Case Study : a three-generation-means analysis in barley

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 create a short and simple script to explore a barley trial. Two parental lines, their F1 and F2 offsprings were sown in a same place. Three phenotypes are recorded on each plant:

1. number of grains per ear;
2. the presence/absence of long awns on the ear
3. resistance to *Puccinia hordei*

The goals are:

1. to test for putative differences between the different generations
2. to test for simple genetic model for both qualitative traits – awn and resistance ; and to evaluate if they are linked
3. to test if one of the morphologic trait is linked to Grains per ear.

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

```
## In this example, we will use  R-base graphics.
## We will use the newer 'ggplot2' graphic package in other examples

library(Hmisc)      ## for describe()
library(openxlsx)   ## to import Excel files
library(agricolae)  ## for Newman-Keuls
```

[*] l.gentzbittel@skoltech.ru

[†] c.ben@skoltech.ru

STARTING THE ANALYSIS

```r
#####################################
# Import of data
#####################################

## before loading data, open the excel file. Inspect organisation.
## Understand what are the factors, what are the variables.

## import from Excel to R - We will see other methods later in Regular Training
Barley <- read.xlsx("01_UsingR_BarleyPreBreeding_YieldRustEarAwns.xlsx", sheet = 1, startRow = 1, colNames

# The data
Barley
```

```
##    Population GrainPerEar  Resistance EarAwn
## 1       MU302          29   resistant    yes
## 2       MU302          33   resistant    yes
## 3       MU302          31   resistant    yes
## 4       MU302          29   resistant    yes
## 5       MU302          28   resistant    yes
## 6       MU302          28   resistant    yes
## 7       MU302          27   resistant    yes
## 8       MU302          31   resistant    yes
## 9       MU302          27   resistant    yes
## 10      MU302          30   resistant    yes
## 11      MU302          28   resistant    yes
## 12      MU302          31   resistant    yes
## 13      MU302          30   resistant    yes
## 14      MU302          27   resistant    yes
## 15      MU302          32   resistant    yes
## 16      MU302          28   resistant    yes
## 17      MU302          30   resistant    yes
## 18      MU302          29   resistant    yes
## 19      MU302          30   resistant    yes
## 20      MU302          29   resistant    yes
## 21      MU302          29   resistant    yes
## 22      MU302          29   resistant    yes
## 23      MU302          29   resistant    yes
## 24   Thibault          35 susceptible     no
## 25   Thibault          35 susceptible     no
## 26   Thibault          37 susceptible     no
## 27   Thibault          38 susceptible     no
## 28   Thibault          33 susceptible     no
## 29   Thibault          34 susceptible     no
## 30   Thibault          33 susceptible     no
## 31   Thibault          34 susceptible     no
## 32   Thibault          34 susceptible     no
## 33   Thibault          39 susceptible     no
## 34   Thibault          37 susceptible     no
## 35   Thibault          36 susceptible     no
## 36   Thibault          35 susceptible     no
## 37   Thibault          34 susceptible     no
## 38   Thibault          35 susceptible     no
## 39   Thibault          34 susceptible     no
## 40   Thibault          36 susceptible     no
## 41         F1          37   resistant     no
## 42         F1          38   resistant     no
## 43         F1          38   resistant     no
## 44         F1          41   resistant     no
## 45         F1          38   resistant     no
## 46         F1          39   resistant     no
## 47         F1          42   resistant     no
## 48         F1          37   resistant     no
## 49         F1          42   resistant     no
## 50         F1          39   resistant     no
## 51         F1          38   resistant     no
## 52         F1          38   resistant     no
## 53         F1          39   resistant     no
## 54         F1          39   resistant     no
## 55         F1          37   resistant     no
## 56         F1          35   resistant     no
## 57         F1          34   resistant     no
## 58         F1          40   resistant     no
## 59         F2          29   resistant     no
## 60         F2          31   resistant     no
## 61         F2          31   resistant     no
## 62         F2          32   resistant    yes
## 63         F2          33   resistant     no
## 64         F2          33   resistant     no
## 65         F2          33   resistant     no
## 66         F2          34   resistant    yes
## 67         F2          34   resistant     no
## 68         F2          34   resistant     no
## 69         F2          35   resistant     no
## 70         F2          35   resistant    yes
## 71         F2          35   resistant    yes
## 72         F2          35   resistant     no
## 73         F2          35   resistant     no
## 74         F2          35   resistant     no
## 75         F2          35   resistant     no
## 76         F2          36   resistant     no
## 77         F2          36   resistant    yes
## 78         F2          36   resistant     no
## 79         F2          37   resistant     no
## 80         F2          37 susceptible    yes
## 81         F2          37 susceptible    yes
## 82         F2          38   resistant     no
## 83         F2          38 susceptible     no
## 84         F2          38 susceptible     no
## 85         F2          39   resistant     no
## 86         F2          40 susceptible     no
## 87         F2          40 susceptible     no
## 88         F2          44 susceptible    yes
```

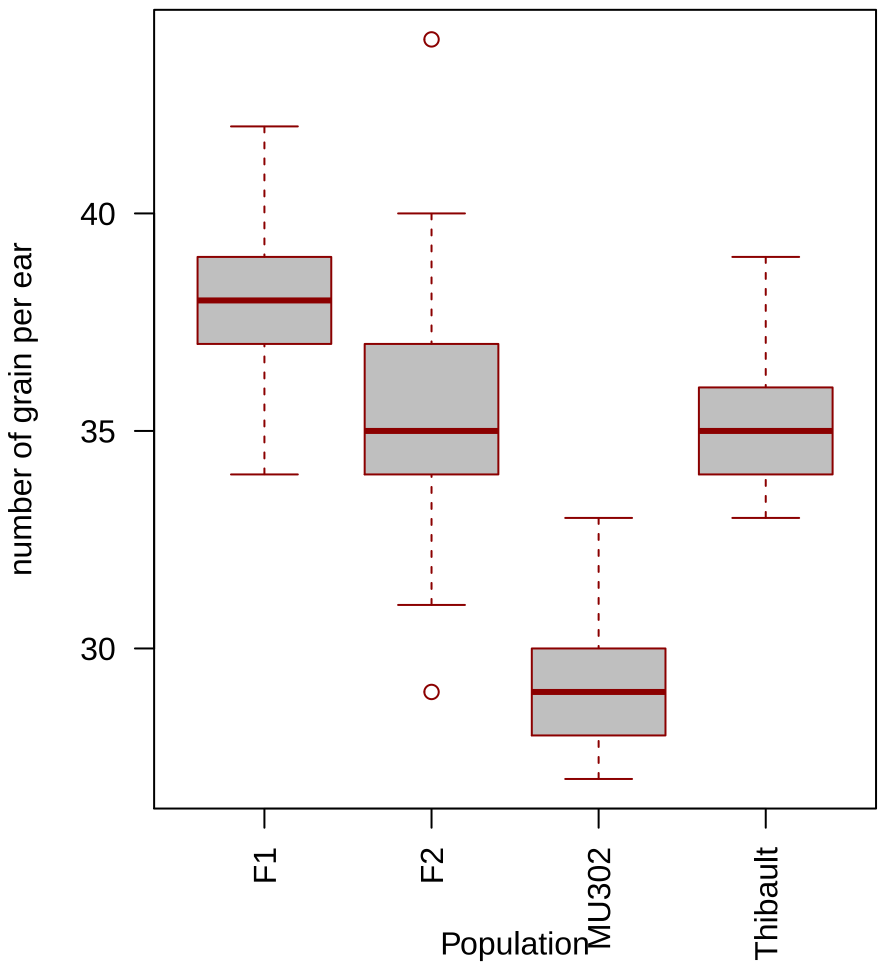
```
# Structure of dataset -- important, to check if data import is OK
str(Barley)     ## important.
```

```
## 'data.frame':    88 obs. of  4 variables:
##  $ Population : chr  "MU302" "MU302" "MU302" "MU302" ...
##  $ GrainPerEar: num  29 33 31 29 28 28 27 31 27 30 ...
##  $ Resistance : chr  "resistant" "resistant" "resistant" "resistant" ...
##  $ EarAwn     : chr  "yes" "yes" "yes" "yes" ...
```

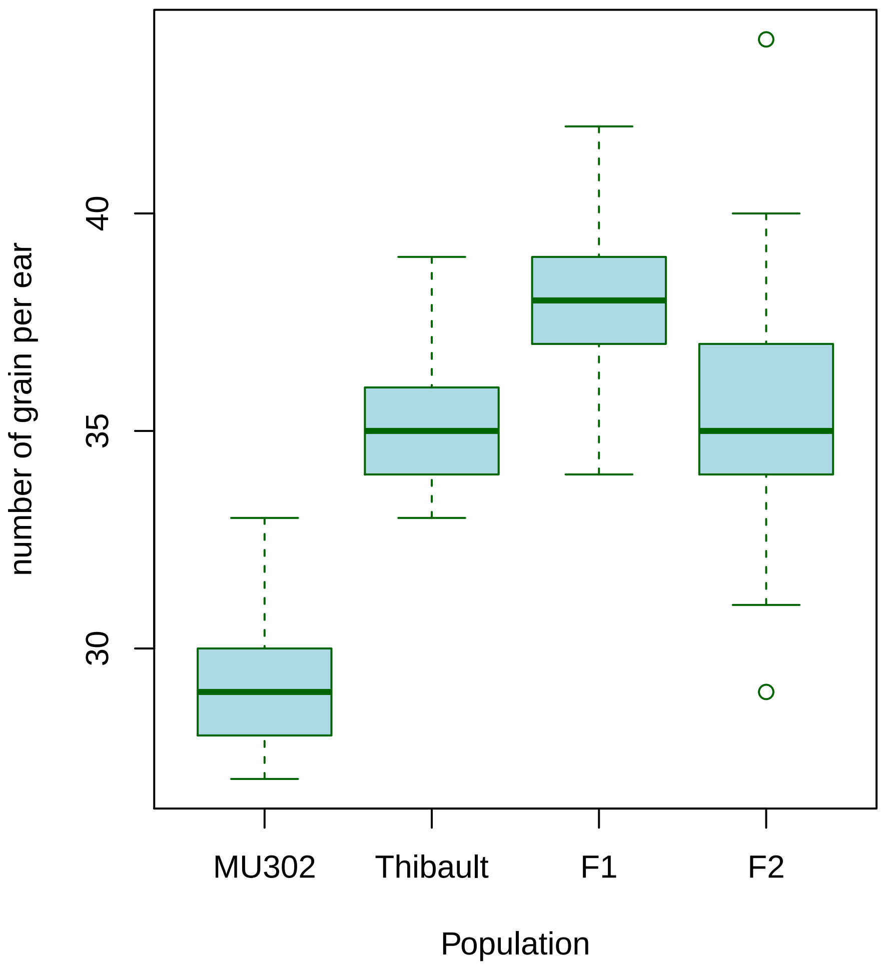
```
# A quick description of all columns of the dataset
describe(Barley)
```

```
## Barley
##
##  4  Variables      88  Observations
## --------------------------------------------------------------------------------
## Population
##        n  missing distinct
##       88        0        4
##
## Value            F1       F2    MU302 Thibault
## Frequency        18       30       23       17
## Proportion    0.205    0.341    0.261    0.193
## --------------------------------------------------------------------------------
## GrainPerEar
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       88        0       17    0.993    34.42    4.558    28.00    29.00
##      .25      .50      .75      .90      .95
##    31.00    35.00    37.25    39.00    40.00
##
## lowest : 27 28 29 30 31, highest: 39 40 41 42 44
##
## Value         27    28    29    30    31    32    33    34    35    36    37
## Frequency      3     4     8     4     5     2     6     9    12     5     8
## Proportion 0.034 0.045 0.091 0.045 0.057 0.023 0.068 0.102 0.136 0.057 0.091
##
## Value         38    39    40    41    42    44
## Frequency      9     6     3     1     2     1
## Proportion 0.102 0.068 0.034 0.011 0.023 0.011
## --------------------------------------------------------------------------------
## Resistance
##        n  missing distinct
##       88        0        2
##
## Value        resistant susceptible
## Frequency           64          24
## Proportion       0.727       0.273
## --------------------------------------------------------------------------------
## EarAwn
##        n  missing distinct
##       88        0        2
##
## Value         no   yes
## Frequency     57    31
## Proportion 0.648 0.352
## --------------------------------------------------------------------------------
```

```r
# A shortcut to avoid typing dataframe name in subsequent analyses
attach(Barley)


#################
# 1.  Graphic Analysis of data
#################

# boxplot per population
x11()  ## opens a graphic window to display the figure
boxplot(GrainPerEar ~ Population,
        las = 2,  # variety names written vertical
        ylab='number of grain per ear',
        col = "grey75", border = "darkred")
```



```
## The order of population is not convenient (alphabetic order)
## We can reorganise the 'population factor in a suitable order
Population <- factor(Population, levels = c("MU302", "Thibault", "F1","F2"))

## Back to the boxplot to view re-ordering of names
x11()
boxplot(GrainPerEar ~ Population,
        ylab='number of grain per ear',
        col = "lightblue", border = "darkgreen")
```



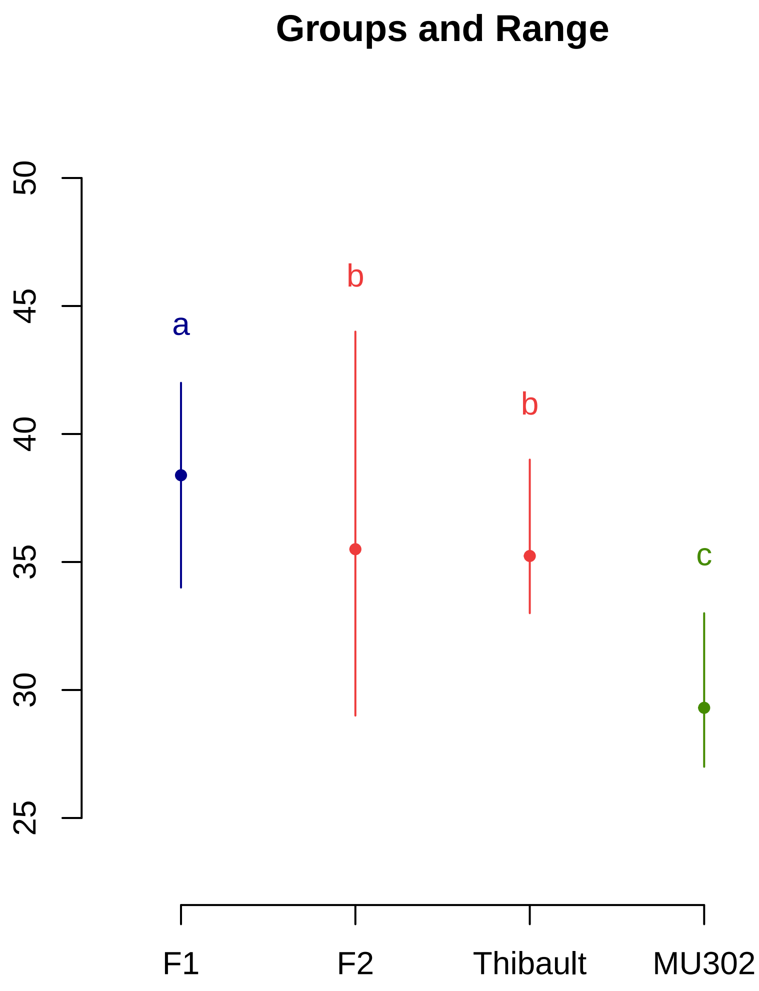
```
## please note that the variability of F2 values is greater than that of parental lines and F1:
## Any ideas why ?

#############
# let's test if it exists a difference among populations
#############
##
################  IMPORTANT : WHY do we have the right to do this comparison ?
##
## aov() is the fonction to adjust an ANOVA model to the data
## if the present case, we will fit a model with ONE factor
##  ResultOfAnova <- aov( VariableToTest ~ Factor )   ##
##
## the model is :
##
## GrainPerEar = mu + Population Effect + residual variability

ana1 <- aov(GrainPerEar ~ Population)
summary(ana1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Population    3  931.7  310.58   58.01 <2e-16 ***
## Residuals   84  449.7    5.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Your conclusions ?

## Post-hoc analysis/ what are the populations which significantly differ
## regarding the number of grains per ear
MultCompTest  <- SNK.test( ana1, trt = "Population", console = TRUE )
```

```
##
## Study: ana1 ~ "Population"
##
## Student Newman Keuls Test
## for GrainPerEar
##
## Mean Square Error:  5.353645
##
## Population,  means
##
##          GrainPerEar      std  r Min Max
## F1          38.38889 2.090283 18  34  42
## F2          35.50000 3.070999 30  29  44
## MU302       29.30435 1.579263 23  27  33
## Thibault    35.23529 1.714986 17  33  39
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Means with the same letter are not significantly different.
##
##          GrainPerEar groups
## F1          38.38889      a
## F2          35.50000      b
## Thibault    35.23529      b
## MU302       29.30435      c
```

```
x11()
plot(MultCompTest)  ## simple but useful figure
```



```
## Just to see if we would get the same results for the difference among the parental lines
## using a t-test for means
GrainP1 <- GrainPerEar[ Population == 'MU302' ]  ## [ ] is the operator to subset among the data
GrainP2 <- GrainPerEar[ Population == 'Thibault']

t.test(GrainP1, GrainP2, var.equal = FALSE)  ## in case the variances among P1 and P2 are different
```

```
##
##  Welch Two Sample t-test
##
## data:  GrainP1 and GrainP2
## t = -11.18, df = 32.933, p-value = 9.486e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.010375 -4.851518
## sample estimates:
## mean of x mean of y
##  29.30435  35.23529
```

```
## if the variances among P1 and P2 are different (or among the other populations) ...
## this is an issue for ANOVA. We will discuss it in a few minutes)
```

```r
# We can also use an ANOVA for a factor with only two levels
## subset() subset a dataframe. It is an alternative to []
toto <- subset( Barley,  subset = Population %in% c('Thibault','MU302') )
toto
```

```
##    Population GrainPerEar  Resistance EarAwn
## 1       MU302          29   resistant    yes
## 2       MU302          33   resistant    yes
## 3       MU302          31   resistant    yes
## 4       MU302          29   resistant    yes
## 5       MU302          28   resistant    yes
## 6       MU302          28   resistant    yes
## 7       MU302          27   resistant    yes
## 8       MU302          31   resistant    yes
## 9       MU302          27   resistant    yes
## 10      MU302          30   resistant    yes
## 11      MU302          28   resistant    yes
## 12      MU302          31   resistant    yes
## 13      MU302          30   resistant    yes
## 14      MU302          27   resistant    yes
## 15      MU302          32   resistant    yes
## 16      MU302          28   resistant    yes
## 17      MU302          30   resistant    yes
## 18      MU302          29   resistant    yes
## 19      MU302          30   resistant    yes
## 20      MU302          29   resistant    yes
## 21      MU302          29   resistant    yes
## 22      MU302          29   resistant    yes
## 23      MU302          29   resistant    yes
## 24   Thibault          35 susceptible     no
## 25   Thibault          35 susceptible     no
## 26   Thibault          37 susceptible     no
## 27   Thibault          38 susceptible     no
## 28   Thibault          33 susceptible     no
## 29   Thibault          34 susceptible     no
## 30   Thibault          33 susceptible     no
## 31   Thibault          34 susceptible     no
## 32   Thibault          34 susceptible     no
## 33   Thibault          39 susceptible     no
## 34   Thibault          37 susceptible     no
## 35   Thibault          36 susceptible     no
## 36   Thibault          35 susceptible     no
## 37   Thibault          34 susceptible     no
## 38   Thibault          35 susceptible     no
## 39   Thibault          34 susceptible     no
## 40   Thibault          36 susceptible     no
```

```r
## the t-test can be related to an ANOVA with only two levels of a factor
anaParents <- aov( GrainPerEar ~ Population, data = toto)
summary(anaParents)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Population   1  343.8   343.8   128.2 9.7e-14 ***
## Residuals   38  101.9     2.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## awness :2 phenotypic categories in the F2 population in segregation ->
## one locus with recessive/dominance relationships
## resistance / 2 phenotypic categories in the F2 population in segregation ->
## one locus  with recessive/dominance relationships

#####################################
#  to test AT THE SAME TIME if our genetic hypothesis is true AND if the locus are linked or not,
## we will test the expected segregation in a F2
#
#  we observe 4 phenotypic classes in F2 ->  our basis hypothesis is :  ?
## what are the expected genetic and phenotypic formulas in F2 given our basis hypothesis ?

TwoTraits <- table( EarAwn[Population == 'F2'], Resistance[Population == 'F2'] )
TwoTraits  ## is a table
```

```
##      
##       resistant susceptible
##   no         18           4
##   yes         5           3
```

```r
segreg <- as.vector( TwoTraits )  ## table as vector for next computations
segreg
```

```
## [1] 18  5  4  3
```

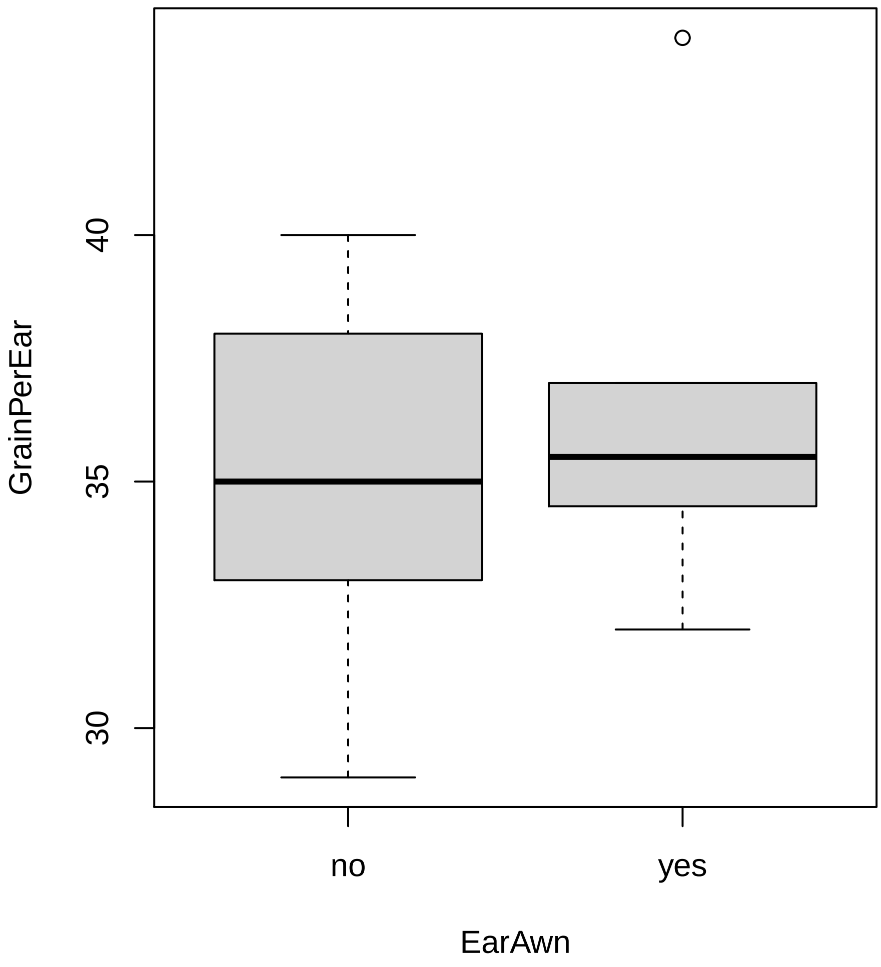
```r
# Test to fit the theoretical distribution, using a ChiSquare test :
chisq.test(segreg,  # observed distribution to test
           p = c( 9/16, 3/16, 3/16, 1/16))    ## expected distribution of segregation of two
```

```
## 
##  Chi-squared test for given probabilities
## 
## data:  segreg
## X-squared = 1.2889, df = 3, p-value = 0.7318
```

```r
                                             ## unlinked loci with Recessive/Dominance

# this test is approximate because one case is less than 5
## conclusions ?  Is Awness a possible marker for resistance:susceptibility to brown rust ?



#####################################
# Linkage between Awness and Grain per Ear ?

## Use only F2 data, so we create a new dataframe with only F2 data

F2Data <- Barley[ Population == 'F2', ] ## select all lines where population equals F2; and all columns
F2Data
```

```
##    Population GrainPerEar  Resistance EarAwn
## 59         F2          29   resistant     no
## 60         F2          31   resistant     no
## 61         F2          31   resistant     no
## 62         F2          32   resistant    yes
## 63         F2          33   resistant     no
## 64         F2          33   resistant     no
## 65         F2          33   resistant     no
## 66         F2          34   resistant    yes
## 67         F2          34   resistant     no
## 68         F2          34   resistant     no
## 69         F2          35   resistant     no
## 70         F2          35   resistant    yes
## 71         F2          35   resistant    yes
## 72          F2          35   resistant     no
## 73          F2          35   resistant     no
## 74          F2          35   resistant     no
## 75          F2          35   resistant     no
## 76          F2          36   resistant     no
## 77          F2          36   resistant    yes
## 78          F2          36   resistant     no
## 79          F2          37   resistant     no
## 80          F2          37 susceptible    yes
## 81          F2          37 susceptible    yes
## 82          F2          38   resistant     no
## 83          F2          38 susceptible     no
## 84          F2          38 susceptible     no
## 85          F2          39   resistant     no
## 86          F2          40 susceptible     no
## 87          F2          40 susceptible     no
## 88          F2          44 susceptible    yes
```

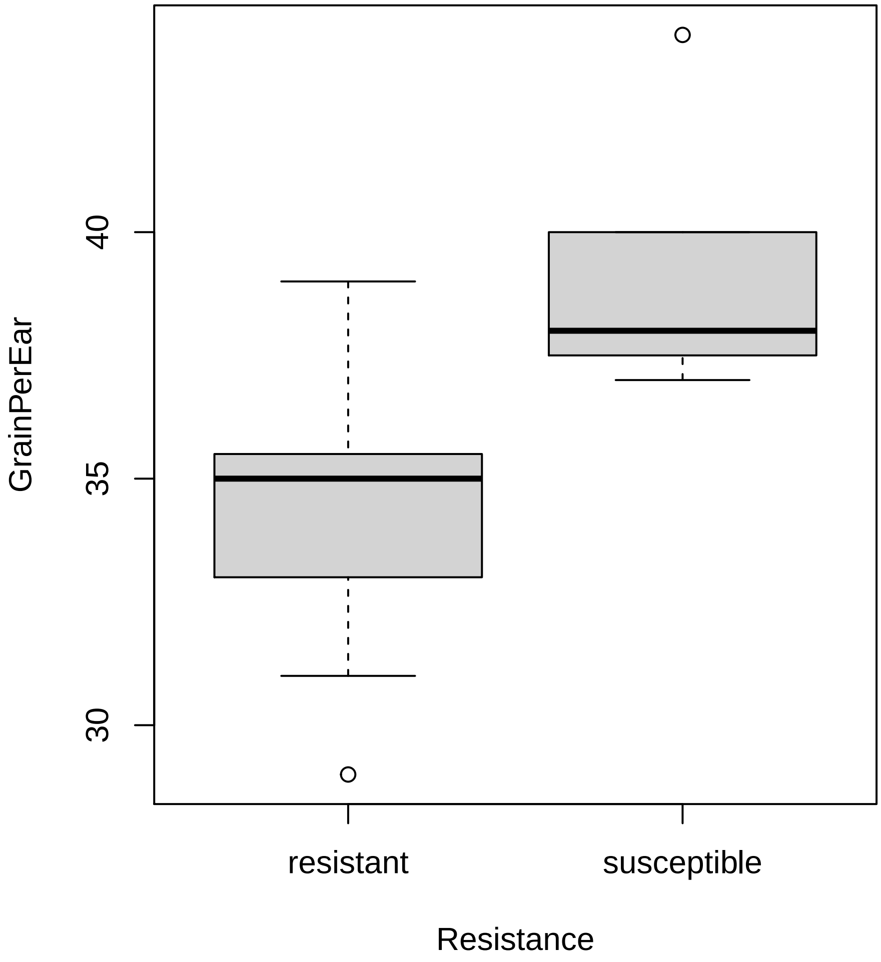
```
# Ear awness and Grain per Ear
# A graphic
x11()
boxplot(GrainPerEar ~ EarAwn, data = F2Data )
```



```
### test for difference using ANOVA - note there are only two levels of EarAwn
## and a student t-test may have been sufficient

GrainAwn  <- aov(GrainPerEar ~ EarAwn, data = F2Data)
summary(GrainAwn)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## EarAwn       1   6.14   6.136   0.643   0.43
## Residuals   28 267.36   9.549
```

```
## conclusions ? Is awness a good marker for high yield ?


####################################
# Linkage between Resistance and Grain per Ear ?

# A graphic
x11()
boxplot(GrainPerEar ~ Resistance, data = F2Data )
```



```
### test for difference using ANOVA - note there are only two levels of Resistance
## and a student t-test may have been sufficient
```

```
GrainResistance  <- aov(GrainPerEar ~ Resistance, data = F2Data)
summary(GrainResistance)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Resistance   1  121.2  121.16   22.27 5.98e-05 ***
## Residuals   28  152.3    5.44                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## conclusions ? Is resistance a good marker for high yield  ?
## Will it be easy to breed for resistance AND high yielding varieties ?


################  in fact :
################  you've done your first QTL detection !!
```