Math 3080 § 1.	**Carapace Example: Confidence Intervals**	Name: Example
Treibergs	**in the Wilcoxon Sign-rank Test**	April 19, 2014

In this discussion, we look at confidence intervals in the Wilcoxon Sign-Rank Test for one sample. The data comes from an article by Jeffries, Voris and Yang, "Diversity and Distribution of the Pedunculate Barnacles ... on the Scillarid Lobster," *Crustaceana,* 1984, as quoted by Mendenhall, Beaver & Beaver, *Introduction to Probability sand Statistics,* 14th ed., Brooks Cole, 2013. The carapace lengths (in mm) of randomly selected lobsters caught in the seas near Singapore are simulated consistent with the given data.

We assume that the sample $X_1, \ldots, X_N$ comes from a continuous distribution that is symmetric about its mean μ. The Wilcoxon signed rank statistic is computed as follows. To test

$$\mathcal{H}_0 : \mu = \mu_0;$$
$$\mathcal{H}_a : \mu \neq \mu_0.$$

we order the values $|X_i - \mu_0|$ from lowest to highest. after throwing out any zero values, we have n nonzero terms. We rank them 1 to n. S_+ is the sum of the ranks corresponding to terms for which $X_i - \mu_0$ is positive. The null hypothesis is rejected at the significance level α if $S_+ \geq c$ or $S_+ \leq \frac{n(n+1)}{2} - c$ where $2P(S_+ \geq c) = \alpha$.

Theorem 1. *Fix μ_0. Assume that there are no ties and no zeros among the values $|X_i - \mu_0|$ for $i = 1 \ldots, n$. Then S_+ is equal to the number of pairs (i, j) such that $i \leq j$ and $X_i + X_j \geq 2\mu_0$.*

Proof. Observe that the number of such pairs is unchanged if the observations are rearranged. Thus we may suppose that the observations are arranged from smallest to largest $X_{(1)} < \cdots < X_{(n)}$. Let r_i denote the rank of $|X_{(i)} - \mu_0|$. Let k be the index where the $X_{(i)} - \mu_0$ change sign, that is

$$X_{(k)} - \mu_0 < 0 < X_{(k+1)} - \mu_0$$

Then the ranks decrease to k and then increase from $k + 1$:

$$r_1 > r_2 > \cdots > r_k, \qquad r_{k+1} < r_{k+2} < \cdots < r_n.$$

For example if the sample is $12.2, 10.0, 11.1, 15.5, 14.4$ then $X_{(1)} = 10.0$, $X_{(2)} = 11.1$, $X_{(3)} = 12.2$, $X_{(4)} = 14.4$ and $X_{(5)} = 17.7$ so if $\mu_0 = 13.9$ then $|X_{(1)} - \mu_0| = 3.9$, $|X_{(2)} - \mu_0| = 2.8$, $|X_{(3)} - \mu_0| = 1.7$, $|X_{(4)} - \mu_0| = .5$ and $|X_{(5)} - \mu_0| = 3.8$ thus $r_1 = 5$, $r_2 = 3$, $r_3 = 2$, $r_4 = 1$ and $r_5 = 4$, and finally $k = 3$ and $S_+ = r_{k+1} + \cdots + r_n = 1 + 4 = 5$.

Let $\chi_{\mathcal{P}}$ be the indicator function which equals 1 if $\mathcal{P}$ is true and 0 otherwise. Then the number of pairs satisfying the condition is

$$\begin{aligned}
\sum_{i \leq j} \chi\{X_i + X_j \geq 2\mu_0\} &= \sum_{i \leq j} \chi\{X_{(i)} + X_{(j)} \geq 2\mu_0\} && (1)\\
&= \sum_{j=k+1}^{n} \sum_{i=1}^{j} \chi\{X_{(j)} - \mu_0 \geq \mu_0 - X_{(i)}\} && (2)\\
&= \sum_{j=k+1}^{n} \sum_{i=1}^{n} \chi\{r_i \leq r_j\} && (3)\\
&= \sum_{j=k+1}^{n} r_j && (4)\\
&= S_+ && (5)
\end{aligned}$$

(1) holds because the number of pairs does not depend on the order of the variables. (2) holds because if $i \le j \le k$ then $X_{(i)} - \mu_0 < X_{(j)} - \mu_0 < 0$ so that $X_{(i)} + X_{(j)} < 2\mu_0$ and does not add to the sum. $k+1 \le i \le j$ if and only if $r_i \le r_j$ and $j < i$ if and only if $r_j < r_i$. On the other hand if $i \le k$ then $r_i \le r_j$ if and only is $\mu_0 - X_{(i)} = |X_{(i)} - \mu_0| \le |X_{(j)} - \mu_0| = X_{(j)} - \mu_0$ so $X_{(i)} + X_{(j)} \ge 2\mu_0$. On the other hand, $r_i \le r_j$ if and only is $\mu_0 - X_{(i)} = |X_{(i)} - \mu_0| > |X_{(j)} - \mu_0| = X_{(j)} - \mu_0$ so $X_{(i)} + X_{(j)} < 2\mu_0$. It means that for a given j and any i, $X_{(i)} + X_{(j)} \ge 2\mu_0$ if and only if $r_i < r_j$ so we may replace the sum in (3). Now there are exactly r_j is such that $r_i \le r_j$ giving (4). But (5) is the definition of S_+. □

The confidence interval associated to the Sign-Rank test $\mathcal{H}_0$ above is the following: we consider the set of values

$$\mathcal{A} = \left\{ \frac{X_i + X_j}{2} : 1 \le i \le j \le n \right\}$$

If there are no ties, then S_+ is the number of elements of $\mathcal{A}$ that are not smaller than μ_0. If we sort $\mathcal{A}$ into $\{A_{(1)} \le A_{(2)} \le \cdots \le A_{(m)}\}$ where $m = \frac{n(n+1)}{2}$.

The one and two-sided critical values for S_+ may be computed as follow. Assuming $\mathcal{H}_0$, the chance that the sign of the ith observation $X_i - \mu_0$ is positive or negative is equally likely. Thus the distribution of S_+ is obtained by looking at the histogram of values

$$\pm 1 \pm 2 \pm 3 \pm \cdots \pm n$$

each of which has an equal chance 2^{-n} of occurring. They range from 0 to m. The pmf $p(x)$ of S_+ is symmetrical about the mean $\mu_{S_+} = \frac{n(n+1)}{4}$. The one-sided upper critical value is $P(S_+ \ge c_1) = \sum_{i \ge c_1} p(i) = \alpha$. The two sided critical value c satisfies $P(S_+ \ge c) = P(S_+ \le m - c) = \frac{\alpha}{2}$. These numbers are tabulated in tables A 13 and A 15 of the text. We show how to compute the values of these tables in the code, albeit inefficiently.

Then the two-sided confidence interval for μ at the level α where c is the two-sided critical value that satisfies $P(S_+ \ge c) = \frac{\alpha}{2}$ is given by

$$\left(A_{(m+1-c)}, A_{(c)} \right).$$

Data Set Used in this Analysis :

```
# Math 3080              Carapace Data           April 19, 2014
# Treibergs
#
# From an article by Jeffries, Voris and Yang, "Diversity and Distribution
# of the Pedunculate Barnacles  on the Scillarid Lobster," Crustaceana,
# 1984  as quoted by Mendenhall, Beaver & Beaver, Introduction to
# Probability sand Statistics, 14th ed., Brooks Cole, 2013.
#
# The carapace lengths (in mm) of randomly selected lobsters caught in
# the seas near Singapore are simulated consistent with the given data.
Length
79.2
81.1
56.4
78.2
64.8
64.4
69.7
69.8
45.3
64.3
65.9
69.4
70.0
68.6
52.6
63.6
54.5
60.6
```

R Session:

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
```

```
Type 'q()' to quit R.

[R.app GUI 1.41 (5874) i386-apple-darwin9.8.0]

[History restored from /Users/andrejstreibergs/.Rapp.history]

>
> tt=read.table("M3082DataCarapace.txt",header=T)
> attach(tt)
> tt
   Length

> t2=read.table("M3082DataCarapace.txt",header=T)
> attach(t2)
   Length
1    79.2
2    81.1
3    56.4
4    78.2
5    64.8
6    64.4
7    69.7
8    69.8
9    45.3
10   64.3
11   65.9
12   69.4
13   70.0
14   68.6
15   52.6
16   63.6
17   54.5
18   60.6
> table(Length)
Length
45.3 52.6 54.5 56.4 60.6 63.6 64.3 64.4 64.8 65.9 68.6 69.4 69.7 69.8   70
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
78.2 79.2 81.1
   1    1    1
```

```
> ###############  CRITICAL VALUES FOR SIGNED-RANK TEST  #################
> ############### ILLUSTRATE DETAILS FOR  N=5  ###########################
> di=5
> #### MATRIX WHOSE COLS ARE ALL POSSIBLE POSITIVE RANK SUMS  ###########
> M=matrix(1:(di*2^di),ncol=2^di)
> for(i in 1:di){M[i,]=rep(i*0:1,each=2^(i-1),length=2^di)}
> M
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]    0    1    0    1    0    1    0    1    0     1     0     1     0
[2,]    0    0    2    2    0    0    2    2    0     0     2     2     0
[3,]    0    0    0    0    3    3    3    3    0     0     0     0     3
[4,]    0    0    0    0    0    0    0    0    4     4     4     4     4
[5,]    0    0    0    0    0    0    0    0    0     0     0     0     0
     [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
[1,]     1     0     1     0     1     0     1     0     1     0     1     0
[2,]     0     2     2     0     0     2     2     0     0     2     2     0
[3,]     3     3     3     0     0     0     0     3     3     3     3     0
[4,]     4     4     4     0     0     0     0     0     0     0     0     4
[5,]     0     0     0     5     5     5     5     5     5     5     5     5
     [,26] [,27] [,28] [,29] [,30] [,31] [,32]
[1,]     1     0     1     0     1     0     1
[2,]     0     2     2     0     0     2     2
[3,]     0     0     0     3     3     3     3
[4,]     4     4     4     4     4     4     4
[5,]     5     5     5     5     5     5     5
> ################  PDF OF  S+  ########################################
> p=table(margin.table(M,2))/2^di;sum(p)
[1] 1
> p

      0       1       2       3       4       5       6       7       8
0.03125 0.03125 0.03125 0.06250 0.06250 0.09375 0.09375 0.09375 0.09375
      9      10      11      12      13      14      15
0.09375 0.09375 0.06250 0.06250 0.03125 0.03125 0.03125

# #############  WORK OUT TAIL PROBABILITIES  #########################
> cump=cumsum(p)
> np=length(p);np;di*(di+1)/2+1
[1] 16
[1] 16
> names(cump)=np-1:np
> ####  N=5  ONE-TAILED CRITICAL FOR SIGNED-RANK TEST (TABLE A13)  ####
> ####  c1  AND  ALPHA TABLE  #########################################
> cump
     15      14      13      12      11      10       9       8       7
0.03125 0.06250 0.09375 0.15625 0.21875 0.31250 0.40625 0.50000 0.59375
      6       5       4       3       2       1       0
0.68750 0.78125 0.84375 0.90625 0.93750 0.96875 1.00000
```

```
> cum2p=1-2*cumsum(p)
> names(cum2p)=np-1:np
> ####  N=5  TWO-TAILED CRITICAL FOR SIGNED-RANK TEST (TABLE A15)  ####
> ####  c  AND  ALPHA TABLE  ##########################################
> cum2p[1:(np/2)]
    15     14     13     12     11     10      9      8
0.9375 0.8750 0.8125 0.6875 0.5625 0.3750 0.1875 0.0000
>
> #### COMPUTE TWO-SIDED CRITICAL VALUES FOR CARAPACE DATA  ###########

> di=length(Length); di
[1] 18
> M=matrix(1:(di*2^di),ncol=2^di)
> for(i in 1:di){M[i,]=rep(i*0:1,each=2^(i-1),length=2^di)}
> p=table(margin.table(M,2))/2^di;sum(p)
[1] 1
> cum2p=1-2*cumsum(p)
> np=length(p);np;di*(di+1)/2+1
[1] 172
[1] 172
> names(cum2p)=np-1:np
> ####  N=18  TWO-TAILED CRITICAL FOR SIGNED-RANK TEST (TABLE A15)  ###
> ####  c  AND  ALPHA TABLE  ##########################################
> cum2p[1:(np/2)]
       171        170        169        168        167        166        165
0.99999237 0.99998474 0.99997711 0.99996185 0.99994659 0.99992371 0.99989319
       164        163        162        161        160        159        158
0.99985504 0.99980927 0.99974823 0.99967194 0.99958038 0.99946594 0.99932861
       157        156        155        154        153        152        151
0.99916077 0.99895477 0.99871063 0.99842072 0.99806976 0.99766541 0.99719238
       150        149        148        147        146        145        144
0.99663544 0.99599457 0.99525452 0.99440002 0.99342346 0.99230957 0.99103546
       143        142        141        140        139        138        137
0.98959351 0.98796844 0.98612976 0.98406982 0.98176575 0.97918701 0.97632599
       136        135        134        133        132        131        130
0.97315216 0.96963501 0.96576691 0.96150970 0.95683289 0.95172119 0.94614410
       129        128        127        126        125        124        123
0.94006348 0.93346405 0.92631531 0.91857147 0.91023254 0.90126038 0.89161682
       122        121        120        119        118        117        116
0.88129425 0.87026215 0.85848236 0.84595490 0.83264923 0.81853485 0.80361176
       115        114        113        112        111        110        109
0.78785706 0.77124786 0.75379181 0.73547363 0.71627045 0.69620514 0.67527008
       108        107        106        105        104        103        102
0.65345001 0.63078308 0.60726166 0.58288574 0.55770111 0.53170776 0.50492096
       101        100         99         98         97         96         95
0.47738647 0.44912720 0.42015839 0.39054108 0.36030579 0.32947540 0.29811859
        94         93         92         91         90         89         88
0.26627350 0.23397064 0.20129395 0.16827393 0.13495636 0.10142517 0.06771851
        87         86
0.03388214 0.00000000
```

```
>
> Length
 [1] 79.2 81.1 56.4 78.2 64.8 64.4 69.7 69.8 45.3 64.3 65.9 69.4 70.0 68.6 52.6
[16] 63.6 54.5 60.6
> #########  GENERATE ALL (X[I]+X[J])/2  FOR  I \le J  #####################
> xPx=1:(n*(n+1)/2)
> n=length(Length);n
[1] 18
> for(j in 1:n){for(i in 1:j){xPx[j*(j-1)/2+i]=(Length[i]+Length[j])/2}}
> sxPx=sort(xPx)
> ##############  SORTED VALUES  (X[I]+X[J])/2  FOR  I \le J  ##############
> sxPx
  [1] 45.30 48.95 49.90 50.85 52.60 52.95 53.55 54.45 54.50 54.50 54.80 54.85
 [13] 55.05 55.45 55.60 56.40 56.60 56.95 57.35 57.50 57.55 57.55 57.65 58.10
 [25] 58.45 58.50 58.50 58.70 59.05 59.25 59.40 59.45 59.65 60.00 60.20 60.35
 [37] 60.40 60.60 60.60 60.60 61.00 61.15 61.15 61.20 61.30 61.55 61.75 61.95
 [49] 62.10 62.10 62.15 62.25 62.25 62.45 62.50 62.50 62.70 62.90 63.05 63.10
 [61] 63.20 63.20 63.25 63.60 63.95 64.00 64.20 64.30 64.35 64.40 64.55 64.60
 [73] 64.60 64.75 64.80 65.00 65.10 65.15 65.15 65.20 65.30 65.35 65.40 65.90
 [85] 65.90 66.10 66.35 66.45 66.50 66.50 66.65 66.70 66.70 66.80 66.85 66.85
 [97] 66.85 66.90 67.00 67.05 67.05 67.10 67.10 67.15 67.20 67.25 67.25 67.30
[109] 67.30 67.40 67.65 67.80 67.80 67.80 67.85 67.95 68.60 68.75 69.00 69.15
[121] 69.20 69.30 69.40 69.40 69.55 69.60 69.70 69.70 69.75 69.80 69.85 69.90
[133] 69.90 70.00 70.85 70.90 71.25 71.30 71.40 71.50 71.75 71.80 72.00 72.05
[145] 72.35 72.55 72.70 72.75 72.95 73.40 73.50 73.80 73.90 73.95 74.00 74.10
[157] 74.30 74.45 74.50 74.60 74.85 75.25 75.40 75.45 75.55 78.20 78.70 79.20
[169] 79.65 80.15 81.10

> ##############  TWO-SIDED CI ON  MU  FOR CARAPACE DATA  ##################
> ##############  ALPHA = .05  ##############################################
> c05=131;n1=di*(di+1)/2-c05+1;c(sxPx[n1],sxPx[c05])
[1] 61.00 69.85
> ##############  CANNED CI FOR  MU  IN  SIGNED-RANK TEST  #################
> wilcox.test(Length,conf.int=T)

Wilcoxon signed rank test

data:  Length
V = 171, p-value = 7.629e-06
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 61.00 69.85
sample estimates:
(pseudo)median
          66.1
```

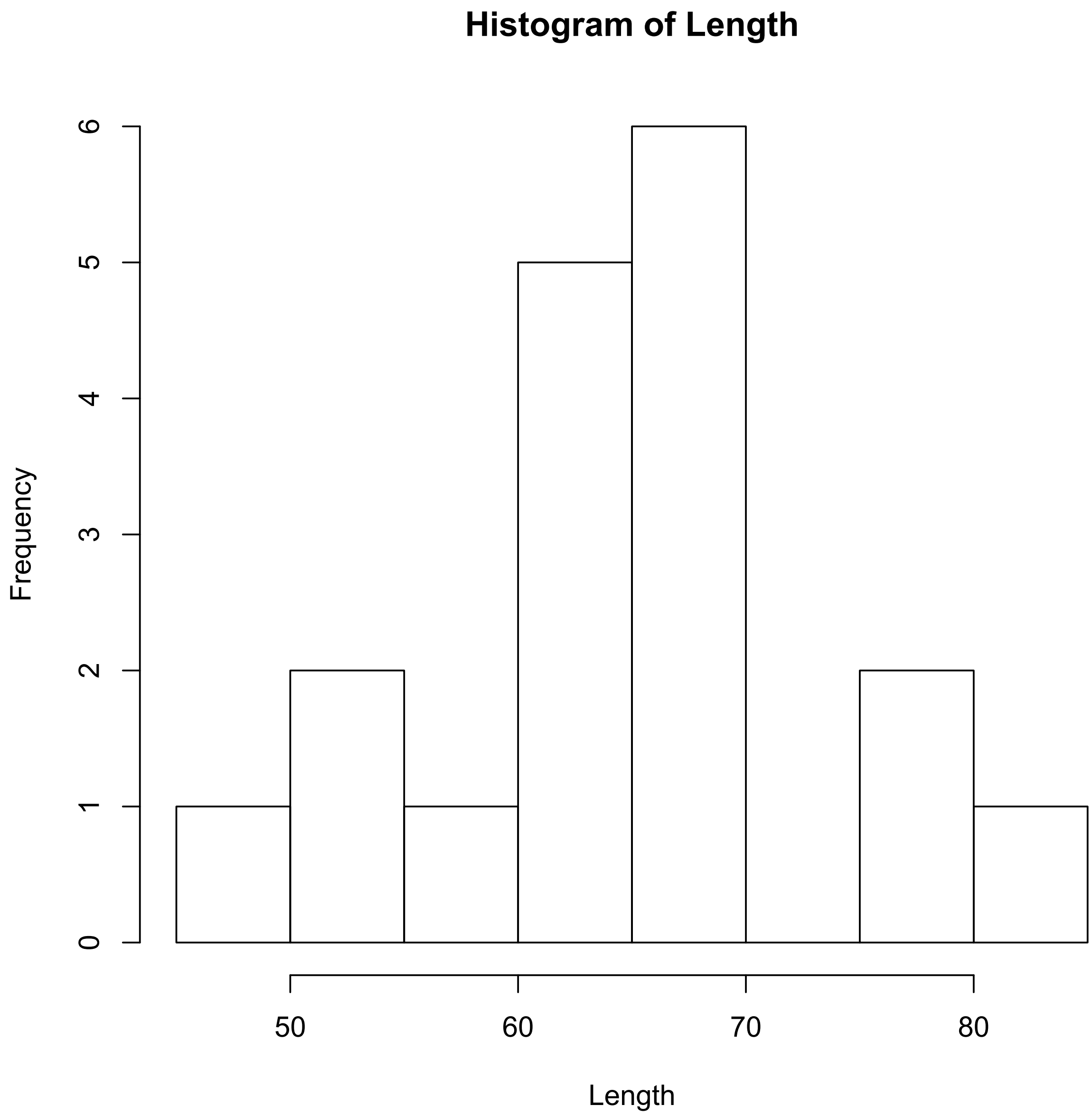
```
> ##############  ALPHA = .01  #############################################
> c01=143;n1=di*(di+1)/2-c01+1;c(sxPx[n1],sxPx[c01])
[1] 59.05 72.00

> wilcox.test(Length,conf.int=T,conf.level=.99)

Wilcoxon signed rank test

data:  Length
V = 171, p-value = 7.629e-06
alternative hypothesis: true location is not equal to 0
99 percent confidence interval:
 58.70 72.05
sample estimates:
(pseudo)median
          66.1

> ###  CANNED CI DOES NOT USE  C=143  OF TABLE A15 FOR  99.0 %  TEST  #####
> c01=144;n1=di*(di+1)/2-c01+1;c(sxPx[n1],sxPx[c01])
[1] 58.70 72.05
```


Histogram of Length
Frequency
0
1
2
3
4
5
6
50
60
70
80
Length