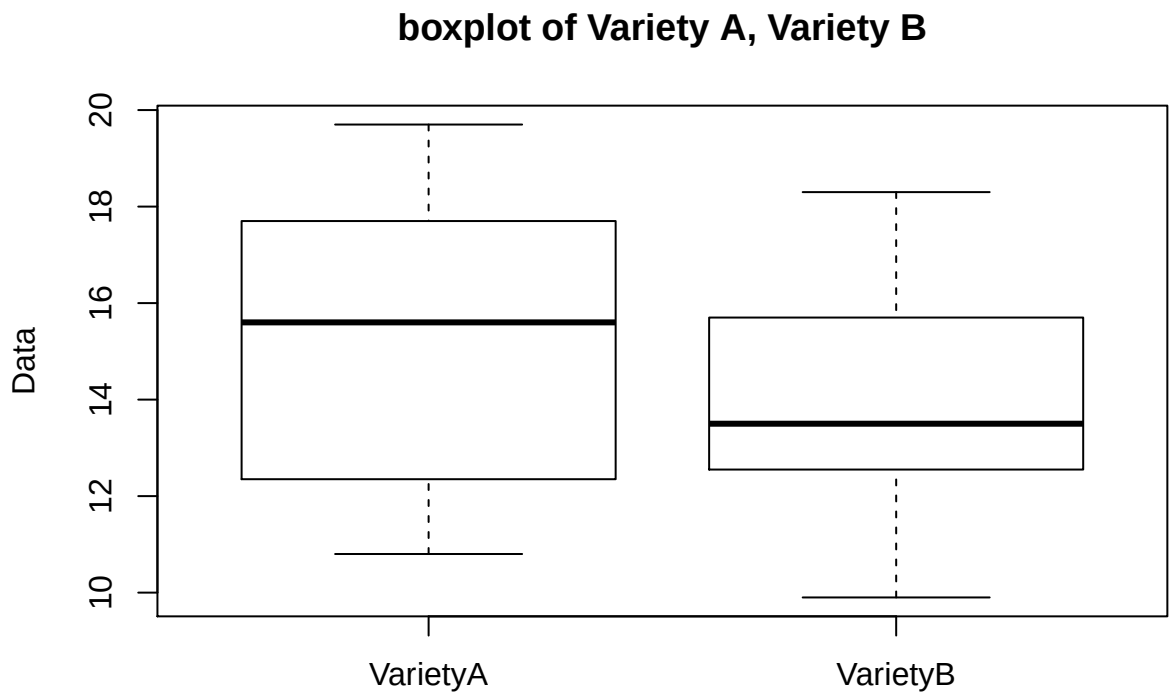
EDA.R

agmmortl

Sun Jun 25 00:00:08 2017

```
#######################
# Exploratory data analysis
# EDA.R
# Miranda Mortlock
# Example of looking at data, using a small sample
# Enter the data for two varieties
# we use a small data set to show some principles

VarietyA<-c(17.8,18.5,12.2,19.7,10.8,11.9,15.6,12.5,14.7,16.3,17.6)
VarietyB<-c(14.7,15.2,12.9,18.3,10.1,12.2,13.5,9.9,16.2,13.5,17.0)

# Check they are stored VarietyA and VarietyB

# A quick visual comparison
# How are they distributed?


boxplot(as.data.frame(cbind(VarietyA,VarietyB)), main = "boxplot of Variety A, Variety B",ylab="Data")
```



```
######cbind:  take a sequence of vector, matrix or data frames' arguments and combine by columns.
####### as.data.frame:  functions to check if an object is a data frame, or coerce it if possible.

# look at the medians
# Look at box (which encloses the middle 50% of the data)
# Look at the Whiskers
# Check if there there any outliers (stars)


myvar <-  as.data.frame(cbind(VarietyA,VarietyB))


# Extract some summary statistics

summary(VarietyA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.80   12.35   15.60   15.24   17.70   19.70
```

```
summary(VarietyB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.90   12.55   13.50   13.95   15.70   18.30
```

```
length(VarietyA)  # number in the dataset
```

```
## [1] 11
```

```
length(VarietyB)
```

```
## [1] 11
```

```
sd(VarietyA)      # standard deviation
```

```
## [1] 3.029941
```

```
sd(VarietyB)
```

```
## [1] 2.668469
```

```
sd(VarietyA)/sqrt(11)   ####SE mean
```

```
## [1] 0.9135617
```

```
sd(VarietyB)/sqrt(11)
```

```
## [1] 0.8045737
```

```
sd(VarietyA)/mean(VarietyA)   ######Coefficient of Variance
```

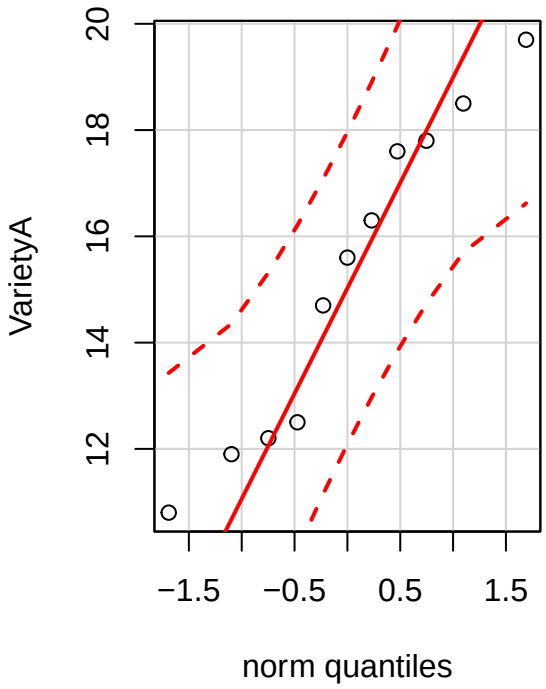
```
## [1] 0.1988625
```

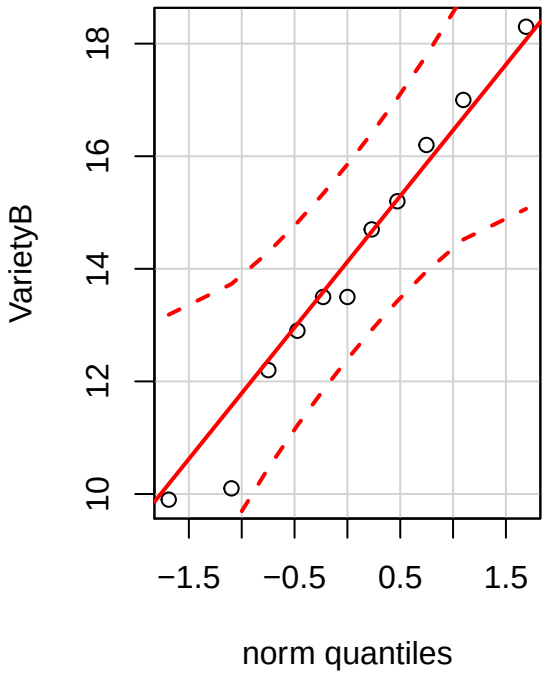
```
sd(VarietyB)/mean(VarietyB)
```

```
## [1] 0.1912258
```

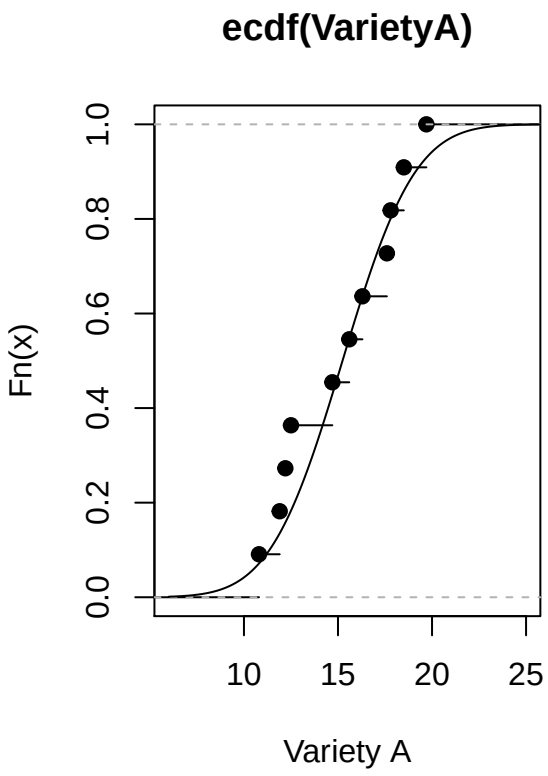
```
################################

layout(matrix(1:2, nrow = 1))
library(car)
```

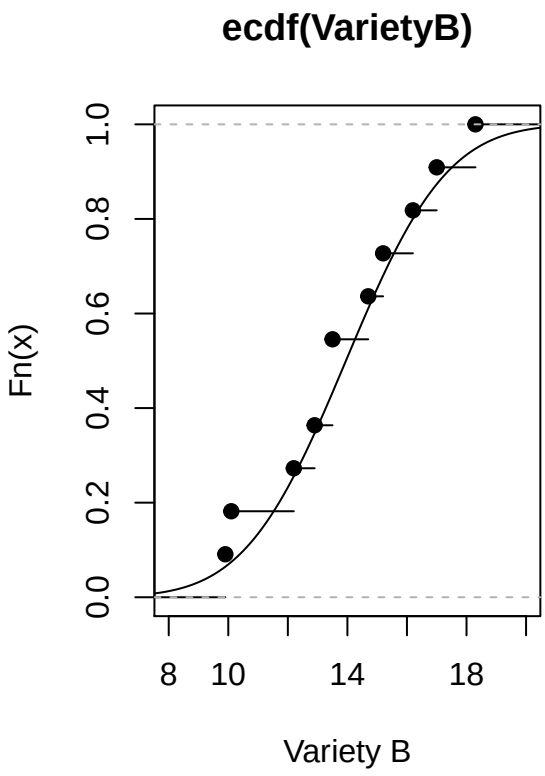
```r
qqPlot(VarietyA)
qqPlot(VarietyB)
```






```r
layout(matrix(1:2, nrow = 1))
# First set up a vector Z using seq() command
# It requires a min amd max and a step

z<-seq(6,30,0.1)

## Then make a pnormal dsitribution from your data.
# pnorm requires a vector of quantiles
# and the mean and sd
cdfA <- pnorm(z,15.24,3.03)
cdfB <- pnorm(z,13.95,2.668)

# ecdf is the empirical distribution

F10 <- ecdf(VarietyA)  #####empirical cdf of Variety A######
F12 <- ecdf(VarietyB)  #####empirical cdf of Variety B######


plot(F10,xlab="Variety A",xlim=c(6,25))  #####xlim: the range of x-axis
lines(z, cdfA, type="l")

plot(F12,xlab="Variety B",xlim=c(8,20))
lines(z, cdfB, type="l")
```






```
 # We could do a test for normality

 # the question is what do we test- the whole Variable
 # or within a group ?

 library(agricolae)
# Kolmogorov-Smirnov Tests
# only for one or two sample data
# look at help if you wish to use it
 ?ks.test()
```

```
## starting httpd help server ...
```

```
##  done
```

```
 ####### Shapiro  test
 # The null hypothesis for this test is that the
 # data are normally distributed. The Prob < W value
 # listed in the output is the p-value. If the chosen
 # alpha level is 0.05 and the p-value is less than 0.05,
 # then the null hypothesis that the data are normally
 # distributed is rejected.


 shapiro.test(VarietyA)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  VarietyA
## W = 0.93809, p-value = 0.4983
```

```
shapiro.test(VarietyB)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  VarietyB
## W = 0.96953, p-value = 0.8817
```

```
# We could also log transform the data
# Run the analysis on transformed data and see what happens

varAlog <- log(VarietyA)

## end of session

# Any questions add to yieldingresults.org
```