Implementing the Wilcoxon Test in R

My definitions (matching McClave & Sincich)

Two samples: sizes $n_1 > n_2$ (sample 2 is **smaller** group).

- **Wilcoxon statistic :** sum of ranks for group 2 (with size n_2.

$$W = R_2$$

- **Mann-Whitney statistic :** sum of ranks for group 2 with **minimum possible rank sum subtracted**

$$U = R_2 - \frac{n_2(n_2+1)}{2}$$

For example, if $n_1 = 4, n_2 = 3$, minimum possible sum is

$$1 + 2 + 3 = 6 = \frac{3 \times 4}{2}$$

i.e. this happens when the three lowest ranks belong to group 2.

R definitions

Two samples: sizes n_1 and n_2 (group 1 is the **first** listed group).

- **Mann-Whitney-Wilcoxon statistic :** sum of ranks for sample 1 with minimum possible rank sum subtracted

$$W = R_1 - \frac{n_1(n_1+1)}{2}$$

Example

Red Group	:	0.67	1.18	2.27	2.65	0.82	1.35	0.99	2.65
Blue Group	:	0.84	0.21	1.39	-3.06	0.54	1.64		

According to McClave & Sincich, the Blue Group is Group 2 (smaller sample size), and the rank sum is

$$R_2 = 6 + 2 + 10 + 1 + 3 + 11 = 33$$

However, if the data are supplied to R with Red first, then Blue, then the rank statistic is computed as

$$W = (4 + 8 + 12 + 14 + 5 + 9 + 7 + 13) - \frac{8 \times 9}{2} = 36$$

Had Blue been supplied first, the test statistic would be $W = 33 - (6 \times 7/2) = 12$.

R code: Simulation

```
> n1<-6
> n2<-8
> Y1<-rnorm(n1)
> Y2<-rnorm(n2)+1
> Y<-c(Y1,Y2)
> G<-rep(c(1,2),c(n1,n2))

> (Blue<-Y1)
[1]  1.6382091 -1.1355267  1.2103083  0.4298248
0.2353172  -0.8588152
> (Red<-Y2)
[1]  2.0429021 -1.2744787  1.2135546  2.0253638
3.6128987 0.9564887  0.3988407 2.5625341
```

R code: Test with two-sided alternative

```
> wilcox.test(Blue,Red)

        Wilcoxon rank sum test

data:  Blue and Red
W = 12, p-value = 0.1419
alternative hypothesis: true location shift is not

> wilcox.test(Red,Blue)

        Wilcoxon rank sum test

data:  Red and Blue
W = 36, p-value = 0.1419
alternative hypothesis: true location shift is not
```

R code: Test with one-sided alternative

```
> wilcox.test(Blue,Red,alternative="greater")

        Wilcoxon rank sum test

data:  Blue and Red
W = 12, p-value = 0.9461
alternative hypothesis: true location shift is
greater than 0

> wilcox.test(Red,Blue,alternative="greater")

        Wilcoxon rank sum test

data:  Red and Blue
W = 36, p-value = 0.07093
alternative hypothesis: true location shift is
greater than 0
```

R code: Test with two-sided alternative

```
> G
 [1] 1 1 1 1 1 1 2 2 2 2 2 2 2 2
> wilcox.test(c(Blue,Red) ~ G)

        Wilcoxon rank sum test

data:  c(Blue, Red) by G
W = 12, p-value = 0.1419
alternative hypothesis: true location shift is
not equal to 0
```

R code: Test using a factor

Note: when the factor `Colour` is used, the statistic is computed for the **first** level (in this case, Blue is chosen on the basis of alphabetical ordering)

```
> Colour<-as.factor(rep(c("Red","Blue"),c(n1,n2))
> Colour
 [1] Red  Red  Red  Red  Red  Red  Blue Blue
 Blue Blue Blue Blue Blue Blue
Levels: Blue Red

> wilcox.test(c(Red,Blue)~Colour)

        Wilcoxon rank sum test

data:  c(Red, Blue) by Colour
W = 36, p-value = 0.1419
alternative hypothesis: true location shift is
not equal to 0
```