University of Washington

STATISTICS

Stat 425

Introduction to Nonparametric Statistics

Blocked Comparisons for Two Treatments

Fritz Scholz

Spring Quarter 2009[*]

[*] May 20, 2009

Paired Comparisons

So far we compared treatment and control over two subject groups or samples.

Sometimes the background variation among study subjects or within the samples is quite large and it becomes difficult to detect treatment effects.

We dealt with this already in Example 3 (Cultural Influences on IQ).

The effectiveness of treatment/control comparison can be increased if we make such comparisons separately within several homogeneous subgroups.

The responses from subgroup to subgroup are allowed to vary substantially.

The smallest type of subgroup consists of two subjects.

Natural Subgroups of Size Two

Twins or identical twins are good homogeneous subgroups of size two.

The feet, legs, hands or eyes of the human body are natural “twins” or pairs.

Subject can serve as control and treatment by applying both in some random order.

We can also create pairs of subjects by matching many of their background characteristics, e.g., age, sex, health history, severity of disease under study, geographic region, community size, etc.

Randomization Model for Paired Comparisons

We have N paired subjects (homogeneous within, possibly quite variable between).

Assign the treatment at random, i.e., with probability $1/2$, to one subject in each pair and use the other as a control.

At this point we treat the subjects as given and not as a random selection from some population.

Randomness only enters through the treatment assignment coin flip.

If a subject serves both for treatment and control, then the order is randomized.

This is called the randomization model for paired comparisons.

A corresponding population version of this model is discussed in Chapter 4.

The Sign Test

Let S_N be the number of pairs in which the response of the treated subject is “better” than the response obtained for the corresponding control subject.

The hypothesis H_0 of no difference between treatment and control is rejected whenever S_N is too large.

Often the designation of “better” is based on the difference of some score under treatment and control, and a positive difference is interpreted as better.

That’s is the reason for calling this test the sign test.

At this points we assume just $+$ and $-$ judgments. Zeros will be addressed later.

Null Distribution of the Sign Test

Under the hypothesis H_0 we perform N independent coin flips.

The responses of both subjects within any given pair would be the same, regardless of the assignment of treatment and control.

Thus S_N has a binomial distribution with parameters N and $p = 1/2$, i.e.,

$$P_{H_0}(S_N = k) = \binom{N}{k}\frac{1}{2^N} \qquad \text{for} \quad k = 0, 1, 2, \ldots, N.$$

It is easy to find p-values for any observed value s_N of S_N, i.e.,

$$P_{H_0}(S_N \geq s_N) = \sum_{i=s_N}^{N} \binom{N}{i}\frac{1}{2^N}$$

Working with p-values to judge level α significance is more effective than using the critical points available for a finite number α's.

Example 1: A Headache Remedy

15 subjects are each given two two bottles, one containing the standard drug for tension headache and the other the new treatment.

Each bottle had been randomly labeled A and B. Only the experimenter knows whether A or B corresponds to the new treatment for any given patient.

The patients are asked to alternate taking pills from the two bottles at the onset of any tension headache and to make an overall judgment as to which bottle generally gave better results, A or B.

It turned out that 10 of 15 subjects preferred the new drug.

$\Longrightarrow$ p-value $= P_{H_0}(S_{15} \geq 10) = 1 - P_{H_0}(S_{15} \leq 9) =$`1-pbinom(9,15,.5)=0.15088`

Large Sample Approximation

Under H_0 : no difference ($\Longrightarrow p = .5$) the mean and variance of S_N are

$$E_{H_0}(S_N) = Np = \frac{N}{2} \qquad \text{and} \qquad \text{var}_{H_0}(S_N) = Np(1-p) = \frac{N}{4}\,.$$

For large N the null distribution of S_N is well approximated by a normal distribution (the quality of the normal approximation to Binomial(N, p) is best when $p = .5$)

$$\frac{S_N - N/2}{\sqrt{N/4}} \longrightarrow \mathcal{N}(0,1)$$

Again, the continuity correction greatly improves the approximation.

For our previous example we get

$$P_{H_0}(S_N \geq 10) = P_{H_0}\left(\frac{S_N - N/2}{\sqrt{N/4}} \geq \frac{9.5-7.5}{\sqrt{15/4}}\right) = \Phi\left(\frac{-2}{1.936492}\right) = 0.15085$$

remarkably close to 0.15088.

Zero Differences

So far we acted as though we will have just $+$ and $-$ responses.

However, in some cases we may get a zero or a tie,

neither control nor treatment comes out ahead.

In such situations we will have to track three counts N_+, N_0, and N_-.

If we attach scores of $1, \frac{1}{2}$, and 0 to each subject pair response and then proceed analogous to our tie version of the Mann-Whitney statistic W^*_{XY}, we would take $N_+ + N_0/2$ as our test statistic.

Note that under H_0 the count N_0 will always be fixed, since the nature of a tie is preordained under H_0. Thus $N_+ + N_0/2$ and N_+ are equivalent test statistics. Under H_0 we have $N_+ \sim \text{Binom}(N - N_0, 1/2)$ and we reject H_0 for large N_+. We discard the tie subjects and work with fewer subjects, suffering a loss of power.

Using More Than Just the Signs

The sign test only uses the signs of differences.

A more effective comparison would use the magnitudes of the differences as well.

Example 2: Testing a new fertilizer

Three strawberry fields are divided into two parts each. Two fertilizers, a new one and the traditional one, are randomly assigned to one of the parts in each field.

The yields on the fields are: 76 and 78 lb for field 1, 82 and 91 lb for field 2, and 80 and 86 lb for field 3, with corresponding differences of 2, 9 and 6 lb.

The sign test uses only the signs of the differences based on fertilizer assignments.

Testing H_0: no fertilizer difference, two $+$ signs would seem to present a stronger case against H_0 when the $+$ signs go with 6 and 9 rather than with 2 and 6.

Signed Differences and Signed Ranks

signed differences	$-2, -6, -9$	$-2, -6, +9$	$-2, +6, -9$	$-2, +6, +9$
signed ranks	$-1, -2, -3$	$-1, -2, +3$	$-1, +2, -3$	$-1, +2, +3$
signed differences	$+2, -6, -9$	$+2, -6, +9$	$+2, +6, -9$	$+2, +6, +9$
signed ranks	$+1, -2, -3$	$+1, -2, +3$	$+1, +2, -3$	$+1, +2, +3$

Under H_0 all yields would be the same no matter which fertilizer was used.

The randomized ferilizer assignment simply determines whether we evaluate the yields of 76 and 78 lb on the first field as $76 - 78 = -2$ lb or $78 - 76 = 2$ lb, and similarly for the other fields, i.e., -9 or $+9$ and -6 or $+6$.

There $2 \times 2 \times 2 = 8$ possible fertilizer assignments, i.e., sign combinations, and thus signed rank sets. This gives us the null distribution of the signed rank sets.

Signed Ranks in General

The previous, deliberately simple example easily generalizes to N pairs of subjects.

We have N differences of scores, treatment score $-$ control score.

$N_+ = n$ with a $+$ sign and $N_- = m$ with a $-$ sign, with $N_+ + N_- = N$.

The absolute differences are ranked $1, 2, \ldots, N$. We assume no ties among absolute ranks and also no zero differences. We will revisit that issue later.

The ranks corresponding to the $+$ signs are denoted by $S_1 < \ldots < S_n$ and those corresponding to $-$ signs are $R_1 < \ldots < R_m$.

There are 2^N possible sign combinations $\pm 1, \pm 2, \ldots, \pm N$.

There is a one-to-one correspondence between sign combinations and ranks sets $(S_1, \ldots, S_n)$ (including also the empty set, corresponding to N $-$ signs).

Binomial Identity and 1-1 Correspondence

For any given n there are $\binom{N}{n}$ ways to select any ordered subset $S_1 < \ldots < S_n$ from $1, 2, \ldots, N$.

This covers $n = 0$ as well, since $\binom{N}{0} = 1$ and there is just one way to choose $n = 0$ ranks from $1, 2, \ldots, N$, just as there is $\binom{N}{N} = 1$ way to choose all when $n = N$.

Our previous equivalence is just a reformulation of the following binomial identity

$$2^N = (1+1)^N = \binom{N}{0} + \binom{N}{1} + \ldots + \binom{N}{N-1} + \binom{N}{N}$$

which in more general form is written as

$$\begin{aligned}(x+y)^N &= \binom{N}{0} x^0 y^{N-0} + \binom{N}{1} x^1 y^{N-1} + \ldots + \binom{N}{N-1} x^{N-1} y^1 + \binom{N}{N} x^N y^0 \\ &= \sum_{i=0}^{N} \binom{N}{i} x^i y^{N-i}\end{aligned}$$

The Wilcoxon Signed Rank Test

Under H_0 each sign vector has same chance $1/2^N$ of occurring.

Another way to express this is by

$$P_{H_0}(N_+ = n, S_1 = s_1, \ldots, S_n = s_n) = \frac{1}{2^N}$$

If we test H_0 against the alternative of a beneficial treatment effect we would consider a lot of positive score differences and in addition high positive differences as strong evidence against H_0.

These two criteria can be combined in the following signed-rank test statistic

$$V_s = S_1 + \ldots + S_n \qquad \text{with } V_s = 0 \text{ when } n = 0.$$

We reject H_0 when $V_s \geq c$. This is the Wilcoxon signed-rank test.

Wilcoxon Rank-Sum and Signed-Rank Tests

We note the formal similarity in the two Wilcoxon type tests

$$W_s = S_1 + \ldots + S_n \geq c \qquad \text{and} \qquad V_s = S_1 + \ldots + S_n \geq c$$

In the first (rank-sum) test the S_i are the ranks (among $1, 2, \ldots, N$) of the treated subjects scores. In that case n is fixed a priori.

In the second (signed rank) test the S_i are the ranks of the positive signed ranks, among $\pm 1, \ldots, \pm N$. In that case n is random and can take any value from $0, 1, \ldots, N$.

The two rank statistics W_s and V_s have quite different null distributions and thus different critical points c for any given α.

When no confusion is possible we refer to the Wilcoxon signed-rank test simply as the Wilcoxon test, just as we did in discussing the Wilcoxon rank-sum test.

The Null Distribution of V_s

We illustrate the calculation of the null distribution of V_s by using the previous fertilizer example. We simply compute V_s for each realization of signed ranks and aggregate probabilities for common results.

signed ranks	v
$(-1,-2,-3)$	0
$(-1,-2,+3)$	3
$(-1,+2,-3)$	2
$(-1,+2,+3)$	5
$(+1,-2,-3)$	1
$(+1,-2,+3)$	4
$(+1,+2,-3)$	3
$(+1,+2,+3)$	6

v	$P_{H_0}(V_s = v)$
0	$1/8$
1	$1/8$
2	$1/8$
3	$2/8$
4	$1/8$
5	$1/8$
6	$1/8$

Another Example

The following table gives the tensile strengths of tape-closed and suture-closed wounds. These results were obtained on 10 rats, 40 days after incisions on the rats' backs were treated by tape or suture.

Rat	1	2	3	4	5	6	7	8	9	10
Tape	659	984	397	574	447	479	676	761	647	577
Suture	452	587	460	787	351	277	234	516	577	513
Difference	207	397	-63	-213	96	202	442	245	70	64
signed rank	6	9	-1	-7	4	5	10	8	3	2

Suppose we test the hypothesis H_0 of no differences in method against the alternative that tape-closed wounds tend to be stronger.

Computing Effort

If we were to repeat the full enumeration for all signed rank vectors as in the previous example, where $2^N = 2^3 = 8$, we would now face $2^N = 2^{10} = 1024$, a more substantial undertaking.

Since $V_s + V_r = N(N+1)/2$ and since there are only few negative signs, it is computationally more effective to use $V_r = R_1 + \ldots + R_m$ as test statistic and reject H_0 when V_r is too small, since that is equivalent to V_s being too large.

The observed value of V_r is $V_r = 1 + 7 = 8$ and we won't have to gather too many signed rank vectors to get the p-value or observed significance level for $v = 8$.

The Calculation

To obtain the p-value of $v = 8$ we need to list all m and rank vectors $r_1 < \ldots < r_m$ for which $V_r \leq 8$.

$m = 0$ empty set

$m = 1$ $r_1 = 1, 2, \ldots, 8$

$m = 2$ $(r_1, r_2) = (1,2), (1,3), \ldots, (1,7), (2,3), \ldots, (2,6), (3,4), (3,5)$

$m = 3$ $(r_1, r_2, r_3) = (1,2,3), (1,2,4), (1,2,5), (1,3,4)$

with a total number of signed-rank cases of $1 + 8 + (6 + 4 + 2) + (3 + 1) = 25$

$$\Longrightarrow \quad p\text{-value}(8) \;= P_{H_0}(V_r \leq 8) = \frac{25}{1024} = 0.0244$$

In comparison, the sign test would have given us a less significant p-value of

$$P_{H_0}(S_{10} \leq 2) = \left[\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right] \frac{1}{2^{10}} = \frac{1 + 10 + 45}{1024} = .0547$$

Sign Test & Signed-Rank Test

The Wilcoxon test is typically more powerful than the sign test.

The sign test statistic in the case $N = 3$ has just 4 possible values, namely $0, 1, 2, 3$.

The Wilcoxon test statistic for $N = 3$ had 7 possible values $0, 1, 2, \ldots, 6$.

This gives it more discriminatory possibilities and thus more power.

Why then use the sign test?

There are situations where only $+$ and $-$ can be obtained in paired comparisons.

Computational Issues

The two examples treated so far were either for

very small $N = 3$, where we gave the full null distribution of V_s, or

for moderate $N = 10$, where a full null distribution was not attempted manually, since $2^{10} = 1024$, but a p-value was easily managed by taking advantage of circumstances and organized listing of more extreme cases than the observed one.

2^N grows very rapidly, in fact more rapidly than $\binom{N}{n}$, see previous binomial identity.

R provides the function `psignrank` and other associates (→ documentation).

```
psignrank(8,10) = 0.02441406
```

I don't know its limitations. It may switch to a normal approximation.

V_s and V_r have the same distribution, symmetric around its mean $N(N+1)/4$, and with variance $\text{var}(V_s) = N(N+1)(2N+1)/24$.

Normal Approximation

$V_s = \sum_{i=1}^{N} i \cdot B_i$ with B_i independent Bernoulli random variables with $p = 1/2$

suggests a normal approximation should be reasonable for moderate or large N.

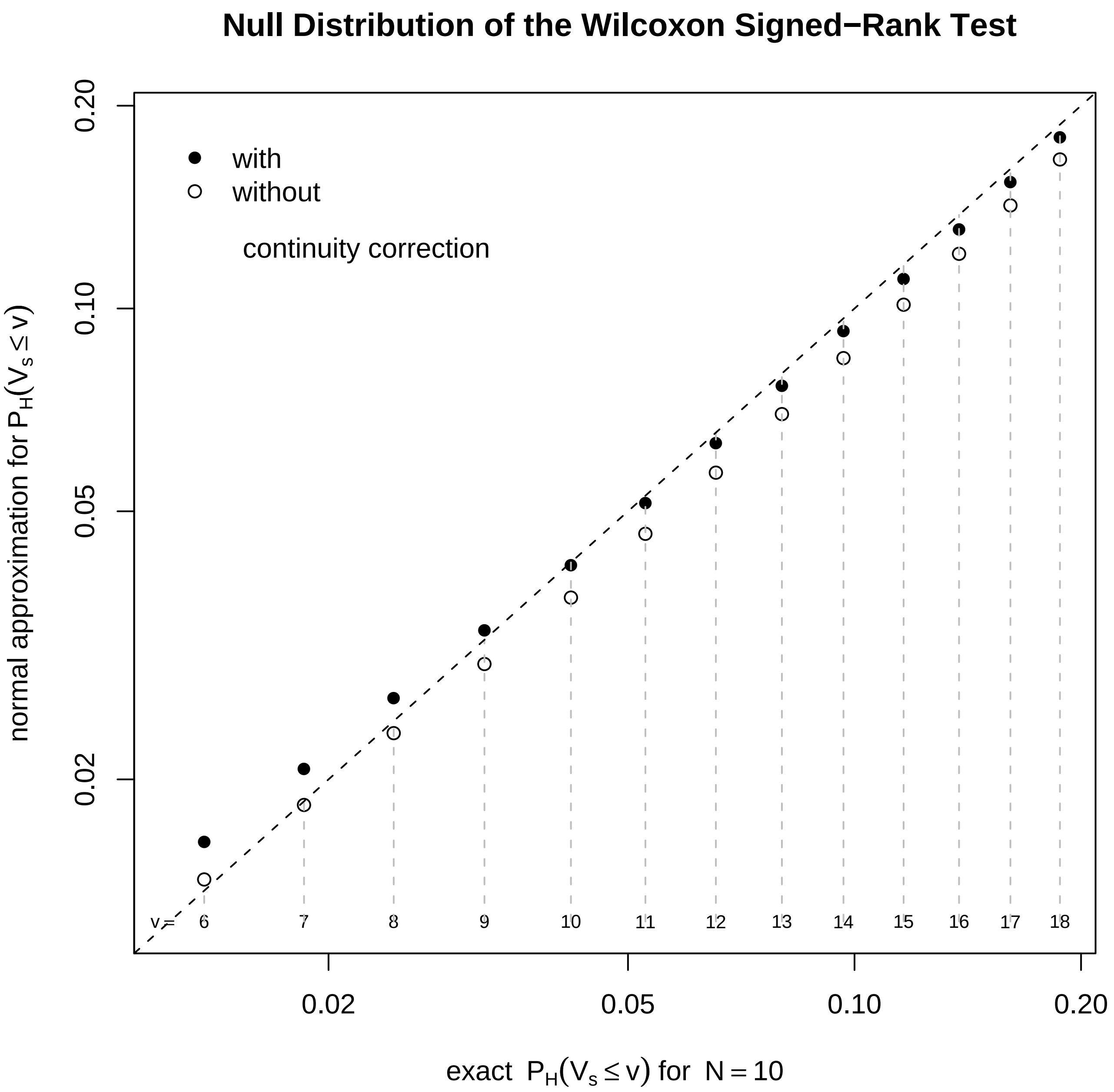
The above representation of V_s easily leads to the previous mean and variance formulas, using $\sum_{i=1}^{N} = N(N+1)/2$ and $\sum_{i=1}^{N} i^2 = N(N+1)(2N+1)/6$.

For our previous wound treatment example involving $N = 10$ rats we get

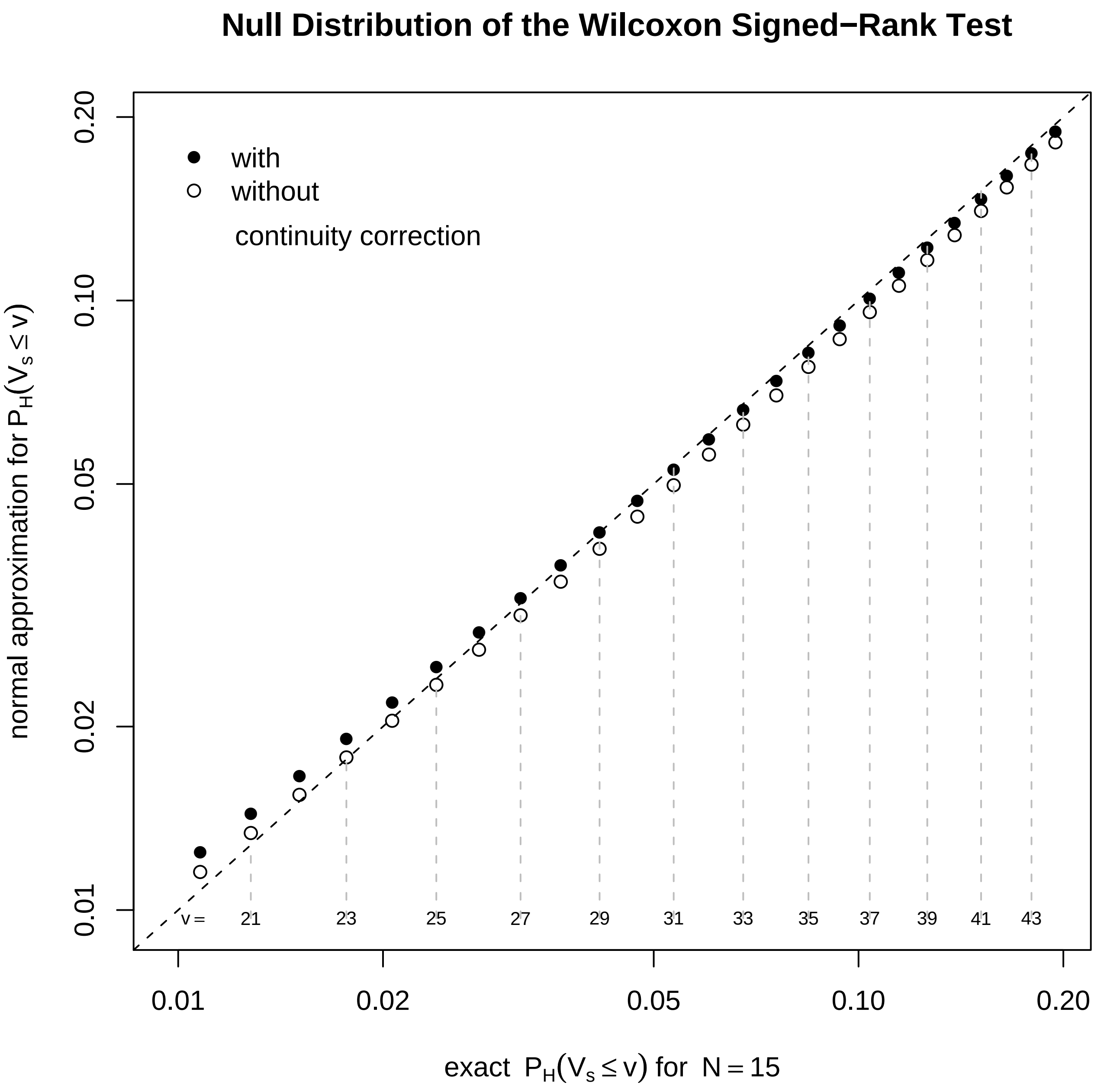
$$\begin{aligned} P_{H_0}(V_r \le 8) &= P_{H_0}\left(\frac{V_r - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \le \frac{8 + .5 - 10 \cdot 11/4}{\sqrt{10 \cdot 11 \cdot 21/24}}\right) \\ &\approx \Phi(-1.9367) = 0.0264 \qquad \text{(reasonably close to } .0244\text{)} \end{aligned}$$

Note again the use of the continuity correction $8 + .5$ in place of just 8.

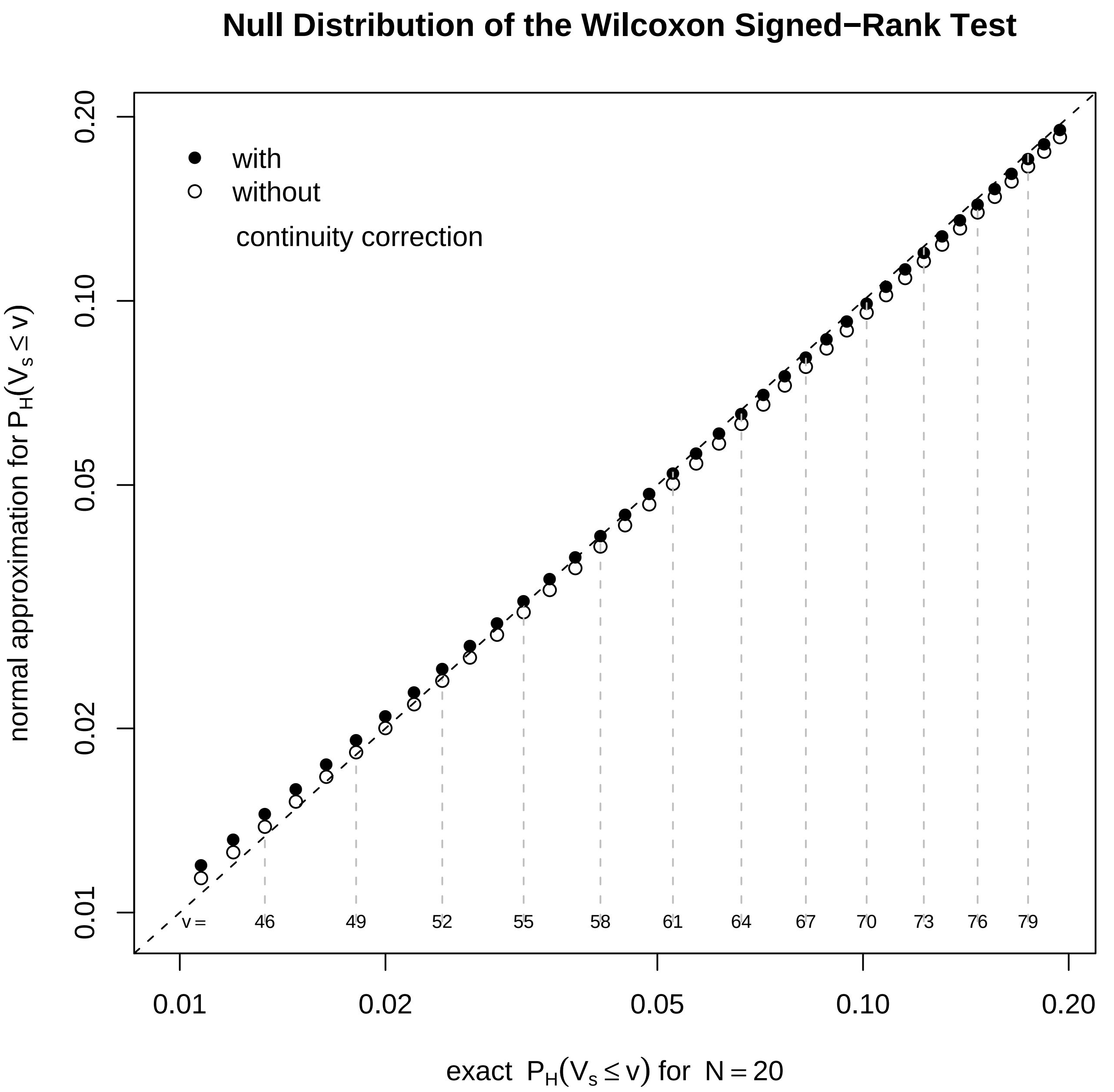
Approximation Quality $N = 10$



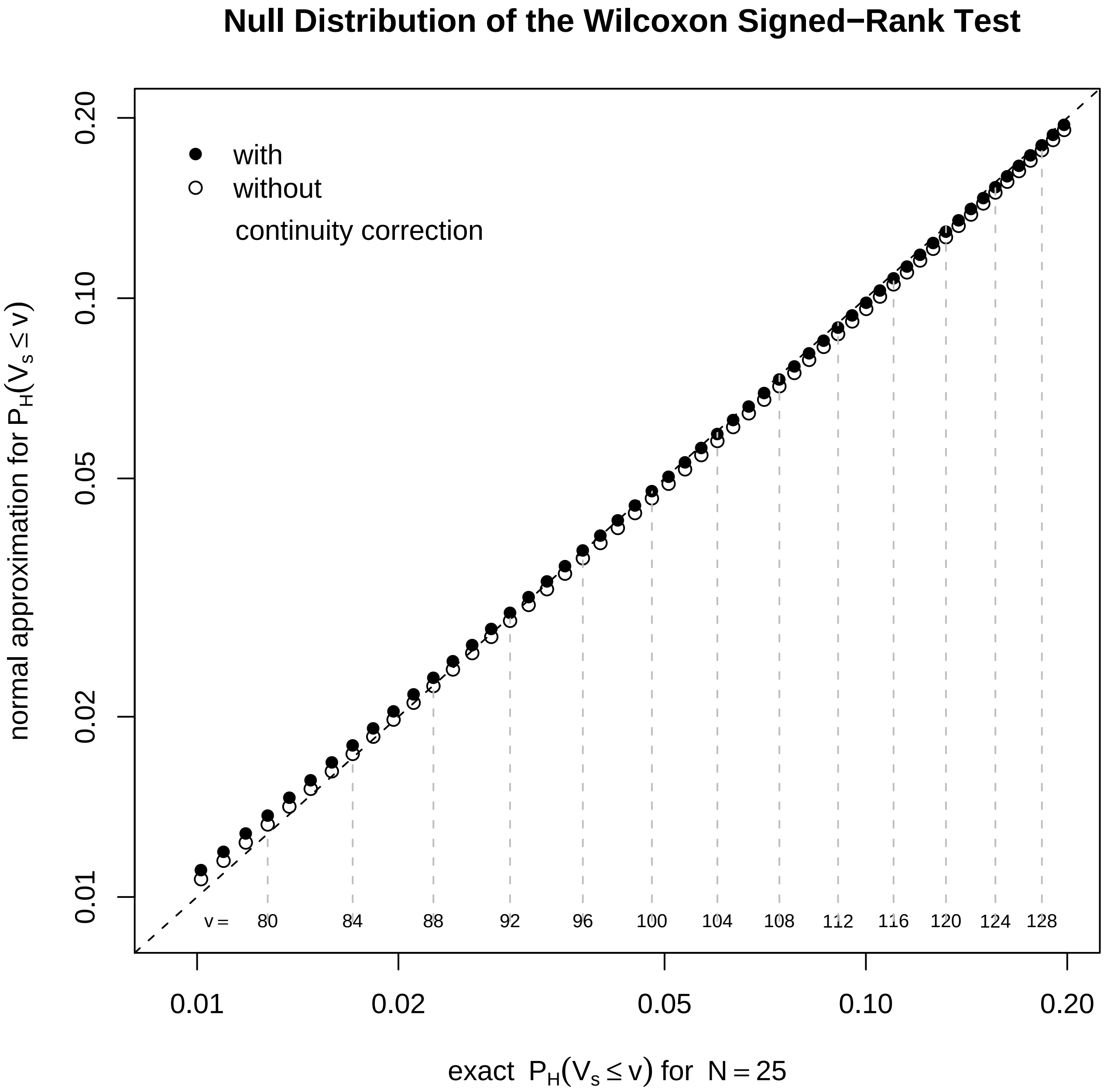
Approximation Quality $N = 15$



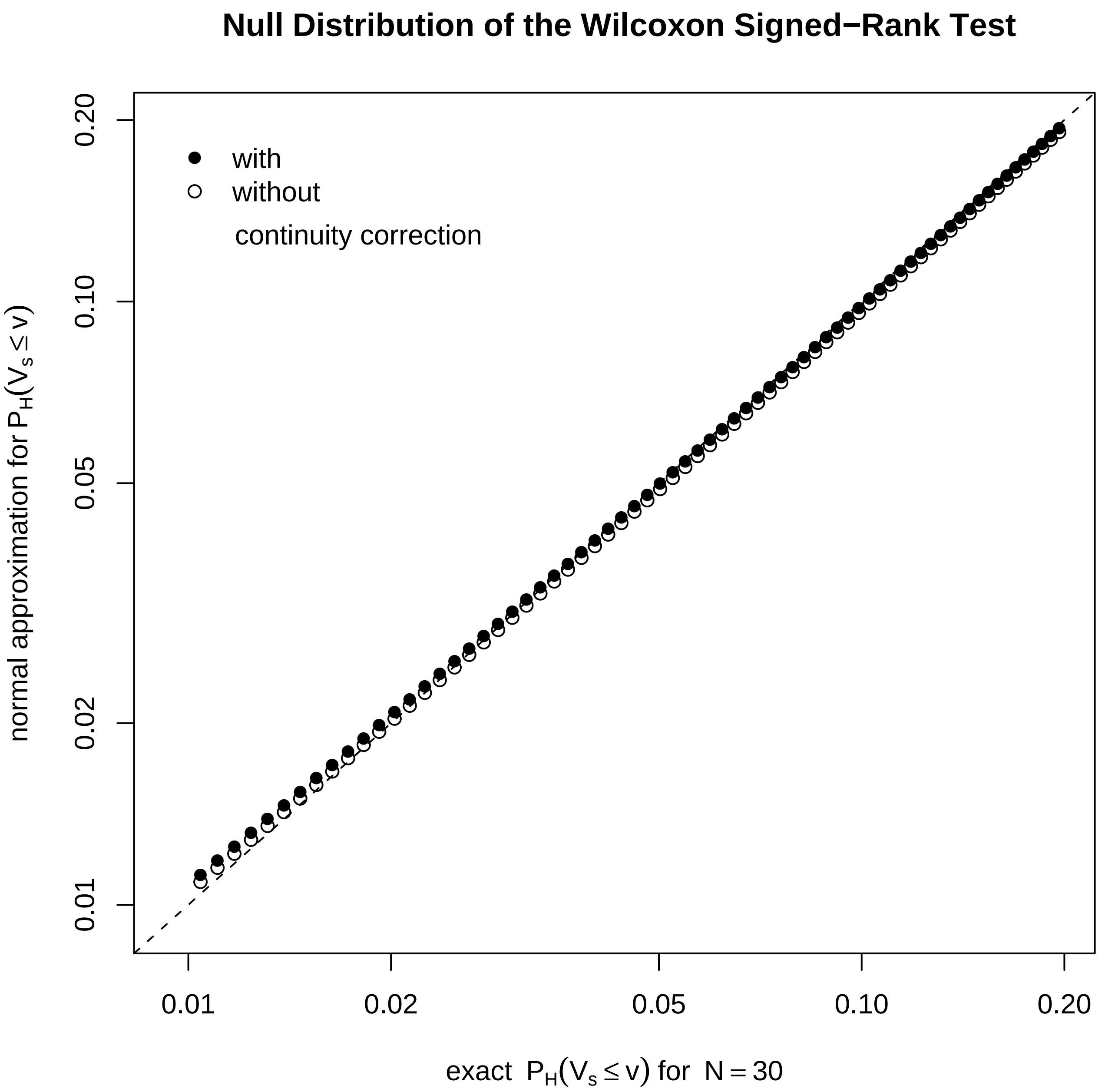
Approximation Quality $N = 20$



Approximation Quality $N = 25$



Approximation Quality $N = 30$



Comments on Approximation Quality

The previous plots compare left tail probabilities $p = P_{H_0}(V_s \leq v)$ over the range $[.01, .2]$ on a logarithmic scale.

The continuity correction appears to be better than the straight normal approximation for $p \geq .04$, for $p < .04$ the situation appears reversed.

However, the differences become smaller as N gets larger.

Since we have `psignrank` at our disposal, this is no great issue for us.

Using wilcoxsign_test

```
tape=c(659 , 984 , 397 , 574 , 447 , 479 , 676 , 761 , 647 , 577)
suture=c(452 , 587 , 460 , 787 , 351 , 277 , 234 , 516 , 577 , 513)
wilcoxsign_test(tape~suture,alter="greater",dist=exact())

Exact Wilcoxon-Signed-Rank Test

data:  y by x (neg, pos)
 stratified by block
Z = 1.9876, p-value = 0.02441
alternative hypothesis: true mu is greater than 0
```

Some Comments

Previously we computed the p-value via $P_0(V_r \leq 8)$ by careful enumeration or using $\texttt{psignrank}(8,10) = 0.02441$.

The previous slide rejects H_0 when V_s is too large.

Since $V_s + V_r = N(N+1)/2 = 55$ we have $\quad V_r \leq 8 \quad \Longleftrightarrow \quad V_s \geq 55 - 8 = 47$ with p-value $1 - \texttt{psignrank}(46,10) = 1 - 0.9756 = 0.02441$.

```
wilcoxsign_test(tape~suture,alter="less",dist=exact())
```

will give you a p-value of $P_0(V_r \leq 47) = 0.9814$ which includes $P_0(V_r = 47)$ just as $P_0(V_r \geq 47)$ did.

That is why the two p-values don't add to 1.

Alternative Interpretation of V_s

Let $Z_1, \ldots, Z_N$ denote the comparison differences for the N subjects.

Consider the $\binom{N}{2} + N$ averages $(Z_i + Z_j)/2$ with $i \le j$. Then

$$V_s = \text{number of positive averages } (Z_i + Z_j)/2 \text{ with } i \le j$$

Proof: Assume that the Z_i are indexed such that $0 < |Z_1| < |Z_2| < \ldots < |Z_N|$

$$\sum_{i \le j} I_{[Z_i + Z_j > 0]} = \sum_{j=1}^{N} \sum_{i=1}^{j} I_{[Z_i + Z_j > 0]} = \sum_{j=1}^{N} j \times I_{[Z_j > 0]} = V_s$$

where the indicator function I_B is 1 when B is true and 0 otherwise.

For the second $=$ in the above equation note that

$$Z_j > 0 \quad \Longrightarrow \quad \pm Z_i < Z_j = |Z_j| \text{ for } i \le j\,, \text{ i.e., } Z_i + Z_j > 0 \text{ for all } i \le j$$

$$Z_j < 0 \quad \Longrightarrow \quad \pm Z_i > Z_j = -|Z_j| \text{ for } i \le j\,, \text{ i.e., } Z_i + Z_j < 0 \text{ for all } i \le j \quad \square$$

Possibility of Ties and Zeros

Our previous treatment assumed that there are no ties among the absolute differences $|Z_i|$ and that no differences Z_i are zero.

Example: Suppose that we have $N = 7$ pairs of scores, each score on a scale of $-2, -1, 0, 1, 2$, corresponding to an assessment of very poor, poor, indifferent, good, and very good.

Suppose the 7 observed pairs (control,treatment) of scores are

$$(-1,0), \quad (-2,0), \quad (1,0), \quad (2,2), \quad (0,0), \quad (-1,1), \quad (0,0)$$

with corresponding score differences, treatment $-$ control, of $1,\ 2,\ -1,\ 0,\ 0,\ 2,\ 0$,

with absolute values in increasing order given by $0,\ 0,\ 0,\ 1,\ 1,\ 2,\ 2$

with midranks $2,\ 2,\ 2,\ 4.5,\ 4.5,\ 6.5,\ 6.5$.

The example covers all contingencies of violating our previous assumption:

we have zeros, even several zeros, we have tied absolute values,

some with same sign, some not.

Signed Midrank Statistic with Ties and Zeros

Multiplying each midrank by $+1$, -1, or 0, as the corresponding difference is positive, negative or zero, we get the signed midranks shown in the table below

Difference	-1	0	0	0	$+1$	$+2$	$+2$
Signed Midrank	-4.5	0	0	0	$+4.5$	$+6.5$	$+6.5$

The sum of positive signed midranks is then

$$V_s^* = 4.5 + 6.5 + 6.5 = 17.5$$

To assess the significance of the observed value 17.5 we need to get the null distribution of V_s^*.

The Null Distribution of V_s^* (Special Case)

Under H_0 the zeros are not affected by treatment/control assignments, thus we disregard them after midranking of the $|Z_i|$. Due to Pratt (1959) JASA 655-667.

$(s_1^*,\ldots,s_n^*)$	None	4.5	6.5	4.5,4.5	4.5, 6.5	6.5, 6.5	4.5, 4.5, 6.5	4.5, 6.5, 6.5	4.5, 4.5, 6.5, 6.5
V_s^*	0	4.5	6.5	9	11	13	15.5	17.5	22
Probability	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

For midranking after discarding zeros → Text under Further Developments 5B.

The null distributions are not necessarily equivalent.

If a priori large values of V_s^* are considered to be significant evidence against H_0, then we get as observed significance level or p-value from the above table

$$P_{H_0}(V_s^* \geq 17.5) = \frac{3}{16} = .1875$$

The Null Distribution of V_s^* (General Case)

It becomes quickly impossible to obtain the null distribution of V_s^* manually.

`psignrank` does not apply in such situations (zeros and ties).

We can use `combn` to accumulate all possible sums of positive signed midranks.

This may work when $M = N - d_0$ is not too large, where $d_0 =$ # of zero differences.

We can use

```
wilcoxsign_test(y1~x1,alter="greater",dist=exact(), ties = "Pratt")
```

where `y` is the vector of treatment scores and `x` is the vector of control scores.

We could estimate the null distribution by simulation, i.e., simulating multipliers ± 1 according to independent fair coin flips. `2*rbinom(10,1,.5)-1` produces such a vector of length `10`. The multipliers are used to create the signed rank statistics.

We can use a normal approximation for moderate or large N.

Normal Approximation for the V_s^* Null Distribution

The mean and variance of V_s^* under H_0 are given by

$$E_{H_0}(V_s^*) = \frac{N(N+1) - d_0(d_0+1)}{4}\,, \quad \text{where } d_0 = \text{number of zero differences,}$$

and

$$\mathrm{var}_{H_0}(V_s^*) = \frac{N(N+1)(2N+1) - d_0(d_0+1)(2d_0+1)}{24} - \frac{1}{48}\sum_{i=1}^{e} d_i(d_i-1)(d_i+1)$$

where $d_1, \ldots, d_e$ are the numbers of ties for each of the e distinct nonzero absolute differences.

A limit theorem shows that

$$\left(V_s^* - E_{H_0}(V_s^*)\right) \Big/ \sqrt{\mathrm{var}_{H_0}(V_s^*)} \quad \approx \quad \mathcal{N}(0,1) \qquad \text{as} \quad N - d_0 \longrightarrow \infty$$

Again we do not use a continuity correction in the approximation.

Symmetry of the V_s^* Null Distribution

The distribution of V_s^* is symmetric around its mean, i.e.,

$$V_s^* - E_{H_0}(V_s^*) \overset{\mathcal{D}}{=} -[V_s^* - E_{H_0}(V_s^*)] = [E_{H_0}(V_s^*) - V_s^*]$$

We will use the following representation of V_s^*

$$V_s^* = \sum_{i=1}^{M} B_i a_i \quad \Longrightarrow \quad E_{H_0}(V_s^*) = \frac{1}{2}\sum_{i=1}^{M} a_i \quad \text{and} \quad \mathrm{var}_{H_0}(V_s^*) = \frac{1}{4}\sum_{i=1}^{M} a_i^2$$

Here the B_i are independent Bernoulli random variables with $p = .5$ and the a_i are the remaining midranks of the absolute $M \leq N$ differences (after omission of those corresponding to zeros).

$$V_s^* = \sum_{i=1}^{M} B_i a_i \;\overset{\mathcal{D}}{=}\; \sum_{i=1}^{M} (1-B_i)a_i = \sum_{i=1}^{M} a_i - \sum_{i=1}^{M} B_i a_i = \sum_{i=1}^{M} a_i - V_s^*$$

$$V_s^* - \frac{1}{2}\sum_{i=1}^{M} a_i \;\overset{\mathcal{D}}{=}\; \frac{1}{2}\sum_{i=1}^{M} a_i - V_s^* \qquad \square$$

Example 4: Vitamin B and IQ

74 children in an orphanage were divided into 37 matched pairs, a randomly chosen child in each pair got the Vitamin B pill each day while the other got a placebo.

After 6 weeks the gains in IQ were obtained. The table shows results for 12 pairs

Pair	2	5	8	11	14	17	20	23	26	29	32	35
Treated	14	18	2	4	-5	14	-3	-1	1	6	3	3
Control	8	26	-7	-1	2	9	0	-4	13	3	3	4
Difference	6	-8	9	5	-7	5	-3	3	-12	3	0	-1
Signed Midrank	8	-10	11	6.5	-9	6.5	-4	4	-12	4	0	-2

with $V_s^* = 40$, $E_{H_0}(V_s^*) = 38.5$, $\text{var}_{H_0}(V_s^*) = 161.625$ and

$$p\text{-value}(40) = P_{H_0}(V_s^* \geq 40) = 1 - \Phi\left(\frac{40 - 38.5}{\sqrt{161.625}}\right) = .45304$$

$\Longrightarrow$ vitamin treatment clearly is not significant, at least not over this time span.

Example 4: Vitamin B and IQ (Simulation)

```
VitaminBsim=function(Nsim=10000){
y=c(14, 18, 2, 4, -5, 14, -3, -1, 1, 6, 3, 3)
x=c(8, 26, -7, -1, 2, 9, 0, -4, 13, 3, 3, 4)
dyx=y-x ; adyx=abs(dyx)
signyx=rep(0,length(dyx)) ; signyx[dyx>0]=1 ; signyx[dyx<0]=-1
rd=rank(adyx)*signyx
Vstar=sum(rd[rd>0])
adyxr=adyx[signyx!=0] ; Nr=length(adyxr)
rdr=abs(rd[signyx!=0])
Vvec=NULL
for(i in 1:Nsim){
   signyxr=rbinom(Nr,1,.5)*2-1
   Vvec[i]=sum(rdr[signyxr>0])
}
pval.sim=mean(Vvec>=Vstar)
pval.sim
}
```

Simulation Result for $N_{\text{sim}} = 100,000$

Running

$$\texttt{VitaminBsim}(\texttt{Nsim} = \texttt{100000})$$

yielded $46138/100000 =$`0.46138` (≈ 80 seconds)

A 99% confidence interval $(.4573, .4654)$ for the true p-value can be computed via

```
 > qbeta(.005,46138,100000+1-46138) # 99.5% lower bound
[1] 0.4573163
> qbeta(.995,46138+1,100000-46138) # 99.5% upper bound
[1] 0.4654473
```

This interval just misses the normal approximation value `.45304`.

This reflects on the quality of the normal approximation, not on the interval.

Another run gave `.46123` with 99% confidence interval $(0.4572, 0.4653) \not\ni .45304$.

Example 4: Vitamin B and IQ (Exact)

```
> y=c(14 , 18 , 2 , 4 , -5 , 14 , -3 , -1 , 1 , 6 , 3 , 3)
> x=c(8 , 26 , -7 , -1 , 2 , 9 , 0 , -4 , 13 , 3 , 3 , 4)
> wilcoxsign_test(y~x,alternative="greater",distr=exact())

Exact Wilcoxon-Signed-Rank Test

data:  y by x (neg, pos)
 stratified by block
Z = 0.0891, p-value = 0.4741
alternative hypothesis: true mu is greater than 0
> wilcoxsign_test(y~x,alternative="greater",distr=exact())
```

The exact value falls outside the previous confidence interval by an even larger margin. Given that it is exact, there is no excuse. What gives?

SignedRankExact

We can use combn to go through all 2^M possibilities of summing k of the midranks to form V_s^*, for $k = 0, 1, 2, \ldots, M$.
$M = N - d_0$ is the number of midranks left after those midranks corresponding to zero differences are deleted.

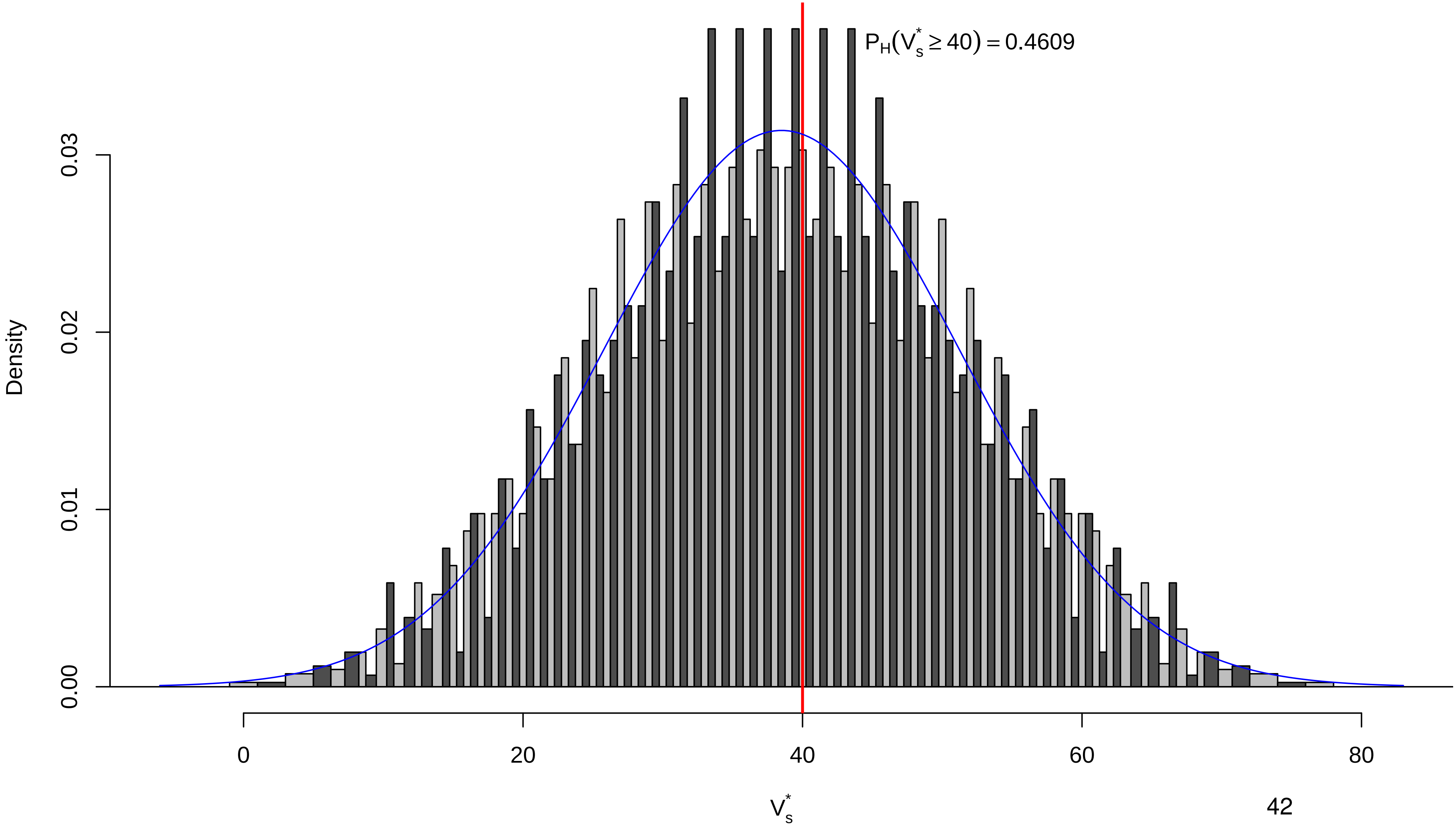
The R function SignedRankExact (⟶ class web site) computes the exact null distribution of V_s^*, returns its mean and standard deviation, its observed value $V_{s,\text{obs}}^*$, the corresponding Z value, and the p-value, assuming that Y will tend to be larger than X under the alternative.

For the opposite alternative reverse the roles of X and Y.

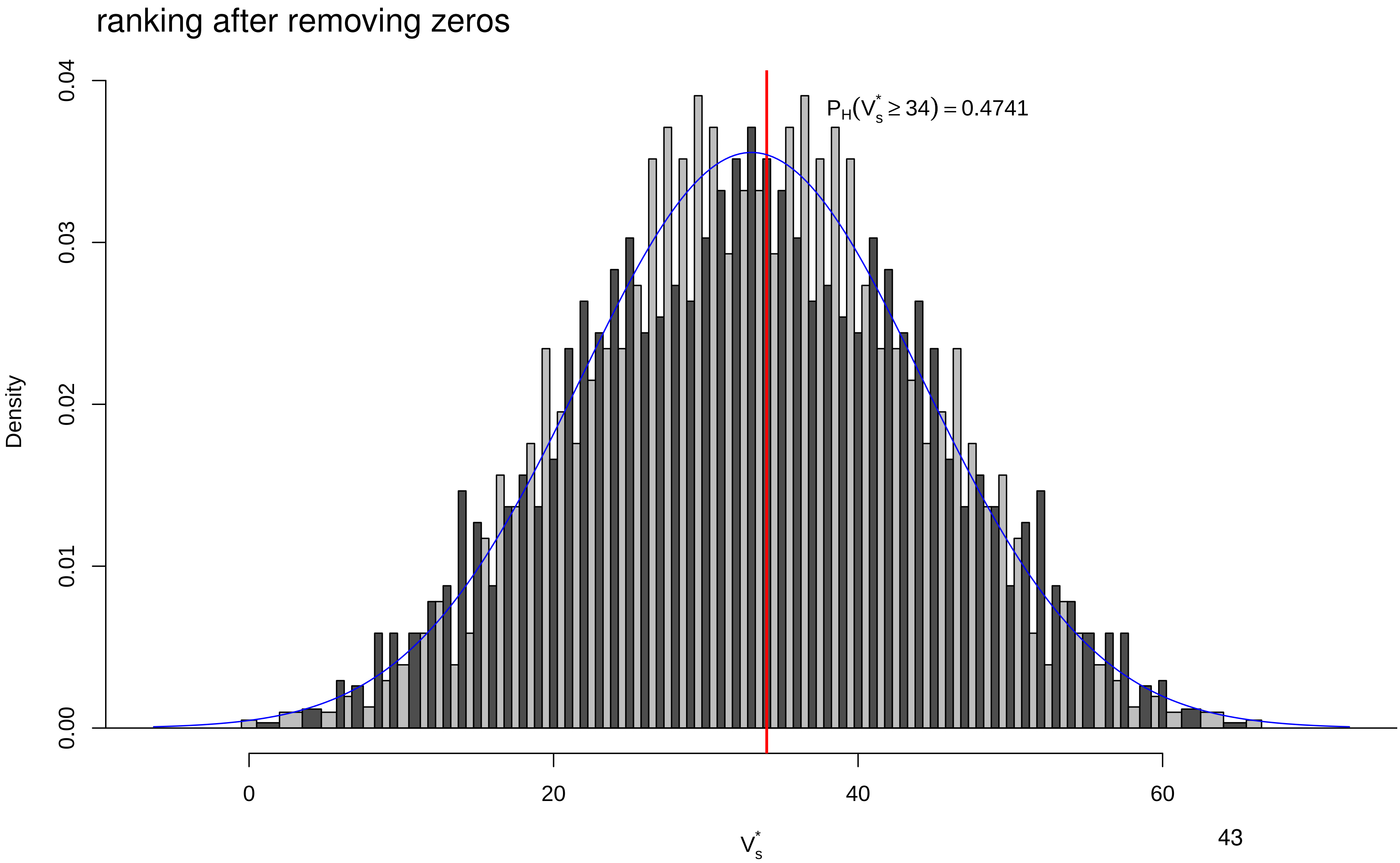
SignedRankExact has a flag argument. If flag=TRUE (FALSE) the ranking is done before (after) the removal of the zero cases. The next two plots show the results for flag=TRUE and flag=FALSE for the Vitamin B-IQ data.

Exact Null Distribution (Vitamin B-IQ)

remove zeros after ranking



Exact Null Distribution (Vitamin B-IQ)



Some Comments

It appears that `wilcoxsign_test` uses the `flag=FALSE` option.

This was the version of the test originally proposed by Wilcoxon.
It is also the choice in Hollander and Wolfe, Nonparametric Statistical Methods, (1999, 2nd edition). However, Pratt (1958) makes a strong case for using the version in the Text. This limits our use of `wilcoxsign_test` to the non-zero case.

I have probed `SignedRankExact` a little to find out how large a value of M could still be used, but I have not pushed it to the limit.

For $M = 20$ we have $2^M = 1,048,576$
It took about about 57 seconds on this old 900 MHz laptop with 500MB RAM.
A 1.73 GHz laptop with 2GB of RAM took 21 seconds. Both are on Unbuntu Linux.
For $M = 25$, i.e., $2^M = 33,554,432$, the latter took 555 seconds or 9.25 minutes.

Revised Version of `wilcoxsign_test`

These findings on the treatment of zero rankings $\Longrightarrow$ updated `coin` package.

`wilcoxsign_test` in `coin 1.0-3` now has an additional argument `ties`.

```
> y1 <- c(14 , 18 , 2 , 4 , -5 , 14 , -3 , -1 , 1 , 6 , 3 , 3)
> x1 <- c(8 , 26 , -7 , -1 , 2 , 9 , 0 , -4 , 13 , 3 , 3 , 4)
> pvalue(wilcoxsign_test(y1~x1,alter="greater",dist=exact()))
[1] 0.4741211
> pvalue(wilcoxsign_test(y1~x1,alter="greater",dist=exact(),
+               ties = "Pratt"))
[1] 0.4609375
```

Note the use of `pvalue(...)` which extracts the p-value from the structure returned by `wilcoxsign_test(...)`

The other value for `ties` is `ties="HollanderWolfe"` which also is the default when `ties` is omitted.

Alternate Form of V_s^*

Suppose the N differences are denoted by $Z_1, \ldots, Z_N$.

Then

$$\begin{aligned} V_s^* &= \Big[\text{number of positive averages } (Z_i + Z_j)/2 \text{ with } i \le j\Big] \\ &\quad + \tfrac{1}{2}\Big[\text{number of averages } (Z_i + Z_j)/2 \text{ with } i \le j \text{ that are zero}\Big] \\ &\quad - \tfrac{1}{4} d_0(d_0 + 1) \end{aligned}$$

The proof is similar to the previous one for V_s.

Combining Data from Several Experiments or Blocks

Sometimes it is desirable to combine results from several comparative studies concerning the same treatment, with the hope of increasing the power of detecting a treatment effect.

Example 5: Televised vs. Live Instruction

12 students of widely varying background and ability are divided into 6 more homogeneous pairs. One student in each pair is randomly selected to be part of the live instruction class, the other is put in the televised version.

Then it became known that another instructor had carried out a similar experiment with 5 matched student pairs.

Televised vs. Live Instruction Results

TV	70	77	80	80	84	73
Live	73	75	80	83	85	74
Difference	-3	2	0	-3	-1	-1
Signed midrank	-5.5	$+4$	0	-5.5	-2.5	-2.5

TV	85	93	90	91	89
Live	89	92	90	98	87
Difference	-4	1	0	-7	2

Difference	-7	-4	-3	-3	-1	-1	0	0	1	2	2
Signed midrank	-11	-10	-8.5	-8.5	-4	-4	0	0	4	6.5	6.5

$$\Longrightarrow \qquad V_s^* = 4 + 6.5 + 6.5 = 17$$

Comments

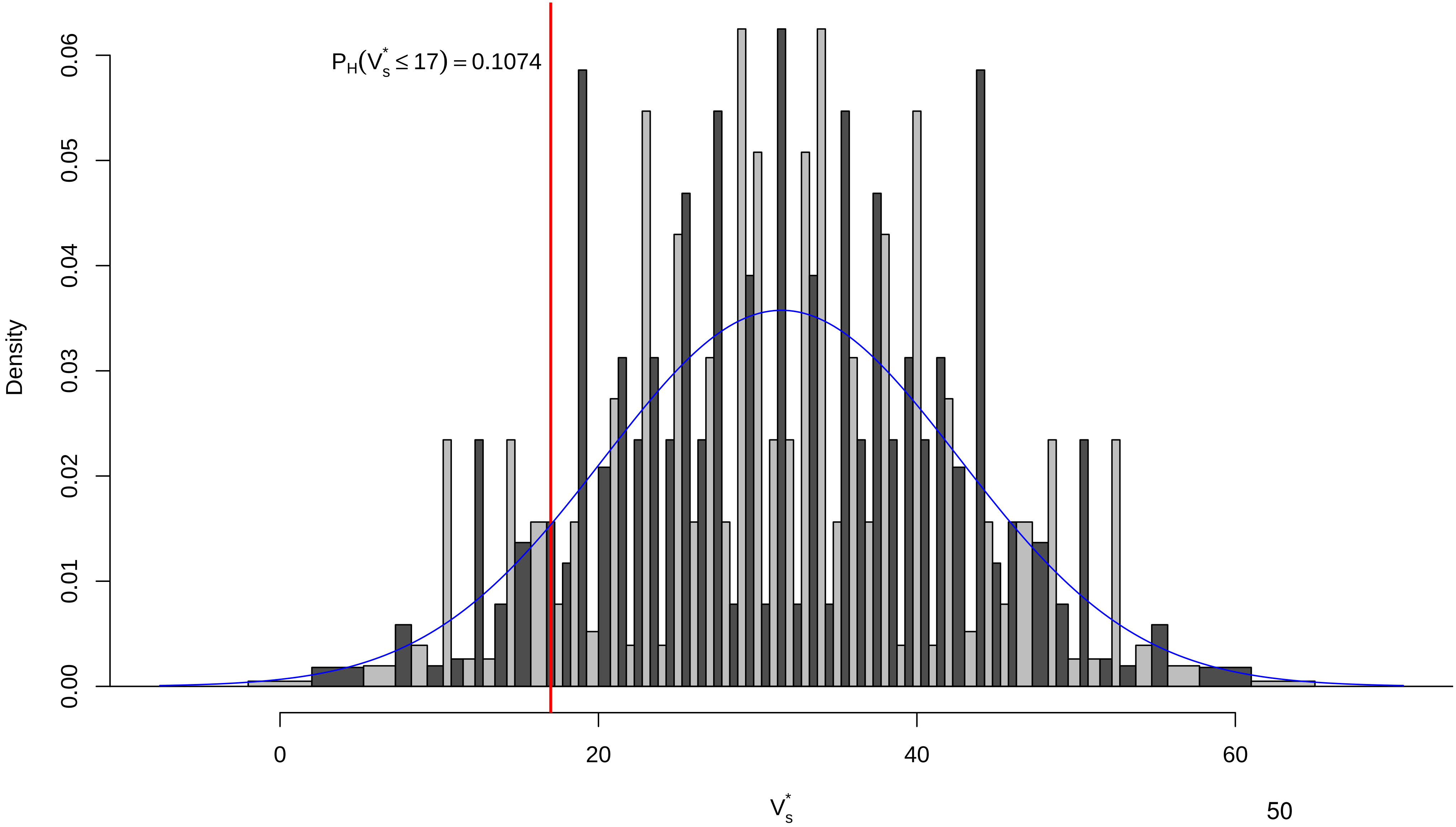
Under H_0 of no difference between live and TV instruction the scores for all 11 pairwise comparisons would have been the same, with equal chance of $\pm$ assigned to each absolute score.

We took the signed rank sum of midranks, as obtained before discarding zeros.

```
> SignedRankExact(TV,Live,alternative="less")
 Vs.star.obs  meanVs.star  sigVs.star  Z           p.val    p.val.normal
 17.0000      31.5000      11.1580    -1.2995    0.1074   0.0969
```

The normal approximation gives a reasonable result.

Exact Null Distribution (Live-TV Instruction)



Different Experimental Scenario for Live-TV Instruction

Without pairing any of the 10 students we randomly select 5 to get TV instruction, the other 5 getting live instruction. We obtain the following scores

$68 \quad 69 \quad 74 \quad 82 \quad 93$ and $72 \quad 75 \quad 83 \quad 95 \quad 100$

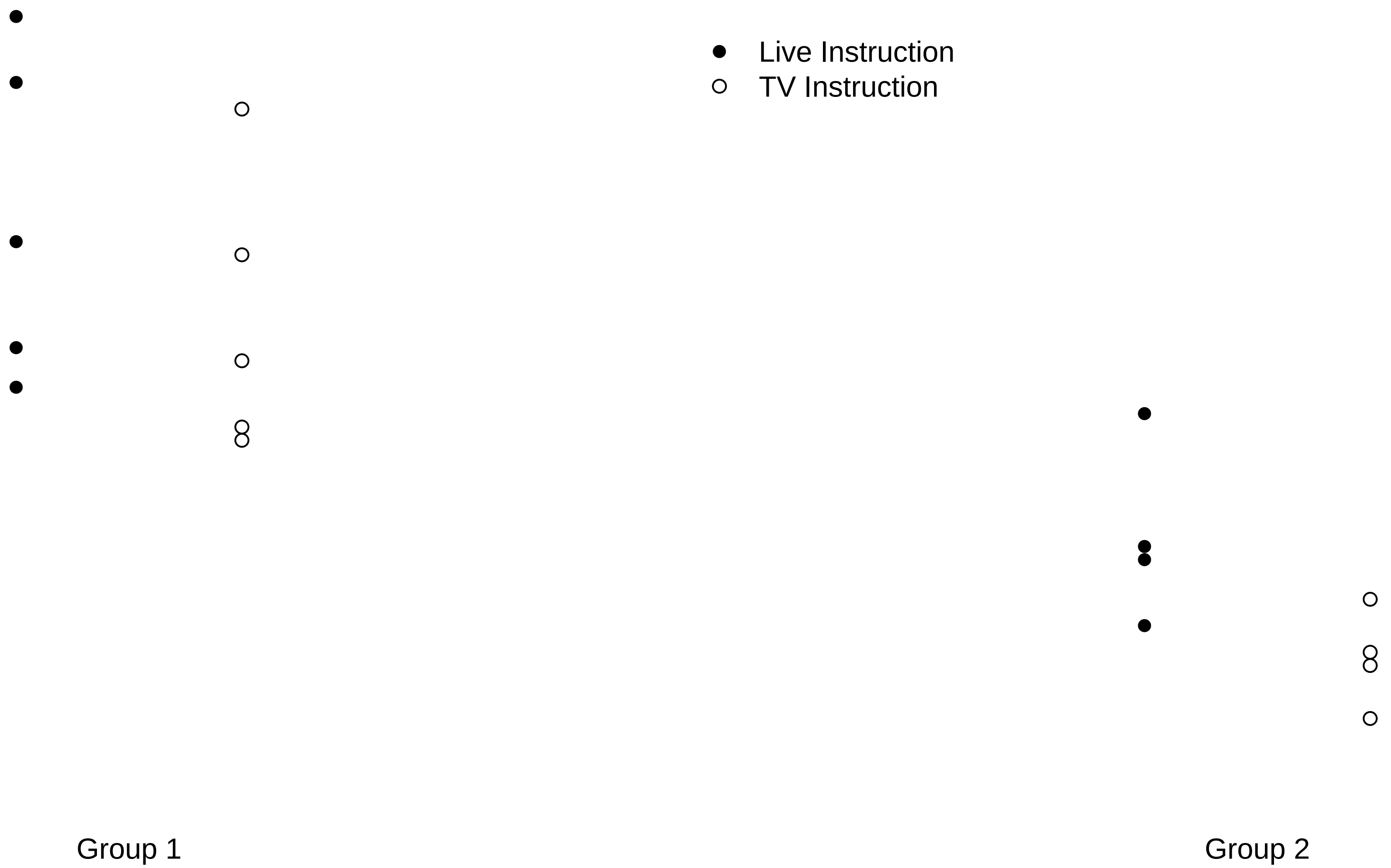
for televised and live instruction, respectively.

Assume the comparison is repeated in another semester with 8 students, split randomly into groups of 4 and 4 with corresponding scores

$47 \quad 51 \quad 52 \quad 56$ and $54 \quad 59 \quad 60 \quad 70$

Can we combine the scores for each teaching method into 9 and 9 scores and apply the Wilcoxon rank-sum test?

Block to Block Variation



Joint Ranking?

As the previous plot made clear, there is a substantial difference between the two groups or blocks, the scores of the second group being generally lower.

Such block effects may be due to a different teacher, harder tests, stiffer grading, different groups of students.

In performing a joint ranking we would compare treatment scores from the second group with control scores from the first.

The fact that these control scores would be higher than the second group's treatment scores would confuse the group difference with possible score differences due to treatment.

Blocked Comparison Situations

Blocking of subjects occurs frequently: animal litters, observations taken on the same day, in the same clinic or school, composite material produced from same chemical batch, and so on.

Blocks can also be created by matching subjects on extraneous variables

such as age, sex, income, etc.

Let there be b blocks of N_i respective experimental subjects, $i = 1, \ldots, b$.

n_i randomly chosen subjects within the i^{th} block receive the treatment,

the other $m_i = N_i - n_i$ act as controls. Let

$$N = N_1 + \ldots + N_b \qquad n = n_1 + \ldots + n_b \qquad m = m_1 + \ldots + m_b = N - n$$

By design, the n treatment assignments are no longer completely random, since

not all $\binom{N}{n}$ treatment assignment are possible.

Instead we have $\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_b}{n_b}$ possible treatment assignments, all equally likely.

Ranking within Blocks

In order to avoid having block effects getting tangled up with treatment effects we should rank the subjects separately within each block.

Let $S_{i1} < \ldots < S_{in_i}$ be the ranks of the treated subjects in the i^{th} block and denote their rank-sum by

$$W_s^{(i)} = S_{i1} + \ldots + S_{in_i}$$

Before we examined a set of treatment ranks for treatment effect by taking their rank-sum as a single univariate criterion.

Similarly we could take the sum of block rank-sums as a single univariate criterion.

Better yet, since different block sizes N_i are involved with each block rank-sum, it might make sense to take a weighted linear combination, i.e., $\sum_i c_i W_c^{(i)}$.

Optimally Weighted Sum of Rank-Sums

It turns out that taking the coefficients $c_i = 1/(N_i + 1)$ gives optimal power in certain settings, see the Text, Further Developments 5D.

This leads to the blocked comparison Wilcoxon test statistic

$$W_s = \sum_{i=1}^{b} \frac{W_s^{(i)}}{N_i + 1} \quad \text{note} \quad W_s + W_r = \sum_{i=1}^{b} \frac{W_s^{(i)} + W_r^{(i)}}{N_i + 1} = \sum_{i=1}^{b} \frac{N_i(N_i + 1)/2}{N_i + 1} = \sum_{i=1}^{b} \frac{N_i}{2}$$

W_s is equivalent to the straight sum of block rank-sums when all N_i are the same.

Due to the complexity of the possible blocking structures $((N_i, n_i),\ i = 1 \ldots, b)$ it is no longer feasible to tabulate the null distribution of W_s.

However, the normal approximation can be applied quite easily and R allows computation of the exact null distribution by successive uses of `combn` and `outer`, provided $\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_b}{n_b}$ does not get too large.
These two approaches balance each other.

Comments on Symmetry

If the rankings in each block avoid ties then the resulting symmetry of the $W_s^{(i)}$ distributions implies symmetry of the W_s distribution.

$$W_s^{(i)} - a_i \overset{\mathcal{D}}{=} a_i - W_s^{(i)} \quad \text{for } i = 1, \ldots, b$$

$$\Longrightarrow \sum_{i=1}^{b} c_i W_s^{(i)} - \sum_{i=1}^{b} c_i a_i = \sum_{i=1}^{b} c_i (W_s^{(i)} - a_i) \overset{\mathcal{D}}{=} \sum_{i=1}^{b} c_i (a_i - W_s^{(i)}) = \sum_{i=1}^{b} c_i a_i - \sum_{i=1}^{b} c_i W_s^{(i)}$$

With ties in the within block rankings such symmetry may no longer hold.

Example 4: Advertising Methods

Two methods of advertising are to be compared for the same product.

One method is aggressive and obnoxious (treatment),
the other is pleasing (control). The treatment is conjectured to improve sales.

The success is measured by the consumption/sales of that product.

Test market cities vary greatly in size. If consumption were proportional to size,
one could adjust for that. But that assumption is doubtful.

Hence we block by the size of the test market cities.

2 large cities, 2 groups of 3 intermediate size cities, and one group of 6 small towns.

The treatments are randomly assigned within each block.

Results of Advertising Experiment

	consumption figures		ranks	
Block	Control	Treatment	Control	Treatment
1	236	255	1	2
2	183	179, 193	2	1, 3
3	115, 128	132	1, 2	3
4	61, 70, 79	67, 84, 88	1, 3, 4	2, 5, 6

For convenience we work with the smaller numbers and use

$$W_r = \sum_{i=1}^{4} \frac{W_r^{(i)}}{N_i+1} = \frac{W_r^{(1)}}{3} + \frac{W_r^{(2)}}{4} + \frac{W_r^{(3)}}{4} + \frac{W_r^{(4)}}{7}$$

We reject H_0 : no treatment effect, when $W_r \leq c$. Since it is easier to work with integers, we use the smallest common denominator multiple of W_s, i.e.,

$$84W_r = 28W_r^{(1)} + 21W_r^{(2)} + 21W_r^{(3)} + 12W_r^{(4)}$$

The Calculation

The observed value of $84W_r$ is

$$84W_r = 28 \times 1 + 21 \times 2 + 21 \times (1+2) + 12 \times (1+3+4) = 229$$

For the p-values we need to find $P_{H_0}(84W_r \leq 229)$.

$28W_r^{(1)}$		$21W_r^{(2)}$			$21W_r^{(3)}$			$12W_r^{(4)}$					
28	56	21	42	63	63	84	105	72	84	96	108	120	$\cdots$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{3}{20}$	$\cdots$

Note that the 63's come about as $21W_r^{(2)} = 21 \times 3$ and $21W_r^{(3)} = 21 \times (1+2)$.

Fix $28W_r^{(1)}, 21W_r^{(2)}, 21W_r^{(3)}$ at their lowest values $28, 21, 63$ with sum 112, with probability $\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{18}$.

To get $84W_r \leq 229$ we need $12W_r^{(4)} \leq 229 - 112 = 117$ with probability $\frac{7}{20}$, with total probability $\frac{7}{360}$ for all 4 conditions.

The Calculation (continued)

The next smallest sum for $28W_r^{(1)} + 21W_r^{(2)} + 21W_r^{(3)}$ has to be 21 higher

$$28 + 21 + 84 = 28 + 42 + 63 = 133 \qquad \text{with probability } \tfrac{2}{18}.$$

To have $84W_r \leq 229$ we need $12W_r^{(4)} \leq 229 - 133 = 96$, with probability $\frac{4}{20}$, thus with overall probability $\frac{8}{360}$ for these cases.

After dealing with a few more cases in similar fashion we arrive at (Problem 46)

$$P_{H_0}\left(W_r \leq \frac{229}{84}\right) = P_{H_0}(84W_r \leq 229) = \frac{20}{360} = .0556$$

This calculation was tedious but manageable, because the problem was of small size and we used a systematic process.

Means & Variances for $W_r^{(i)}/(N_i+1)$ & $W_s^{(i)}/(N_i+1)$

Means and variances of $W_r^{(i)}/(N_i+1)$ & $W_s^{(i)}/(N_i+1)$ are given (without ties) by

$$E_{H_0}\frac{W_r^{(i)}}{N_i+1}=\frac{m_i}{2} \qquad E_{H_0}\frac{W_s^{(i)}}{N_i+1}=\frac{n_i}{2}$$

$$\mathrm{Var}_{H_0}\frac{W_r^{(i)}}{N_i+1}=\mathrm{Var}_{H_0}\frac{W_s^{(i)}}{N_i+1}=\frac{m_i n_i}{12(N_i+1)}$$

i	m_i	n_i	N_i	$E_{H_0}[(W_r^{(i)}/(N_i+1)]$	$\mathrm{Var}_{H_0}[W_r^{(i)}/N_i+1)]$
1	1	1	2	$\frac{1}{2}$	$\frac{1}{36}$
2	1	2	3	$\frac{1}{2}$	$\frac{1}{24}$
3	2	1	3	1	$\frac{1}{24}$
4	3	3	6	$\frac{3}{2}$	$\frac{3}{28}$

Means & Variances for Normal Approximation

$$E_{H_0}(W_r) = \sum_{i=1}^{4} E_{H_0}\left(\frac{W_r^{(i)}}{N_i+1}\right) = \sum_{i=1}^{4} \frac{m_i}{2} = \frac{1}{2} + \frac{1}{2} + 1 + \frac{3}{2} = \frac{7}{2}$$

and because of the independent randomizations from block to block

$$\begin{aligned} \mathrm{var}_{H_0}(W_r) &= \sum_{i=1}^{4} \mathrm{var}_{H_0}\left(\frac{W_r^{(i)}}{N_i+1}\right) = \sum_{i=1}^{4} \frac{m_i n_i}{12(N_i+1)} \\ &= \frac{1}{36} + \frac{1}{24} + \frac{1}{24} + \frac{3}{28} = \frac{55}{252} = 0.218254 \end{aligned}$$

As normal approximation we thus get

$$P_{H_0}\left(W_r \le \frac{229}{84}\right) \approx \Phi\left(\frac{229/84 - 7/2}{\sqrt{55/252}}\right) = \Phi(-1.6564) = 0.0488$$

as compared to the exact value of $.0556$ obtained earlier. The approximation quality is not too bad given that we deal with very small block sizes.

BlockedWilcoxon

```
BlockedWilcoxon=function(datlist,alternative="greater",
                                         Nsim=10000,PDF=F){
# This function needs as input a list of lists, say b lists.
# Each of the b lists should consist of two vectors x and y
# in that order, where y represents the treatment scores
# and x represents the control scores for the block
# represented by this sublist.
# It is assumed that the random assigment of treatments
# from block to block are independent.
# alternative="greater" means that we expect that the y-scores
# will tend to be larger than the x-scores under the alternative.
# Other values for alternative are "less" and "two.sided", with
# corresponding meanings.
# This function evaluates the exact null distribution and
# exact p-value when the number of combined randomization
# combinations over all blocks does not exceed Nsim. Otherwise it
# estimates the null distribution by Nsim simulations.
```

Input List for `BlockedWilcoxon`

The following shows how to construct the input list for the advertising example.

```
> advertising=list()
> advertising[[1]]=list(x=236,y=255)
> advertising[[2]]=list(x=183,y=c(179,193))
> advertising[[3]]=list(x=c(115,128),y=132)
> advertising[[4]]=list(x=c(61,70,79),y=c(67,84,88))
```

The call `BlockedWilcoxon(advertising)` produces the plot on the next slide.

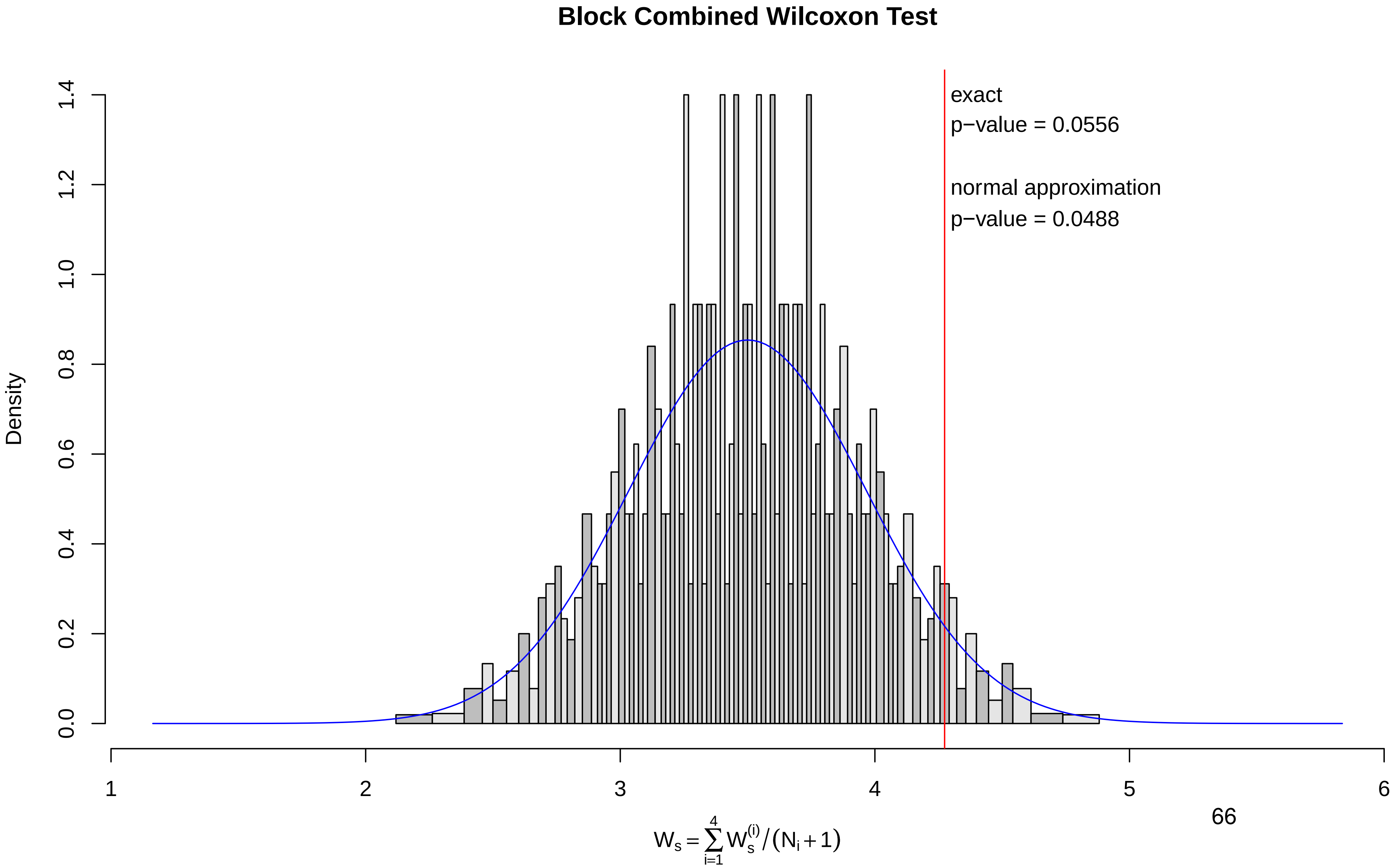
Note that we show there the distribution for W_s and not that of the equivalent W_r.

However, for W_s we would reject H_0 when W_s is too large. Recall

$$W_s + W_r = \sum_{i=1}^{b} \frac{W_s^{(i)}}{N_i + 1} + \sum_{i=1}^{b} \frac{W_r^{(i)}}{N_i + 1} = \sum_{i=1}^{b} \frac{N_i}{2}$$

The p-values are the same.

Null Distribution of $W_s = \sum_{i=1}^{4} W_s^{(i)} / (N_i + 1)$



Comments on Normal Approximation

Note the symmetry of the W_s distribution (no ties in within block ranking).

It can be shown that $$\frac{W_r - E_{H_0}(W_r)}{\sqrt{\mathrm{var}_{H_0}(W_r)}} \longrightarrow \mathcal{N}(0,1)$$

when either the $N_i \longrightarrow \infty$ for $i = 1, \ldots, b$

or when the number b of blocks $\longrightarrow \infty$.

Thus we can use the normal approximation when either b is large or when the N_i are large.

Considering that in our example we had $b = 4$ and $\min(N_1, \ldots, N_4) = 2$, not exactly large, the approximation was not too bad.

Special Case of Blocked Comparison

Let $N_i = 2$ for $i = 1, \ldots, b$, with $m_i = n_i = 1$.

Then each $W_s^{(i)}$ has either rank 1 or rank 2, depending on whether $Y_i < X_i$ or $Y_i > X_i$, which translates to $Y_i - X_i < 0$ or $Y_i - X_i > 0$.

Assuming no ties we thus have

$$W_s = \sum_{i=1}^{b} W_s^{(i)} = \sum_{i=1}^{b} \left(1 \times I_{[Y_i - X_i < 0]} + 2 I_{[Y_i - X_i > 0]}\right) = b + \sum_{i=1}^{b} I_{[Y_i - X_i > 0]}$$

$\Longrightarrow$ The blocked comparison Wilcoxon test is equivalent to the sign test.

Some Discussion of the Special Case

We remarked previously that the sign test typically has low power.

Thus the blocked comparison Wilcoxon test has relatively low power when $N_i = 2,\ i = 1,\ldots,b$.

For small block sizes subjects are typically easy to rank without scores.

The deficiency in power decreases as b increases.

The deficiency and advantage of the blocked comparison Wilcoxon test derive from the same circumstance:

Within each block we need to make only few comparisons (easy ranking)

That reduces the total number of possible comparisons $\binom{N_i}{2}$ per block, compared to doing all $\binom{N}{2}$. We get less comparison information, thus less power.

Aligning Scores

The scores in the 4 blocks were compared (ranked) within each block.
The differences in treatment and control ranks were assessed within blocks and then accumulated over all blocks.

From block to block we refrained from such comparisons in order not to entangle treatment and block effects.

Now we will remove the block effect to some extent by subtracting the block averages from all scores within respective blocks.

This should bring all the adjusted scores into the same ballpark, i.e., the scores will be aligned. They are now jointly comparable. Any differences will be mainly due to treatment affects alone or due to the randomization under H_0.

The aligned scores are ranked jointly across all blocks.

$\Longrightarrow$ We get a more extensive ranking scale for all subjects $\Longrightarrow$ more power.

Results of Advertising Experiment

	consumption figures		ranks	
Block	Control	Treatment	Control	Treatment
1	236	255	1	2
2	183	179, 193	2	1, 3
3	115, 128	132	1, 2	3
4	61, 70, 79	67, 84, 88	1, 3, 4	2, 5, 6

Example 6: Advertising (Continued)

The average of scores in the 3^{rd} block is $(115+128+132)/3 = 125$.

To align the scores we subtract it from all scores in that block

$$115-125=-10, \qquad 128-125=3, \qquad 132-125=7$$

Proceeding in this way with each of the four blocks we get these aligned scores

	aligned scores		aligned ranks	
Block	Control	Treatment	Control	Treatment
1	$-9\frac{1}{2}$	$9\frac{1}{2}$	3	13
2	-2	$-6,\ 8$	7	5, 11
3	$-10,\ 3$	7	2, 8	10
4	$-13\frac{5}{6},\ -4\frac{5}{6},\ 4\frac{1}{6}$	$-7\frac{5}{6},\ 9\frac{1}{6},\ 13\frac{1}{6}$	1, 6, 9	4, 12, 14

Since the scores are aligned they are now comparable. It makes some sense to rank all 14 observations jointly, as shown on the right side under aligned ranks.

Comments on Aligned Ranks

Note the generally lower control ranks and higher treatment ranks.

Block	Control	Treatment
1	3	13
2	7	5, 11
3	2, 8	10
4	1, 6, 9	4, 12, 14

Ranking across all 4 blocks allows for a wider ranking spectrum $1, 2, \ldots, 13, 14$ rather than the shorter ranking spectra for each block.

This allows greater expression depth for the strength of the treatment effect.

$\Longrightarrow$ greater discrimination power.

Joint Null Distribution of Aligned Ranks

Under H_0 treatment and control have the same effect.

The assignment of treatment or control label to subjects has no effect on the scores.

Randomization of treatment/control labels is done separately within each block.

The block averages are unchanged under all treatment/control assignments.

Although the alignment of blocks is different from block to block it does not matter whether we randomly assign treatment labels before or after aligning the blocks. Either way, all assignments are equally likely.

The set of aligned scores and thus the set of aligned ranks for each block is completely determined by the subjects alone and not by the treatment/control label.

Randomly selecting subjects for treatment amounts to randomly selecting aligned ranks to be associated with the treatment label.

Do this selection separately and independently within each block.

Joint Null Distribution in Advertising Example

The single aligned treatment rank $\hat{S}_{11}$ in block 1 can be one of 3 and 13, with probability $1/\binom{2}{1} = 1/2$ each.

The two aligned treatment ranks $(\hat{S}_{21}, \hat{S}_{22})$ from block 2 can be any ordered pair taken from $7, 5, 11$, with probability $1/\binom{3}{2} = 1/3$ each.

The single aligned treatment rank $\hat{S}_{31}$ from block 3 can be any one of $2, 8, 10$, with probability $1/\binom{3}{1} = 1/3$ each.

The three aligned treatment ranks $(\hat{S}_{41}, \hat{S}_{42}, \hat{S}_{43})$ from block 4 can be any ordered triple taken from $1, 6, 9, 4, 12, 14$ with probability $1/\binom{6}{3} = 1/20$ each.

Jointly, each set of such four choices has probability

$$\frac{1}{\binom{2}{1}} \times \frac{1}{\binom{3}{2}} \times \frac{1}{\binom{3}{1}} \times \frac{1}{\binom{6}{3}} = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{20} = \frac{1}{360}$$

Null Distribution of Aligned Rank-Sum

Smaller numbers are easier to deal with. Thus we use aligned control ranks $\hat{R}_{ij}$.

Using the joint null distribution for these aligned control ranks we can obtain the null distribution of any statistic derived from them, in particular their sum $\hat{W}_r$,

with observed value $\quad 3+7+2+8+1+6+9=36$

Since we would reject H_0 for small values of $\hat{W}_r$, our p-value is $P_{H_0}(\hat{W}_r \leq 36)$.

As in our previous treatment it is organizationally convenient to view $\hat{W}_r$ as the sum of the aligned rank sums over each respective block, i.e., $\hat{W}_r = \hat{W}_r^{(1)} + \ldots + \hat{W}_r^{(4)}$.

The smallest value of $\hat{W}_r^{(1)} + \hat{W}_r^{(2)} + \hat{W}_r^{(3)}$ is $3+5+10=18$.

To have $\hat{W}_r \leq 36$ we then need $\hat{W}_r^{(4)} \leq 18$.

The Calculation

i	1		2			3			4				
w	3	13	5	7	11	10	12	18	11	14	16	17	19
$P_H(\hat{W}_r^{(i)} = w)$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$

Table $\Longrightarrow$ $P_{H_0}(\hat{W}_r^{(4)} \leq 18) = 4/20$, so that the combined probability of

$\hat{W}_r^{(1)} = 3, \hat{W}_r^{(2)} = 5, \hat{W}_r^{(3)} = 10, \hat{W}_r^{(4)} \leq 18$ is ${}^{1}/_{2} \times {}^{1}/_{3} \times {}^{1}/_{3} \times {}^{4}/_{20} = {}^{4}/_{360}$

$\hat{W}_r^{(1)} = 3, \hat{W}_r^{(2)} = 7, \hat{W}_r^{(3)} = 10, \hat{W}_r^{(4)} \leq 16$ is ${}^{1}/_{2} \times {}^{1}/_{3} \times {}^{1}/_{3} \times {}^{3}/_{20} = {}^{3}/_{360}$

$\hat{W}_r^{(1)} = 3, \hat{W}_r^{(2)} = 5, \hat{W}_r^{(3)} = 12, \hat{W}_r^{(4)} \leq 16$ is ${}^{1}/_{2} \times {}^{1}/_{3} \times {}^{1}/_{3} \times {}^{3}/_{20} = {}^{3}/_{360}$

and so on

$$\Longrightarrow \qquad P_{H_0}(\hat{W}_r \leq 36) = \frac{4+3+3+2+1}{360} = \frac{13}{360} = 0.03611 \quad \text{as compared to } .056$$

AlignedBlockedWilcoxon

The class web page has an R function `AlignedBlockedWilcoxon` that carries out the previous tedious calculations.

For its usage read the internal documentation.

The input data list is of the same form as for `BlockedWilcoxon`.

It has an additional argument `align` that specifies the alignment process, implemented internally via an alignment function `AlignFun=function(z)`$\{\ldots\}$

Alignment Options

The argument `align` allows for three different aligments.

The alignment is supposed to make scores from different blocks more comparable.

The crucial aspect of any alignment operation within a block is that it should yield the same aligned scores no matter how we permute all scores `z` within the block, i.e., no matter how the treatment labels are assigned to scores within each block.

Thus the aligned rank set corresponding to a block score vector `z` is not affected by the treatment/control assignment within each block. All are equally likely under H_0

If `align="mean"` we use as alignment function

```
AlignFun=function(z){z-mean(z)}
```

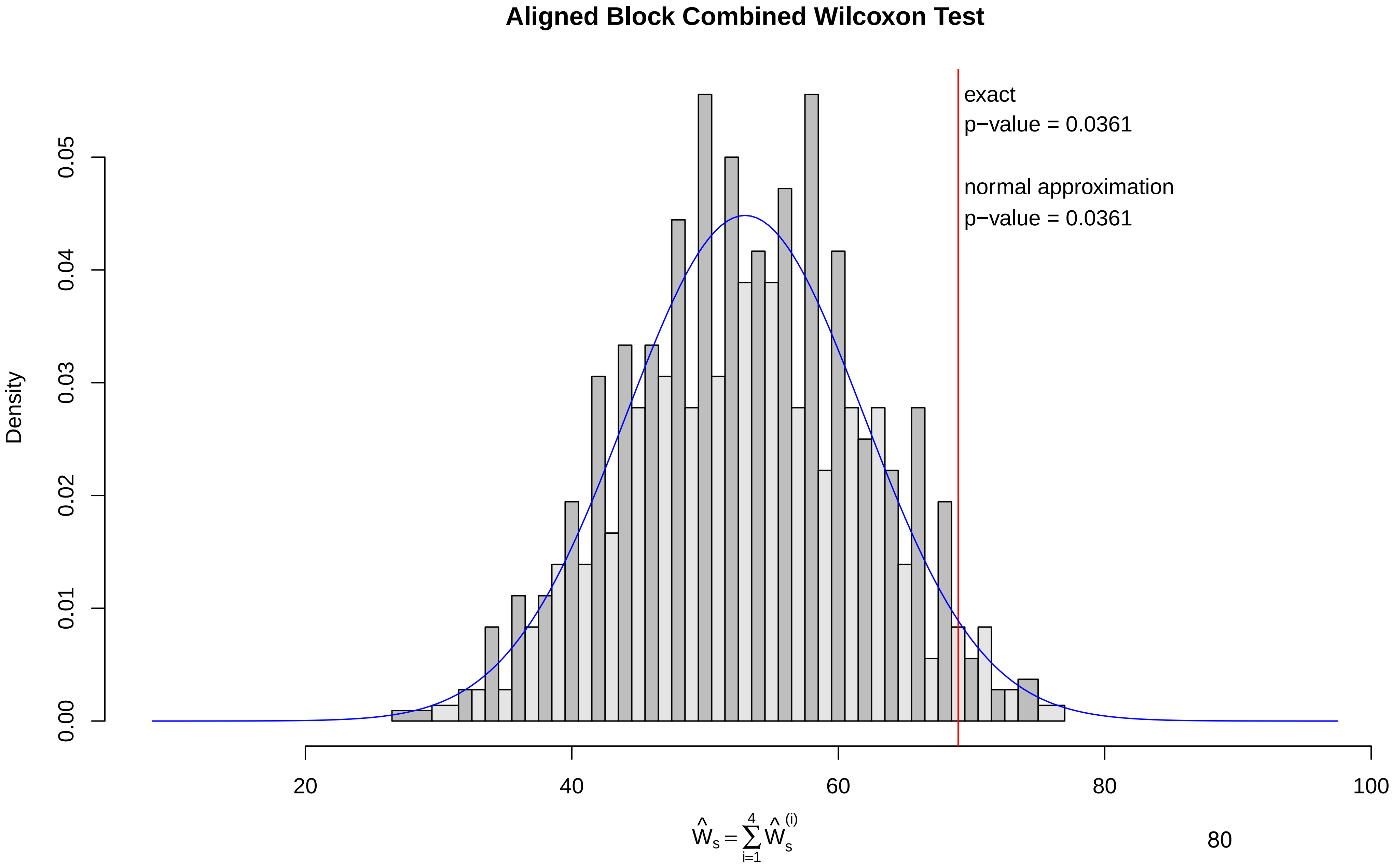
If `align="median"` we use as alignment function

```
AlignFun=function(z){z-median(z)}
```

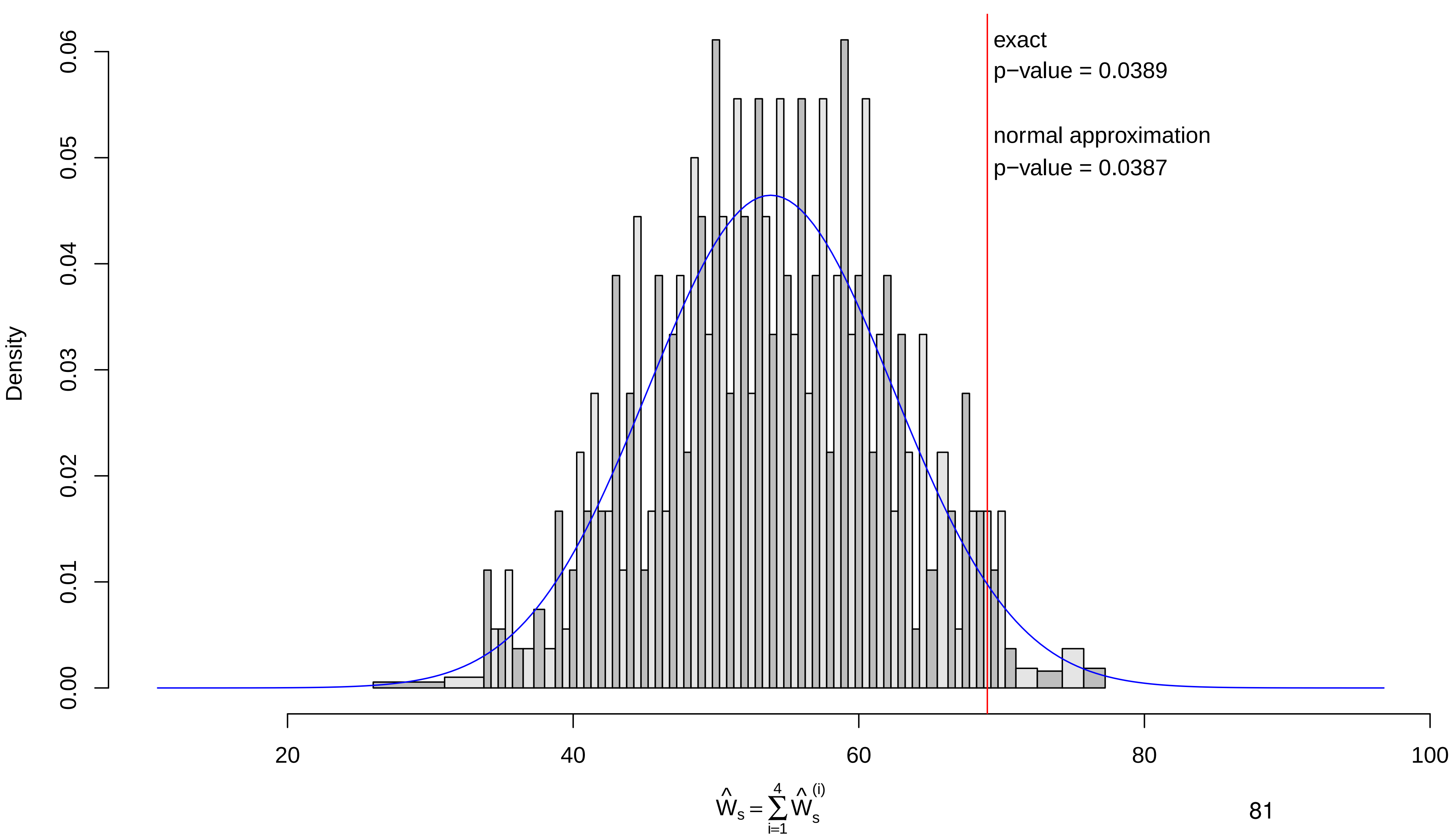
If `align="std.residual"` we use as alignment function

```
AlignFun=function(z){(z-mean(z))/sqrt(var(z))}
```

Aligning by Block Mean Subtraction



Aligning by Block Median Subtraction

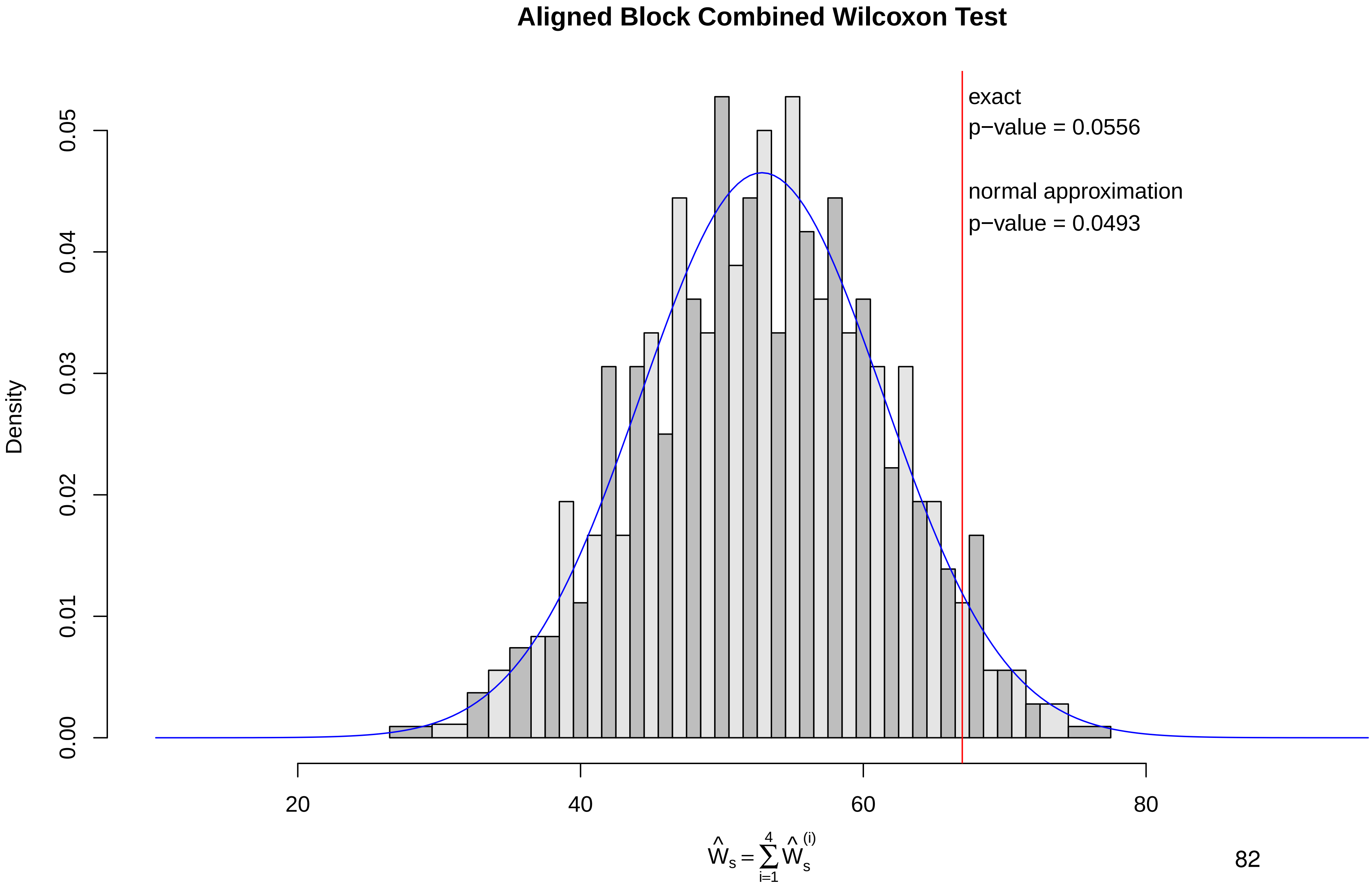


Aligned Block Combined Wilcoxon Test

exact
p−value = 0.0389

normal approximation
p−value = 0.0387

Aligned Using Standardized Block Residual



Alignment: General Case

The previous specific advertising example generalizes easily.

We have b blocks, with $m_i + n_i = N_i$ subject scores in the i^{th} block.

Align the observations within each block (e.g., by subtracting the block mean).

Rank the $N_1 + \ldots + N_b$ scores, using midranks if needed.

Denote the aligned treatment midranks in the i^{th} block by $\hat{S}_{i1}, \ldots, \hat{S}_{in_i}$, $i = 1, \ldots, b$.

The joint null distribution of all these $n_1 + \ldots + n_b$ midranks is

$$P_{H_0}\left(\hat{S}_{11} = s_{11}, \ldots, \hat{S}_{1n_i} = s_{1n_1}, \ldots, \hat{S}_{b1} = s_{b1}, \ldots, \hat{S}_{bn_b} = s_{bn_b}\right) = \frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_b}{n_b}}$$

Null Distribution of $\hat{W}_s$: General Case

The sum $\hat{W}_s$ of all aligned treatment midranks can be viewed as

$$\hat{W}_s = \hat{W}_s^{(1)} + \ldots + \hat{W}_s^{(b)}$$

where $\hat{W}_s^{(i)}$ is the sum of aligned treatment midranks from the i^{th} block.

The distribution vector `z.i` of $\hat{W}_s^{(i)}$ can be obtained as before by using `combn`.

The distribution vector `z.ij` of a sum of independent $\hat{W}_s^{(i)}$ and $\hat{W}_s^{(j)}$ can be obtained by using the `outer(z.i,z.j,"+")` function call on their respective distribution vectors `z.i` and `z.j`.

This is implemented repeatedly in the previously introduced function `AlignedBlockedWilcoxon`.

Comments on `AlignedBlockedWilcoxon`

`AlignedBlockedWilcoxon` works as long as $\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_b}{n_b}$ is not too large.

I tried $m_i = n_i = 5$ for $i = 1, 2, 3$ with $\binom{10}{5} = 252$ and $252^3 = 16,003,008$.

On this laptop it ran in 76 seconds and produced the result on the next slide.

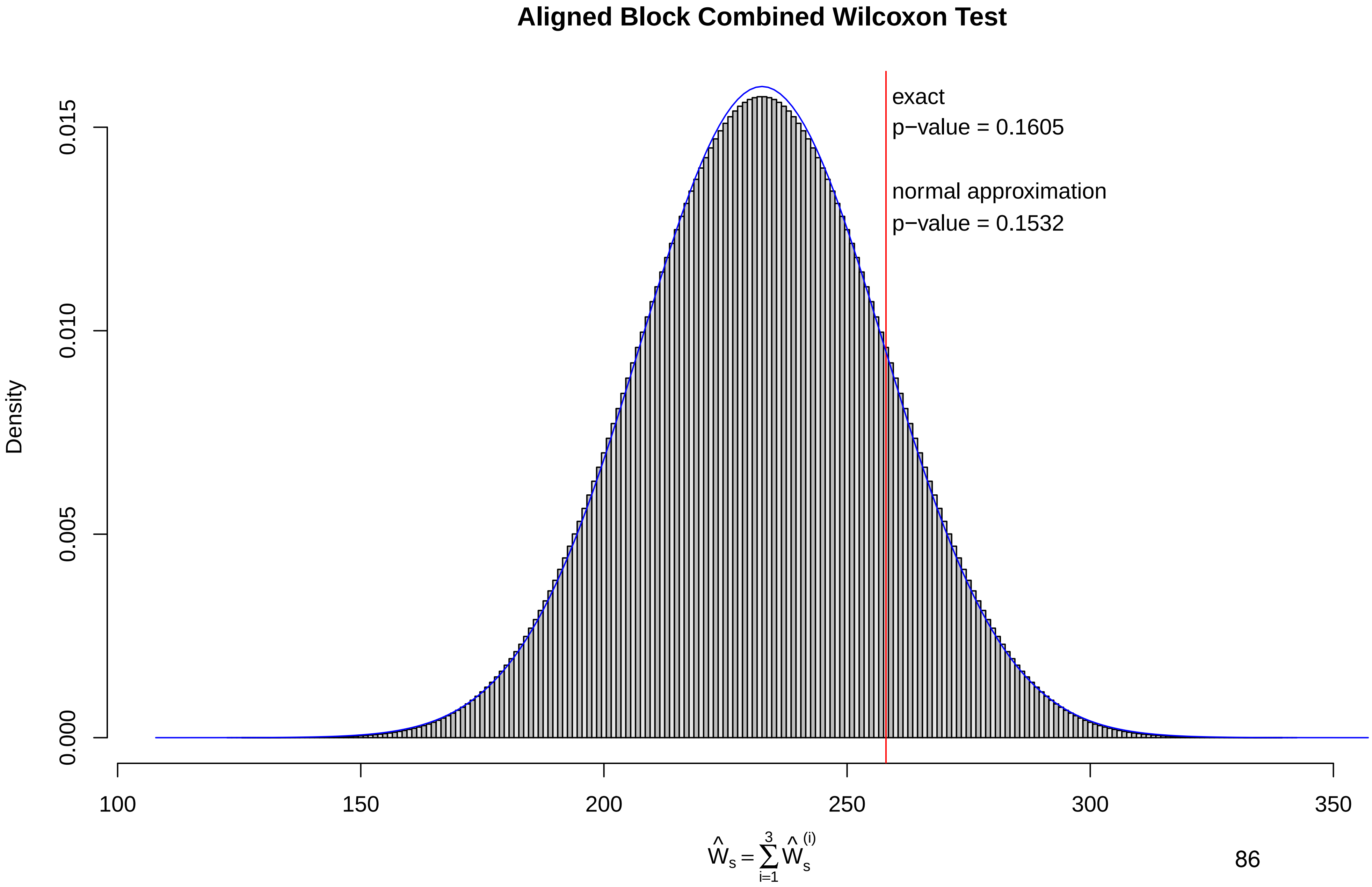
On my other laptop it took 15 seconds.

When I tried $b = 4$ $(252^4 = 4,032,758,016)$ I got an error message of

`negative length vectors are not allowed`, presumably memory overflow.

The 3×2 control/treatment samples came from a normal distribution with variance 1 and respective means of $1, 2, 3$ under control and $1.5, 2.5, 3.5$ under treatment.

$$m_i = n_i = 5,\ i = 1,2,3$$


Aligned Block Combined Wilcoxon Test
exact
p−value = 0.1605
normal approximation
p−value = 0.1532
Density
0.000
0.005
0.010
0.015
100
150
200
250
300
350
$\hat{W}_s = \sum_{i=1}^{3} \hat{W}_s^{(i)}$

Means and Variances of $\hat{W}_s$

The insurmountable computational size involved in getting the exact null distribution of $\hat{W}_s$ in general situations can be overcome by using a normal approximation.

Since the randomizations are independent from block to block we have

$$E(\hat{W}_s) = E(\hat{W}_s^{(1)}) + \ldots + E(\hat{W}_s^{(b)}) \quad \text{and} \quad \mathrm{var}(\hat{W}_s) = \mathrm{var}(\hat{W}_s^{(1)}) + \ldots + \mathrm{var}(\hat{W}_s^{(b)})$$

Let $k_{i1}, \ldots, k_{iN_i}$ denote the aligned midranks in the i^{th} block. Then

$$E(\hat{W}_s^{(i)}) = n_i \bar{k}_{i\bullet} \qquad \text{and} \qquad E(\hat{W}_r^{(i)}) = m_i \bar{k}_{i\bullet} \qquad \text{with} \qquad \bar{k}_{i\bullet} = \frac{k_{i1} + \ldots + k_{iN_i}}{N_i}$$

$$\mathrm{var}(\hat{W}_s^{(i)}) = \mathrm{var}(\hat{W}_r^{(i)}) = \frac{n_i m_i}{N_i(N_i - 1)} \sum_{j=1}^{N_i} (k_{ij} - \bar{k}_{i\bullet})^2$$

See our previous finite population formulas for mean and variance of a sample sum.

Here we are dealing with b finite populations of midranks $k_{i1}, \ldots, k_{iN_i}$, $i = 1, \ldots, b$.

Comments on Normal Approximation

Under a variety of conditions, as discussed before in the unaligned case, we have

$$\frac{\left(\hat{W}_s - E(\hat{W}_s)\right)}{\sqrt{\mathrm{var}(\hat{W}_s)}} \quad \longrightarrow \quad \mathcal{N}(0,1)$$

as $\quad \min(N_1,\ldots,N_b) \longrightarrow \infty \quad$ or as $\quad b \longrightarrow \infty$.

In the latter case ($b \rightarrow \infty$) each block should contain positive and negative aligned scores, as was the case in all three of our alignment schemes.

As the superimposed normal approximation in the last histogram shows, the histogram has slightly less probability in the center compared to the normal approximation.

This deficieny is mostly compensated by slight excesses of probability in the shoulders. The tails appear to be well approximated.

Special Case of Block Size $N_i = 2$

For alignment on the midpoint between the two observations of each the $b = N$ blocks it can be shown that

$$\hat{W}_s = 2V_s - S_N + \frac{N(N+1)}{2}$$

where V_s is the Wilcoxon signed rank statistic (for N matched pairs) and S_N is the corresponding sign test statistic.

I don't know whether the above identity holds (in modified form) in the case of ties.

In the case of no ties we have

$$\mathrm{var}(S_N) = \frac{N}{4} \qquad \text{and} \qquad \mathrm{var}(V_s) = \frac{N(N+1)(2N+1)}{24}$$

Thus the variability of S_N is dwarfed by the variability of V_s.

This means that tests based on $\hat{W}_s$ and tests based on V_s will come to approximately the same conclusion.

Retrospective

We previously saw that the blocked comparison (non-aligned) Wilcoxon test is equivalent to the sign test when the block sizes are $N_i = 2,\ i = 1,\ldots,b$.

The previous slide showed that the aligned block Wilcoxon test is basically equivalent to the Wilcoxon signed rank test.

Thus we may view the aligned block Wilcoxon test as an extension of the Wilcoxon signed rank test to block sizes $N_i \geq 2$, which beats the sign test.

For efficiency reasons we should thus prefer the aligned block Wilcoxon test over the blocked comparison (non-aligned) Wilcoxon test.

Blocking or No Blocking?

Without block to block variation a completely randomized treatment design over the full set of subjects would usually be more efficient than treatment randomization within each block, even with alignment.

Recall: Response variability has a detrimental effect on detecting treatment effects.

The power is a function of Δ/σ.

We block to reduce this variability to within block variability.

Confounding in Paired Comparisons

Sometimes paired subjects are distinguishable as being of type A or B.

First and second born twin, left and right hand, order of task performance by same subject.

Even in deliberate matching of pairs based on other factors one could still distinguish within each pair a high and a low level (A and B) of such factors

If treatment and control are randomly assigned (probability $1/2$) to such pairs, it is possible that all or an undue preponderance of A subjects get the treatment.

Then it would be unclear wether any seen effect is due to treatment or due to the type A that occurred most often together with the treatment.

In such situations treatment and subject type become confounded.

Balanced Design in Paired Comparisons

The previous confounding difficulty can and should be avoided by a different random assignment of treatment and control.

Instead of having the number of treatment cases of type A be random we will fix this number to some number a, typically $a = N/2$, when the number of available pairs is even. The assigment is called balanced.

When N is odd, say $N = 2k+1$ we would choose $a = k$ (or $a = k+1$).

Now we randomly choose a of the N available pairs, with equal chance $1/\binom{N}{a}$, and assign the treatment to the type A member of the pair and let the type B member act as control.

For the remaining $N-a$ pairs we assign the treatment to the type B subjects and let the A subjects act as controls.

Analysis of Balanced Design Paired Comparisons

For each of the N pairs calculate the differences of $A-B$, i.e.,

response of type A subject $-$ response of type B subject.

View the N pairs as N "subjects" and the differences $A-B$ as their "responses."

These "subjects" were randomly divided into two groups.

In group 1 we had the A-subjects treated and the B subjects act as control,

in group 2 we had the B-subjects treated and the A subjects act as control.

Under the hypothesis H_0 of no difference between treatment and control there is no difference between the two groups.

If we rank the $A-B$ "responses" and take the rank sum for group 1, then we have the Wilcoxon rank-sum test statistic, with known null distribution, with $m=n=a$ (when N is even, otherwise $m=a, n=a+1$ or $m=a+1, n=a$).

Treatment Effect

If there is a beneficial treatment effect, then we would expext the $A-B$ “response” in group 1 to be higher than experienced under H_0, and we would expect the $A-B$ “response” in group 2 to be lower than experienced under H_0.

Taking both effects together, we would thus expect the $A-B$ “response” in group 1 to be higher than the $A-B$ “response” in group 2.

High values of the rank-sum W_s for group 1 would be judged significant.
The p-value of the observed w_s is $P_{H_0}(W_s \geq w_s)$.

If the treatment lowers the response relative to H_0 responses then low values of the rank-sum W_s for group 1 would be judged significant.
The p-value of the observed w_s is $P_{H_0}(W_s \leq w_s)$.

Example 7: Effect of Hypnosis on Speech

To study the effect of hypnosis on speech (measured in number of words over a given time span), 10 subjects were observed both under hypnosis (H) and waking state (W).

The hypothesis of no hypnosis effect was to be tested against the alternative of fewer words under hypnosis.

The subjects were randomly split into 2 groups of 5
(waking state 1^{st}, hypnosis 2^{nd}) (WH) and (hypnosis 1^{st}, waking state 2^{nd}) (HW)

	Group 1: WH					*Group 2: HW*				
W	255	1250	126	480	371	308	688	345	264	306
H	67	67	89	129	491	304	49	281	131	107
Difference	188	1183	37	251	−120	4	639	64	133	199

Analysis

The role of A and B here is the order of measurement on each subject.

$A \equiv 1^{\text{st}}$ measurement and $B \equiv 2^{\text{nd}}$ measurement.

$A - B = W - H$ in Group 1 and $A - B = H - W$ in Group 2.

According to the previous table we have

$A - B = 188, 1183, 37, 251, -120, -4, -639, -64, -133, -199$ with respective ranks

$R_1, \ldots, R_5, S_1, \ldots, S_5 = 8, 10, 7, 9, 4, 6, 1, 5, 3, 2 \Longrightarrow W_s = 1 + 2 + 3 + 5 + 6 = 17$.

Fewer words under hypnosis would be indicated in Group 1 by $A - B = W - H$ high and in Group 2 by $A - B = H - W$ low.

Thus we should reject H_0 for low W_s. Hence the p-value is

$$\begin{aligned} P_{H_0}(W_s \le 17) &= P_{H_0}\left(W_{XY} \le 17 - \frac{5(5+1)}{2}\right) = P_{H_0}(W_{XY} \le 2) \\ &= \texttt{pwilcox}(2, 5, 5) = 0.01587302 \end{aligned}$$

Final Comments

There is a bit of a difference to the Wilcoxon rank-sum test as we used it previously.

In both cases the hypothesis is the same:

There is no difference between treatment and control.

Previously the alternative of a treatment effect would act on just one group of subjects, namely those that were treated.

Now the alternative affects both groups in opposite directions, since the treatment H affects the response differences $W - H$ and $H - W$ through opposite signs in the respective groups.