Case Study : Normality of data? No ! Normality of the residuals of ANOVA.

Prof L. Gentzbittel Skoltech, Digital Agriculture Laboratory [*]

Prof C. Ben, Skoltech, Digital Agriculture Laboratory [†]

April, 2nd 2021 - Skoltech

CASE STUDY PRESENTATION

The objective of this script is to exemplify the requirement of normality of the residuals of the ANOVA model – NOT of the raw data

PREPARATION OF THE WORKING INTERFACE IN R

```
### I. Set working directory
# On RStudio: tab 'Session'-> Set Working Directory -> Choose Directory.
# Choose the directory containing the datafile and the associated R script.

### II. Possibly, installation of new R packages needed for the analysis on RStudio:
# Click on the 'Packages' tab in the bottom-right window of R Studio interface->'Install Packages'
# Comment #1: R package installation requires a connection to internet
# Comment #2: Once packages have been installed, no need to re-install them
# again when you close-open again RStudio.

### III. Initialisation of the working space
# To erase all graphs
graphics.off()
# To erase objects from the working space - Clean up of the memory
rm(list = ls())
```

LOADING REQUIRED METHODS FOR ANALYSIS

```
library(ggplot2)    # a new graphic library - Will be presented in details in 'Regular' course
library(car)        # Levene's test for homogeneity of variances
library(agricolae)  # the Newman-Keuls test for multiple mean comparisons
```

LOADING/CREATING THE DATA and STARTING ANALYSIS

```
set.seed(145)  # set random seed generator of computer.
               # All participants will get the 'same' random data set
```

we generate yield data for 100 plants of three hybrid maize varieties (G1, G2 and G3). The environmental variance σ^2_E equals to 1.5. The phenotypic means of G1 equals 120kg/ha, of G2 equals 115kg/ha and of G3 equals 113kg/ha.

```
## generate data :
Yield <- data.frame( yield = c(rnorm(100, 120, 1.5),
                               rnorm(100, 115, 1.5),
```

[*] l.gentzbittel@skoltech.ru

[†] c.ben@skoltech.ru

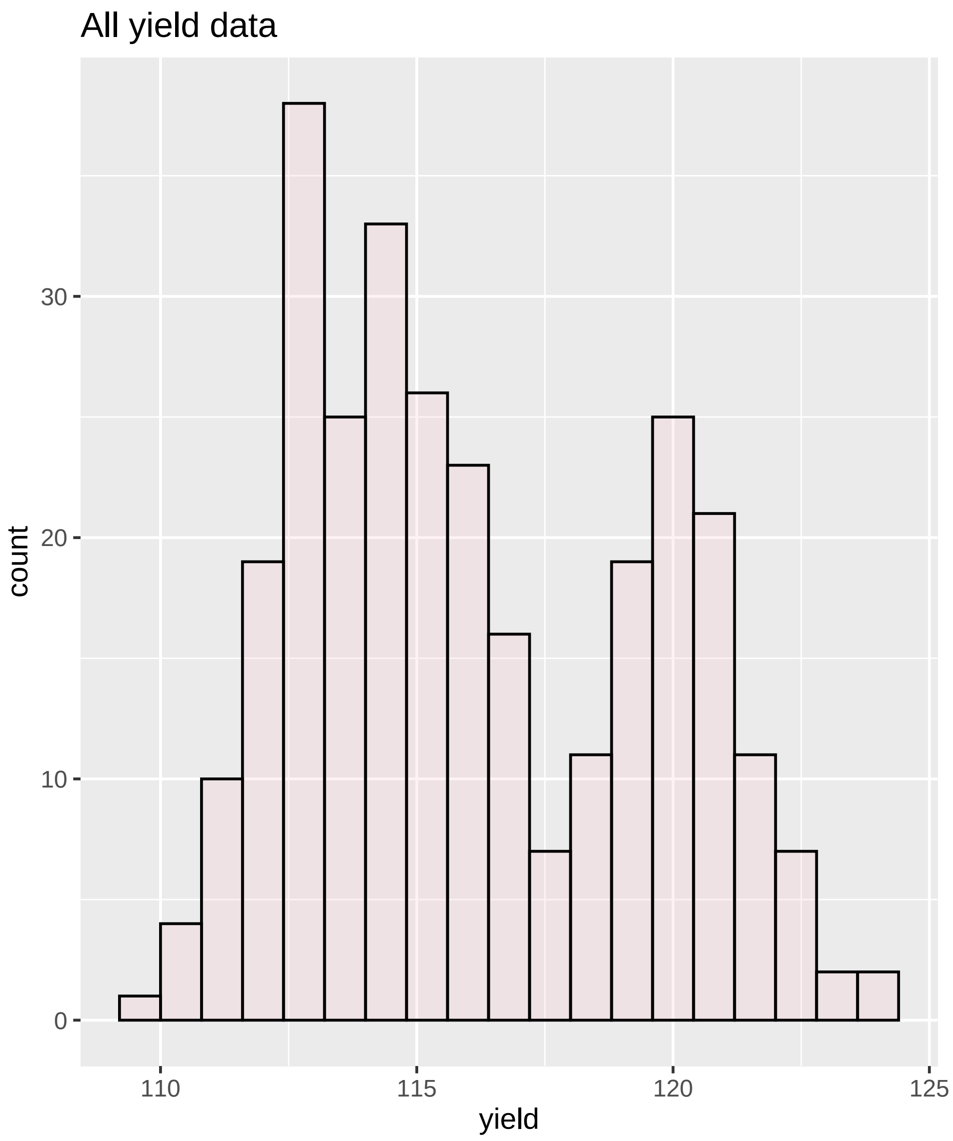
```
                              rnorm(100, 113, 1.5)),
                    genotype = rep(c("G1", "G2", "G3"), each = 100)
                    )
( head(Yield, 12) )  ## print 12 first lines of Yield on screen
```

```
##       yield genotype
## 1  121.0304       G1
## 2  121.5995       G1
## 3  120.8051       G1
## 4  122.8590       G1
## 5  121.5947       G1
## 6  122.0555       G1
## 7  120.7917       G1
## 8  120.6046       G1
## 9  121.7516       G1
## 10 121.1890       G1
## 11 118.1797       G1
## 12 118.4916       G1
```

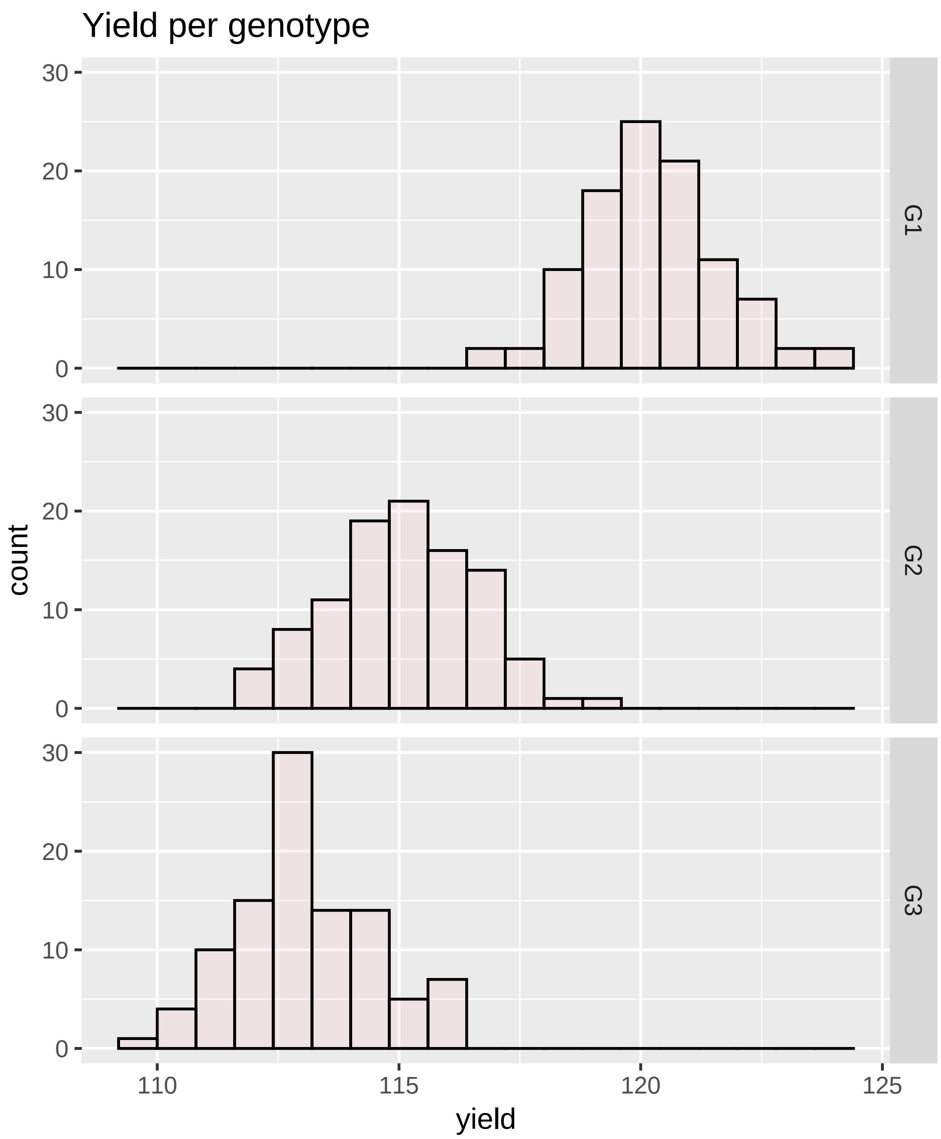
```
( tail(Yield, 12) )  ## print 12 last lines of Yield on screen
```

```
##        yield genotype
## 289 113.3233       G3
## 290 114.1899       G3
## 291 113.1149       G3
## 292 112.3984       G3
## 293 112.9925       G3
## 294 111.0509       G3
## 295 112.1010       G3
## 296 115.4383       G3
## 297 113.5930       G3
## 298 110.4583       G3
## 299 116.0888       G3
## 300 112.9643       G3
```

```
## histograms of yield
x11()  ## open a graphic window
## syntax for ggplot will be explained later - do not spend time to understand/recall for now
ggplot(Yield) + aes( yield) +
    geom_histogram(binwidth = 0.8, col = "black", fill = "pink", alpha = 0.2) +
    ggtitle("All yield data")
```



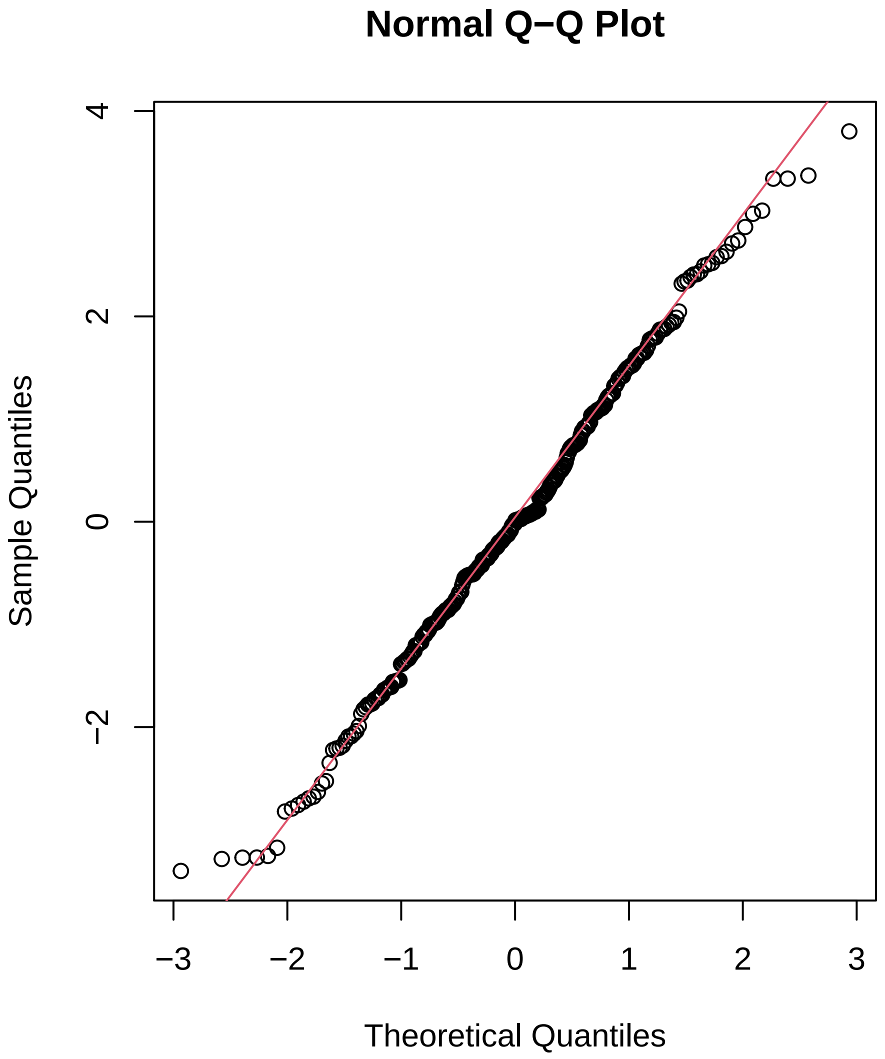
```
## histograms per variety
x11()
ggplot( Yield) + aes( yield) +
        geom_histogram(binwidth = 0.8, col = "black", fill = "pink", alpha = 0.2) +
    facet_grid( genotype ~ .) +
    ggtitle("Yield per genotype")
```



```
# Anova of yield data
model1 <- aov( yield ~ genotype, data = Yield)
summary(model1)  ## your conclusions ?
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## genotype      2 2735.8  1367.9   666.5 <2e-16 ***
## Residuals   297  609.6     2.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
x11()
## quantile - quantile plot are the standard tool to visualise
## observed distribution vs expected distribution
qqnorm(residuals(model1))  ## qqnorm() draws a quantile-quantile plot for normal (gaussian) distribution
qqline(residuals(model1), col = 2)
```



```
# normality test for yield variable of the Yield dataframe
shapiro.test(Yield$yield)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Yield$yield
## W = 0.94824, p-value = 8.772e-09
```

```
# normality test for the residudals of the model
shapiro.test(residuals(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.99519, p-value = 0.477
```

```
# homogeneity of the variances of residuals in the different varieties
## Q: is Sigma2E the same for all varieties ?
leveneTest( aov( yield ~ genotype, data = Yield) )  ## leveneTest automagically test the residuals
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.8289 0.4375
##       297
```

```r
## your conclusions ?

## comparing the varieties using the Neuman-Keuls post-hoc test.
SNK.test(model1, "genotype", group = TRUE, console = TRUE)
```

```
##
## Study: model1 ~ "genotype"
##
## Student Newman Keuls Test
## for yield
##
## Mean Square Error:  2.052467
##
## genotype,  means
##
##       yield      std   r       Min       Max
## G1 120.2716 1.379876 100 116.8708 123.6431
## G2 115.1472 1.520238 100 111.8629 118.9486
## G3 113.0897 1.393636 100 109.9146 116.1193
##
## Alpha: 0.05 ; DF Error: 297
##
## Critical Range
##         2         3
## 0.3987260 0.4772444
##
## Means with the same letter are not significantly different.
##
##       yield groups
## G1 120.2716      a
## G2 115.1472      b
## G3 113.0897      c
```