Wilcoxon rank sum test

Anta at to nasjoner ønsker å finne ut hvem som har de beste skiløperne. Nasjon 1 stiller med 10 løpere, og nasjon 2 stiller med 12 løpere.

Vi tar tiden på løpet, og får følgende tall:

- ✓ Løpere fra nasjon 1: 0,1,1,3,5,6,11,32,180,413
- ✓ Løpere fra nasjon 2: 12,13,14,15,17,24,35,52,72,80,105,239

Fordelingene til løpstidene:

Dette er definitivt skjevfordelt, og gjennomsnittene (65.2 og 56.5) gir ikke en god oppsummering av de to nasjonene.

Antallene i hver gruppe er også så små at gjennomsnittene ikke er stabile: Hvis vi hadde hatt én løper til i hvilken som helst av gruppene, kunne gjennomsnittet endret seg drastisk. Her kan vi ikke anta at data er normalfordelte, og n er for liten til at CLT er til noen hjelp. Hva gjør vi:

Vi ser på rangeringene av tallene i stedet for løpstidene. Fordelingen til løpstidene vises som de øverste prikkene, og plasseringene vises som de nederste prikkene.

Altså:

Disse rangeringene kalles «ranks», og Wilcoxon rank sum test bruker rangeringene av observasjonene i stedet for originalobservasjonene i sin testobservator. Summen av rangeringene i den ene gruppa kalles «Wilcoxon rank sum statistic» (Ch 15.1), og denne har en fordeling som R kan beregne og bruke til å gi oss p-verdien for
H_0: Gruppene har lik fordeling av verdier, mot **H_1:** Gruppene har ikke lik fordeling av verdier, men en av gruppene har systematisk høyere verdier enn den andre.

15.14

Hvis du har lastet ned datasettene fra boka (du finner dem her: http://www.macmillanlearning.com/Catalog/studentresources/ips8e#t_922171) som excel-filer i en mappe på din egen maskin, finner du frem til fila ex15-14talk10.xls, åpner den i excel, og lagrer den som CSV-fil.

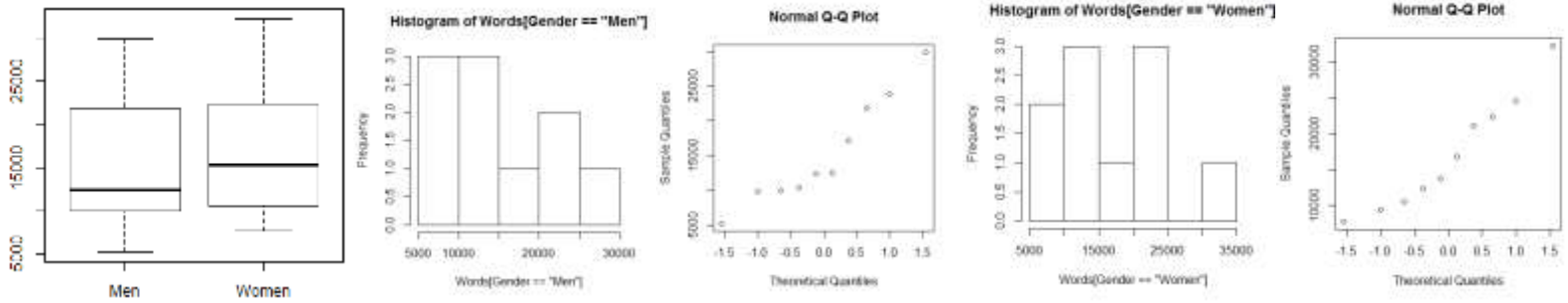
Deretter kan du bruke Import dataset-funksjonen i Rstudio.

Kommandoer:

```
attach(ex15.14talk10)            # Slipper å skrive filnavn$variabelnavn i resten av analysen.
boxplot(Words~Gender)

hist(Words[Gender=="Men"])
qqnorm(Words[Gender=="Men"])
summary(Words[Gender=="Men"])

hist(Words[Gender=="Women"])
qqnorm(Words[Gender=="Women"])
summary(Words[Gender=="Women"])
```




```
> summary(Words[Gender=="Men"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5180   10050   12420   15290   20630   29920
> summary(Words[Gender=="Women"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7694   11020   15280   17080   22050   32290
```

```
t.test(Words~Gender)
wilcox.test(Words~Gender)
```

```
> t.test(Words~Gender)

        Welch Two Sample t-test

data:  Words by Gender
t = -0.51642, df = 17.993, p-value = 0.6118
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9112.358  5516.558
sample estimates:
  mean in group Men mean in group Women
             15287.0              17084.9

> wilcox.test(Words~Gender)

        Wilcoxon rank sum test

data:  Words by Gender
W = 44, p-value = 0.6842
alternative hypothesis: true location shift is not equal to 0
```