Math 3080 § 1. **Hydroxylation Example:** Name: Example
Treibergs **One Sample Wilcoxon Signed-Rank Test** April 12, 2014

Ths program is to do a Wilcoxon Signed-Rank test for location on a single variable. The data set comes from Navidi, *Statistics for Engineers and Scientists,* 2nd ed., MacGraw Hill, New York, problem 6.9.3 which quotes the article by S. Nandi, *et. al,* "Reaction Modeling and Optimization Using Neural Networks and Genetic Algorithms: Case Study Involving TS-1-Catalyzed Hydroxylation of Benzene" in *Industrial and Engineering Chemistry Research,* 2002. The study presents benzene conversions (in mole percent) for 24 benzenehydroxylation reactions. The question is whether the mean conversion rate differs from $\mu_0 = 55$ mole percent.

The Wilcoxon Signed-Rank test assumes that the data comes from a continuous distribution that is left-right symmetric about μ which is both the mean and median of the population. The null and alternative hypotheses in this problem are

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

The absolute deviations from μ_0 are sorted from lowest to highest and $|X_{(i)} - \mu_0|$ and assigned ranks from 1 to n, the number of observations. The statistic V is the sum of ranks corresponding to the positive $X_i - \mu_0$. By the symmetry assumption, the sign of $X_i - \mu_0$ is positive or negative with equal probability under the null hypothesis, so that sum will be near its expected value $\mu_V = E(V) = \dfrac{n(n+1)}{4}$. If this number is high or low compared to μ_V we reject the null hypothesis. In this problem, $V = 71$ compared to the expected 150. The p-value, computed by **R**© from the exact distribution of V (assuming no ties) ended up being .02293, thus we reject the null hypothesis: thus there is significant evidence that μ is not μ_0.

When n is large, ($n > 20$) then the statistic is distributed approximately normally and we may use normal distribution to compute the p-value. Since $\sigma_V^2 = \dfrac{n(n+1)(2n+1)}{24}$, the standardized variable is approximately normal

$$Z = \frac{V - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Hence the p-value is $2\Phi(-|Z|)$ if Z is the observed value. **R**© will compute this number if the exact calculation is turned off and the continuity correction is not used. If the continuity correction is included (in case of no ties as in this data) then the corrected formula (assuming $V < \mu_V$) is used

$$Z = \frac{V + 0.5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

In case of ties, the exact calculation cannot be performed, and **R**© returns the normal approximation to the p-value.

This data turned out not to have any ties. We redid an analysis on the same data x_i except that, arbitrarily, x_{10} and x_{11} were replaced by x_{12} and x_{15} was replaced by x_{16}. Then the $|X_i - \mu_0|$ have ties and so the tied observations are all given their average rank. In this case, the variance in the approximation must be corrected. The correction in the normalization is

$$Z = \frac{V + .5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{T}{48}}} \tag{1}$$

where

$$T = \sum (\tau_i - 1)\tau_i(\tau_i + 1)$$

where τ_i is the frequency of the ith value of $|X_i - \mu_0|$ or of the rank. Thus if the ith datum is not tied, $\tau_i = 1$ and contributes nothing to T. If X_i is tied then τ_i is the number of occurences of this value of $|X_i - \mu_0|$. In the example above, there are two sets of ties: $|x_{10} - \mu_0| = |x_{11} - \mu_0| = |x_{12} - \mu_0|$ and $|x_{15} - \mu_0| = |x_{16} - \mu_0|$ and no others. Thus one $\tau_i = 3$ and another $\tau_j = 2$. The nonzero terms of T give $T = (3-1)3(3+1) + (2-1)2(2+1) = 30$. Note that τ_i is added once for any tied value, not three times for the three i's that are tied. The points removed changed the data enough so that the text does not reject $\mathcal{H}_0$.

Note that when there are ties, **R**© is not able to compute the exact p-value. It uses both the continuity correction and the variance correction to compute an approzimate p-value using formula (1).

Data Set Used in this Analysis :

```
# Math 3080-1                Hydroxylation Data             April 12, 2014
# Treibergs
#
# From Navidi, "Statistics for Engineers and Scientists," 2nd ed., MacGraw
# Hill, New York, problem 6.9.3 who quotes the article by S. Nandi, et. al,
# "Reaction Modeling and Optimization Using Neural Networks and Genetic
# Algorithms: Case Study Involving TS-1-Catalyzed Hydroxylation of
# Benzene" in Industrial and Engineering Chemistry Research, 2002.
# The study presents benzene conversions (in mole percent) for 24
# benzenehydroxylation reactions.
Conversion
52.3
41.1
28.8
67.8
78.6
72.3
9.1
19
30.3
41
63
80.8
26.8
37.3
38.1
33.6
14.3
30.1
33.4
36.2
34.6
40
81.2
59.4
```

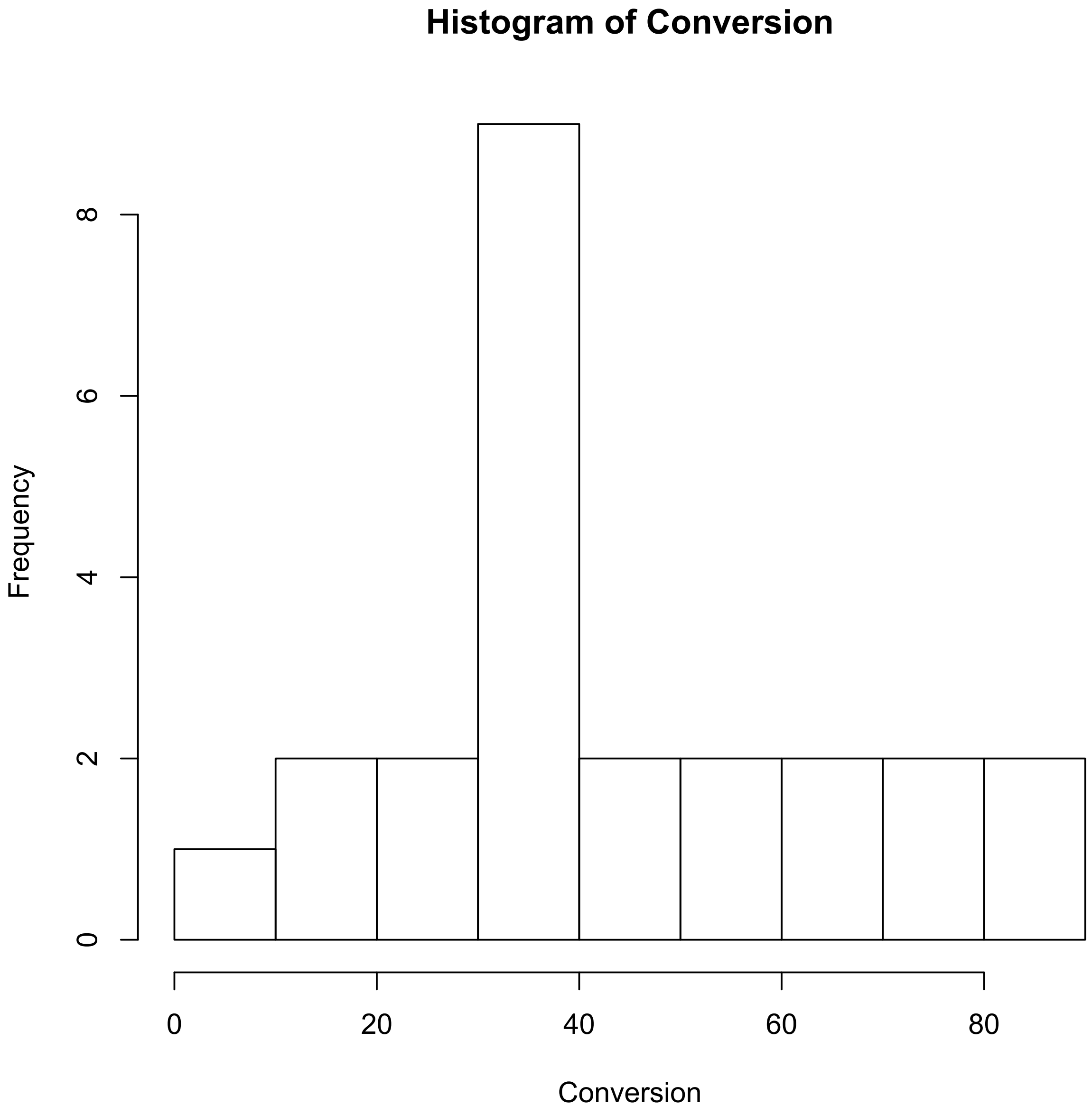
R Session:

```
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("M3082DataHydroxylation.txt",header=T)
> tt
   Conversion
1        52.3
2        41.1
3        28.8
4        67.8
5        78.6
6        72.3
7         9.1
8        19.0
9        30.3
10       41.0
11       63.0
12       80.8
13       26.8
14       37.3
15       38.1
16       33.6
17       14.3
18       30.1
19       33.4
20       36.2
21       34.6
22       40.0
23       81.2
24       59.4
>
```

```
> attach(tt)
> ################## HISTOGRAM TO CHECK ASSUMPTIONS ##############
> hist(Conversion)
> ################## HISTOGRAM LOOKS SYMMETRIC ###################
>
> ########  RUN CANNED WILCOXON SIGN-RANK TEST  ##################

> wilcox.test(Conversion,mu=55,alternative="two.sided")

Wilcoxon signed rank test

data:  Conversion
V = 71, p-value = 0.02293
alternative hypothesis: true location is not equal to 55

> #############  DO SIGNED-RANK TEST BY HAND  ####################
> mu=55
> n=length(Conversion); n
[1] 24
> #############  RANK FUNCTION ASSIGNS RANKS "IN PLACE"  #########
> ra=rank(abs(Conversion-mu));ra
 [1]  1  5 19  4 15  9 24 22 16  6  3 18 21 10  8 13 23 17 14 11 12  7 20  2

> #############  ADD UP RANKS OF POSITIVE X-mu0’S  ###############
> SP=sum(ra[Conversion>mu]);SP
[1] 71

> #############  UNCORRECTED NORMALIZED STATISTIC  ###############
> z=(SP-n*(n+1)/4)/sqrt(n*(n+1)*(2*n+1)/24); z
[1] -2.257143
> ############## P-VALUE  ########################################
> 2*pnorm(z)
[1] 0.02399915

> ### SAME AS CANNED WITH NO EXACT P-VALUE, NO CONT. CORR. #######

> wilcox.test(Conversion,mu=55,exact = FALSE, correct = FALSE)

Wilcoxon signed rank test

data:  Conversion
V = 71, p-value = 0.024
alternative hypothesis: true location is not equal to 55
```

```
> ########## WITH CONTINUITY CORRECTION  #########################
> zc=(SP+.5-n*(n+1)/4)/sqrt(n*(n+1)*(2*n+1)/24); zc
[1] -2.242857
> 2*pnorm(zc)
[1] 0.02490603


> wilcox.test(Conversion,mu=55,exact = FALSE, correct = TRUE)

Wilcoxon signed rank test with continuity correction

data:  Conversion
V = 71, p-value = 0.02491
alternative hypothesis: true location is not equal to 55

> ##########  DATA WITH TIES. CONSTRUCT TIED DATA SET.  ##########
> x=Conversion; x[10]=x[11]=x[12]; x[15]=x[16]
> x
 [1] 52.3 41.1 28.8 67.8 78.6 72.3  9.1 19.0 30.3 80.8 80.8 80.8 26.8 37.3 33.6
[16] 33.6 14.3 30.1 33.4 36.2 34.6 40.0 81.2 59.4
> ra2 = rank(abs(x-mu))
> ra2
 [1]  1.0  4.0 19.0  3.0 13.0  6.0 24.0 22.0 14.0 17.0 17.0 17.0 21.0  7.0 10.5
[16] 10.5 23.0 15.0 12.0  8.0  9.0  5.0 20.0  2.0
> sort(ra2)
 [1]  1.0  2.0  3.0  4.0  5.0  6.0  7.0  8.0  9.0 10.5 10.5 12.0 13.0 14.0 15.0
[16] 17.0 17.0 17.0 19.0 20.0 21.0 22.0 23.0 24.0

> ##############  TABULATE FREQUENCIES OF  |X - mu0|  ################
> t2=table(ra2); t2
ra2
   1    2    3    4    5    6    7    8    9 10.5   12   13   14   15   17
   1    1    1    1    1    1    1    1    1    2    1    1    1    1    3
  19   20   21   22   23   24
   1    1    1    1    1    1
> ### NOTE THAT |X - mu0| = 80.8  AND  |X - mu0| = 33.6  ARE TIED  ###
> ####  CORRECTION FOR VARIANCE  ####################################
> fixtie=function(t){(t-1)*t*(t+1)}
> T=sum(fixtie(t2)); T
[1] 30

> ############  CANNED TEST ON  X  ###################################
> wilcox.test(x,mu=55,exact = FALSE, correct = FALSE)

Wilcoxon signed rank test

data:  x
V = 95, p-value = 0.116
alternative hypothesis: true location is not equal to 55
```

```
> ###########  TEST X  BY HAND  #####################################

> V2 = sum(ra2[x>mu]); V2
[1] 95
> z2=(V2-n*(n+1)/4)/sqrt(n*(n+1)*(2*n+1)/24); z2
[1] -1.571429
> 2*pnorm(z2)
[1] 0.1160831


> ###########  CANNED TEST WITH CONTINUITY CORRECTION  ##############

> wilcox.test(x,mu=55,exact = FALSE, correct = TRUE)

Wilcoxon signed rank test with continuity correction

data:  x
V = 95, p-value = 0.1193
alternative hypothesis: true location is not equal to 55


> #### TEST BY HAND WITH CONTINUITY AND VARIANCE CORRECTION  #######

> z3 = (V2+.5-n*(n+1)/4)/sqrt(n*(n+1)*(2*n+1)/24-T/48); z3
[1] -1.55754
> pvalue3 = 2*pnorm(z3); pvalue3
[1] 0.1193423


> #####  IN TEST DATA WITH TIES, EXACT P-VALUE FAILS  ##############

> wilcox.test(x,mu=55,exact = TRUE)

Wilcoxon signed rank test with continuity correction

data:  x
V = 95, p-value = 0.1193
alternative hypothesis: true location is not equal to 55

Warning message:
In wilcox.test.default(x, mu = 55, exact = TRUE) :
  cannot compute exact p-value with ties
>
```


Histogram of Conversion
Frequency
0
2
4
6
8
0
20
40
60
80
Conversion