Math 3080 § 1.	**Baseball Game Time Example:**	Name: Example
Treibergs	**Wilcoxon Rank-Sum Test**	April 13, 2014

This program does a Wilcoxon Rank-Sum Test for comparing locations of two samples. The data comes from Larsen & Marx, *An Introduction to Mathematical Statistics and its Applications,* 4th ed., Pearson, Upper Saddle River, 2006. Rule differences among other factors may affect the length of major league baseball games differently in the American and National Leagues. Is there a difference in mean game times? To study this question, the average home game completion times (in minutes) are recorded for all teams for the 1992 season.

The Wilcoxon Rank-Sum Test, also called the Mann-Whitney Test, assumes that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent random samples that come from continuous distribution that have the same shape and spread, but may have possibly different means μ_X and μ_Y, respectively. The null and alternative hypotheses in this problem are

$$\mathcal{H}_0 : \mu_X - \mu_Y = 0;$$
$$\mathcal{H}_a : \mu_X - \mu_Y \neq 0.$$

We may arrange that $m \leq n$ to agree with the text. The samples are lumped together, sorted from lowest to highest and assigned ranks from 1 to $m + n$, the number of observations. The statistic W is the sum of ranks corresponding to the X observations. If both distributions are the same, X and Y values will intermingle and W will be near than the expected sum $\mu_W = E(W) = \dfrac{m(m+n+1)}{2}$ of m randomly chosen ranks from 1 to $m+n$. If W is high or low compared to μ_W we reject the null hypothesis.

In this problem, X_i are National League times with $m = 12$ teams and Y_j are American League Times with $n = 14$ teams. The side-by-side histograms show that both samples plausibly come from distributions with the same shape and spread. The statistic works out to be $W = 110.5$ compared to the expected $\mu_W = 162$. Note that **R**© uses another equivalent statistic. The continuity-corrected p-value estimated by **R**© ended up being 0.008405, thus we reject the null hypothesis: there is significant evidence that $\mu_X \neq \mu_Y$. Looking at the average times, the senior circuit game times are on average over ten minutes shorter than those of the junior circuit.

When m and n is large, $(m, n > 8)$ then the statistic is distributed approximately normally and we may use normal distribution to compute the p-value. Since $\sigma_W^2 = \dfrac{mn(m+n+1)}{12}$, the standardized variable is approximately normal

$$Z = \frac{W - m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}$$

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. In case there are no ties, W takes integer values and the continuity correction may be applied to estimate the p-value (assuming $W < \mu_W$) using

$$Z_c = \frac{W + .5 - m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}$$

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

This data has many ties. For any set of tied observations, the average rank is assigned to each. In this case, the statistic may take fractional values. The variance in the approximation must be corrected when there are ties. The correction in the normalization is

$$Z_{ct} = \frac{W + .5 - \dfrac{m(m+n+1)}{2}}{\sqrt{\dfrac{mn(m+n+1)}{12} - \dfrac{mnT}{12(m+n)(m+n+1)}}} \tag{1}$$

where

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

where τ_i is the frequency of the ith value of the lumped X_i's and Y_j's or of their ranks. Thus if the ith value is not tied, $\tau_i = 1$ and contributes nothing to T. If X_i or Y_j is tied then τ_i is the number of occurences of this value. In the example above, ranks 3.5, 7.5, 9.5, 12.5, 14.5, 16.5, 24.5 are each tied with another so have $\tau_i = 2$ and 21 is a five-fold tie with $\tau_i = 5$. The nonzero terms of T give $T = 7(2-1)2(2+1) + (5-1)5(5+1) = 162$. Note that τ_i is added once for any tied value, not three times for the three i's that are tied.

Note that when there are ties, **R**© is not able to compute the exact p-value. It uses a slightly different approximation than this variance correction with continuity correction given by formula (1).

Data Set Used in this Analysis :

```
# Math 3080              Baseball Game Time Data            April 13, 2014
# Treibergs
#
# From Larsen & Marx, "An Introduction to Mathematical Statistics and its
# Applications," 4th ed.,  Pearson, Upper Saddle River, 2006.
#
# Do the rule differences in the American and National League affect the
# length of major league baseball games? To study this question, the average
# home game completion times (in minutes) are recorded for all teams for the
# 1992 season.
Team         League   Time
Baltimore    American  177
Boston       American  177
California   American  165
Chicago(AL)  American  172
Cleveland    American  172
Detroit      American  179
KansasCity   American  163
Milwaukee    American  175
Minnesota    American  166
NewYork(AL)  American  182
Oakland      American  177
Seattle      American  168
Texas        American  179
Toronto      American  177
Atlanta      National  166
Chicago(NL)  National  154
Cincinnati   National  159
Houston      National  168
LosAngeles   National  174
Montreal     National  174
NewYork(NL)  National  177
Philadelphia National  167
Pittsburg    National  165
SanDiego     National  161
SanFrancisco National  164
SaintLouis   National  161
```

R Session:

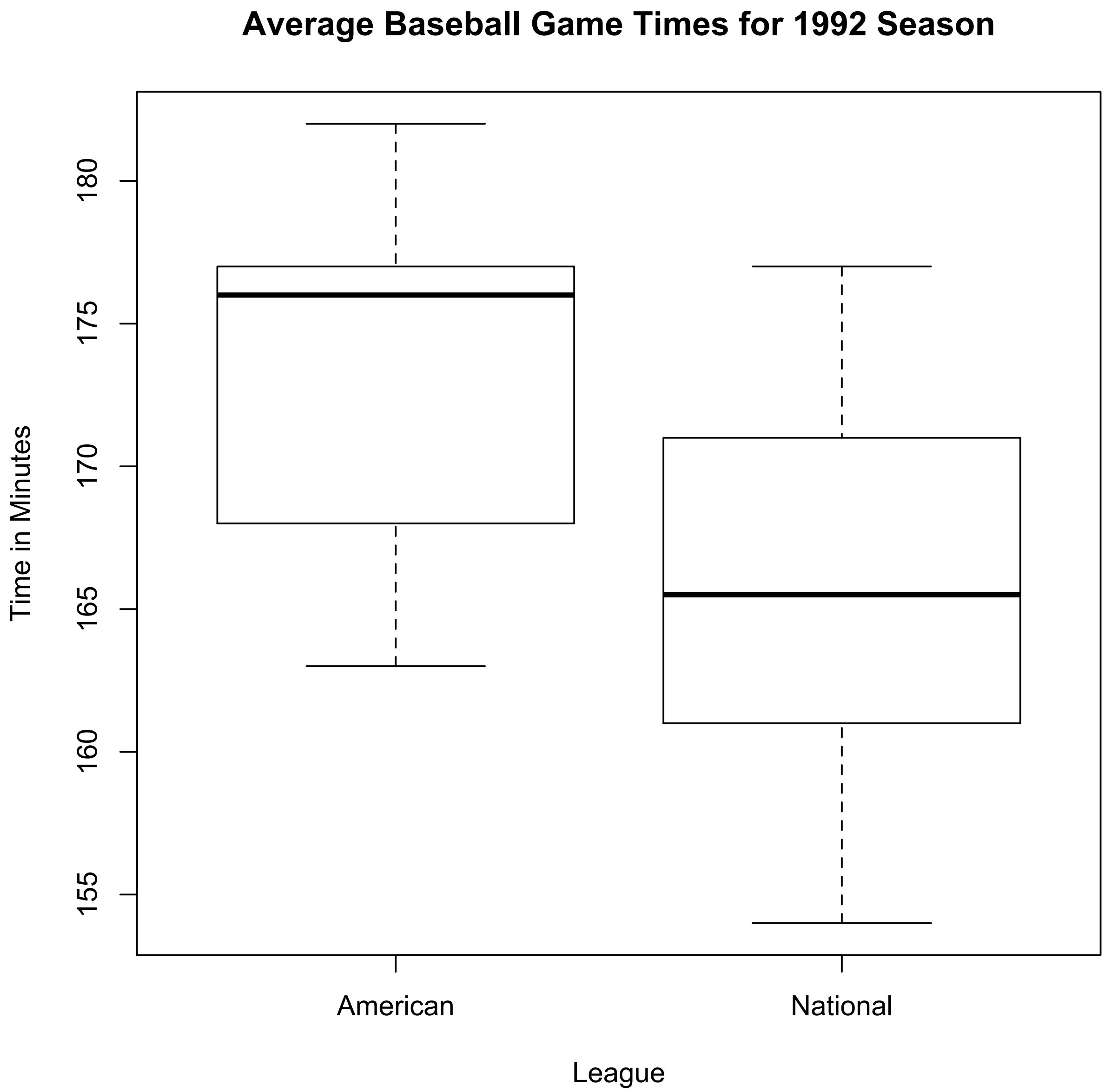
```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

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.

> ############  READ THE DATA  ###############################
> tt=read.table("M3082DataBaseballTime.txt",header=T)
> attach(tt)
> tt
           Team   League Time
1     Baltimore American  177
2        Boston American  177
3    California American  165
4   Chicago(AL) American  172
5     Cleveland American  172
6       Detroit American  179
7    KansasCity American  163
8     Milwaukee American  175
9     Minnesota American  166
10  NewYork(AL) American  182
11      Oakland American  177
12      Seattle American  168
13        Texas American  179
14      Toronto American  177
15      Atlanta National  166
16  Chicago(NL) National  154
17   Cincinnati National  159
18      Houston National  168
19   LosAngeles National  174
20     Montreal National  174
21  NewYork(NL) National  177
22 Philadelphia National  167
23    Pittsburg National  165
24     SanDiego National  161
25 SanFrancisco National  164
26   SaintLouis National  161
> League=factor(League)
```

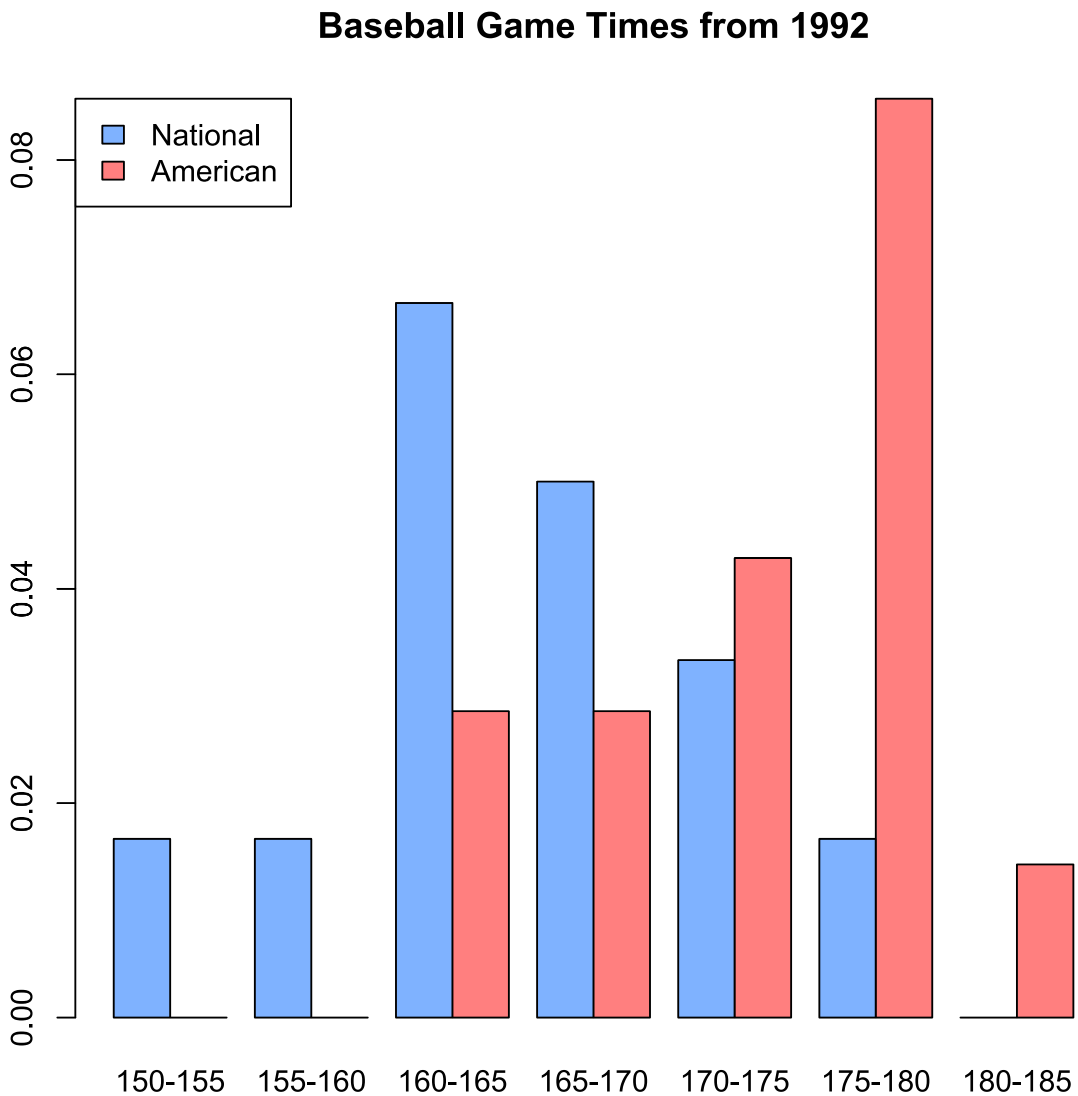
```
> ############  SUMMARY BY LEAGUE  ###########################
> tapply(Time,League,summary)
$American
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  163.0   169.0   176.0   173.5   177.0   182.0 

$National
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  154.0   161.0   165.5   165.8   169.5   177.0 

> ############  PICK OFF NL AND AL TIMES  ##################
> X=Time[League=="National"]; m=length(X); m
[1] 12
> Y=Time[League=="American"]; n=length(Y); n
[1] 14
> X
 [1] 166 154 159 168 174 174 177 167 165 161 164 161
> Y
 [1] 177 177 165 172 172 179 163 175 166 182 177 168 179 177

> ############  PLOT SIDE-BY-SIDE HISTOGRAMS OF TIMES  #####
> plot(Time~League, main="Average Baseball Game Times for 1992 Season"
  ,ylab="Time in Minutes")
> hx=hist(X,breaks=seq(150,185,5),freq=F)
> hy=hist(Y,breaks=seq(150,185,5),freq=F)
> mx=t(cbind(hx$density,hy$density))
> colnames(mx)=c("150-155","155-160","160-165","165-170",
                 "170-175","175-180","180-185")
> colo=c(rainbow(10,alpha=.5)[7],rainbow(10,alpha=.5)[1])
> b=barplot(mx,beside=T,col=colo,main="Baseball Game Times from 1992",
    legend.text=c("National","American"),args.legend=list(x="topleft"),
    space=c(0,.5))

> ############  RUN CANNED RANK-SUM TEST  ##################
> wilcox.test(X,Y)

Wilcoxon rank sum test with continuity correction
data:  X and Y
W = 32.5, p-value = 0.008405
alternative hypothesis: true location shift is not equal to 0
Warning message:
In wilcox.test.default(X, Y) : cannot compute exact p-value with ties

> wilcox.test(Y,X)

Wilcoxon rank sum test with continuity correction
data:  Y and X
W = 135.5, p-value = 0.008405
alternative hypothesis: true location shift is not equal to 0
Warning message:
In wilcox.test.default(Y, X) : cannot compute exact p-value with ties
```

```
> #############  COMPUTE RANK-SUM TEST "BY HAND"  ############
> rank(Time)
 [1] 21.0 21.0  7.5 14.5 14.5 24.5  5.0 18.0  9.5 26.0 21.0 12.5 24.5 21.0  9.5
[16]  1.0  2.0 12.5 16.5 16.5 21.0 11.0  7.5  3.5  6.0  3.5
> rt=rank(Time); rt
 [1] 21.0 21.0  7.5 14.5 14.5 24.5  5.0 18.0  9.5 26.0 21.0 12.5 24.5 21.0  9.5
[16]  1.0  2.0 12.5 16.5 16.5 21.0 11.0  7.5  3.5  6.0  3.5
> League
 [1] American American American American American American American American
 [9] American American American American American American National National
[17] National National National National National National National National
[25] National National
Levels: American National

 > ##############  TOTAL OF RANKS  ##############################
> sum(rt)
[1] 351
> (m+n)*(m+n+1)/2
[1] 351

> ##############  COMPTE  W  ##################################
> W = sum(rt[League=="National"]); W
[1] 110.5

> ###############  EXPECTED  muW  AND  sig2W  #################
> m;n
[1] 12
[1] 14
> muW=m*(m+n+1)/2;muW
[1] 162
> sig2W=m*n*(m+n+1)/12;sig2W
[1] 378

> ############  UNCORRECTED  z, CRITICAL VALUE, P-VALUE  ######
> W
[1] 110.5
> muW
[1] 162
> z=(W-muW)/sqrt(sig2W);z
[1] -2.648874
> alpha=.05
> z2tailcrit = qnorm(alpha/2,lower.tail=F); z2tailcrit
[1] 1.959964
> pvalue = 2*pnorm(z); pvalue
[1] 0.008076039
```

```
> ##############  CANNED UNCORRECTED WILCOXON TEST  ###########
> wilcox.test(X,Y,paired=F,alternative="two.sided",exact=F,correct=F)

Wilcoxon rank sum test

data:  X and Y
W = 32.5, p-value = 0.007787
alternative hypothesis: true location shift is not equal to 0
```

```
> ##############  VARIANCE CORRECTION FOR TIES  #################
> ###############  COUNT NUMBER OF TIES  #######################
> xt=table(rt); xt
rt
   1    2  3.5    5    6  7.5  9.5   11 12.5 14.5 16.5   18   21 24.5   26
   1    1    2    1    1    2    2    1    2    2    2    1    5    2    1

> ############  P-VALUES WITH VARIANCE CORRECTION FOR TIES  #####

> fixtie=function(t){(t-1)*t*(t+1)}
> T=sum(fixtie(xt)); T
[1] 162
> Tp=n*m*T/(12*(m+n)*(m+n+1));Tp
[1] 3.230769

> z=(W+.5-muW)/sqrt(sig2W-Tp);z
[1] -2.634439
> pvalue = 2*pnorm(z); pvalue
[1] 0.008427635

> z=(W-muW)/sqrt(sig2W-Tp);z
[1] -2.660267
> pvalue = 2*pnorm(z); pvalue
[1] 0.007807867
> ######  NOTE THAT THESE DISAGREE WITH CANNED RESULTS  #########
>
```


Average Baseball Game Times for 1992 Season
Time in Minutes
155
160
165
170
175
180
American
National
League



Baseball Game Times from 1992
National
American
0.00
0.02
0.04
0.06
0.08
150-155
155-160
160-165
165-170
170-175
175-180
180-185