STA258H5

Al Nosedal
and Alison Weir

Winter 2017

NONPARAMETRIC TESTS

Example

The managing director of a mail-order company is concerned about the possible slow progress of orders through her administration department. She tells the administration manager to ensure that at least half of the orders received are processed within one day (eight working hours). Some weeks later she times 18 orders selected at random to check whether her demand has been met.

Hypotheses

Being somewhat distrustful of her administration manager's ability, the managing director's main aim in performing the test is to check for evidence that her instructions **have not** been met. This happens when less than half of the orders are being processed with eight working hours, or in other words, that the median processing time ϕ exceeds 8 hours. The hypotheses may be summarized as follows:

$$H_0 : \phi = 8$$

$$H_a : \phi > 8$$

Data

The times spent by the 18 orders in the administration process were recorded as:

16 h 30 min	14 h 00 min	5 h 40 min	9h 10 min	11 h 45 min
4 h 20 min	7 h 55 min	10 h 15 min	7 h 45 min	16 h 05 min
10 h 05 min	7 h 30 min	9 h 15 min	11 h 55 min	9 h 25 min
10 h 35 min	8 h 20 min	10 h 10 min		

Statistic

The test statistic is called the **sign test** because the statistic used is computed from the data that are in (or have been reduced to) the form of "+" and "-" signs. So here let us write a "+" for all those times greater than 8 h and a "-" for those times less than 8 h.

+	+	-	+	+
-	-	+	-	+
+	-	+	+	+
+	+	+		

Statistic

The test statistic, which we will denote by S, may be either the number of "$+$" signs or the number of "-" signs, according to the context. Clearly, if H_0 is true we would expect the number of "$+$"s and "-"s to be roughly equal, whereas if $H_a : \phi > 8$ is true, we would expect a relatively large number of "$+$"s and correspondingly few "-"s.

Statistic

Let us define S as the number of "+" signs. With S defined in this way, the larger the value of S then the more evidence there is against H_0. Thus rejection regions are of the form:

$$S \geq \text{critical value}$$

P-value

Significance level $= \alpha = 0.05$.
P-value $= P(S \geq 13) = 1 - P(S < 13) = 1 - P(S \leq 12)$
(Note that S has a Binomial distribution with parameters $n = 18$ and $p = 1/2$, under H_0)

P-value (R code)

```
p.value = 1 - pbinom(12,18,0.5);

p.value

## [1] 0.04812622
```

Conclusion

The 18 signs include 13 "$+$" signs. This result is significant at the $\alpha = 5\%$ level, and consequently the managing director has obtained evidence against H_0. It seems quite likely that her requirement is not being met by the administration department.

Example

We shall illustrate Wilcoxon's signed-rank test with the same problem as for the sign test, i.e. using the times that 18 orders took to be processed by an administration system.

Hypotheses

As before, we have

$$H_0 : \phi = 8$$

$$H_a : \phi > 8$$

Statistic

For each observation we compute the difference between the time taken and the hypothesized median value of 8 h.

+8 h 30 min	+6 h 00 min	-2 h 20 min	+1h 10 min	+3 h 45 min
-3 h 40 min	-0 h 05 min	+2 h 15 min	-0 h 15 min	+8 h 05 min
+2 h 05 min	-0 h 30 min	+1 h 15 min	+3 h 55 min	+1 h 25 min
+2 h 35 min	+0 h 20 min	+2 h 10 min		

Statistic

Next, these differences are ranked, **ignoring whether they are positive or negative**, from 1 for the smallest to 18 for the largest.

Difference	Rank	Difference	Rank
+8 h 30 min	18	+6 h 00 min	16
-2 h 20 min	11	+1h 10 min	5
+3 h 45 min	14	-3 h 40 min	13
-0 h 05 min	1	+2 h 15 min	10
-0 h 15 min	2	+8 h 05 min	17
+2 h 05 min	8	-0 h 30 min	4
+1 h 15 min	6	+3 h 55 min	15
+1 h 25 min	7	+2 h 35 min	12
+0 h 20 min	3	+2 h 10 min	9

Statistic

The Wilcoxon signed-rank statistic T is then calculated as the sum of the ranks of either the positive differences or the negative differences, **whichever the one-sided alternative hypothesis suggests should be the smaller**. In our problem H_a suggests that there should be a smaller number of **negative** differences than positive ones, and so we take T to be the sum of the ranks of the negative differences.

Rejection Regions

The critical values in our Table are given in terms of **small** values of the signed-rank statistic, critical regions for the Wilcoxon signed-rank test are of the form

$$T \leq \text{critical value}$$

For $n = 18$ and significance level $\alpha = 0.05$, the critical region is $T \leq 47$. (Significance levels relevant to one-sided tests are denoted by α_1 and those relevant to two-sided tests by α_2.)
The sum of the ranks for the negative differences (-2 h 20 min, -3 h 40 min, -0 h 05 min, -0 h 30 min, and -0 h 15 min) is given by

$$T = 11 + 13 + 1 + 2 + 4 = 31$$

Conclusion

The result is in the 5% critical region (also in the 1% critical region), and so we have strong evidence to support the alternative hypothesis that the median processing time is greater than 8 h.

Null distributions

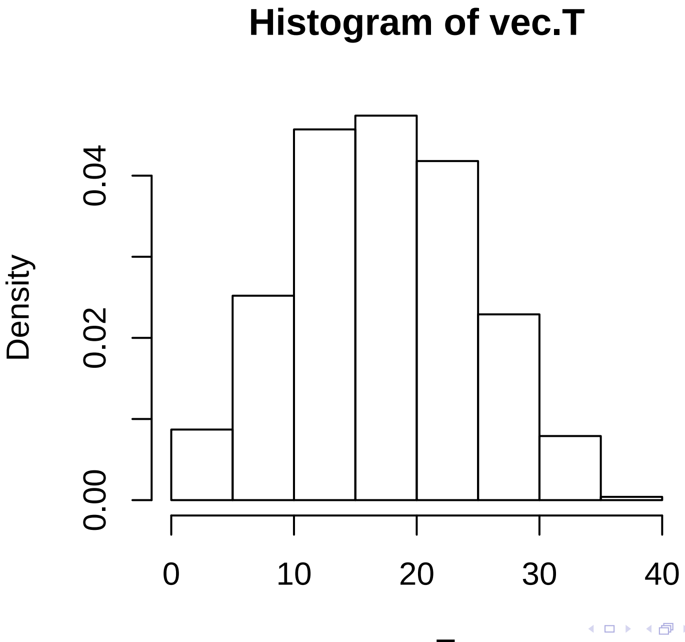
Under the null hypothesis, the true median of the differences is zero. Thus, assuming that the distribution of the differences is **symmetric** about zero, the chances are even that the smallest difference (with rank 1) is positive or negative; the same is true for the next smallest difference (with rank 2); and similarly for every one of the ranked differences. (The theory assumes that there are no ties.) It follows that all possible allocations (2^n in all) of signs to ranks are equally likely. So if one generates these 2^n allocations and calculates the value of T for each, a frequency distribution for T can be constructed and then the null probability distribution of T is obtained by simply dividing the frequencies by 2^n.

Null distributions

We now outline this procedure for $n = 8$. We will consider the one-sided case where T is the rank sum of negative differences.

```
### one simulation;
matrix.signs<-rbinom(8,1,0.5);
matrix.signs

## [1] 0 1 0 0 1 1 0 0

#ranks = vector that contains ranks;
ranks<-matrix(1:8,nrow=8,ncol=1);
t(ranks)

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1    2    3    4    5    6    7    8

# vec.T = vector with values of T;
vec.T<-matrix.signs%*%ranks;
vec.T

##      [,1]
## [1,]   13
```

```
### a bunch of simulations;

# sim = number of simulations
sim<-2000



# matrix.signs = possible allocations;
matrix.signs<-matrix(rbinom(8*sim,1,0.5),sim,ncol=8)

#ranks = vector that contains ranks;
ranks<-matrix(1:8,nrow=8,ncol=1)

# vec.T = vector with values of T;
vec.T<-matrix.signs%*%ranks;

hist(vec.T,freq=FALSE)
```


Histogram of vec.T
Density
0.00
0.02
0.04
0
10
20
30
40

```
mean(vec.T);

## [1] 17.6865

sd(vec.T);

## [1] 7.313957
```

Large sample sizes

As the sample size increases, the shape becomes ever closer to that of the Normal distribution. This enables us to obtain approximate critical values for T when n is large. It can be shown that T has mean $\mu = \frac{n(n+1)}{4}$ and standard deviation $\sigma = \sqrt{n(n+1)(2n+1)/24}$

Example

Does the presence of small numbers of weeds reduce the yield of corn? Lamb's-quarter is a common weed in corn fields. A researcher planted corn at the same rate in 8 small plots of ground, then weeded the corn rows by hand to allow no weeds in 4 randomly selected plots and exactly 3 lamb's-quarter plants per meter of row in the other 4 plots. Here are the yields of corn (bushels per acre) in each of the plots.

Weeds per meter	Yield (bu/acre)			
0	166.7	172.2	165.0	176.9
3	158.6	176.4	153.1	156.0

The samples are too small to rely on the robustness of the two-sample t test. We may prefer to use a test that does not require Normality.

Example (cont.)

We first rank all 8 observations together. To do this, arrange them in order from smallest to largest:
153.1 156.0 158.6 **165.0 166.7 172.2** 176.4 1 **176.9**
The boldface entries in the list are the yields with no weeds present.

Example (cont.)

We see that four of the five highest yields come from that group, suggesting that yields are higher with no weeds. The idea of rank tests is to look just at position in this ordered list. To do this, replace each observation by its order, from 1 (smallest) to 8 (largest). These numbers are the ranks:

Yield	153.1	156.0	158.6	**165.0**	**166.7**	**172.2**	176.4	**176.9**
Rank	1	2	3	**4**	**5**	**6**	7	**8**

Ranks

To rank observations, first arrange them in order from smallest to largest. The **rank** of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

Moving from the original observations to their ranks retains only the ordering of the observations and makes no other use of their numerical values. Working with ranks allows us to dispense with specific assumptions about the shape of the distribution, such as Normality.

If the presence of weeds reduces corn yields, we expect the ranks of the yields from plots with weeds to be smaller as a group than the ranks from plots without weeds. We might compare the sums of the ranks from the two treatments:

Treatment	Sum of ranks
No weeds	23
Weeds	13

Note that the sum of the ranks from 1 to 8 is always equal to 36, so it is enough to report the sum for one of the two groups. If the weeds have no effect, we would expect the sum of the ranks in either group to be 18 (half of 36).

THE WILCOXON RANK SUM TEST

Draw an SRS of size n_1 from one population and draw an independent SRS of size n_2 from a second population. There are N observations in all, where $N = n_1 + n_2$. Rank all N observations. The sum W of the ranks for the first sample is the **Wilcoxon rank sum statistic**. If the two populations have the same continuous distribution, then W has mean

$$\mu_W = \frac{n_1(N+1)}{2}$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

The **Wilcoxon rank sum test** rejects the hypothesis that the two populations have identical distributions when the rank sum W is far from its mean.

In the corn yield study of our Example, we want to test
H_0 : no difference in distribution of yields
against the one-sided alternative
H_a : yields are systematically higher in weed-free plots.
Our test statistic is the rank sum $W = 23$ for the weed-free plots.

Example (cont.)

In our Example, $n_1 = 4$, $n_2 = 4$, and there are $N = 8$ observations in all. The sum of ranks for the weed-free plots has mean

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{(4)(9)}{2} = 18$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2(N+1)}{12}} = \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464$$

Example (cont.)

Although the observed rank sum $W = 23$ is higher than the mean, it is only about 1.4 standard deviations high. We now suspect that the data do not give strong evidence that yields are higher in the population of weed-free corn.
The P-value for our one-sided alternative is $P(W \geq 23)$, the probability that W is at least as large as the value for our data when H_0 is true.

To calculate the P-value $P(W \geq 23)$ for our Example, we need to know the sampling distribution of the rank sum W when the null hypothesis is true. This distribution depends on the two sample sizes n_1 and n_2. Most statistical software will give you P-values, as well as carry out the ranking and calculate W. However, many software packages give only approximate P-values. You must learn what your software offers. With or without software, P-values for the Wilcoxon test are often based on the fact that **the rank sum statistic W becomes approximately Normal as the two sample sizes increase**.

We can then form yet another z^* statistic by standardizing W:

$$z^* = \frac{W - \mu_W}{\sigma_W} = \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}}$$

Use standard Normal probability calculations to find P-values for this statistic. Because W takes only whole-number values, we use a trick called **continuity correction** to improve the accuracy of the approximation.

P-value (without correction)

The standardized rank sum statistic W in our corn yield example is

$$z^* = \frac{W - \mu_W}{\sigma_W} = \frac{23 - 18}{3.464} = 1.44$$

We expect W to be larger when the alternative hypothesis is true, so the approximate P-value is

$$P(Z \geq 1.44) = 0.0749$$

P-value (WITH continuity correction)

We can improve this approximation by using the continuity correction. To do this, act as if the whole number 23 occupies the entire interval from 22.5 to 23.5. Calculate the P-value $P(W \geq 23)$ as $P(W \geq 22.5)$ because the value 23 is included in the range whose probability we want.

$$P(W \geq 22.5) = P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{22.5 - 18}{3.464}\right)$$

$$= P(Z \geq 1.30) = 0.0968$$

R Code

```
# Entering data;

zero<-c(166.7,172.2,165,176.9);

three<-c(158.6,176.4,153.1,156);

# Doing HT;
wilcox.test(zero,three,alternative="greater",
exact=TRUE,paired=FALSE);

##
##  Wilcoxon rank sum test
##
## data:  zero and three
## W = 13, p-value = 0.1
## alternative hypothesis: true location shift is greater thar
```

R Code

```
wilcox.test(zero,three,alternative="greater",
exact=FALSE,paired=FALSE);

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  zero and three
## W = 13, p-value = 0.09697
## alternative hypothesis: true location shift is greater than
```

What hypotheses does Wilcoxon test?

When the distributions may not be Normal, we might restate the hypotheses in terms of population medians rather than means:
H_0: $\text{median}_1 = \text{median}_2$

H_a: $\text{median}_1 > \text{median}_2$
The Wilcoxon rank sum test provides a significance test for these hypotheses, but only if an additional condition is met: both populations must have distributions of the same shape. That is, the density curve for corn yields with 3 weeds per meter looks exactly like that for no weeds except that it may slide to a different location on the scale of yields.

What hypotheses does Wilcoxon test?

The same-shape condition is too strict to be reasonable in practice. Fortunately, the Wilcoxon test also applies in a much more general and more useful setting. It tests hypotheses that we can state in words as

H_0 : two distributions are the same

H_a : one has values that are systematically larger

Dealing with ties in rank tests

The exact distribution for the Wilcoxon rank sum is obtained assuming that all observations in both samples take different values. This allows us to rank them all. In practice, however, we often find observations tied at the same value. What shall we do? The usual practice is to assign all tied values the **average** of the ranks they occupy. Here is an example with 6 observations:

Observation	153	155	158	158	161	164
Rank	1	2	3.5	3.5	5	6

Dealing with ties in rank tests

The tied observations occupy the third and fourth places in the ordered list, so they share rank 3.5. The exact distribution for the Wilcoxon rank sum W applies only to data without ties. Moreover, the standard deviation σ_W must be adjusted if ties are present. The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation. In practice, software is required if you want to use rank tests when the data contain tied values.

Example

A study of early childhood education asked kindergarten students to tell fairy tales that had been read to them earlier in the week. Each child told two stories. The first had been read to them and the second had been read but also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data for five low-progress readers in a pilot study:

Child	1	2	3	4	5
Story 2	0.77	0.49	0.66	0.28	0.38
Story 1	0.40	0.72	0.00	0.36	0.55
Difference	0.37	-0.23	0.66	-0.08	-0.17

We wonder if illustrations improve how the children retell a story. We would like to test the hypotheses
H_0: scores have the same distribution for both stories

H_a : scores are systematically higher for Story 2
Because this is a matched pairs design, we base our inference on the differences. The matched pairs t test gives $t^* = 0.635$ with one-sided P-value $P = 0.280$. We cannot assess Normality from so few observations. We would therefore like to use a rank test.

Positive differences in our Example indicate that the child performed better telling Story 2. If scores are generally higher with illustrations, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. We therefore compare the absolute values of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:
0.37 0.23 **0.66** 0.08 0.17

Arrange these in increasing order and assign ranks, keeping track of which values were originally positive. Tied values receive the average of their ranks. If there are zero differences, discard them before ranking.

Absolute value	0.08	0.17	0.23	**0.37**	**0.66**
Rank	1	2	3	**4**	**5**

The test statistic is the sum of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.) This is the Wilcoxon signed rank statistic. Its value here is $W^+ = 9$.

THE WILCOXON SIGNED RANK TEST FOR MATCHED PAIRS

Draw an SRS of size n from a population for a matched pairs study and take the differences in responses within pairs. Rank the absolute values of these differences. The sum W^+ of the ranks for the positive differences is the **Wilcoxon signed rank statistic**. If the distribution of the responses is not affected by the different treatments within pairs, then W^+ has mean

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

and standard deviation

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The **Wilcoxon signed rank test** rejects the hypothesis that there are no systematic differences within pairs when the rank sum W^+ is far from its mean.

Example (cont.)

In the storytelling study of our Example, $n = 5$. If the null hypothesis (no systematic effect of illustrations) is true, the mean of the signed rank statistic is

$$\mu_{W^+} = \frac{n(n+1)}{4} = \frac{(5)(6)}{4} = 7.5$$

The standard deviation of W^+ under the null hypothesis is

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{(5)(6)(11)}{24}} = \sqrt{13.75} = 3.708$$

The observed value $W^+ = 9$ is only slightly larger than the mean. We now expect that the data are not statistically significant.

The distribution of the signed rank statistic when the null hypothesis (no difference) is true becomes approximately Normal as the sample size becomes large. We can then use Normal probability calculations (with the continuity correction) to obtain approximate P-values for W^+. Let's see how this works in the storytelling example, even though $n = 5$ is certainly not a large sample.

Example (P-value)

We observed $W = 9$, so the one-sided P-value is $P(W^+ \geq 9)$. The continuity correction calculates this as $P(W^+ \geq 8.5)$.

$$P(W^+ \geq 8.5) = P\left(\frac{W^+ - 7.5}{3.708} \geq \frac{8.5 - 7.5}{3.708}\right)$$

$$= P(Z \geq 0.27) = 0.394$$

Conclusion

Our test tells us that this very small sample gives no evidence that seeing illustrations improves the storytelling of low-progress readers.

R Code

```
story2<-c(0.77, 0.49, 0.66, 0.28, 0.38);

story1<-c(0.40, 0.72, 0.00, 0.36, 0.55);

wilcox.test(story2,story1,alternative="greater",
exact=TRUE,paired=TRUE);

##
##  Wilcoxon signed rank test
##
## data:  story2 and story1
## V = 9, p-value = 0.4062
## alternative hypothesis: true location shift is greater than
```

R Code

```
story2<-c(0.77, 0.49, 0.66, 0.28, 0.38);

story1<-c(0.40, 0.72, 0.00, 0.36, 0.55);

wilcox.test(story2,story1,alternative="greater",
exact=FALSE,paired=TRUE);

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  story2 and story1
## V = 9, p-value = 0.3937
## alternative hypothesis: true location shift is greater than
```