Undergrad Biostatistics - R Training - Week X

Ege Ulgen

Chi-squared Test

We'll read data from an online resource and perform Chi-squared test to see whether treatment is associated with improvement or not.

```
treatment_df <- read.csv("https://goo.gl/j6lRXD")

head(treatment_df)
```

```
##   id   treatment  improvement
## 1  1     treated     improved
## 2  2     treated     improved
## 3  3 not-treated     improved
## 4  4     treated     improved
## 5  5     treated not-improved
## 6  6     treated not-improved
```

```
dim(treatment_df)
```

```
## [1] 105   3
```

```
table(treatment_df$treatment, treatment_df$improvement)
```

```
##
##               improved not-improved
##   not-treated       26           29
##   treated           35           15
```

```
chisq.test(table(treatment_df$treatment, treatment_df$improvement))
```

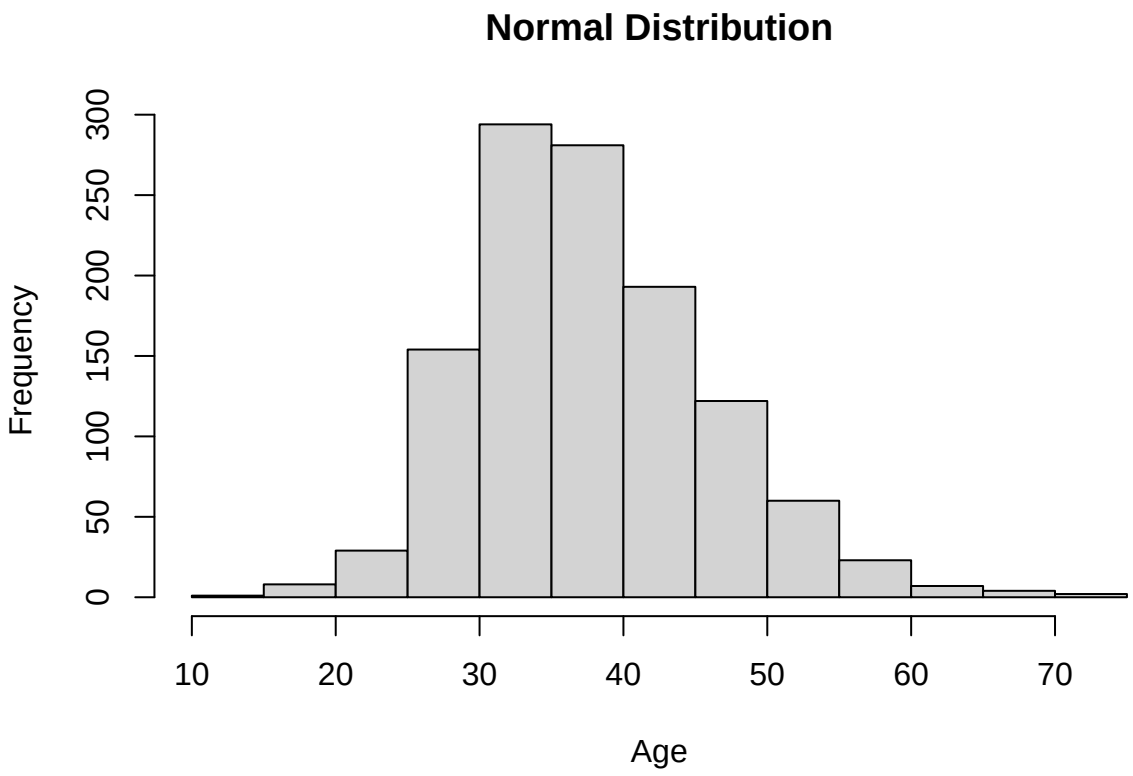
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(treatment_df$treatment, treatment_df$improvement)
## X-squared = 4.66, df = 1, p-value = 0.031
```

```
# we reject the null hypothesis and conclude that the two variables are associated
```

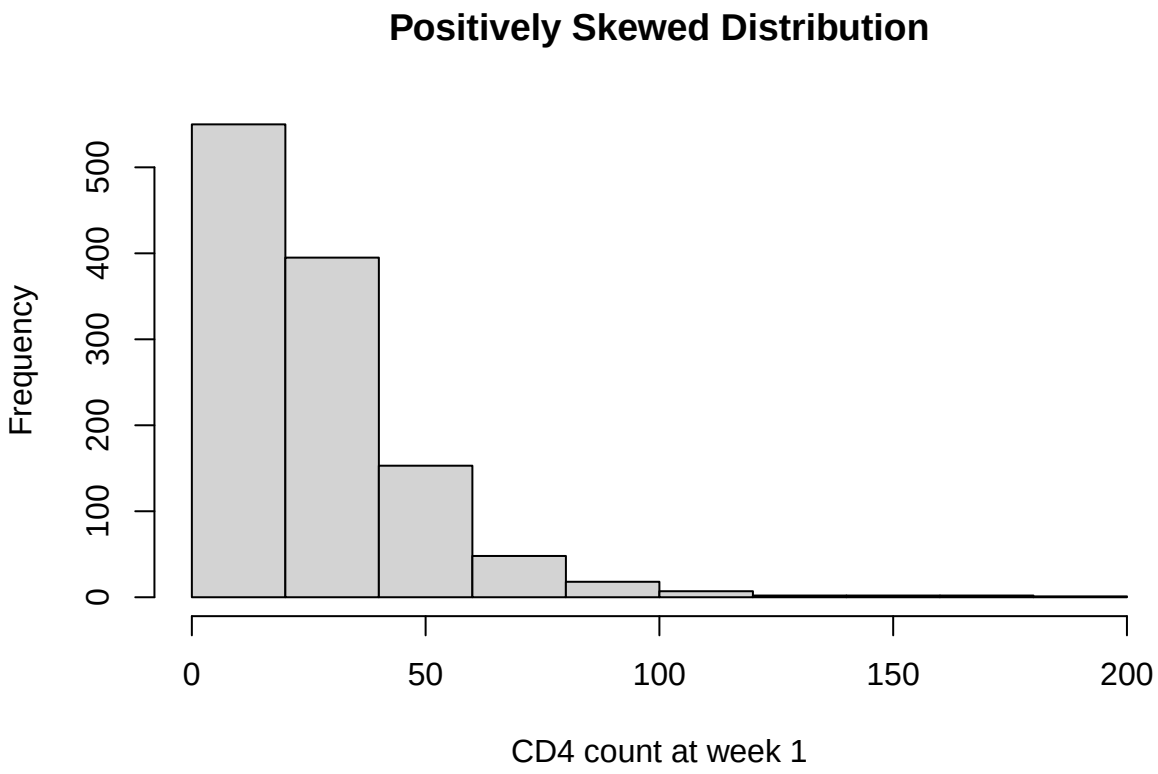
A detailed tutorial can be found on: http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

Assessment of normality

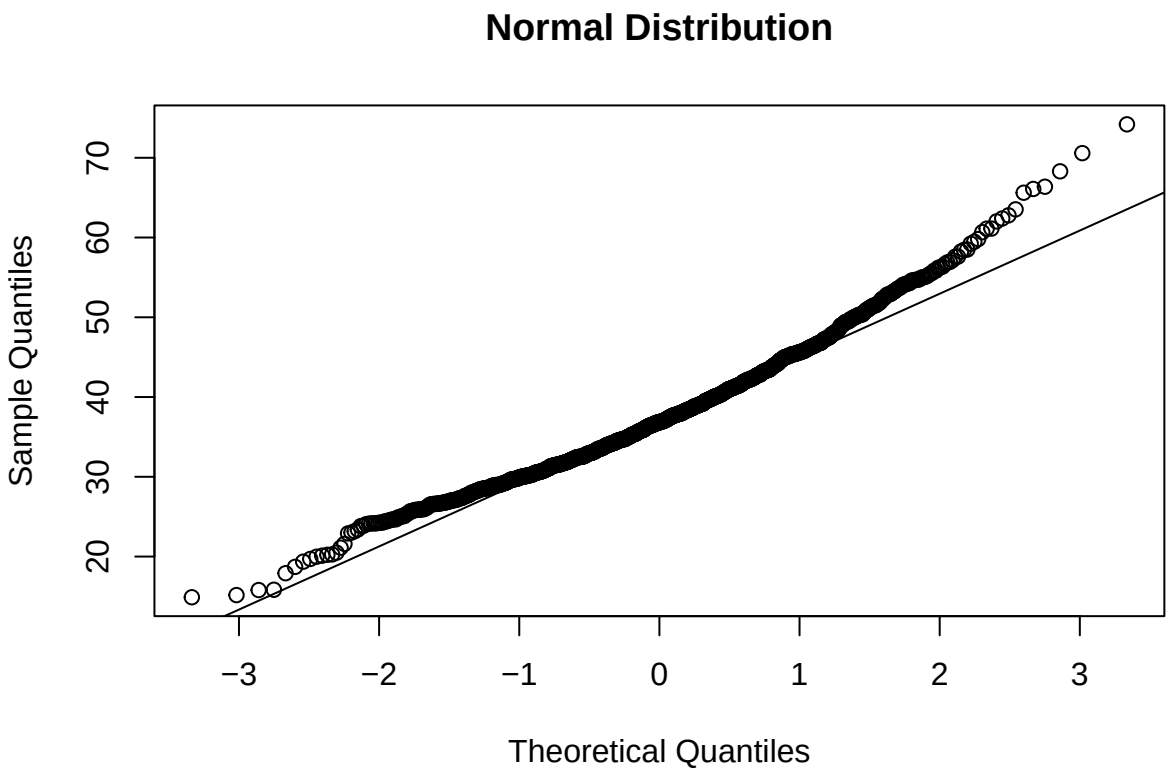
```
aids_df <- read.delim("../data/aids_dataset.txt", sep = " ")

### inspect the distribution of age and CD4 at week 1
hist(aids_df$age, xlab = "Age", main = "Normal Distribution")
```



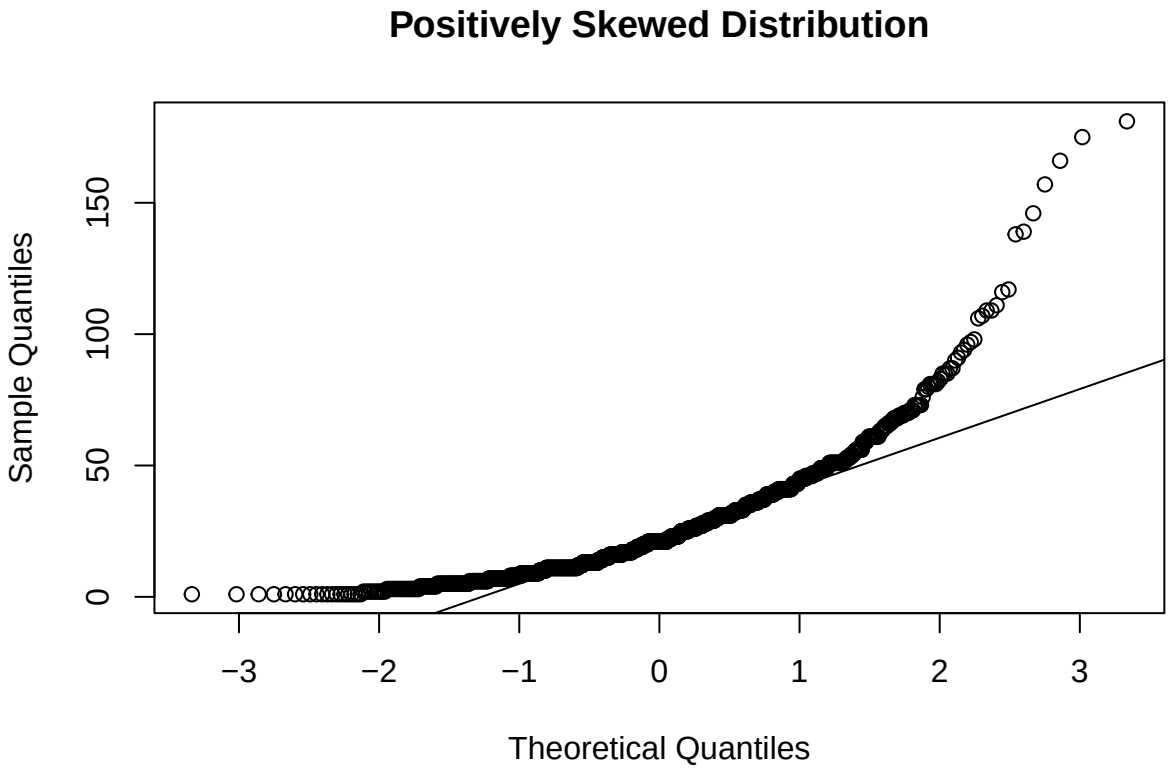

```
hist(aids_df$cd4_1, xlab = "CD4 count at week 1", main = "Positively Skewed Distribution")
```




```
## QQ plot
qqnorm(aids_df$age, main = "Normal Distribution")
qqline(aids_df$age)
```




```
qqnorm(aids_df$cd4_1, main = "Positively Skewed Distribution")
qqline(aids_df$cd4_1)
```



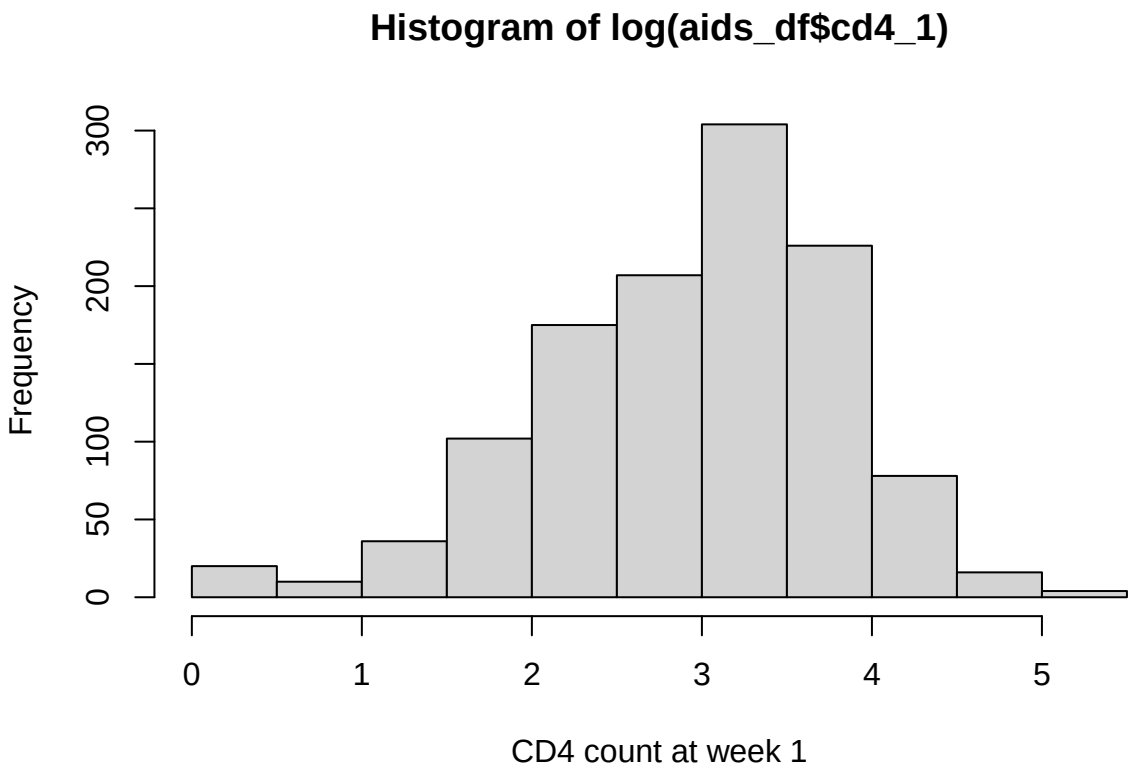

```
## Shapiro test
shapiro.test(aids_df$cd4_1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aids_df$cd4_1
## W = 0.826, p-value <2e-16
```

```
shapiro.test(aids_df$age)
```

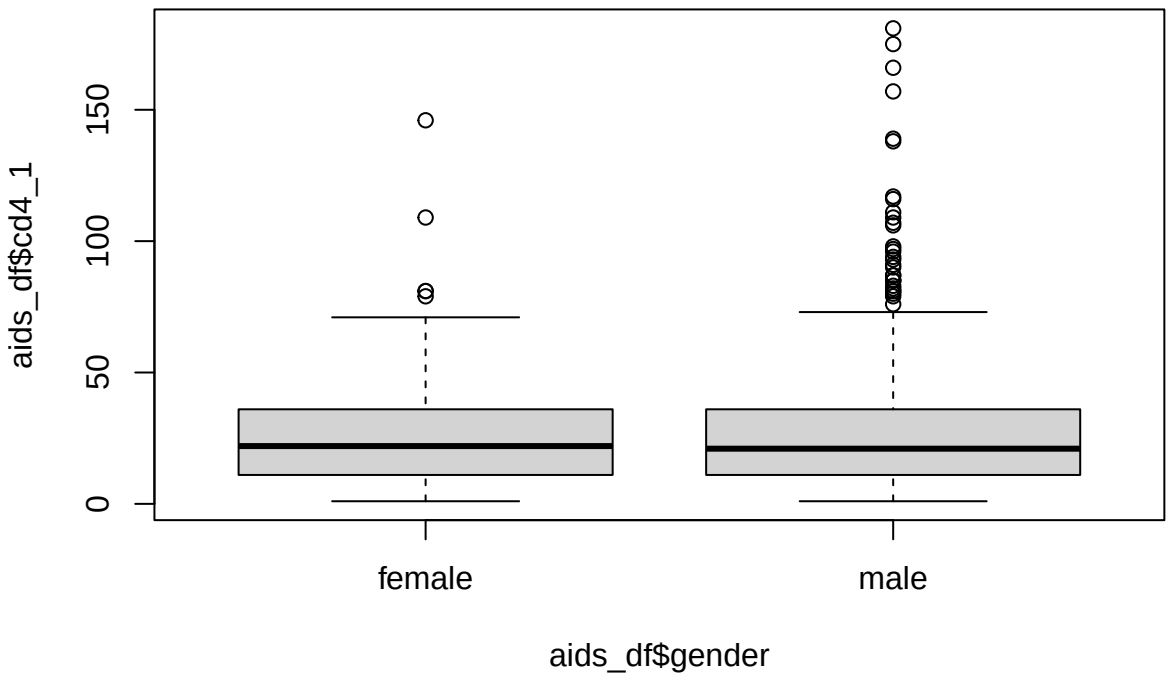
```
##
##  Shapiro-Wilk normality test
##
## data:  aids_df$age
## W = 0.978, p-value = 2.2e-12
```

```
### "normalize" CD4 at week 1
hist(log(aids_df$cd4_1), xlab = "CD4 count at week 1")
```



Non-parametric Tests

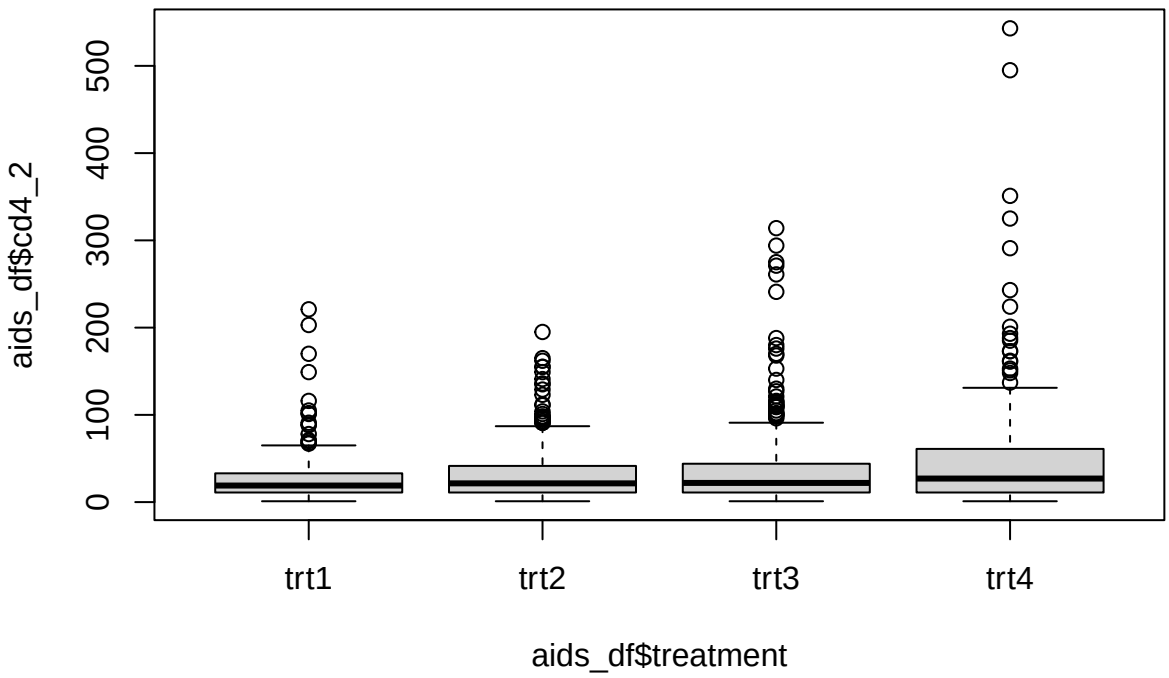
```
# is the CD4 at week 1 different between genders?
boxplot(aids_df$cd4_1~aids_df$gender)
```




```
wilcox.test(aids_df$cd4_1~aids_df$gender)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  aids_df$cd4_1 by aids_df$gender
## W = 75228, p-value = 0.66
## alternative hypothesis: true location shift is not equal to 0
```

```
# is the CD4 at week 2 different between treatment groups?
boxplot(aids_df$cd4_2~aids_df$treatment)
```




```
kruskal.test(aids_df$cd4_2~aids_df$treatment)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  aids_df$cd4_2 by aids_df$treatment
## Kruskal-Wallis chi-squared = 20, df = 3, p-value = 0.00017
```

```
pairwise.wilcox.test(aids_df$cd4_2, aids_df$treatment)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  aids_df$cd4_2 and aids_df$treatment
##
##      trt1  trt2 trt3
## trt2 0.20  -    -
## trt3 0.05  0.45 -
## trt4 5e-05 0.04 0.20
##
## P value adjustment method: holm
```