18 Non-parametric Methods

The ordered values of a sample of observations are called the order statistics of the sample, and the smallest and the largest called the extremes. Order statistics and extremes are among the most important functions of a set of random variables that we study in probability and statistics. There is natural interest in studying the highs and lows of a sequence, and the other order statistics help in understanding concentration of probability in a distribution, or equivalently, the diversity in the population represented by the distribution. Order statistics are also useful in statistical inference, where estimates of parameters are often based on some suitable functions of the order statistics.

18.1 Order Statistics

Definition. Let $\boldsymbol{X_1, X_2, \cdots, X_n}$ be any $\boldsymbol{n}$ real valued random variables. Let $\boldsymbol{X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}}$ denote the ordered values of $\boldsymbol{X_1, X_2, \cdots, X_n}$. Then, $\boldsymbol{X_{(1)}, X_{(2)}, \cdots, X_{(n)}}$ are called the order statistics of $\boldsymbol{X_1, X_2, \cdots, X_n}$.

Remark: Thus, the minimum among $\boldsymbol{X_1, X_2, \cdots, X_n}$ is the first order statistic $\boldsymbol{X_{(1)}}$

$$\boldsymbol{X_{(1)} = \min\{X_1, X_2, \cdots, X_n\}}$$

and the maximum the $\boldsymbol{n}$th order statistic

$$\boldsymbol{X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}}$$

. The range is Range$\boldsymbol{\{X_1, X_2, \cdots, X_n\} = X_{(n)} - X_{(1)}}$. The middle value among $\boldsymbol{X_1, X_2, \cdots, X_n}$ is called the median. But it needs to be defined precisely, because there is really no middle value when n is an even integer. Here is our definition.

Definition. Let $\boldsymbol{X_1, X_2, \cdots, X_n}$ be any $\boldsymbol{n}$ real valued random variables. Then, the median of $\boldsymbol{X_1, X_2, \cdots, X_n}$ is defined to be

$$\boldsymbol{M_n = X_{(m+1)}} \text{ if } \boldsymbol{n = 2m + 1} \text{ (an odd integer)},$$

and

$$\boldsymbol{Mn = X(m)} \text{ if } \boldsymbol{n = 2m} \text{ (an even integer)}.$$

That is, in either case, the median is the order statistic $\boldsymbol{X_{(k)}}$ where $\boldsymbol{k}$ is the smallest integer $\geq \frac{\boldsymbol{n}}{\boldsymbol{2}}$.

Example. Suppose 0.3, 0.53, 0.68, 0.06, 0.73, 0.48, 0.87, 0.42, 0.89, 0.44 are ten independent observations from the $\boldsymbol{U[0, 1]}$ distribution. Then, the order statistics are 0.06, 0.3, 0.42, 0.44, 0.48, 0.53, 0.68, 0.73, 0.87, 0.89. Thus, $\boldsymbol{X_{(1)} = 0.06}$, $\boldsymbol{X_{(n)} = 0.89}$, and since $\frac{\boldsymbol{n}}{\boldsymbol{2}} = \boldsymbol{5}, \boldsymbol{M_n = X_{(5)} = 0.48}$.

18.1.1 Distribution of Order Statistics

Let $\boldsymbol{X_1, X_2, \cdots, X_n}$ be i.i.d. absolutely continuouslt distributed variables, and $\boldsymbol{X_{(1)}, X_{(2)}, \cdots, X_{(n)}}$ be the corresponding order statistics. Let $\boldsymbol{f(x)}$ be the probability density function and $\boldsymbol{F(x)}$ be the cumulative distribution function of $\boldsymbol{X_i}$. Then the probability density of the $\boldsymbol{k^{th}}$ statistic can be found as follows.

Theorem:

The marginal cdf of $\boldsymbol{Y = X_{(i)}}$ is

$$\boldsymbol{F_Y(y) = \sum_{k=i}^{n} \binom{n}{k} F^k(y)(1 - F(y))^{n-k}, \quad \text{for any } y \in R}$$

and the marginal pdf of $\boldsymbol{Y = X_{(i)}}$is

$$\boldsymbol{f_Y(y) = f(y)n\binom{n-1}{i-1} F^{i-1}(y)(1 - F(y))^{n-i}, \quad \text{for any } y \in R}$$

Results:

Following the above Theorem, we can get the distribution of the largest Order Statistic (i.e., $\boldsymbol{X_{(n)}}$)and the smallest Order Statistic(i.e., $\boldsymbol{X_{(1)}}$), respectively,

$$\boldsymbol{P(X_{(n)} \le y) = F_Y(y) = (P(X \le y))^n = F^n(y)}$$
$$\boldsymbol{P(X_{(1)} \le y) = \sum_{k=1}^{n} \binom{n}{k} F^k(y)(1 - F(y))^{n-k} = 1 - P(k = 0) = 1 - (1 - F(y))^n.}$$

Actually, the above results can be proved easily by

$$\begin{aligned}
\boldsymbol{P(X_{(n)} \le y)} &= \boldsymbol{P(\max\{X_1, X_2, \cdots, X_n\} \le y) = P(X_1 \le y, X_2 \le y, \cdots, X_n \le y)} \\
&= \boldsymbol{P(X_1 \le y)P(X_2 \le y) \cdots P(X_n \le y) = F^n(y),} \\
\boldsymbol{P(X_{(1)} \le y)} &= \boldsymbol{1 - P(X_{(1)} > y)} \\
&= \boldsymbol{1 - P(\min\{X_1, X_2, \cdots, X_n\} > y) = 1 - P(X_1 > y, X_2 > y, \cdots, X_n > y)} \\
&= \boldsymbol{1 - P(X_1 > y)P(X_2 > y) \cdots P(X_n > y) = 1 - (1 - F(y))^n,}
\end{aligned}$$

Example:

If $\boldsymbol{X_1, X_2, \cdots, X_n \sim U(0, b)}$, then we have

$$\boldsymbol{f_X(x) = \begin{cases} \frac{1}{b} & 0 \le x \le b, \\ 0 & \text{o.w.} \end{cases}}$$

and

$$\boldsymbol{F_X(x) = \begin{cases} 0 & x < 0, \\ \frac{x}{b} & 0 \le x < b, \\ 1 & x \ge b. \end{cases}}$$

Find the cdf and pdf of $\boldsymbol{X_{(1)}}$ and $\boldsymbol{X_{(n)}}$.

Let $\boldsymbol{Y = X_{(n)}}$, then

$$\begin{aligned} \boldsymbol{F_Y(y) = P(Y \le y) = P(X_{(n)} \le y) = P(X \le y) = [F_X(y)]^n} \\ \boldsymbol{= \begin{cases} 0 & y < 0, \\ (\frac{y}{b})^n & 0 \le y < b, \\ 1 & y \ge b. \end{cases}} \end{aligned}$$

The corresponding pdf can be obtained by

$$\boldsymbol{f_Y(y) = \frac{d}{dy}[F_Y(y)] = \begin{cases} \frac{n}{b}\left(\frac{y}{b}\right)^{n-1} & 0 \le y \le b, \\ 0 & \text{o.w.} \end{cases}}$$

Let $\boldsymbol{Y = X_{(1)}}$, then

$$\begin{aligned} \boldsymbol{F_Z(z) = P(Z \le z) = P(X_{(1)} \le z) = 1 - P(X_{(1)} > z)} \\ \boldsymbol{= 1 - [1 - P(X \le z))]^n = 1 - [1 - F_X(z)]^n} \\ \boldsymbol{= \begin{cases} 0 & z < 0, \\ 1 - (1 - \frac{z}{b})^n & 0 \le z < b, \\ 1 & z \ge b. \end{cases}} \end{aligned}$$

The corresponding pdf can be obtained by

$$\boldsymbol{f_Z(z) = \frac{d}{dz}[F_Z(z)] = \begin{cases} \frac{n}{b}\left(1 - \frac{z}{b}\right)^{n-1} & 0 \le z \le b, \\ 0 & \text{o.w.} \end{cases}}$$

Statistical models are always, at best, an approximation. Methods that make few assumptions about the data are therefore desirable. However, there is no free lunch: gaining robustness by dropping assumptions always comes at the cost of a loss of efficiency. Notably, this loss can be quite modest.

18.2 Sign Test

This is a simple alternative to the one-sample t-test, this time targeting the median of a distribution. Let $X_1, \cdots, X_n$ be an i.i.d. random sample from a distribution with median M. Assume this median is uniquely defined, in particular assume

$$P(X_i = M) = 0.$$

We wish to test

$$H_0 : M = M_0 \text{ against } H_1 : M > M_0.$$

We will also consider the two-sided alternative $H_1 : M \neq M_0$. A suitable test statistic is just

$$T = \sum_{i=1}^{n} 1_{\{X_i > M_0\}},$$

which means that T be the number of X_i which satisfies $X_i > M_0$. Under H_0, $T \sim$Bin$(n, 1/2)$, if we let p be the probability of the Binomial distribution, we may re-write down the hypothesis as

$$H_0 : p = 1/2 \text{ against } H_1 : p > 1/2.$$

If testing against the one-sided alternative $H_1 : M > M_0$ we should reject the null hypothesis is T is significantly large.

Example

The table below shows the grades obtained by each of a random sample of ten students in two pieces of Statistic Coursework.

student	1	2	3	4	5	6	7	8	9	10
Coursework 1 grade	A	B	B	C	D	C	C	A	B	C
Coursework 2 grade	B	C	B	D	C	D	C	B	C	D

What distinguishes these data from those considered so far?

- Category - grades in letters not numbers.
- Paired - same ten students constituted both samples.

– Can we use the Paired sample t-test?

In fact, difference can be reasonably listed as

$$+ \quad + \quad \mathbf{0} \quad + \quad - \quad + \quad \mathbf{0} \quad + \quad + \quad +$$

where

+ denotes that coursework 1 grade is better than coursework 2 grade.

0 denotes that coursework 1 grade is equal to coursework 2 grade.

- denotes that coursework 1 grade is worse than coursework 2 grade.

The number of non-zero differences is $\mathbf{10 - 2 = 8}$. Let X be the number of positive signs (+ signs) with observed value of 7, then $\boldsymbol{X \sim}$ Bin$(\boldsymbol{n, p})$. If $\boldsymbol{H_0}$ is true, the $\boldsymbol{X \sim}$ Bin$(\mathbf{8, 0.5})$.

For a two-sided test ($\boldsymbol{H_1 : p \neq 1/2}$) with $\boldsymbol{\alpha = 0.05}$, we can calculate the p-value by

$$\boldsymbol{P(X \geq 7) = 1 - P(X \leq 6) = 0.0352 > 0.025 = \alpha/2.}$$

Thus, we do not reject the $\boldsymbol{H_0}$ at 5% significant level. There is no significant evidence to suggest that there is a difference between the grades achieved in the two statistics coursework.

For a one-sided test (It is reasonable to make the alternative hypothesis as $\boldsymbol{H_1 : p > 1/2}$ since we've got more "+"'s than "-"'s.) with $\boldsymbol{\alpha = 0.05}$, we can calculate the p-value by

$$\boldsymbol{P(X \geq 7) = 1 - P(X \leq 6) = 0.0352 < 0.05 = \alpha.}$$

Thus, we do reject the $\boldsymbol{H_0}$ at 5% significant level. There is significant evidence to suggest that the grades achieved in coursework 1 are better than those in coursework 2.

Finally, If n is large, binomial probabilities can be approximated by normal probabilities, due to the Central Limit Theorem. More precisely, asymptotically $\boldsymbol{T \sim N(n/2, n/4)}$. Note that this median test also allows you to get confidence intervals for the median.

Remark:

Zero differences occur if $\boldsymbol{X_i = M_0}$ for at least one $\boldsymbol{i}$. It is common practice to ignore such observations and adjust $\boldsymbol{n}$ accordingly. An alternative is to count half of the zero differences as positive differences. Yet another possibility of handling the zeros is to assign to all zeros that sign which is least conductive to rejection of $\boldsymbol{H_0}$; this is a strictly conservative approach. Is this notes however we will assume there are no zero-differences for simplicity.

18.3 Ranks

Sign tests provide a very simple, but crude approach to testing. Note that all the magnitude information is being ignored, and even the relative ordering of the samples (which provides quite a lot of information) is completely neglected. This is too dramatic, and we should hope to do better by retaining some of this information.

The rank of $\boldsymbol{X_i}$ within the random sample $\boldsymbol{X_1, \cdots, X_n}$ is given by

$$\boldsymbol{R(X_i) = \sum_{j=1}^{n} 1_{X_i - X_j \geq 0}.}$$

That is, the rank of $\boldsymbol{X_i}$ is the number of elements in the sample that are smaller or equal to $\boldsymbol{X_i}$. Therefore the rank assigns an integer number to the ordered sample. If there are no repetitions in the sample than the sample element $\boldsymbol{X_i}$ corresponding to $\boldsymbol{X_{(1)}}$ is assigned rank $\boldsymbol{R(X_i) = 1}$ and so on.

Important:

Note that the notation for the rank is slightly incomplete, as it doesn't explicitly specify the complete sample used for the rank calculation. For this reason it is important to clearly state with respect to which sample is the rank computed upon, even if relatively clear from the context.

If the $\boldsymbol{X_j}$ is a sample from a continuous distribution, then the random variable $\boldsymbol{R(X_i)}$ follows a discrete uniform distribution on $\boldsymbol{\{1, 2, \cdots, n\}}$ and so

$$\boldsymbol{P(R(X_i) = j) = \frac{1}{n},} \text{ for } \boldsymbol{j = 1, 2, \cdots, n.}$$

Although $\boldsymbol{X_1, \cdots, X_n}$ are independent, $\boldsymbol{R(X_1), \cdots, R(X_n)}$ are of course not. The first and second moments of $\boldsymbol{R(X_i)}$ are given by

$$\boldsymbol{E[R(X_i)] = \frac{1}{n} \sum_{j=1}^{n} j = \frac{n+1}{2},}$$

$$\boldsymbol{E[R^2(X_i)] = \frac{1}{n} \sum_{j=1}^{n} j^2 = \frac{(n+1)(2n+1)}{6},}$$

which implies

$$\boldsymbol{Var[R(X_i)] = \frac{n^2 - 1}{12}}$$

$$\boldsymbol{Cov(R(X_i), R(X_j)) = -\frac{n+1}{12}} \text{ when } \boldsymbol{i \neq j.}$$

Rank based tests are permutation tests that are applied to the ranks of the observations, instead of the observations themselves. Rank tests do not depend on any distributional assump- tions of the

observations. Instead, the relative magnitude and ordering of the observations is utilized. This is particularly useful if observations are not on a numerical but on an ordinal scale.

Advantages of rank tests over classical tests include

- rank tests are less sensitive to outliers;
- rank tests often possess reasonable efficiency;
- rank tests are easy to explain;
- rank tests can be applied without a population model (a randomization model suffices);
- rank tests are applicable to ordinal data (whereas classical tests are strictly speaking not).

18.3.1 The Wilcoxon Signed-Rank Test

If we are willing to make the extra assumption that the data comes from a symmetric distribution under the null hypothesis we can take advantage of the extra information in the ranks effectively (more on this assumption later). Recall that the sign-test only utilizes the sign of the differences between each observation and the hypothesized median $\boldsymbol{M_0}$. The magnitudes of these observations relative to $\boldsymbol{M_0}$ are ignored. If these magnitudes are available and taken into account, better testing procedures can be expected. The Wilcoxon signed rank test is one of the best known examples of such a procedure.

Assume that the distribution of the observations is continuous and symmetric. We wish to test that the median equals a prescribed value $\boldsymbol{M_0}$: $\boldsymbol{H_0 : M = M_0}$. Define $\boldsymbol{D_i = X_i - M_0}$, and let $\boldsymbol{R}(|\boldsymbol{D_i}|)$ denote the rank of $|\boldsymbol{D_i}|$ among $\{|\boldsymbol{D_1}|, \cdots, |\boldsymbol{D_n}|\}$, the WSiR statistic is defined by

$$\boldsymbol{W^+ = \sum_{i=1}^{n} R(|D_i|)1_{\{D_i>0\}}.}$$

Large values of $\boldsymbol{W^+}$ favor the alternative hypothesis $\boldsymbol{H_1 : M > M_0}$. This is the positive rank sum of the absolute differences. Notice that if $\boldsymbol{R(.) = 1}$ this would be simply the sign test. In effect, we order the absolute differences $|\boldsymbol{D_1}|, |\boldsymbol{D_2}|, \cdots, |\boldsymbol{D_n}|$ from smallest to largest and assign them ranks $\boldsymbol{1, 2, \cdots, n}$, while keeping track of the signs of the differences $\boldsymbol{D_i}$.

If the alternative hypothesis is $\boldsymbol{H_1 : M < M_0}$, then we may use

$$\boldsymbol{W^- = \sum_{i=1}^{n} R(|D_i|)1_{\{D_i<0\}}.}$$

instead. However, since $\boldsymbol{W^+ + W^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}}$ tests based on $\boldsymbol{W^+}$, $\boldsymbol{W^-}$ or $\boldsymbol{W^+ - W^-}$ are all equivalent. Now large or small values of

$$\boldsymbol{W = W^+ - W^- = \sum_{i=1}^{n} R(|D_i|)\text{sign}(D_i) = 2W^+ - \frac{n(n+1)}{2}}$$

favor the alternative hypothesis $\boldsymbol{H_1 : M \neq M_0}$. In the above

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

A natural question arises to which is the distribution of W^+ under the null hypothesis. Under H_0, the random variables $1_{\{D_1>0\}}$, $1_{\{D_2>0\}}$,...,$1_{\{D_n>0\}}$ are a random sample from a Bernoulli distribution with success probability $1/2$. The following fact is going to be extremely useful.

Result

Suppose that $D_1, \cdots, D_n$ are mutually independent random variables, symmetric around 0 and with continuous distribution function. The vectors $(R(|D_1|), \cdots, R(|D_n|))$ and $(1_{\{D_1>0\}}, 1_{\{D_2>0\}}, ..., 1_{\{D_n>0\}})$ are independent.

The result above implies that

$$\begin{aligned} E(W^+) &= E\left[\sum_{i=1}^{n} R(|D_i|)1_{\{D_i>0\}}\right] \\ &= \sum_{i=1}^{n} E\left[R(|D_i|)1_{\{D_i>0\}}\right] \\ &= \frac{n(n+1)}{4}. \end{aligned}$$

In a similar way, under H_0,

$$Var(W^+) = \frac{n(n+1)(2n+1)}{24}.$$

An application of Lyapunov's central limit theorem (omitting details here) yields

$$\frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}} \sim N(0,1),$$

and 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.

So, in applying a test using the statistic W^+ we can either use the normal approximation to compute the rejection region (if n is large), or compute the exact test statistic by carefully enumerating all the possibilities and compare with the critical value from the Statistical Table.

Example

A study of early childhood education asked kindergarten students to retell two 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:

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 hypothesis

$\boldsymbol{H_0}$ **:** Scores have the same distribution for both stories.

$\boldsymbol{H_1}$ **:** Scores have systematically higher for story 2.

It is clear this is a matched pairs design, we base our inference on the differences. The matched pairs t test gives $t = \mathbf{0.635}$ with one-sided p-value = 0.280. Displays of the data suggest some lack of normality. We would therefore like to use a nonparametric test: the Wilcoxon Signed-Rank Test.

The difference are 0.37, -0.23, 0.66, -0.08, -0.17. We consider absolute values of the differences: 0.37, 0.23, 0.66, 0.08, 0.17. If scores are generally higher with Story 2, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. Now we arrange absolute values in increasing order and assign ranks

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

The sum of the ranks of the positive differences with ignoring the zeros

$$W^+ = 4 + 5 = 9.$$

We then can compare the observed test statistic with critical value from table or get p-value by Software.

The above table can be used to find the critical values of the test statistic W^+ for both one and two tail tests for samples of $n \leq 20$.

For a two-tail test, you reject the null hypothesis if the computed W^+ test statistic equals or is greater than the upper critical value or is equal to or less than the lower critical value. For a one-tail test in the lower tail, you reject the null hypothesis if the computed W^+ test statistic is less than or equal to the

ONE-TAIL	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
TWO-TAIL	$\alpha = .10$	$\alpha = .05$	$\alpha = .02$	$\alpha = .01$
n		*(Lower, Upper)*		
5	0,15	—,—	—,—	—,—
6	2,19	0,21	—,—	—,—
7	3,25	2,26	0,28	—,—
8	5,31	3,33	1,35	0,36
9	8,37	5,40	3,42	1,44
10	10,45	8,47	5,50	3,52
11	13,53	10,56	7,59	5,61
12	17,61	13,65	10,68	7,71
13	21,70	17,74	12,79	10,81
14	25,80	21,84	16,89	13,92
15	30,90	25,95	19,101	16,104
16	35,101	29,107	23,113	19,117
17	41,112	34,119	27,126	23,130
18	47,124	40,131	32,139	27,144
19	53,137	46,144	37,153	32,158
20	60,150	52,158	43,167	37,173

Source: Adapted from Table 2 of F. Wilcoxon and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures* (Pearl River, NY: Lederle Laboratories, 1964), with permission of the American Cyanamid Company.

Figure 12: Lower and Upper Critical Values, W, of Wilcoxon Signed Ranks Test

lower critical value. For a one tail test in the upper tail, the decision rule is to reject the null hypothesis if the computed $\boldsymbol{W^+}$ test statistic equals or is greater than the upper critical value.

In our example, a one sided test in the upper tail ($\boldsymbol{D > 0}$) is run and $\boldsymbol{W^+ = 9 < 15}$, thus we fail to reject $\boldsymbol{H_0}$ at 5% significant level.

If sample size is large enough ($\boldsymbol{n \geq 20}$), normal approximation can be used and $\boldsymbol{p}$-value can be calculated. Here is the result by Minitab. The $\boldsymbol{p}$-value=0.394 indicates that there is no evidence against $\boldsymbol{H_0}$ at 5% significant level.

```
Wilcoxon Signed Rank Test: difference


Test of median = 0.000000 versus median > 0.000000

               N for   Wilcoxon         Estimated
           N   Test  Statistic      P     Median
difference 5      5        9.0  0.394     0.1000
```

In this case, $\boldsymbol{E(W^+) = \frac{5\times 6}{4} = 7.5}$ and $\boldsymbol{Var(W^+) = \sqrt{\frac{5\times 6\times 11}{24}} = 3.708}$. Our observed $\boldsymbol{W^+ = 9}$ slightly $\boldsymbol{> 7.5}$ indicates that we can not reject $\boldsymbol{H_0}$. If the sample size is large enough, we can then use normal probability calculations (with the continuity correction) to obtain approximate $\boldsymbol{p}$—values for $\boldsymbol{W^+}$ and we find the normal approximation for the $\boldsymbol{p}$-value by standardizing and using the standard normal table,

$$\begin{aligned}\boldsymbol{P(W^+ \geq 9) \approx P(W^+ \geq 8.5) = P(\frac{W^+ - 7.5}{3.708} \geq \frac{8.5 - 7.5}{3.708})} \\ \boldsymbol{= P(Z \geq 0.27) = 0.394.}\end{aligned}$$

Again, we do not reject $\boldsymbol{H_0}$ at 5% significant level.

Example(Treatment of Ties):

Here are the golf scores of 12 members of a college womenąŕs golf team in two rounds of tournament play. (A golf score is the number of strokes required to complete the course, so that low scores are better.)

Player	1	2	3	4	5	6	7	8	9	10	11	12
Round 2	94	85	89	89	81	76	107	89	87	91	88	80
Round 1	89	90	87	95	86	81	102	105	83	88	91	79
Difference	5	-5	2	-6	-5	-5	5	-16	4	3	-3	1

Negative differences indicate better (lower) scores on the second round (6 of the 12 golfers improved their scores.). We would like to test the hypotheses that in a large population of collegiate women golfers

$\boldsymbol{H_0}$ **:** Scores have the same distribution for Rounds 1 and 2.

$\boldsymbol{H_1}$ **:** Scores have systematically lower or higher in Round 2.

we arrange absolute values of differences in increasing order and assign ranks

Absolute value of the differences	5	**5**	2	**6**	**5**	**5**	5	**16**	4	3	**3**	1
Order the absolute value	1	2	**3**	3	4	**5**	5	**5**	5	**5**	**6**	**16**
Rank	1	2	**3.5**	3.5	5	**8**	8	**8**	8	**8**	**11**	**12**

Here, tied values receive the average of their ranks and we have $\boldsymbol{W^+ = 50.5}$ and $\boldsymbol{W^- = 27.5}$. This can be checked by $\boldsymbol{W^+ = W^- = 1/2n(n+1) = 78}$. For a two-sided test,

- we can compare with the critical value, $\boldsymbol{W^+ = 50.5 \in (13, 65)(5\%)}$, thus we do not reject $\boldsymbol{H_0}$ at 5% significant level.

- Statistical software gives p-value=0.388, also suggests that H_0 is not rejected at 5% significant level.
- For these data, the matched pairs t-test gives $t = 0.9314$ with $p = 0.3716$. Once again, t and W^+ lead to the same conclusion.

Assumption underlies the Wilcoxon signed rank test

The term non-parametric tests has generally come to mean tests where there is no assumption that the underlying population has a Normal distribution population. A common mistake is to think such tests require no assumptions, but this is not true. The assumption made is that the variable being tested is **symmetrically** distributed about the median, which would also be the mean. Remember too that it is still vitally important that your sample has been randomly chosen from the population.

18.3.2 The Wilcoxon Rank-Sum Test

Where you have two unrelated samples, which you wish to compare, the Wilcoxon rank-sum test is used. The Wilcoxon rank-sum test is a nonparametric alternative to the two sample t-test which is based solely on the order in which the observations from the two samples fall. We will use the following as a running example.

Example

In a genetic inheritance study discussed by Margolin [1988], samples of individuals from several ethnic groups were taken. Blood samples were collected from each individual and several variables measured. We shall compare the groups labeled 'Native American' and 'Caucasian' with respect to the variable MSCE (mean sister chromatid exchange). The data is as follows:

Native American:	8.50	9.48	8.65	8.16	8.83	7.76	8.63		
Caucasian:	8.27	8.20	8.25	8.14	9.00	8.10	7.20	8.32	7.70

Looking at the data sets for the two groups, several questions come to mind. Firstly, do the data come from Normal distributions? Unfortunately we can't say much about the distributions as the samples are too small. However there does not seem to be any clear lack of symmetry. Secondly, are the two distributions similar in shape? Again it is hard to say much with such small samples, though the Caucasian data seems to have longer tails. Finally, is there any difference in the centers of location? We shall now put this type of problem in a more general context and come back to this example later.

Suppose, more generally, that we have samples of observations from each of two populations $\boldsymbol{A}$ and $\boldsymbol{B}$ containing $\boldsymbol{n_A}$ and $\boldsymbol{n_B}$ observations respectively. We wish to test the hypothesis that the distribution of measurements in population A is the same as that in B, which we will write symbolically as $\boldsymbol{H_0 : A = B}$. The departures from $\boldsymbol{H_0}$ that the Wilcoxon test tries to detect are location shifts. If we expect to detect that the distribution of $\boldsymbol{A}$ is shifted to the right of distribution $\boldsymbol{B}$ as in Figure 13 we will write this as $\boldsymbol{H_1 : A > B}$. The other two possibilities are $\boldsymbol{H_1 : A < B}$ (A is shifted to the left of B), and the two sided-alternative, which we will write as $\boldsymbol{H_1 : A \neq B}$, for situations in which we have no strong prior reason for expecting a shift in a particular direction.

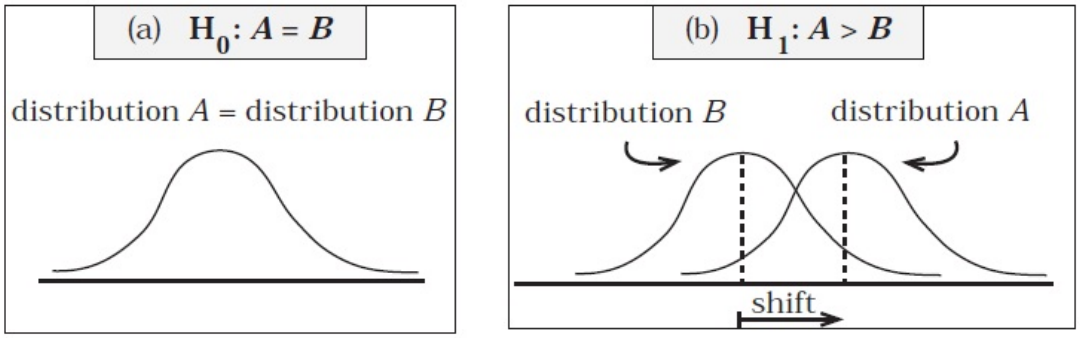

Figure 13: Illustration of $\boldsymbol{H_0 : A = B}$ versus $\boldsymbol{H_1 : A > B}$.

The Wilcoxon test is based upon ranking the $\boldsymbol{n_A + n_B}$ observations of the combined sample. Each observation has a rank: the smallest has rank 1, the 2nd smallest rank 2, and so on. The Wilcoxon rank-sum test statistic is the sum of the ranks for observations from one of the samples. Let us use sample $\boldsymbol{A}$ here and use $\boldsymbol{w_A}$ to denote the observed rank sum and $\boldsymbol{W_A}$ to represent the corresponding random variable.

$$\boldsymbol{w_A} = \text{ sum of the ranks for observations from A.}$$

Example cont

We have sorted the combined data set into ascending order and used vertical displacement as well as ethnic group labels to make very clear which sample an observation comes from ('NA' for the Native American group and 'Ca' for the Caucasian group). The rank of an observation in the combined sample appears immediately below the label.

	7.20	7.70	7.76	8.10	8.14	8.16	8.20	8.25	8.27	8.32	8.50	8.63	8.65	8.83	9.00	9.48
Race	Ca	Ca	NA	Ca	Ca	NA	Ca	Ca	Ca	Ca	NA	NA	NA	NA	Ca	NA
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The sum of the ranks for the Native American group is

$$w_{NA} = 3 + 6 + 11 + 12 + 13 + 14 + 16 = 75.$$

Suppose that $H_1 : A > B$ is true: In this case we would expect behavior more like that in Fig. 13(b) which results in sample A containing more of the larger ranks. Evidence against H_0 which confirms $H_1 : A > B$ is thus provided by an observed rank sum w_A which is unusually large according to the distribution of rank sums when H_0 is true. Thus the p-value for the test is

$$p - value = P(W_A \geq w_A),$$

where the probability is calculated using the distribution that W_A would have if H_0 was true. Suppose, on the other hand, that the alternative $H_1 : A < B$ is true. The p-value for the alternative $H_1 : A < B$ is therefore

$$p - value = P(W_A \leq w_A).$$

For the two-sided test, i.e. testing $H_0 : A = B$ versus the alternative $H_1 : A \neq B$, a rank sum that is either too big or too small provides evidence against H_0. We then calculate the probability of falling into the tail of the distribution closest to w_A and double it. Thus if w_A is in the lower tail then $p-value = 2P(W_A \leq w_A)$, whereas if w_A is in the upper tail then $p-value = P(W_A \geq w_A)$.

Example cont

Here, we want to test a null hypothesis H0 which says that the MSCE distribution for Native Americans is the same as that for Caucasians. Although the Native American MSCE values in the data tend to be higher, there was no prior theory to lead us to expect this so we should be 4 doing a two-sided test. The rank sum for for the Native American group was $w_{NA} = 75$. We know from the data that this will be in the upper tail of the distribution. The p-value is thus

$$p - value = 2P(W_{NA} \geq 75) = 0.1123(software).$$

Minitab confirms the result.

```
Mann-Whitney Test: NA, Ca
```

```
    N  Median
NA  7  8.6300
Ca  9  8.2000

W = 75.0
Test of NA = Ca vs NA not = Ca is significant at 0.1123
```

The evidence against $\boldsymbol{H_0}$ which suggests that median MSCE measurements are higher for Native Americans than for Caucasians is, at best, weak. In fact we can't be sure that this evidence points to a difference in the shapes of the two distributions rather than a difference in the centers of location.

When one performs a Wilcoxon test by hand, Tables are required to find $\boldsymbol{p}$-values. For small sample sizes, tables for Wilcoxon rank-sum test are given. We supplement this with a Normal approximation for use with larger samples. All probabilities discussed relate to the distribution of $\boldsymbol{W_A}$ when $\boldsymbol{H_0}$ is true.

Small sample Tables

Tables for the Wilcoxon rank-sum test are given. When the two samples have different sizes, the tables are set up for use with the rank sum for the smaller of the two samples, so that we define sample $\boldsymbol{A}$ to be the smaller of the two samples. One chooses the row of the table corresponding to the combination of sample sizes, $\boldsymbol{n_A}$ and $\boldsymbol{n_B}$, that one has.

For given $\boldsymbol{n_A}$, $\boldsymbol{n_B}$ and $\boldsymbol{prob}$, the tabulated value for the lower prob-tail is the largest value of $\boldsymbol{w_A}$ for which $\boldsymbol{P(W_A \leq w_A) \leq prob}$. For example, when $\boldsymbol{n_A = 7}$ and $\boldsymbol{n_B = 9}$, the tabulated value for $\boldsymbol{prob = 0.2}$ is $\boldsymbol{w_A = 50}$. Thus,

$$\boldsymbol{P(W_A \leq 50) \leq 0.2,} \text{ but } \boldsymbol{P(W_A \leq 51) > 0.2.}$$

In other words, the values in the lower 0.2 (or 20%) tail are those $\boldsymbol{\leq 50}$.

Normal approximation for larger samples

Our Wilcoxon Tables cater for sample sizes up to $\boldsymbol{n_A = n_B = 12}$. When both sample sizes are 10 or greater, we can treat the distribution of $\boldsymbol{W_A}$ as if it were Normal$(\boldsymbol{\mu_A, \sigma_A})$, where

$$\boldsymbol{\mu_A = \frac{n_A(n_A + n_B + 1)}{2}} \text{ and } \boldsymbol{\sigma_A = \sqrt{\frac{n_A n_B(n_A + n_B + 1)}{12}}.}$$

More precisely,

$$P(W_A \geq w_A) \approx P(Z \geq z), \quad \text{where } z = \frac{w_A - \mu_A}{\sigma_A}$$

and $Z \sim N(0,1)$. For example, suppose that $n_A = 10$, $n_B = 12$, and we want $P(W_A \geq 145)$. Then, $\mu = 10 \times (10 + 12 + 1)/2 = 115$ and

$$\sigma_A = \sqrt{\frac{10 \times 12 \times (10 + 12 + 1)}{12}} = 15.16575$$

so that

$$P(W_A \geq 145) \approx P(Z \geq \frac{145 - 115}{15.16575}) = P(Z \geq 1.978) = 0.024.$$

notes

1. The Wilcoxon test is still valid for data from any distribution, whether Normal or not, and is much less sensitive to outliers than the two-sample t-test.
2. If one is primarily interested in differences in location between the two distributions, the Wilcoxon test has the disadvantage of also reacting to other differences between the distributions such as differences in shape.
3. When the assumptions of the two-sample t-test hold, the Wilcoxon test is somewhat less likely to detect a location shift than is the two-sample t-test. However, the losses in this regard are usually quite small.