)/2| < c$$

The choice of  $c$  is driven by the common probability

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

computed under the assumption of no difference between treatments.

- $2\alpha'_c$  is the probability of deciding that one of the two treatments is better than the other when in fact they are equal. (Our original two-sided test.)

# Discussion of Alternate Three Decision Approach

- The previous procedure, allowing suspension of judgment, can run into two errors: Deciding  $A$  is better than  $B$ , when in fact the opposite is true and vice versa.
- Suppose  $B$  is better than  $A$ , but we observe  $W_B \leq n(N+1)/2 - c$  and thus decide erroneously that  $A$  is better. What is the maximal chance for this error?
- On intuitive grounds this chance for error should increase as  $B$  and  $A$  become more and more alike while still maintaining  $B \succ B_1 \succ B_2 \succ \dots \succ A$ .
- $\lim P_{B_\ell \succ A}(W_B \leq n(N+1)/2 - c) = \alpha'_c$  as  $A = \lim_\ell B_\ell$ .
- The reverse case (deciding  $B \succ A$  erroneously, when in fact  $A \succ B$ ) is treated analogously, with same maximum chance  $\alpha'_c$  of making this error.

- 12 rats were randomly assigned to two diets  $A$  and  $B$ , 7 with diet  $A$  and 5 with diet  $B$ .

A: 156 183 120 113 138 145 142

B: 130 148 117 133 140

- With  $B$ -ranks 2,4,5,7,10 we get  $W_B = 28$  which is less than  $n(N+1)/2 = 32.5$ .
- Thus we would decide in favor of  $A$ .
- What is the smallest  $\alpha'_c$  at which we would still decide  $A \succ B$ ?

$$\alpha'_{28} = P_{H_0}(W_B \leq 28) = 0.2652 \quad (\text{normal approximation } 0.2580)$$

- Any  $\alpha'_c < 0.2652$  would have caused us to suspend judgment at that level of error rate  $\alpha'_c$ .

## Contradiction to Previous Warning in Two-Sided Test?

- It seems that we go against our previous warning not to let the data tell us which side to take in a one-sided procedure when in fact we have a two-sided situation with unknown superiority of  $A$  or  $B$ .
- The answer lies in the meaning of “error.”
- In the two-sided test the error consist in declaring a difference when in fact there is none. The decision can come about in one of two possible ways in any experiment.
- In the three-decision procedure the error to be controlled is deciding for  $A \succ B$  when in fact  $B \succ A$  or the other way around.
- However, here only one of the two actual superiorities can be active in any given experiment.
- For the asymmetric treatment of the three-decision procedure and also for the treatment of tied ranks and for forcing a judgment see the Text, p.29-31.

## Other Alternatives to $H_0$

- So far we only considered alternatives to  $H_0 : A = B$  that were of the type  $A \succ B$  or  $B \succ A$ .
- These alternatives strongly influenced our choice of test statistic, rejecting  $H_0$  for high or low rank-sum  $W_B$ .
- What if treatment differences show in other ways?
- For example, under one treatment the measurements may be more variable than under the other treatment.  
⇒ Siegel-Tukey test.
- Or the differences in the subject responses can be expressed in yet other ways, which are limitless.  
⇒ Kolmogorov-Smirnov and Anderson-Darling tests.

# Wilcoxon Rank-Sum Test for Variability?

- We measure the same physical or biological quantity by two different methods.
- We want to test whether the two methods perform equally well, or whether one produces more variable results than the other (one-sided or two-sided).
- Again our hypothesis is  $H_0$ . We could consider using the Wilcoxon rank-sum test. Would it be effective?
- If  $B$  produces more variable results, it will result in  $B$  ranks both at the low end and the high end of all ranks.
- These low and high ranks will tend to average to near the middle of all ranks,  $(N + 1)/2$ , so that in the end  $W_B$  will wind up with a value near  $n(N + 1)/2$ , its mean under  $H_0$ .
- It will typically not produce significant results.

# Siegel-Tukey Test for Variability

- The Siegel-Tukey rank-sum test uses a very simple ranking idea and the already known null distribution of the Wilcoxon test.
- Rather than ranking subjects or scores from low to high, i.e.,

$$\mathcal{N} = \{1, 2, 3, \dots, N - 2, N - 1, N\}$$

we rank them

$$\mathcal{N}' = \{1, 4, 5, 8, 9, \dots, 10, 7, 6, 3, 2\}$$

- If the new treatment results in less variability, then the ranks associated with its subjects are mostly higher than those for the control treatment.
- We should reject  $H_0$  when  $W_s \geq c$  for some appropriate  $c$ .
- Here  $W_s$  is again the sum of ranks  $S_1, \dots, S_n$  for the subjects under the new treatment, using the  $\mathcal{N}'$  ranking scheme.

# The Null Distribution of the Siegel-Tukey Test

- The null distribution of the Siegel-Tukey test statistic  $W_s$  is the same as that of the Wilcoxon rank-sum statistic  $W_s$ , hence we use the same symbol.
- The reason is that under  $H_0$  any set of  $n$  ranks taken randomly and without replacement from  $\mathcal{N}$  or  $\mathcal{N}'$  is as likely as any other, namely has chance  $1/\binom{N}{n}$ .
- Thus we reject  $H_0$  in favor of the new treatment resulting in less variability whenever  $W_s$  is too large.
- The determination of null distribution probabilities by exact methods, normal approximation or simulation is as before.

# An Example Calculation

- We have  $m = 5$  control and  $n = 6$  treatment measurements

.8 .1 .14 .6 .34      and      .4 .38 .64 .26 .31 .55

- with dispersion ranks

Observ.	.1	.14	.26	.31	.34	.38	.4	.55	.6	.64	.8
Rank	1	4	5	8	9	11	10	7	6	3	2
Source	C	C	T	T	C	T	T	T	C	T	C

- with  $W_r = 1 + 2 + 4 + 6 + 9 = 22$       and  
 $W_s = 3 + 5 + 7 + 8 + 10 + 11 = 44$ .
- With the a priori position that the treatment might lead to less dispersion if there is any effect at all we would look for large values of  $W_s$  as being significant, or equivalently for small values of  $W_r$ .

- The function `SiegelTukey` is available on the class web page, together with the required auxiliary function `ST.rank`.

```
> SiegelTukey(x=c(.8, .1, .14, .6, .34),  
              y=c(.4, .38, .64, .26, .31, .55))
```

Wr	mean(Wr)	st.dev(Wr)	pval.exact	pval.norm	n.comb
2.20e+01	3.00e+01	5.48e+00	4.33e-03	8.55e-02	4.62e+02

- As alternative to  $H_0$  (no difference in treatment effect) assume that the  $x$ -sample is more dispersed than the  $y$ -sample around a **common** central value, i.e., we expect low values for  $W_r$ .

# Siegel-Tukey Test in Case of Ties

- We have  $m = 5$  control and  $n = 6$  treatment measurements, but with ties

.8 .1 .14 .6 .35      and      .4 .35 .64 .26 .35 .55

Observation	.1	.14	.26	.35	.35	.35	.4	.55	.6	.64	.8
Rank	1	4	5	8	9	11	10	7	6	3	2
Source	C	C	T	T	C	T	T	T	C	T	C
Average Rank	1	4	5	$\frac{28}{3}$	$\frac{28}{3}$	$\frac{28}{3}$	10	7	6	3	2

- $W_r = 1 + 2 + 4 + 6 + \frac{28}{3} = 22\frac{1}{3}$ ,  $W_s = 3 + 5 + 7 + \frac{28}{3} + \frac{28}{3} + 10 = 43\frac{2}{3}$
- Since the averaged ranks are not adjacent (8, 9, 11) we need to use the variance formula given on slide ?? in place of the one on slide ?? when using the normal approximation in case of ties (use no continuity correction in that case).
- See [SiegelTukey](#) code.

# R Function SiegelTukey with Ties

```
> SiegelTukey(x=c(.8, .1, .14, .6, .35),  
              y=c(.4, .35, .64, .26, .35, .55))
```

Wr	mean(Wr)	st.dev(Wr)	pval.exact		
2.23e+01	3.00e+01	5.36e+00	4.33e-03		
				pval.norm	n.comb
				7.63e-02	4.62e+02

# Question of Ranking in the Siegel-Tukey Test

Siegel-Tukey Ranks	1	4	5	8	9	11	10	7	6	3	2
Why not in reverse?	2	3	6	7	10	11	9	8	5	4	1

- The two ranking schemes may result in different values for  $W_s$
- and may even change the statistical significance of a result (slightly). This is disconcerting.
- The advantage of either of the Siegel-Tukey ranking schemes (from left to right or in reverse) is that they can use the same null distribution as the Wilcoxon test.
- Important and very appealing at the time of their creation.

# The Ansari-Bradley Test

- The direction bias of the Siegel-Tukey ranking scheme is avoided in the Ansari-Bradley test.
- Essentially or approximately this amounts to averaging the two Siegel-Tukey schemes for ranking.
- In the context of our example the ranking goes as follows:

Observation	.1	.14	.26	.31	.34	.38	.4	.55	.6	.64	.8
Rank	1	2	3	4	5	6	5	4	3	2	1
Source	C	C	T	T	C	T	T	T	C	T	C

- $W_r = 1 + 1 + 2 + 3 + 5 = 12$ ,  $W_s = 2 + 3 + 4 + 4 + 5 + 6 = 24$ .
- We reject the null hypothesis  $H_0$  of no difference against the alternative of less variability under the treatment when  $W_s \geq c$  (or when  $W_r \leq c'$ ).

# The Distribution of the Ansari-Bradley Test

- For large  $m$  and  $n$  the rank-sum is again approximately normally distributed with means and variances given by

$$E(W_s) = \begin{cases} \frac{n(N+2)}{4} \\ \frac{n(N+1)^2}{4N} \end{cases}, \quad \text{var}(W_s) = \begin{cases} \frac{mn(N^2-4)}{48(N-1)} & N \text{ even} \\ \frac{mn(N+1)[3+N^2]}{48N^2} & N \text{ odd} \end{cases}$$

- The utility of the continuity correction should be examined.
- The exact distribution of  $W_s$  can again be obtained either by full enumeration of all possible samples of  $n$  items taken without replacement from  $\mathcal{N} = \{1, 2, \dots, 2, 1\}$  (using `combn`)
- or it can be approximated by taking  $N_{\text{sim}}$  independent samples of size  $n$  from  $\mathcal{N}$  (each sample without replacement) and evaluating  $W_s$  each time.

# Functions for Scale Tests and Comments

- There is `ansari.test` in `R`
- and `ansari_test` in `coin`, which handles ties by exact method as well.
- Whether `oneway_test` in `coin` can be applied to Siegel-Tukey scores can be explored. (HW?)
- These scale tests are not applicable when there is no change in scale but with an unknown shift under  $H_0$ .
- There is no rank test for the latter situation.
- Attempts at aligning the two samples by appropriate shift estimates based on the separate samples conflict with any ranking of combined samples.
- Thus rank tests for scale alternatives are of limited utility.
- Useful in comparing two methods' variability when measuring the same quantity, no shift differential by either method.

- We have considered two types of alternatives to  $H_0$ : treatment scores
  - tend to be larger (smaller) than control scores
  - tend to be less (more) dispersed than control scores
- However, there are a lot of other possibilities for alternatives.
- An effective way of comparing scores from 2 samples is to use their empirical distribution functions (EDFs).
- For a sample  $a_1, \dots, a_m$  its EDF is defined as

$$F_m(x) = \frac{\text{number of } a_i\text{'s } \leq x}{m} \quad \text{as a function of } x \in R.$$

- $F_m$  is the nonparametric maximum likelihood estimate of  $F$ , the sampled CDF.

# Large Sample Behavior of $F_m(x)$



$$F_m(x) \longrightarrow F(x) \quad \text{as} \quad m \longrightarrow \infty$$

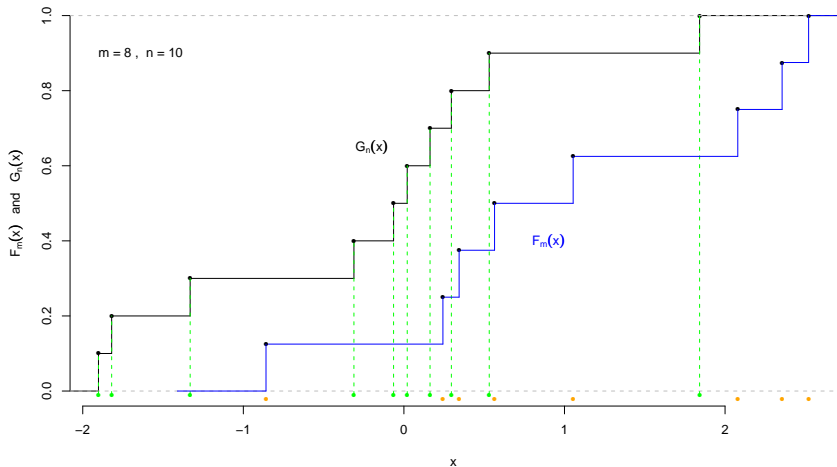
- This follows from the Law of Large Numbers (LLN).
- We can view  $F_m(x)$  as an average of  $m$  independent and identically distributed random variables, namely

$$F_m(x) = \frac{1}{m} \sum_{i=1}^m I_i(x) \quad \text{with} \quad I_i(x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x \end{cases}$$

- The Bernoulli random variables  $I_i(x)$  have mean  $E[I_i(x)] = P(X_i \leq x) = F(x)$  and variance  $\text{var}(I_i(x)) = F(x)(1 - F(x))$ . Thus  $E[F_m(x)] = F(x)$  and

$$\text{var}[F_m(x)] = \frac{F(x)(1 - F(x))}{m} \rightarrow 0 \quad \implies \quad F_m(x) \xrightarrow{m \rightarrow \infty} F(x)$$

# Comparing Two EDFs



# The Smirnov Test

- Compare two samples through some measure of discrepancy between  $F_m(x)$  and  $G_n(x)$  over the full range of  $x$  values.
- Smirnov (1939) proposed to reject  $H_0 : F = G$  whenever

$$D_{m,n} = \max_x |G_n(x) - F_m(x)| \geq c$$

- $c$  is chosen to reject  $H_0$  with probability  $\approx \alpha$  when  $H_0$  is true.
- Thus we need the null distribution of  $D_{m,n}$ .
- Most often this test is also referred to as the **Kolmogorov-Smirnov Two-Sample Test** or **KS Test**, mainly due to parallel or complementing work by Kolmogorov.
- The weakness of the KS test is its focus on the maximum vertical distance between  $F_m$  and  $G_n$ . It does not account for the extent/width of the gap between  $F_m$  and  $G_n$ .

# Smirnov Test is a Rank Test

- From the definition and the previous illustration it is clear that

$$D_{m,n} = \max_{1 \leq j \leq N} |F_m(Z_{(j)}) - G_n(Z_{(j)})|$$

where  $Z_{(1)} < Z_{(2)} < \dots < Z_{(N)}$  represents the ordering of the combined  $m + n = N$  sample values.

- $F_m(x) - G_n(x)$  stays constant between  $Z_{(i)}$  and  $Z_{(i+1)}$
- Let  $R_1 < \dots < R_m$  be the ordered ranks of the  $X$ 's. Then

$$\begin{aligned} F_m(Z_{(j)}) &= \frac{\text{number of } X\text{'s} \leq Z_{(j)}}{m} = \frac{\text{number of } R\text{'s} \leq j}{m} \\ G_n(Z_{(j)}) &= \frac{\text{number of } Y\text{'s} \leq Z_{(j)}}{n} = \frac{j - \text{number of } X\text{'s} \leq Z_{(j)}}{n} \\ &= \frac{j - \text{number of } R\text{'s} \leq j}{n} \end{aligned}$$

- Thus  $D_{m,n}$  depends only on the set of ranks  $R_1 < \dots < R_m$ .

# Null Distribution of the KS Test when $m = n$

- When  $m = n$  the null distribution of the KS test takes the following simple form for  $d = a/n > 0$  (Gnedenko and Korolyuk (1951))

$$P(D_{n,n} \geq d) = \frac{2 \left[ \binom{2n}{n-a} - \binom{2n}{n-2a} + \binom{2n}{n-3a} - \dots \right]}{\binom{2n}{n}}$$

- where the sign alternating summation in the numerator continues as long as  $n - a, n - 2a, n - 3a, \dots \geq 0$ .
- The denominator reflects the equal probability  $1/\binom{2n}{n}$  for each possible ranking  $R_1 < \dots < R_m$ .
- For  $d = 0$ , i.e.,  $a = 0$ , the formula breaks down since

$$\frac{2 \left[ \binom{2n}{n} - \binom{2n}{n} + \binom{2n}{n} - \binom{2n}{n} + \dots \right]}{\binom{2n}{n}}$$

oscillates between 0 and 2

and thus is indeterminate, but we then have  $P(D_{n,n} \geq 0) = 1$ .

## Example 6: Two Drugs for Pain Relief

- For relief from postoperative pain 16 patients were randomly split into 8 getting the standard drug  $A$  and 8 getting an experimental drug  $B$ . The following numbers represent the hours of postoperative relief

A: 3.1 3.3 4.2 4.5 4.7 4.9 5.8 6.8

B: 0.0 2.1 2.3 2.5 2.8 4.4 4.8 6.6

- Again test the hypothesis of no difference between  $A$  and  $B$ .
- Alternatively the  $A$  scores could generally be higher (lower) or more (less) dispersed than the  $B$  scores.
- There may be other ways such differences could show up.
- For example,  $A$  may be more effective for patients with low pain tolerance but  $B$  does not act that way.
- Thus we employ the KS test to test for any type of difference.

The ranks are

A: 6 7 8 10 11 13 14 16

B: 1 2 3 4 5 9 12 15

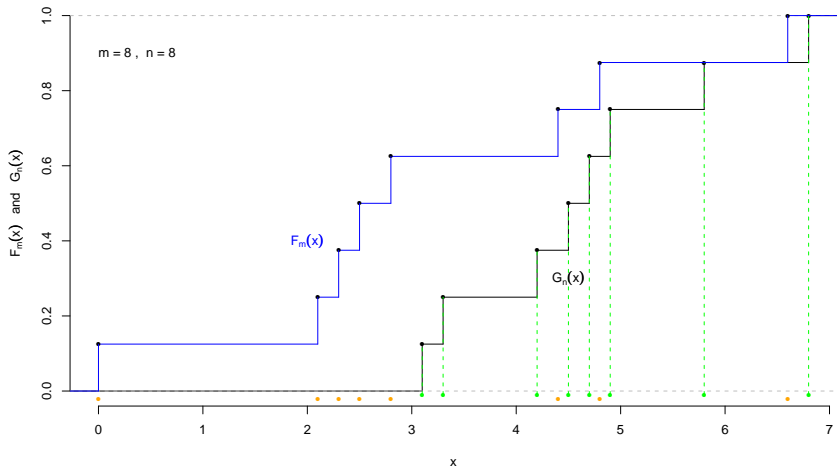
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$nG_n$ :	1	2	3	4	5	5	5	5	6	6	6	7	7	7	8	8
$nF_n$ :	0	0	0	0	0	1	2	3	3	4	5	5	6	7	7	8

- The maximum  $|F_n(x) - G_n(x)|$  is  $5/8$  and occurs at position 5.
- The  $p$ -value of the observed value  $5/8$  for  $D_{8,8}$  can be computed by the previous formula exactly as

$$P_{H_0}(D_{8,8} \geq 5/8) = \frac{2 \left[ \binom{16}{3} \right]}{\binom{16}{8}} = \frac{2 \cdot 560}{12870} = 0.0870$$

- see also Table 7E

# Visual EDF Comparison for Pain Relief Drugs



- As  $m \rightarrow \infty$  and  $n \rightarrow \infty$  we have

$$P \left( \sqrt{\frac{mn}{m+n}} D_{m,n} \geq z \right) \rightarrow K(z)$$

where

$$K(z) = 2 \sum_{j=1}^{\infty} (-1)^{j+1} \exp(-2j^2 z^2) \quad \text{for } z > 0.$$

- This alternating series converges rapidly when  $z$  is not too close to 0, in which case  $K(z)$  approaches 1 anyway, a situation of little interest for  $p$ -values.
- The function `KS.Kfun` on the class web site calculates  $K(z)$ .

# Comments on Approximation Quality

- The approximation quality is easily examined when  $m = n$ , because of the rapid calculation of the Gnedenko-Korolyuk formula.
- It turns out that the approximation tends to be best when  $m = n$ , especially for the small tail probabilities relevant for typical  $p$ -value assessment.
- For very small  $p$ -values the approximation will give somewhat higher  $p$ -values. This leads to more conservative assessments of statistical significance.
- When  $m \neq n$  the exact null distribution requires the evaluation of  $D_{m,n}$  for the full set of  $\binom{m+n}{n}$  sample splits.
- One can always estimate it using  $N_{\text{sim}}$  simulated sample splits.
- When  $m \neq n$  the asymptotic approximation quality is not as good, but deteriorating, and is even more on the conservative side for statistical significance assessment.

## Example Approximation Calculation

- For the previous example of two drugs for pain relief we will show the process of getting a  $p$ -value based on the large sample approximation.

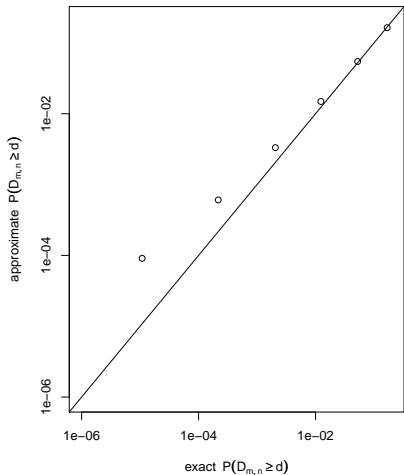
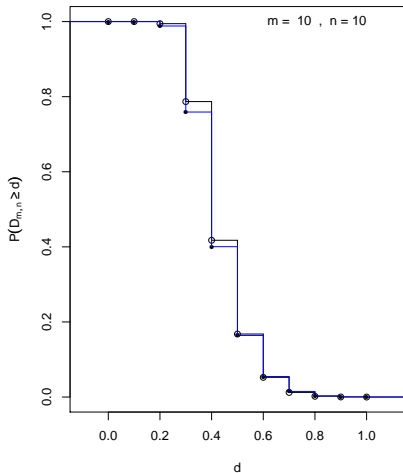
$$\begin{aligned}P(D_{8,8} \geq 5/8) &= P\left(\sqrt{\frac{8 \cdot 8}{16}} D_{8,8} \geq \sqrt{\frac{8 \cdot 8}{16}} \frac{5}{8}\right) \\ &= P(2D_{8,8} \geq 5/4) = .0879\end{aligned}$$

- from Table F or 0.08786641 using `KS.Kfun`, which is reasonably close to .0870.
- Note the conservative nature of the  $p$ -values ( $.0870 < .0879$ ) when using the large sample approximation.

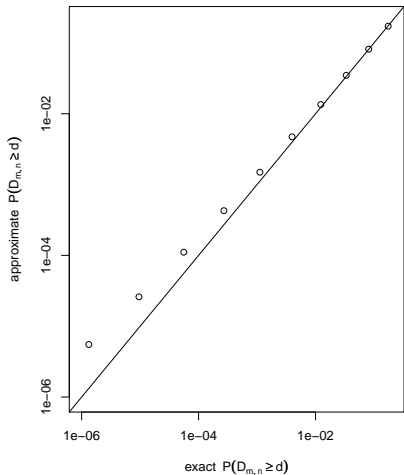
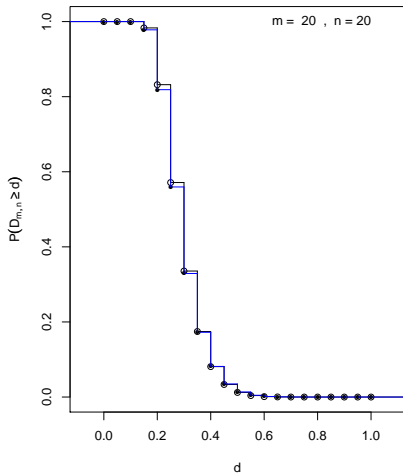
# Approximation Quality (via KSapprox)

- The following slides (obtained via `KSapprox`, class web site) illustrate the quality of the large sample approximation.
- When  $m = n$  we use the Gnedenko-Korolyuk formula.
- When  $m \neq n$  and  $\binom{m+n}{m} \leq 50,000$  we use the exact distribution of the KS-statistic, obtained by full enumeration via `combn`.
- When  $m \neq n$  and  $\binom{m+n}{m} > 50,000$  we use simulation, with  $N_{\text{sim}} = 50,000$ , to estimate the exact distribution of the KS-statistic.
- In the left plot the blue stepfunction shows the approximation to  $P(D_{m,n} \geq d)$  (in black), either exact or by simulation.
- The comparison of the exact (simulated) and approximate  $p$ -values in the right plot is always shown over the interval  $[10^{-6}, .2]$ , i.e., smaller or larger  $p$ -values are clipped.

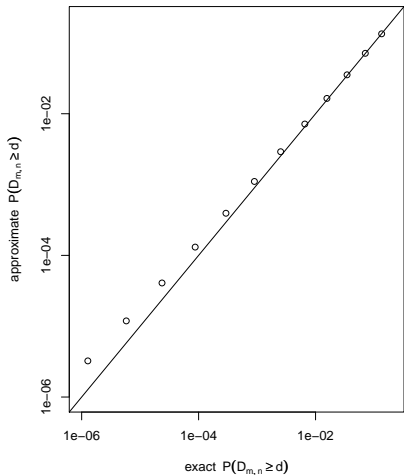
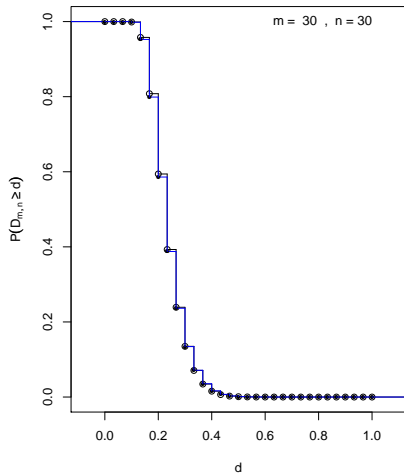
# Approximation Quality $m = 10$ and $n = 10$



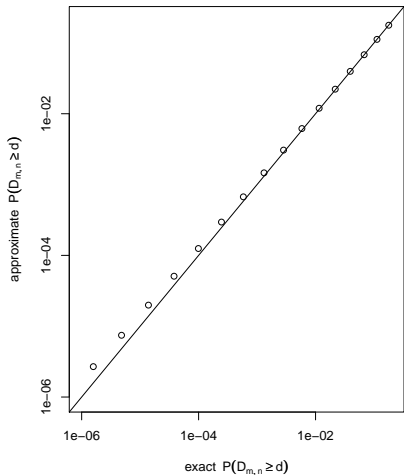
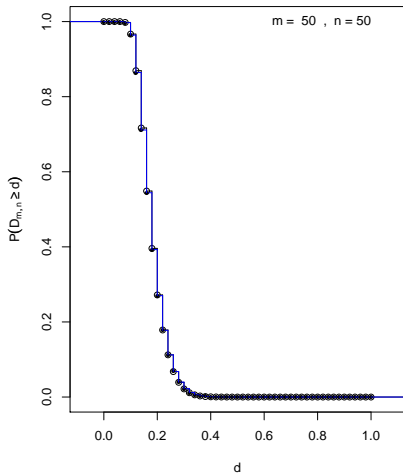
# Approximation Quality $m = 20$ and $n = 20$



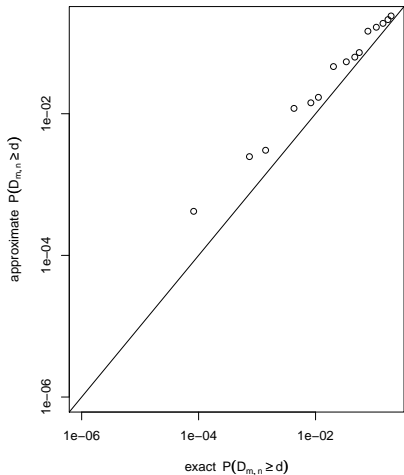
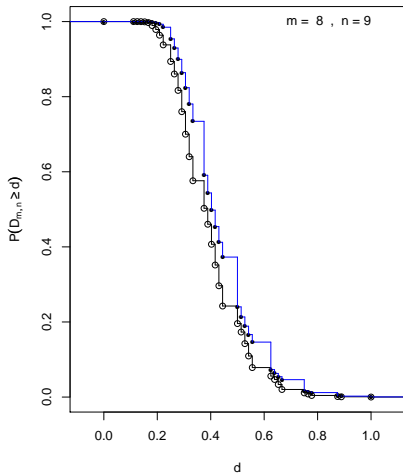
# Approximation Quality $m = 30$ and $n = 30$



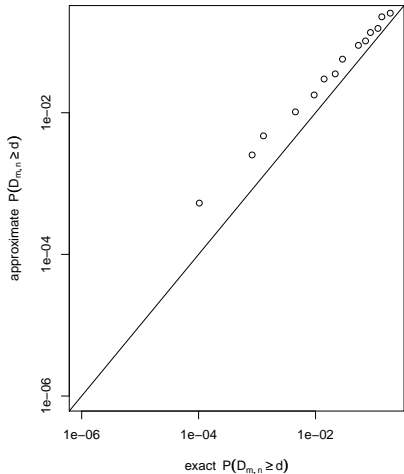
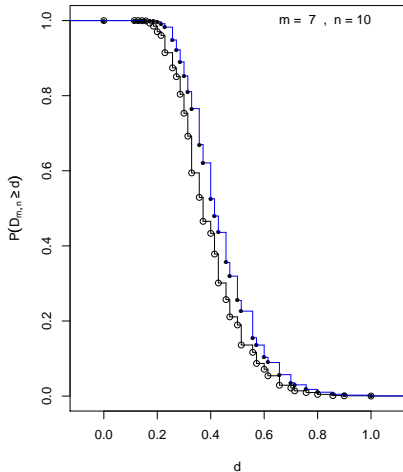
# Approximation Quality $m = 50$ and $n = 50$



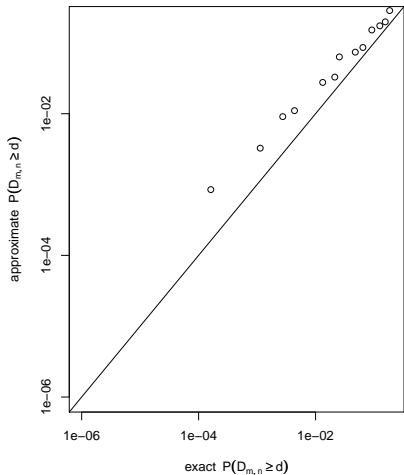
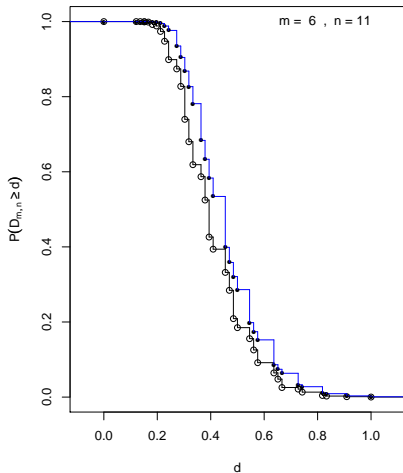
# Approximation Quality $m = 8$ and $n = 9$



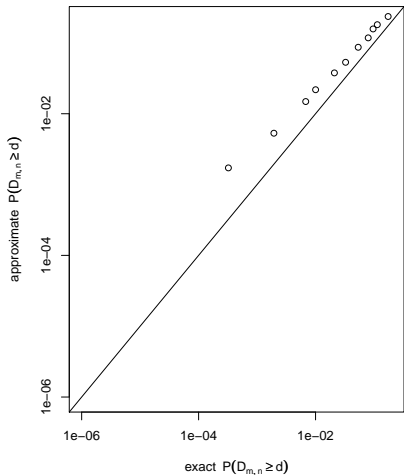
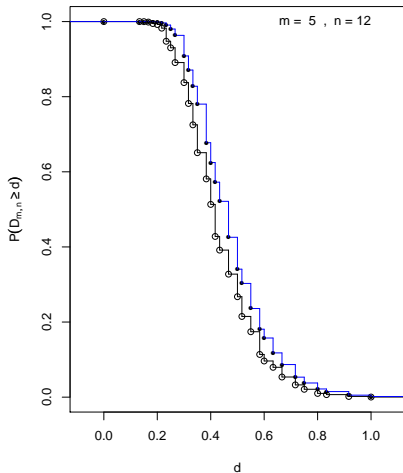
# Approximation Quality $m = 7$ and $n = 10$



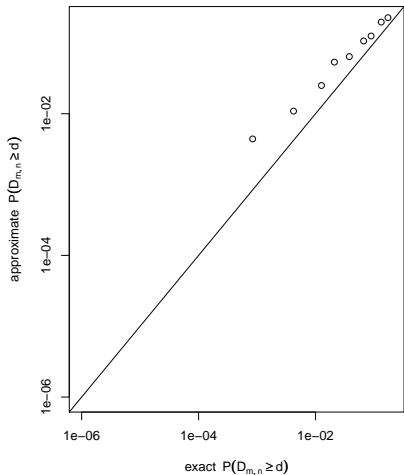
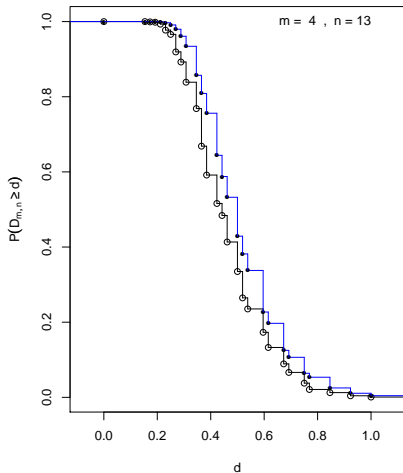
# Approximation Quality $m = 6$ and $n = 11$



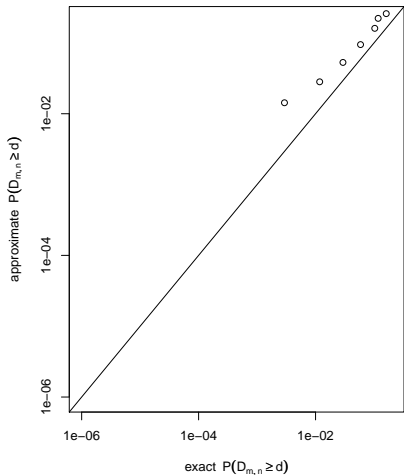
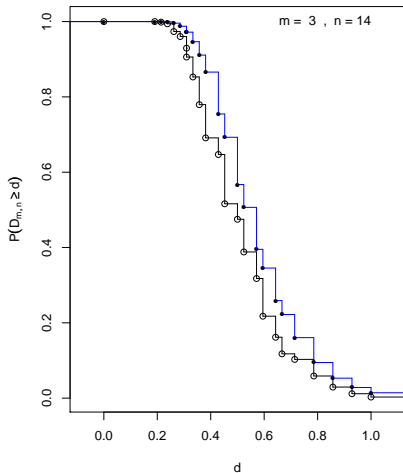
# Approximation Quality $m = 5$ and $n = 12$



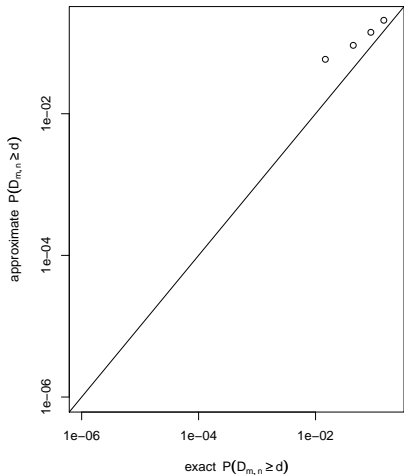
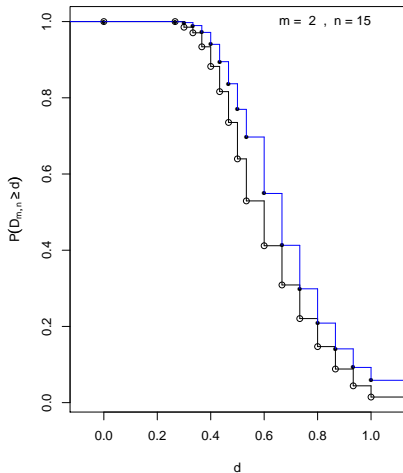
# Approximation Quality $m = 4$ and $n = 13$



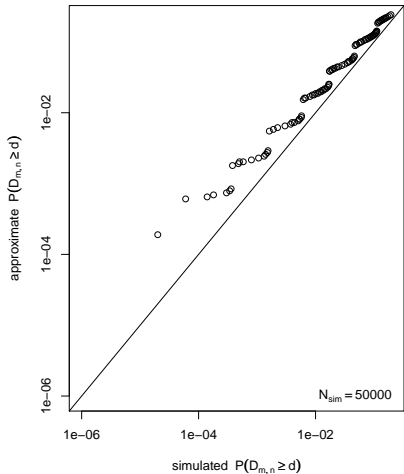
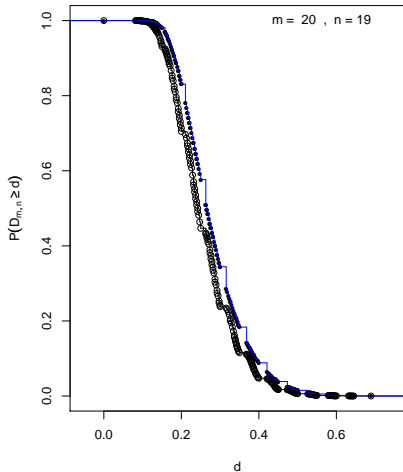
# Approximation Quality $m = 3$ and $n = 14$



# Approximation Quality $m = 2$ and $n = 15$



# Approximation Quality $m = 20$ and $n = 19$



- Aside from writing your own function for evaluating  $D_{m,n}$  and either evaluating it via `combn` for all pooled sample splits of  $m$  and  $n$  observations, or
- alternately sampling such splits  $N_{\text{sim}}$  times, or
- using the large sample approximation  $K(z)$  for large  $m$  and  $n$ ,
- we can also use the R function `ks.test`.
- For details on using it see its documentation.

## ks.test for Pain Relief Example

```
> x=c(3.1,3.3,4.2,4.5,4.7,4.9,5.8,6.8)
> y=c(0.0,2.1,2.3,2.5,2.8,4.4,4.8,6.6)
> ks.test(x,y)
```

Two-sample Kolmogorov-Smirnov test

```
data:  x and y
D = 0.625, p-value = 0.08702
alternative hypothesis: two-sided
```

- $F_m(x)$ ,  $G_n(x)$ ,  $D_{m,n} = \max_x |G_n(x) - F_m(x)|$  remain the same.
- The tabled null distributions (Table E) no longer apply nor does the Gnedenko-Korolyuk formula.
- The large sample approximation no longer applies.
- If manageable, we can still use full enumeration of all splits and evaluate  $D_{m,n}$  for each such split, thus getting the exact null distribution of  $D_{m,n}$ .
- Or we can get  $N_{\text{sim}}$  independent sample splits, evaluate  $D_{m,n}$  for each such split, thus getting an estimate of the exact null distribution of  $D_{m,n}$ .

# The KS Test is a Midrank Test in the Case of Ties

- $D_{m,n}$  is a function of the midranks  $R_1^* \leq \dots \leq R_m^*$  of the  $X$ 's.
- Let  $Z_{(1)} \leq \dots \leq Z_{(N)}$  be the ordered pooled sample values ( $N = m + n$ ) and denote by  $Q_1^* \leq \dots \leq Q_N^*$  the corresponding fixed set of midranks (in case of ties).

$$D_{m,n} = \max_{1 \leq j \leq N} |F_m(Z_{(j)}) - G_n(Z_{(j)})|$$

- Let  $R_1^* \leq \dots \leq R_m^*$  be the ordered midranks of the  $X$ 's. Then

$$\begin{aligned} F_m(Z_{(j)}) &= \frac{\#\{X\text{'s} \leq Z_{(j)}\}}{m} = \frac{\#\{R^*\text{'s} \leq Q_j^*\}}{m} \\ G_n(Z_{(j)}) &= \frac{\#\{Y\text{'s} \leq Z_{(j)}\}}{n} = \frac{\#\{Z\text{'s} \leq Z_{(j)}\} - \#\{X\text{'s} \leq Z_{(j)}\}}{n} \\ &= \frac{\#\{Q^*\text{'s} \leq Q_j^*\} - \#\{R^*\text{'s} \leq Q_j^*\}}{n} \end{aligned}$$

- Thus  $D_{m,n}$  depends only on the set of midranks  $R_1^* \leq \dots \leq R_m^*$  which varies from split to split while  $Q_1^* \leq \dots \leq Q_N^*$  stays fixed. Now write  $D_{m,n}^*$  in place of  $D_{m,n}$ .

$$P_{H_0}(D_{m,n}^* \geq d) \leq P_{H_0}(D_{m,n} \geq d)$$

- We refer to the Text for a nice proof of the above inequality.
- We only discuss the implications.
- Use the inequality to obtain an upper bound for the  $p$ -value observed under ties by just using the corresponding  $p$ -value for  $(m, n)$  without ties as the upper bound.
- Thus if the upper bound is statistically significant ( $\leq \alpha$ ) then that also holds for the actual case with ties.
- However, the actual  $p$ -value might also be substantially less than  $\alpha$ , without us knowing.
- We will be rejecting  $H_0$  less often, whether  $H_0$  is true or not.
- $\implies$  less power, chance of rejecting  $H_0$  when  $H_0$  is false.

# Nonparametric vs Parametric Statistics

- The distinction often arises through the assumed model of the sampled population.
- A population can be characterized by a parametric family of distributions, like the family of normal distributions
$$\mathcal{F}_{\text{normal}} = \{\mathcal{N}(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma > 0\}$$
- “parametric” since parameters index individual members of  $\mathcal{F}$ .
- A nonparametric family is a much wider family of distributions.
- For example, consider the family  $\mathcal{F}_{\text{continuous}}$  of all distributions with continuous cdfs  $F$ .
- Here  $\mathcal{F}_{\text{normal}}$  is just a very thin sliver or subset of  $\mathcal{F}_{\text{continuous}}$ .