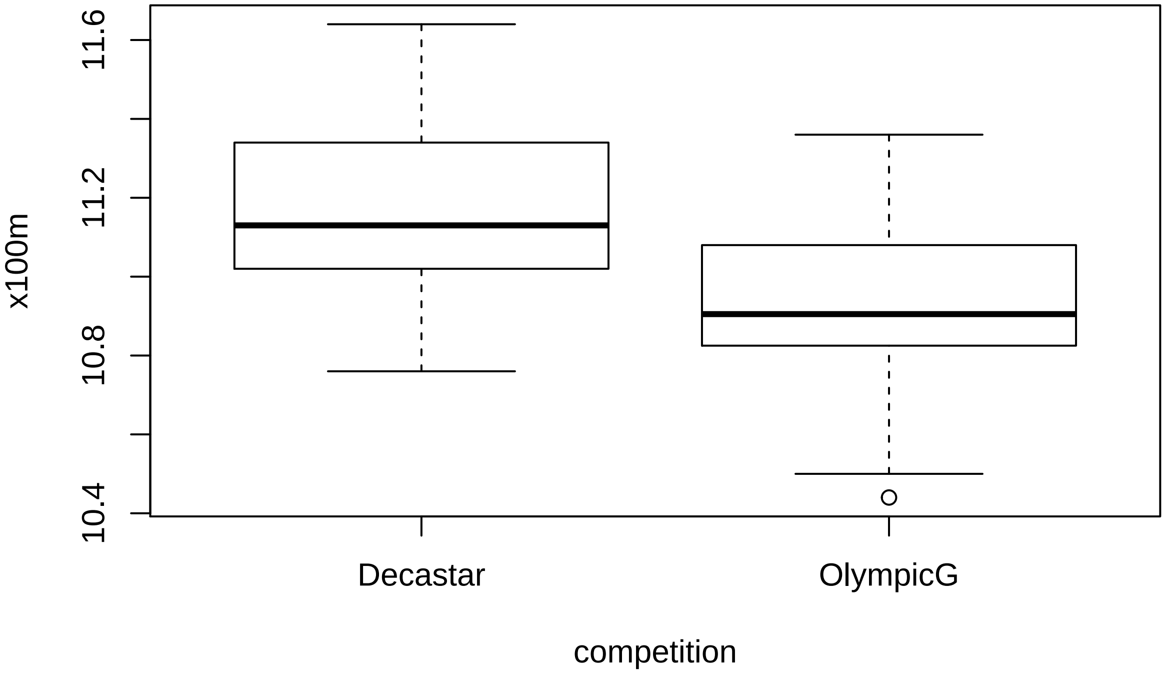
First Question: visualization, Chi-Square and t-test

Arnau Abella

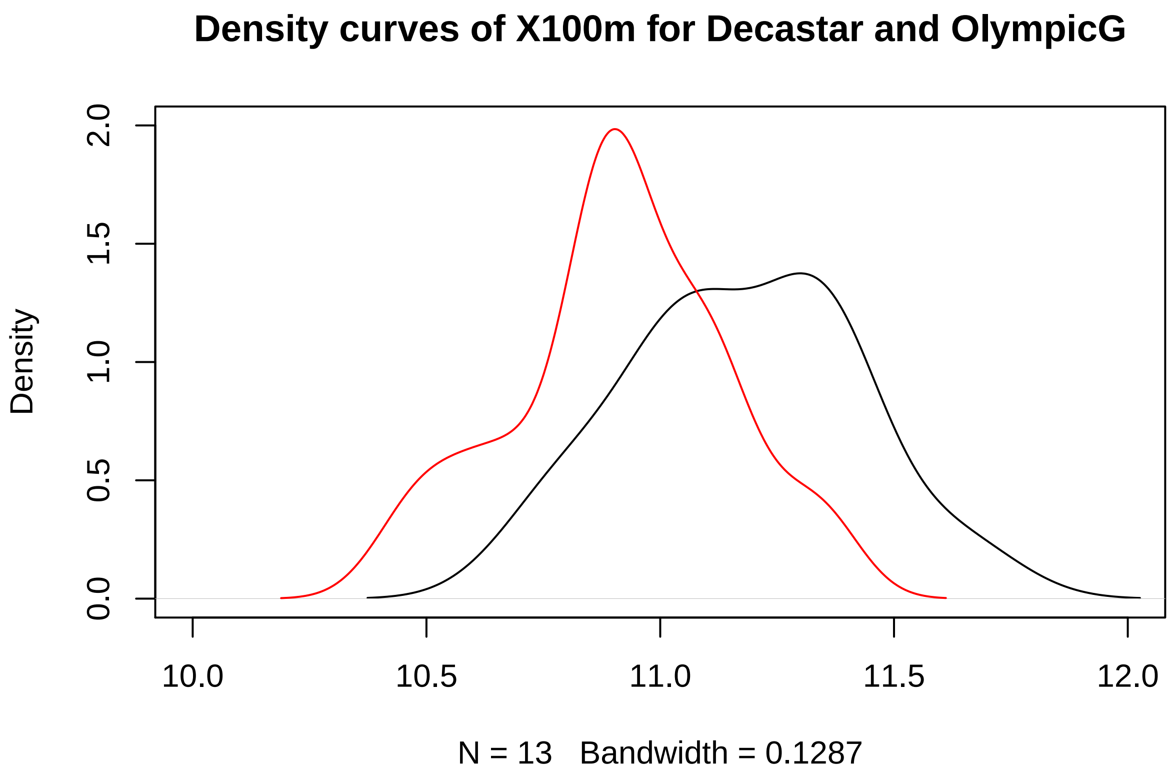
03/03/2020

a. Analyze the distribution of "X100m" according to the type of competition by using boxplot. Write your conclusion.

```
decathlon <- read.csv( "/home/arnau/MIRI/SMDE/hw1/decathlon.csv", header = TRUE, sep = ",")
dim(decathlon)
head(decathlon)
summary(decathlon)
boxplot(X100m~Competition, ylab="x100m", xlab="competition", data=decathlon)
```



```
decastar<-which(decathlon$Competition=="Decastar")
olympic <-which(decathlon$Competition=="OlympicG")
plot(density(decathlon$X100m[decastar]),main="Density curves of X100m for Decastar and OlympicG", xlim=
lines(density(decathlon$X100m[olympic]),col=2)
```



The discret random variable *X100m* follows a normal distribution by the shape of its density function.

Both distributions are very similar but the medians do not coincide.

Notice that we have less samples from Decaster (13) than from OlympicG (28) so we need to make some probabilistic analysis before refusing that the medians are equals.

From the statistics, it is easy to see that, in average, the runners from OlympicG are one second faster than the ones from Decaster.

b. Create a new categorical variable with two categories from the variable “X100m” by using 11 seconds as the cut-off point. Make a cross table from the new categorical variable and the “Competition”. Are these two variables independent? Write your conclusion by checking marginal probabilities and test the independency of two variables by using Chi-Square test.

```
decathlon$X100m_11s<- cut(decathlon$X100m, c(11,0 , 12.0))
levels(decathlon$X100m_11s)<-c("< 11s","> 11s")
tab<-table(decathlon$X100m_11s, decathlon$Competition)
tab
```

```
##
##          Decastar OlympicG
##   < 11s        2       19
##   > 11s       11        9
```

```
prop.table(tab)
```

```
##
##             Decastar   OlympicG
##   < 11s 0.04878049 0.46341463
##   > 11s 0.26829268 0.21951220
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 7.7962, df = 1, p-value = 0.005236
```

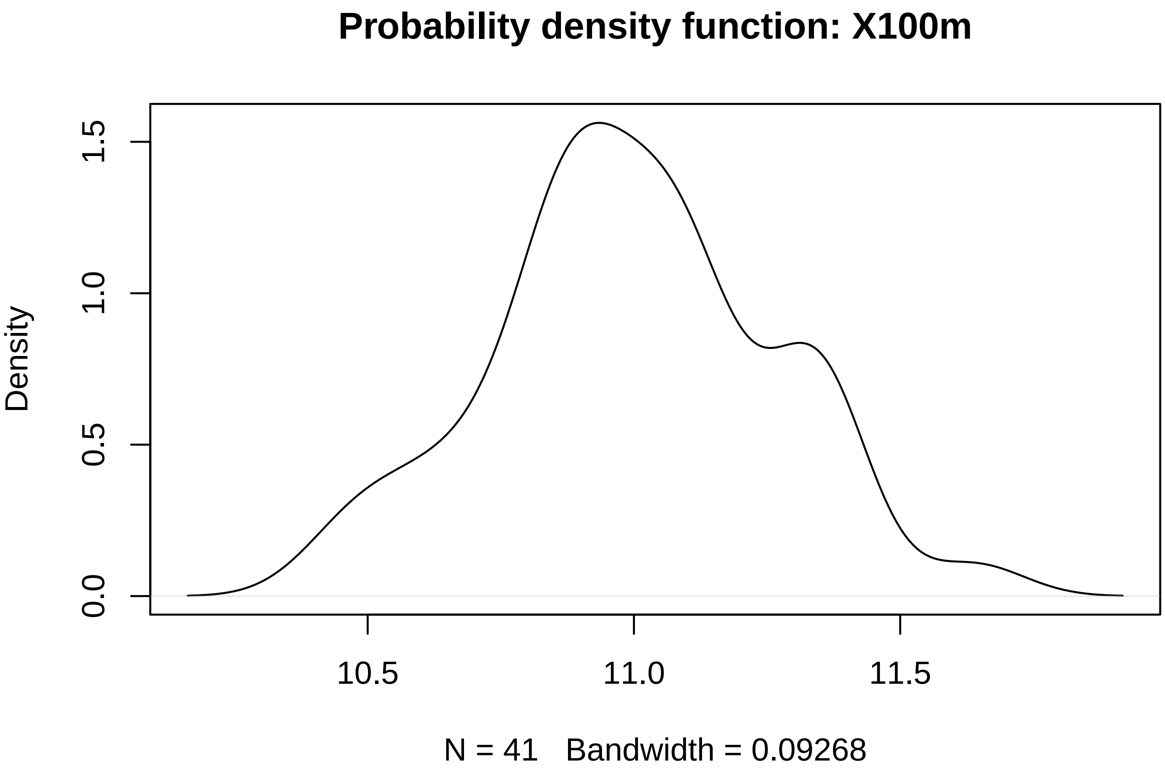
As the p-value 0.0052 is smaller than the .05 significance level, we can reject the null hypothesis that running 100 meters in less/more than 11 seconds is independent of the kind of competition.

Are these two variables independent? We can conclude from the χ^2 test that they are not.

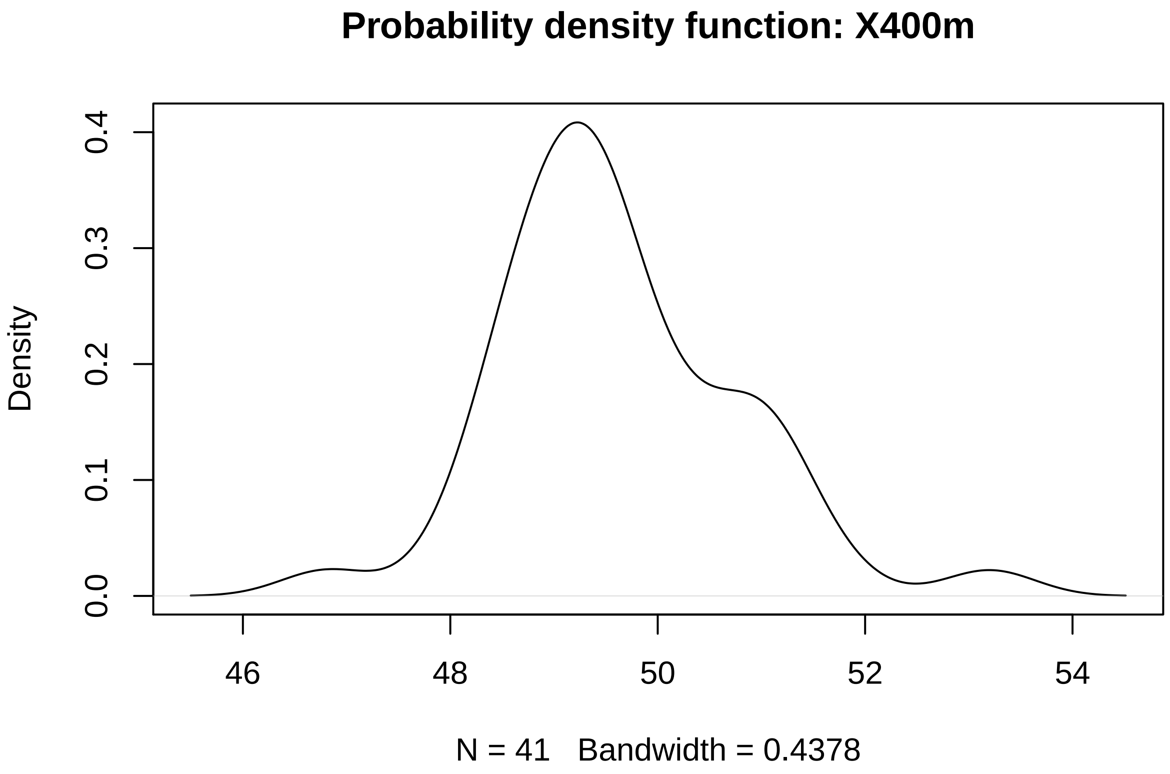
c. Visualize the distribution of quantitative variables by using proper graph. Which of these variables follows a Normal distribution?

All quantitative variables follow a Normal distribution except for **Rank** and **Points**. Here are some sample plots:

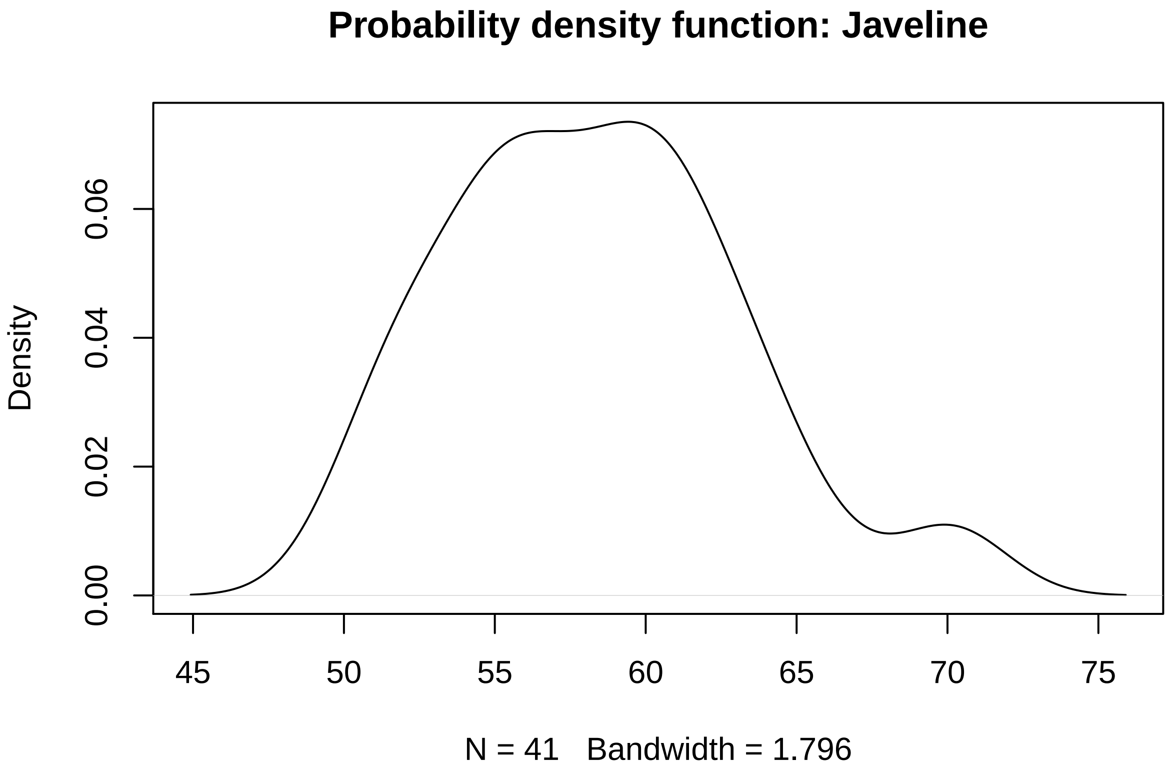
```
plot(density(decathlon$X100m), main="Probability density function: X100m")
```



```
plot(density(decathlon$X400m), main="Probability density function: X400m")
```

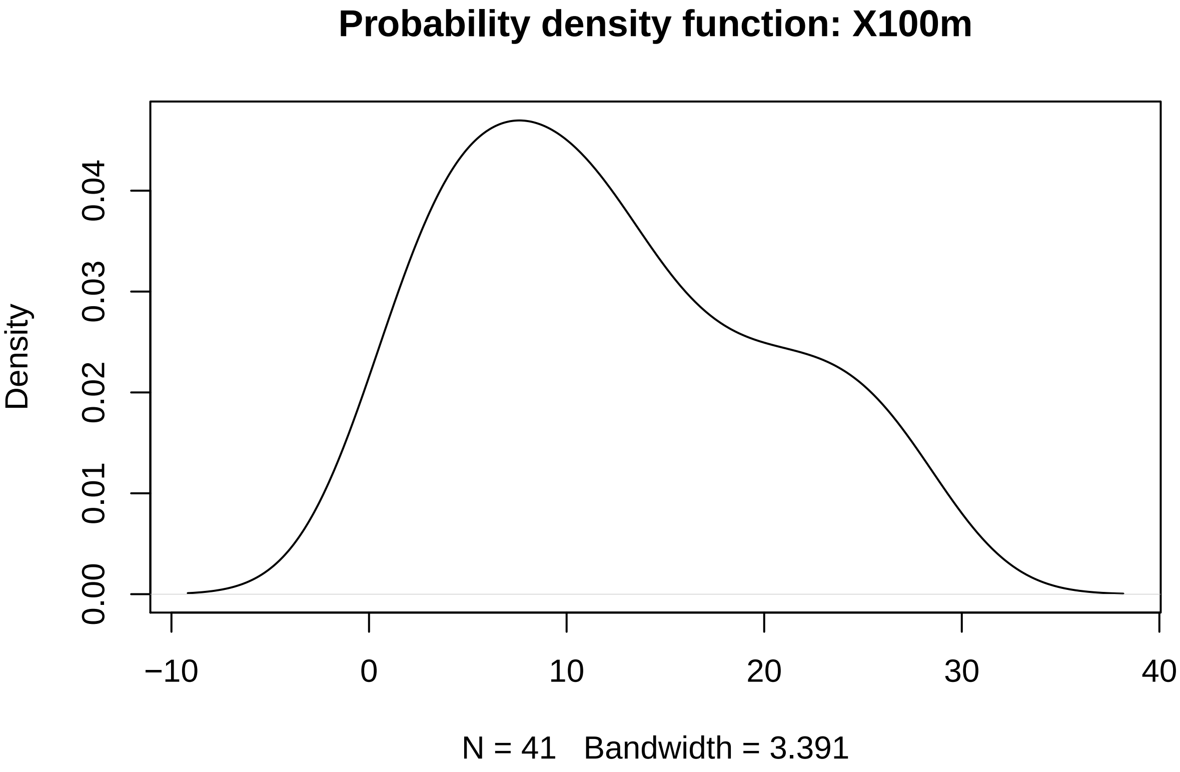


```
plot(density(decathlon$Javeline), main="Probability density function: Javeline")
```

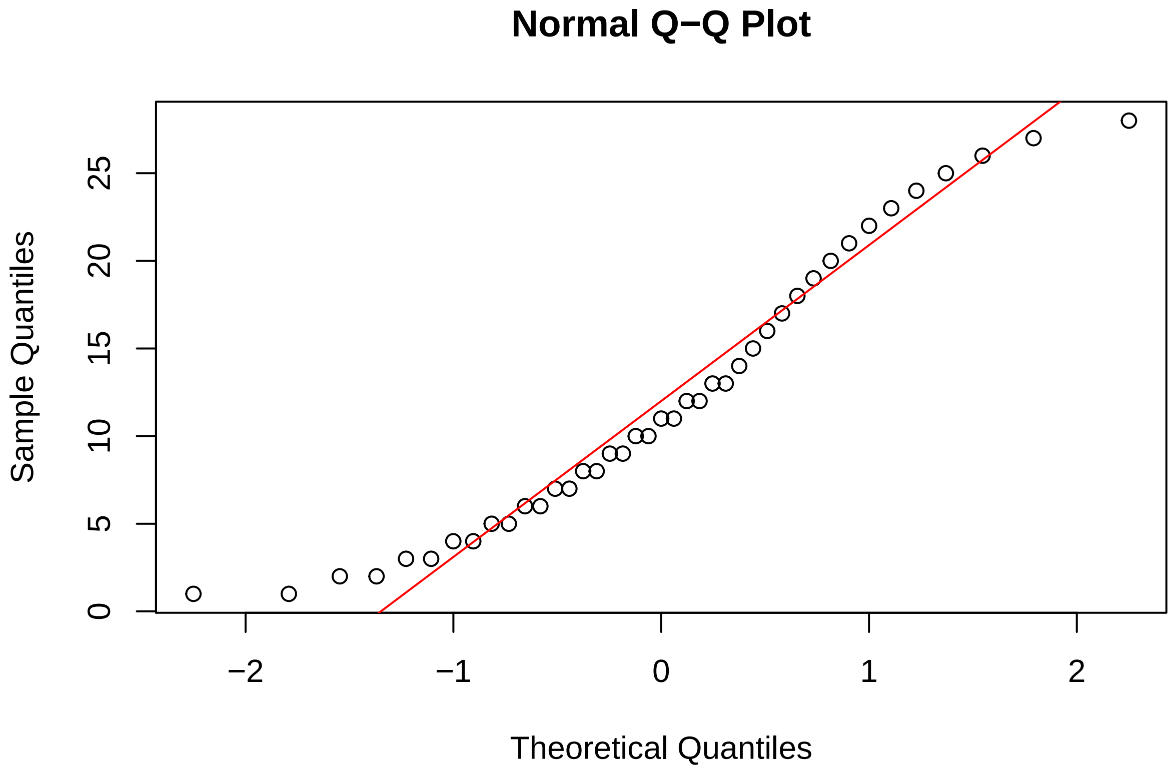


To check that Rank and Position do not follow a Normal distribution we used qqnorm and Shapiro-Wilk normality test.

```r
plot(density(decathlon$Rank), main="Probability density function: X100m")
```



```r
qqnorm(decathlon$Rank); qqline(decathlon$Rank, col = 2)
```



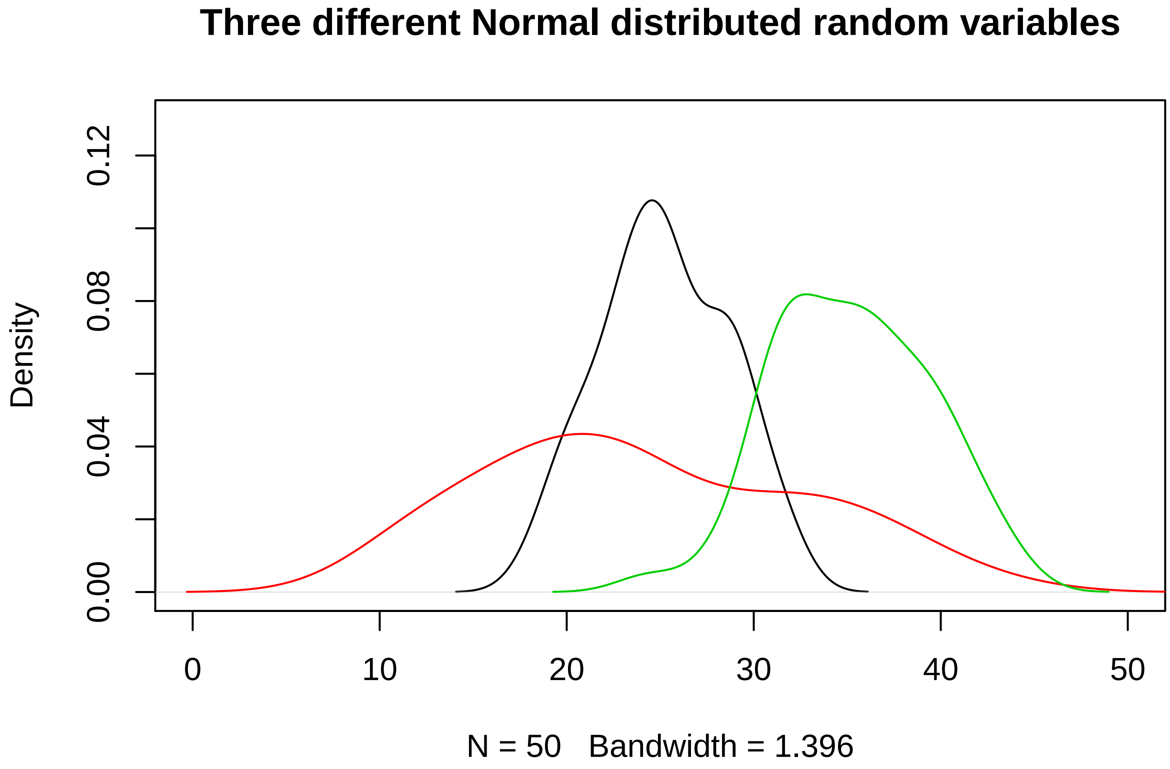
```
shapiro.test(decathlon$Rank)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  decathlon$Rank
## W = 0.94188, p-value = 0.03649
```

d. Generate three Normally distributed random variables of length 50. Two of them should have the same mean, different standard deviations while the third one has a different mean but the same standard deviation with the first distribution. Use t test to compare mean differences between three variables.

```r
s1<-rnorm(50, mean=25,sd=4)
s2<-rnorm(50, mean=25,sd=8)
s3<-rnorm(50, mean=35,sd=4)

plot(density(s1),xlim=c(0,50),ylim=c(0,0.13), main="Three different Normal distributed random variables
lines(density(s2),col=2)
lines(density(s3),col=3)
```



Let's compare the mean difference using t-test:

```r
t.test(s1, s2, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  s1 and s2
## t = 0.81981, df = 98, p-value = 0.4143
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.504051  3.621466
## sample estimates:
## mean of x mean of y
##  25.22125  24.16254
```

We can't refuse that the mean of **s1** and **s2** are different because the p-value 0.56 is greater than α. However, we can refuse this for **s1** and **s3** as expected because the p-value is way slower than α.

```
t.test(s1, s3, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  s1 and s3
## t = -13.187, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.528993  -8.513032
## sample estimates:
## mean of x mean of y
##  25.22125  35.24226
```

Additionaly, we can test the variance of the distributions. The distributions **s1** and **s2** have differente variance as expected.

```
var.test(s1, s2, var.equal=TRUE)
```

```
##
##  F test to compare two variances
##
## data:  s1 and s2
## F = 0.16016, num df = 49, denom df = 49, p-value = 1.943e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.09088914 0.28223881
## sample estimates:
## ratio of variances
##          0.1601638
```

e. Test if there is a difference between two type of competitions according to the variables “X100m” and “X400m” by using t test.

On the *X100m* we can refuse the null hypothesis H_0 because the p-value 0.00407 is smaller than α. So we can state with 95% of confidence that there is difference in average in the times for the 100 meters race dependening on the competition.

```r
t.test(X100m ~ Competition, data = decathlon)
```

```
##
##  Welch Two Sample t-test
##
## data:  X100m by Competition
## t = 3.2037, df = 22.168, p-value = 0.00407
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09164794 0.42769272
## sample estimates:
## mean in group Decastar mean in group OlympicG
##               11.17538               10.91571
```

On the *X400m* we can’t refuse the null hypothesis H_0 because the p-value 0.9543 is greater than α. So there is no significative effidence that the average time on the 400 meters race is different depending on the competition.

```r
t.test(X400m ~ Competition, data = decathlon)
```

```
##
##  Welch Two Sample t-test
##
## data:  X400m by Competition
## t = 0.05771, df = 32.106, p-value = 0.9543
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6858299  0.7258299
## sample estimates:
## mean in group Decastar mean in group OlympicG
##                  49.63                  49.61
```