LGN 5822 - Biometrical Genetics

L07 – Completely Randomized Design (CRD)

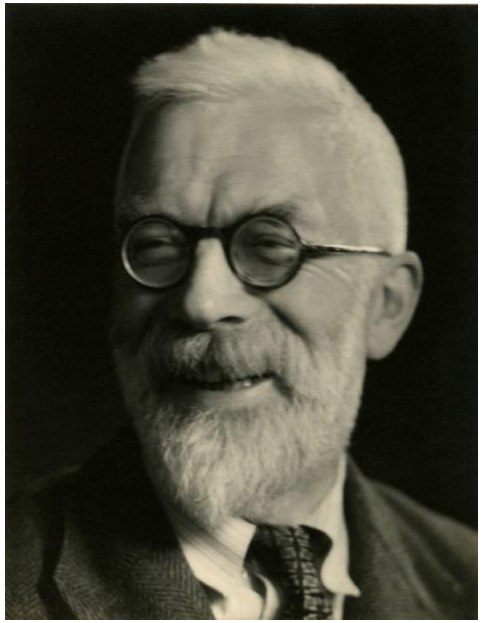
Michele Jorge Silva Siqueira

2023

Introduction

"And so it was … borne in upon me that very often, when the most elaborate statistical refinements possible could increase the **precision** by only a few percent, yet a different design involving little or no additional experimental labour might increase the **precision** two-fold, or five-fold or even more"

R.A. Fisher (1962)



Introduction

The

Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences; and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge.

Fisher, Ronald Aylmer. "The design of experiments." *The design of experiments.* 1st Ed (1935).

Introduction

- Suppose you want to make inferences about a given set of treatments

- The **levels** of the treatment factor can be **randomly sampled** from a larger population or can be chosen based on **specific interest** in each particular level: Condition that will be tested!

Introduction

- You carry out an experiment to compare the effects of t different treatments
- For doing so, you design a simple experiment with r independent replicates of each treatment level to measure these effects
  - There are tr experimental units, in total!

Introduction

Completely Randomized Design (CRD)

Main Features

- In the Completely Randomized Design (CRD), randomization is absolutely necessary throughout the treatments
  - Treatments are distributed to experimental units completely at random
  - Every experimental unit has the same probability of receiving any treatment
    - Experimental units are randomized throughout the experiment
- Consequently, there is no correlation between any two observations

Design Characterization

- The simplest form of experimental scheme
- There is no restriction on the number of treatments examined
- Treatments can have different numbers of replicates, but balance is preferable

CRD uses only the basic principles of repetition and randomization

Design Characterization

Example of Randomization

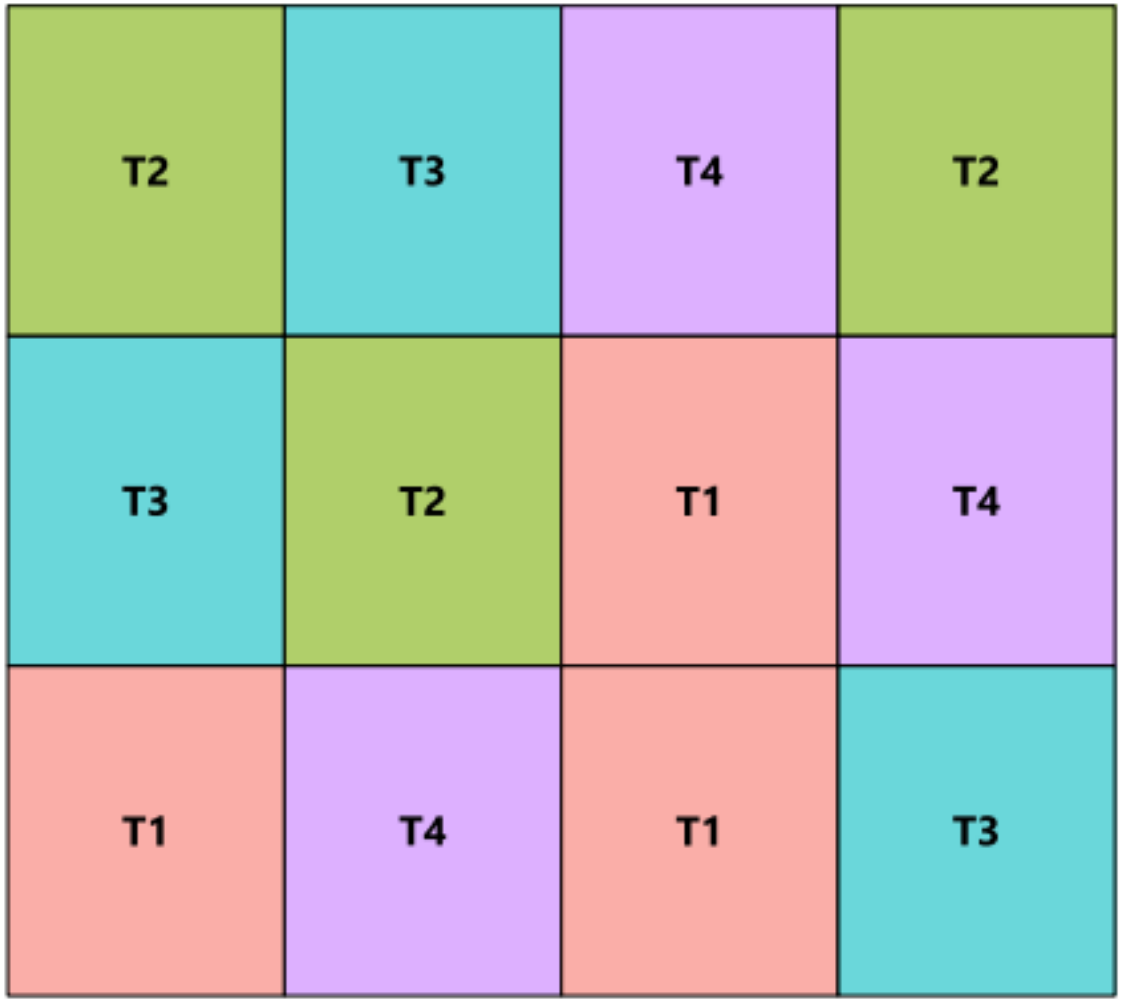
- Given you have 4 treatments (A, B, C, and D) and 5 replicates, how many experimental units would you have?

1 D	2 D	3 B	4 C	5 D	6 C	7 A	8 A	9 B	10 D
11 C	12 B	13 A	14 B	15 C	16 B	17 C	18 D	19 A	20 A

Every experimental unit has the same probability of receiving any treatment!

Design Characterization

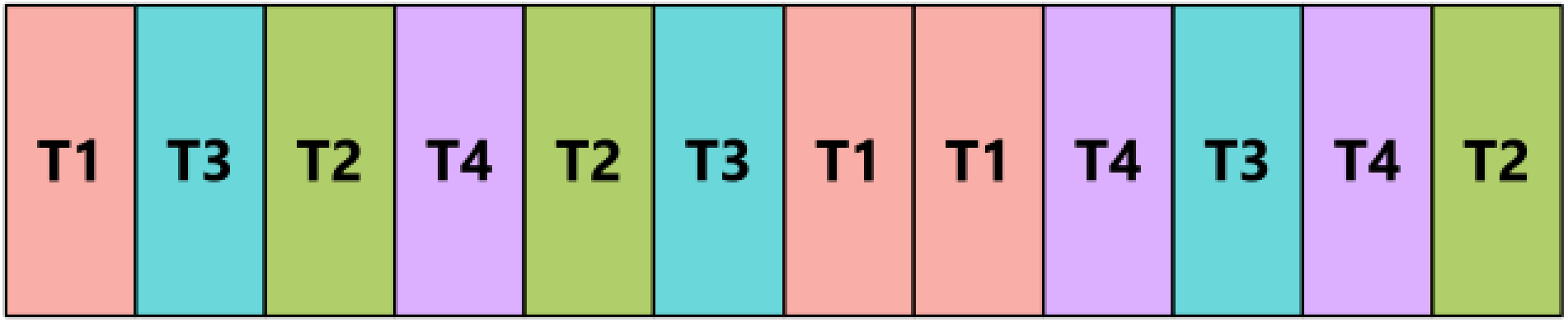
- 4 treatments (A, B, C, and D) and 3 replicates



Design Characterization

- 4 treatments (A, B, C, and D) and 3 replicates

- A different layout for the CRD:



Design Characterization

Note

- Because there is no local control (error), the CRD is appropriate when experimental units are uniform

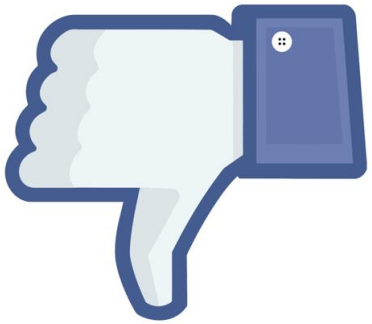
Design Characterization

Advantages of a CRD

- Very flexible design (i.e. number of treatments and replicates is only limited by the available number of experimental units)
- Statistical analysis is simple compared to other designs
- Loss of information due to missing data is small compared to other designs due to the **larger number of degrees of freedom for the error source of variation**

Design Characterization

Disadvantages of a CRD



- The experimental units must be homogeneous
- High estimate of variance due to error can be obtained because all variations (except treatments) are considered as random variation

Design Characterization

Statistical model

- Let y_{ij} represent the observed response of the jth replicate ($j = 1, \ldots, r$) of the ith treatment ($i = 1, \ldots, t$)

Model

- Data from the CRD can be described with the following model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where μ is the intercept (an overall mean), τ_i is the treatment effect and ε_{ij} is the associated random error term

Design Characterization

Statistical model

We assume that:

- $\varepsilon_{ij} \sim N(0, \sigma^2)$, for $i = 1, \ldots, t$ and $j = 1, \ldots, r$
- $cov\left(\varepsilon_{ij}, \varepsilon_{i'j'}\right) = 0$

Errors are independent and identically distributed (i.i.d)

Design Characterization

Statistical model

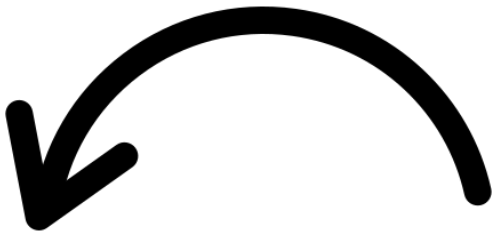
- Consequently, for the observed values:

$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$, for $i = 1, \ldots, t$ and $j = 1, \ldots, r$

Design Characterization

Model Fitting

- Unknown parameters μ, τ_i and σ^2 can be estimated via least squares, as **previously discussed** for standard linear models



Design Characterization

Model Diagnostics

Important

- After fitting the model and **before** making inferences, it is important to check whether the model assumptions are met

Main assumptions of errors:

- Errors ε_{ij} are independent
- Errors ε_{ij} are homoscedastic (have the same variance)
- Errors are normally distributed

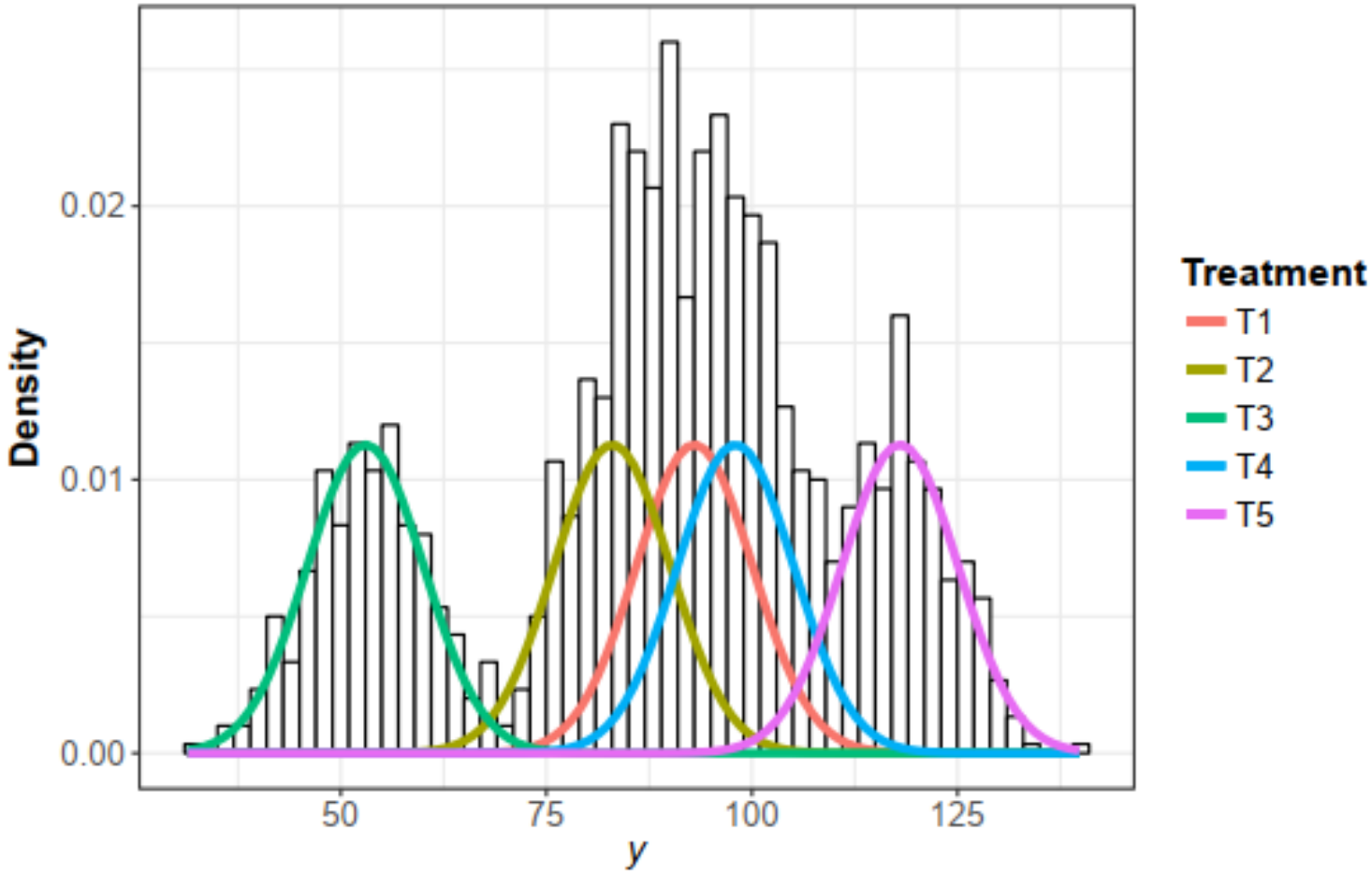
Design Characterization

Model Diagnostics

- Errors ε_{ij} are unknown, so we use the estimated residual errors, or residuals

Design Characterization

Checking Model Assumptions



17 genotypes and 3 replications

Design Characterization

Checking Model Assumptions

- Remember that:
  - $\varepsilon \sim N(0, \sigma^2 I)$
  - $y \sim N(X\beta, \sigma^2 I)$

Design Characterization

Checking Model Assumptions

- We will study the distribution of the residuals to verify validity of the assumptions

Design Characterization

Residual Diagnostics

- We can use the raw residuals, which are deviations between the observed and fitted values:

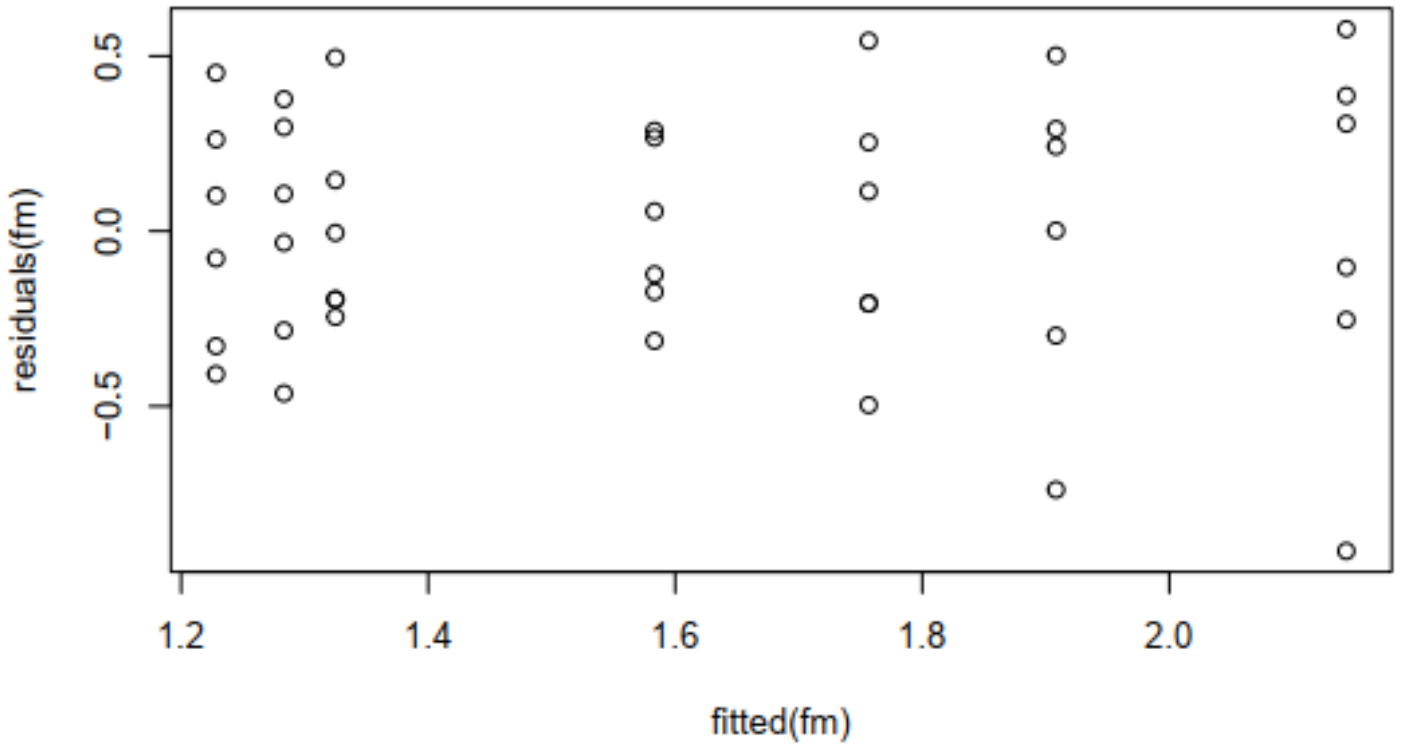
$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij}$$

$$\hat{\varepsilon}_{ij} = y_{ij} - \mu - \tau_i$$

Design Characterization

Residual Diagnostics

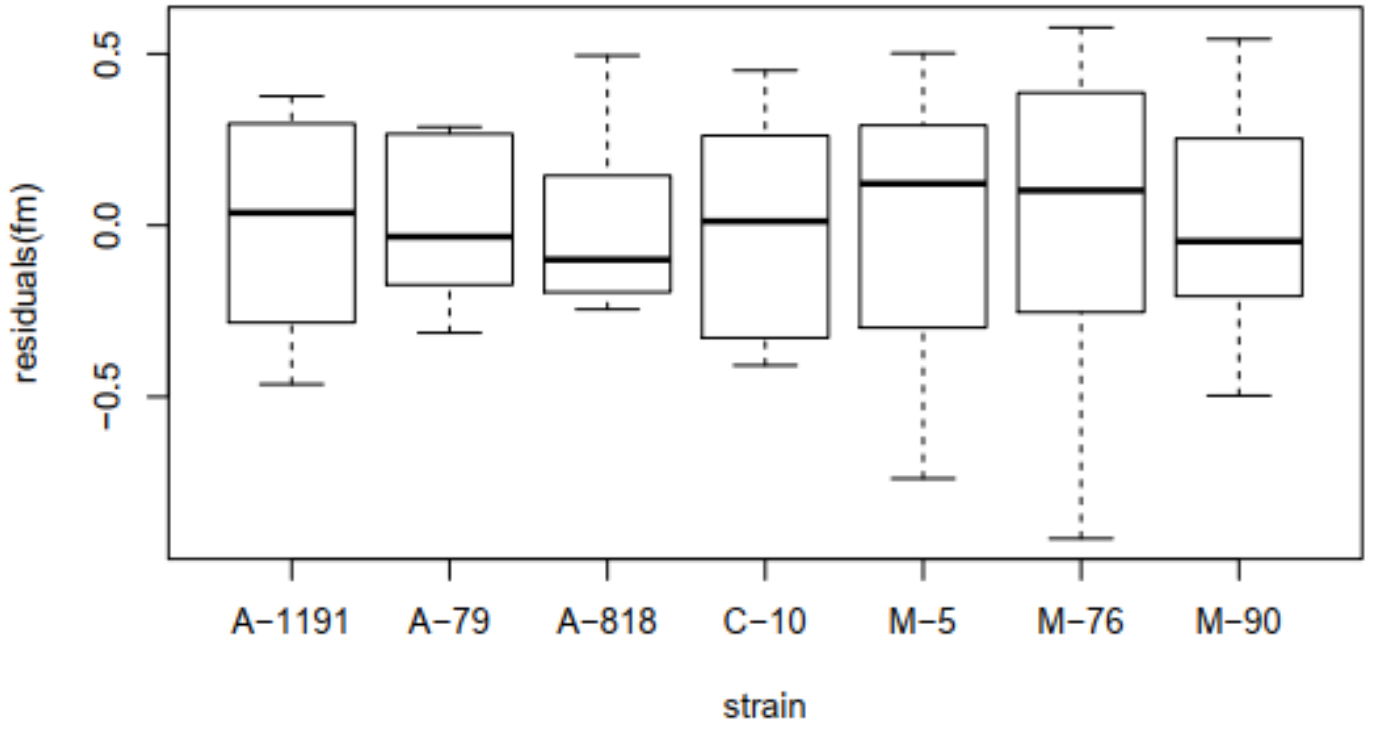
- Raw residuals *vs* fitted values



Design Characterization

Residual Diagnostics

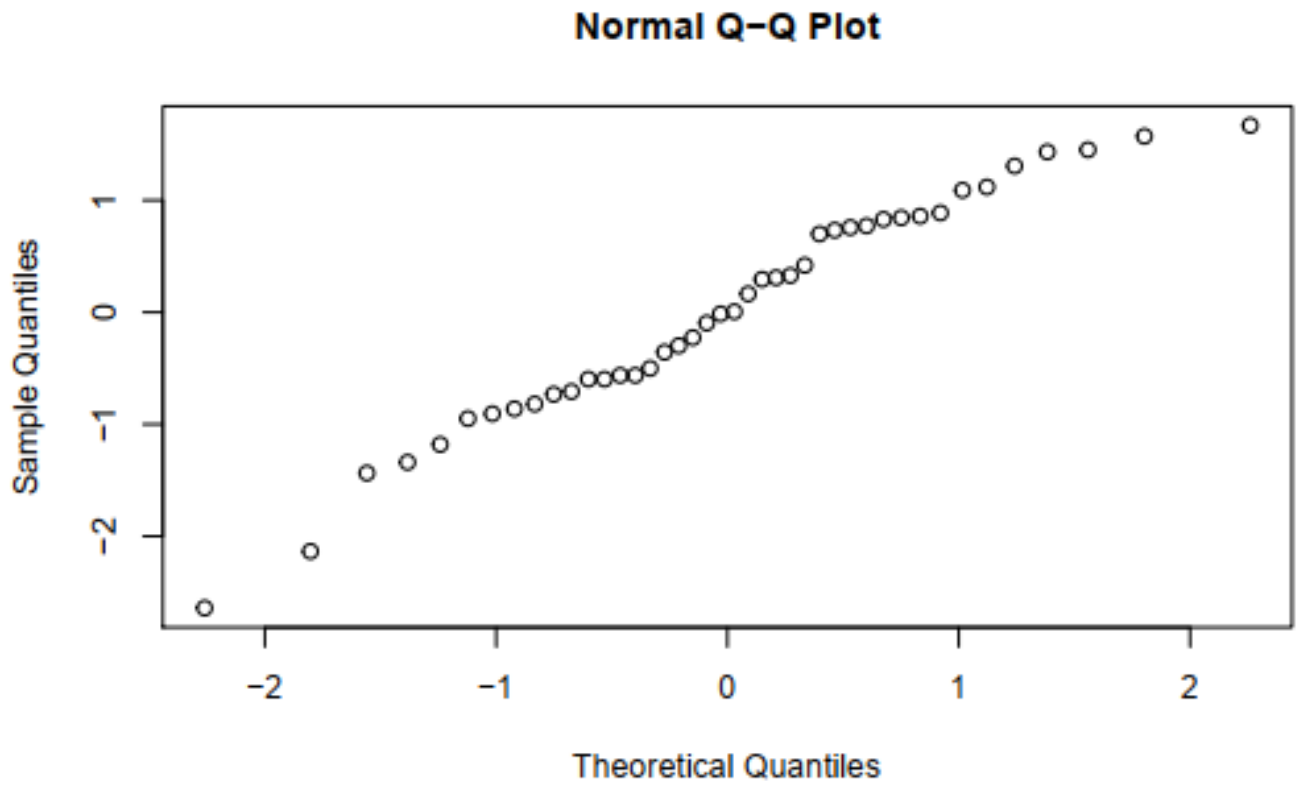
- Raw residuals *vs* treatment levels



Design Characterization

Residual Diagnostics

- Normal quantile-quantile (Q-Q) plots



Design Characterization

Data Transformation

- If (some of) the assumptions are not met, we can transform the observed data so that it is normally distributed and the variance is stabilized

Design Characterization

Data Transformation : Box Cox

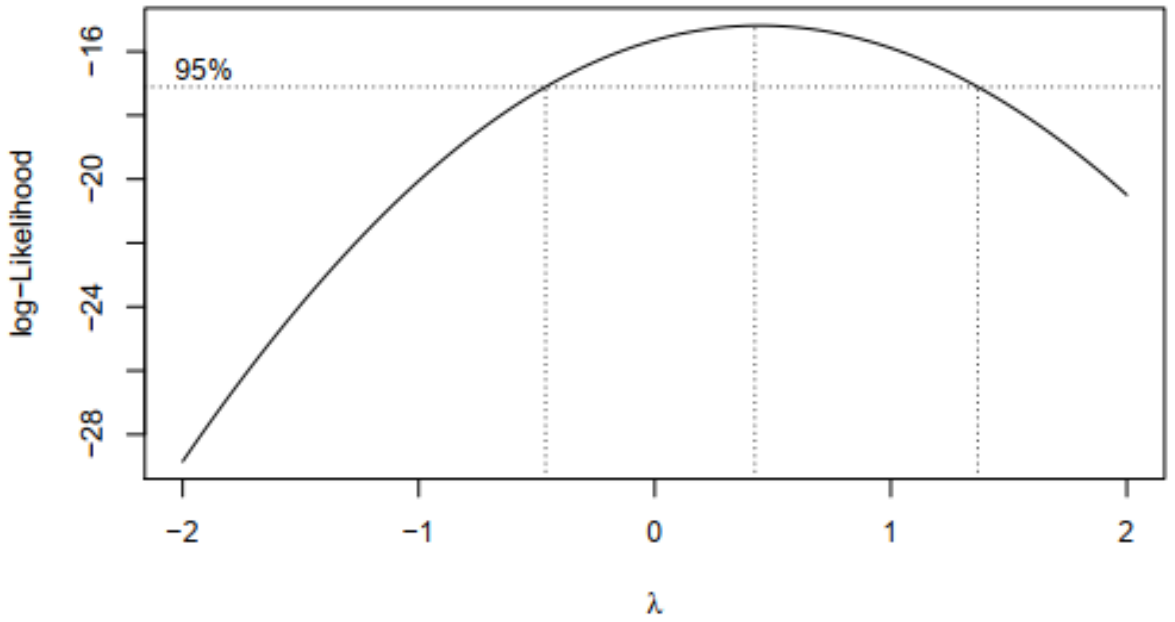
$$y_{ij}(\lambda) = \begin{cases} \dfrac{y_{ij}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_{ij}) & \text{if } \lambda = 0 \end{cases}$$

y represents the original data
$y(\lambda)$ is the transformed data
λ is a parameter that determines the type of transformation to be applied

The idea is to find the value of **λ** that makes the transformed data come closest to a normal distribution

Design Characterization

Data Transformation : Box Cox



The value that maximizes the likelihood, i.e., results in the best approximation to a normal distribution, is used to transform the data

Design Characterization

Data table

Consider an experiment installed at the CRD with i treatments and j replicates

	Treatments			
Replicates	1	2	...	I
1	Y_{11}	Y_{21}	...	Y_{I1}
2	Y_{12}	Y_{22}	...	Y_{I2}
...	...	...	...	...
J	Y_{1J}	Y_{2J}	...	Y_{IJ}
Total	T_1	T_2	...	T_I

Design Characterization

Data table

Replicates	Treatments 1	2	...	I
1	Y_{11}	Y_{21}	...	Y_{I1}
2	Y_{12}	Y_{22}	...	Y_{I2}
...	...	...	...	...
J	Y_{1J}	Y_{2J}	...	Y_{IJ}
Total	T_1	T_2	...	T_I

- Number of experimental units: $\mathrm{N} = \mathrm{I} \times \mathrm{J}$
- Total for treatment i: $\mathrm{T_i} = \sum_{\mathrm{j}=1}^{\mathrm{j}} \mathrm{Y_{ij}} = \mathrm{Y_{i.}}$
- Mean for treatment i: $\hat{m}_i = \frac{T_i}{J}$
- General mean of the experiment: $\hat{m} = \frac{G}{IJ}$

Design Characterization

Analysis of Variance

- Allows you to decompose the total variation in the data into different variation components, a (effects of treatments and residuals)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Treatments	$t-1$	SS_{Trt}	$MS_{Trt} = \frac{SS_{Trt}}{t-1}$
Within treatments	$t(r-1)$	SS_{Within}	$MS_{Within} = \frac{SS_{Within}}{t(r-1)}$
Total	$tr-1$	SS_{Total}	

Remember the error sum of squares (SSE) from past classes

Design Characterization

Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square
Treatments	$t-1$	SS_{Trt}	MS_{Trt}	$\sigma^2 + r\sigma_t^2$
Within	$t(r-1)$	SS_{Within}	MS_{Within}	σ^2
Total	$tr-1$	SS_{Total}		

Variance Components

If we consider the treatment effects to be random, we can estimate the variance component using the method of moments, by setting the mean squares (MS) equal to the expected mean squares (EMS)

Design Characterization

Variance Components

- The residual variance is simply estimated by $\hat{\sigma}^2 = \text{MS}_{\text{Within}}$

- We can then estimate σ_t^2 by $\hat{\sigma}_t^2 = \frac{MS_{trat} - MS_{within}}{r}$

Design Characterization

Hypothesis Testing

- We want to test the null hypothesis of no treatment effects, i.e.,

$$H_0 = \tau_i = 0 \text{ for all } i$$

Design Characterization

Hypothesis Testing

Source of Variation	Degrees of Freedom	Mean Square	Expected Mean Square	F-Statistic
Treatments	$t-1$	MS_{Trt}	$\sigma^2 + r\frac{\sum_i \tau_i^2}{t-1}$	
Within	$t(r-1)$	$\text{MS}_{\text{Within}}$	σ^2	
Total	$tr-1$			

Design Characterization

Hypothesis Testing

Source of Variation	Degrees of Freedom	Mean Square	Expected Mean Square	F-Statistic
Treatments	$t-1$	MS_{Trt}	$\sigma^2 + r\frac{\sum_i \tau_i^2}{t-1}$	$F = \frac{\text{MS}_{\text{Trt}}}{\text{MS}_{\text{Within}}}$
Within	$t(r-1)$	$\text{MS}_{\text{Within}}$	σ^2	
Total	$tr-1$			

F-Statistic

The F-Statistic follows an F distribution with t - 1 and t(r - 1) degrees of freedom

Design Characterization

Multiple Comparions

- The null hypothesis ($H_0 = \tau_i = 0$ for all i) is tested against the alternative that $\tau_i \neq 0$ for at least one of the treatments

- If H_0 is rejected, we need to determine which of the treatments differ

- This can be accomplished via pairwise comparisons between all possible combinations of treatments

Design Characterization

Multiple Comparions

- We thus perform multiple comparisons between the treatment means

- Tukey, Bonferroni, Scheffé, Dunnett's Test

Let´s Practice 01!

- We will work with yield data of 17 rice genotypes, measured in three replications each

- This data set can be found in the R package ***agridat*** and is a subset of the data from Gomez & Gomez (1984) **Statistical Procedures for Agricultural Research**. Wiley-Interscience

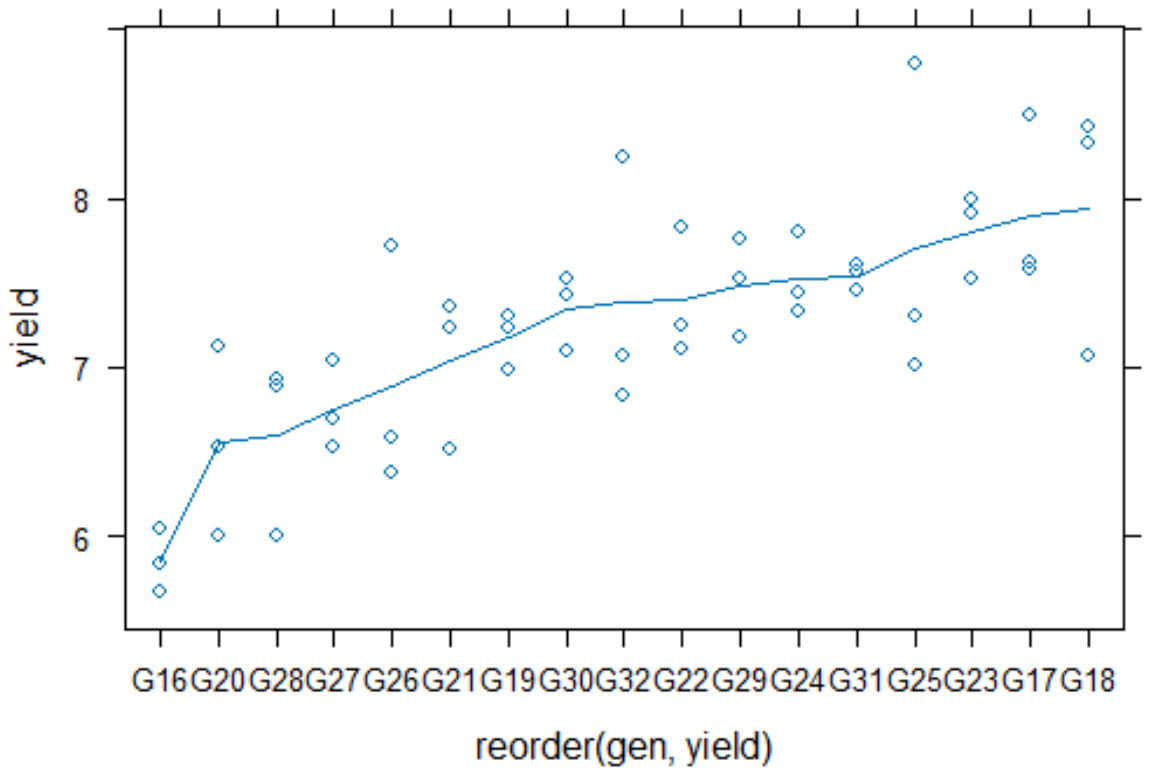
- Let y_{ij} denote the yield of the ith rice hybrid, measured from replicate j

Let´s Practice!

- Use the R function read.csv to import the data
- Fit the model with fixed effects
  #Use lm( ) function and summary( )
- Check if model assumptions are met
  #Check raw residuals vs fitted values
- Build the ANOVA table and test the null hypothesis of no difference between the rice genotypes
- Use multiple pairwise comparisons to assess which genotypes differ
- Fit the model with random effects

Let´s Practice!

```
# Graphic
library(lattice)
xyplot(yield ~ reorder(gen, yield), data = dados, type = c("p", "a"))
```



Let´s Practice!

- Fit the model with fixed effects

  #Use lm( ) function and summary( )

```
# Fit the model with fixed effects
fm <- lm(yield ~ gen, data = dados) # fixed effects
fm
summary(fm)
```

Let´s Practice!

- Fit the model with fixed effects

  #Use lm( ) function and summary( )

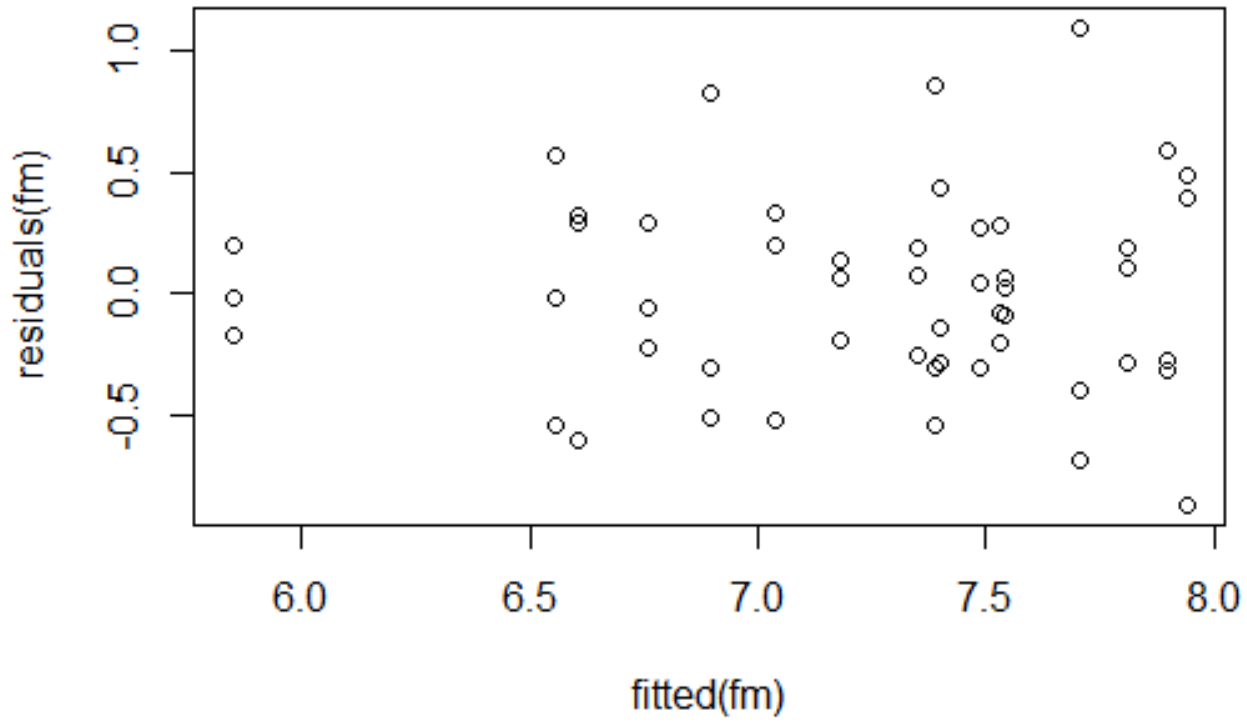
```
> summary(fm)

Call:
lm(formula = yield ~ gen, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87467 -0.28683 -0.01633  0.27500  1.09433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.8483     0.2867  20.402  < 2e-16 ***
genG17        2.0473     0.4054   5.050 1.48e-05 ***
genG18        2.0913     0.4054   5.159 1.07e-05 ***
genG19        1.3300     0.4054   3.281 0.002397 **
genG20        0.7083     0.4054   1.747 0.089615 .
genG21        1.1917     0.4054   2.940 0.005872 **
genG22        1.5513     0.4054   3.827 0.000530 ***
genG23        1.9607     0.4054   4.837 2.80e-05 ***
genG24        1.6793     0.4054   4.143 0.000215 ***
genG25        1.8563     0.4054   4.579 5.99e-05 ***
genG26        1.0453     0.4054   2.579 0.014423 *
genG27        0.9090     0.4054   2.242 0.031574 *
genG28        0.7557     0.4054   1.864 0.070968 .
genG29        1.6380     0.4054   4.041 0.000288 ***
genG30        1.5007     0.4054   3.702 0.000754 ***
genG31        1.6943     0.4054   4.180 0.000193 ***
genG32        1.5370     0.4054   3.791 0.000586 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 on 34 degrees of freedom
Multiple R-squared:  0.6395,    Adjusted R-squared:  0.4699
F-statistic:  3.77 on 16 and 34 DF,  p-value: 0.0005615
```

Let´s Practice!

- Check if model assumptions are met

  #Check raw residuals vs fitted values

```
# Raw residuals vs fitted values:
plot(residuals(fm) ~ fitted(fm))
```

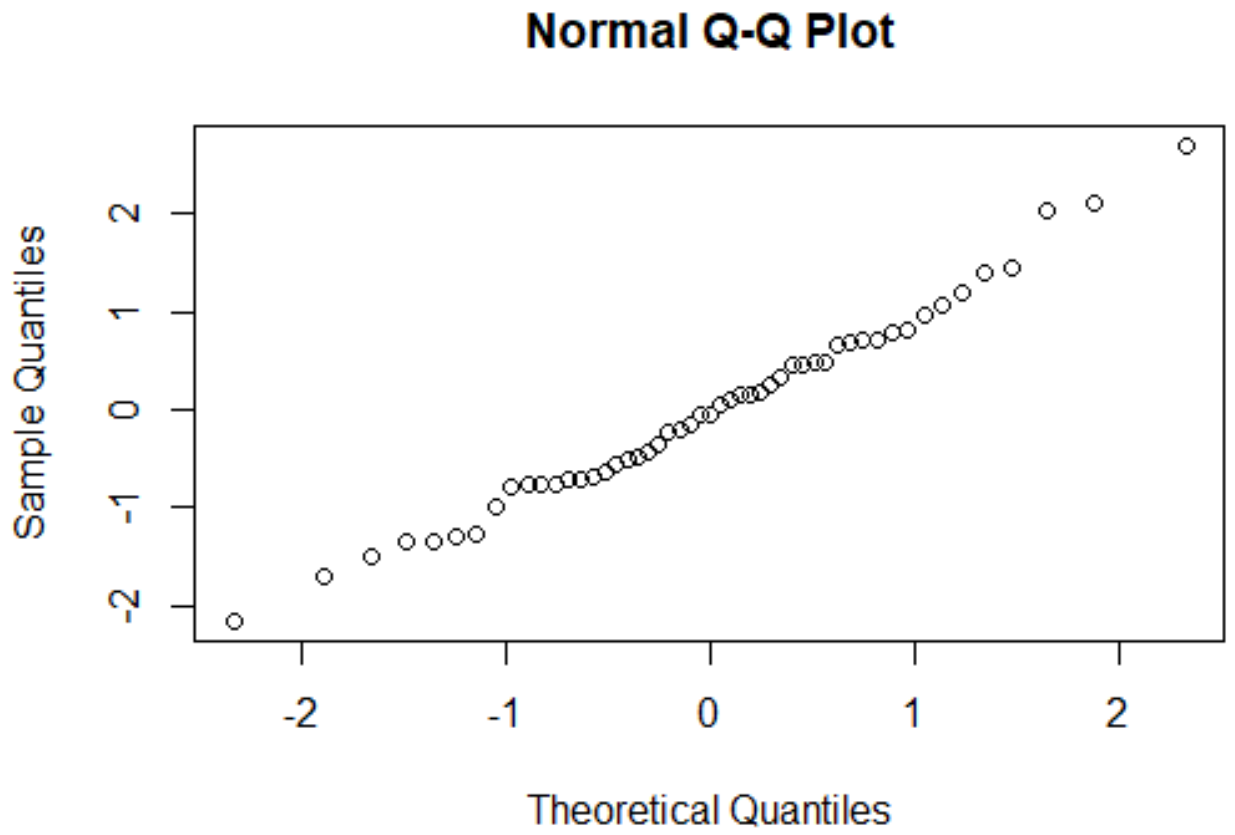


Let´s Practice!

- Check if model assumptions are met

#Check raw residuals vs fitted values

```
# Q-Q plot:
qqnorm(stdres(fm))
```



Let´s Practice!

- Anova

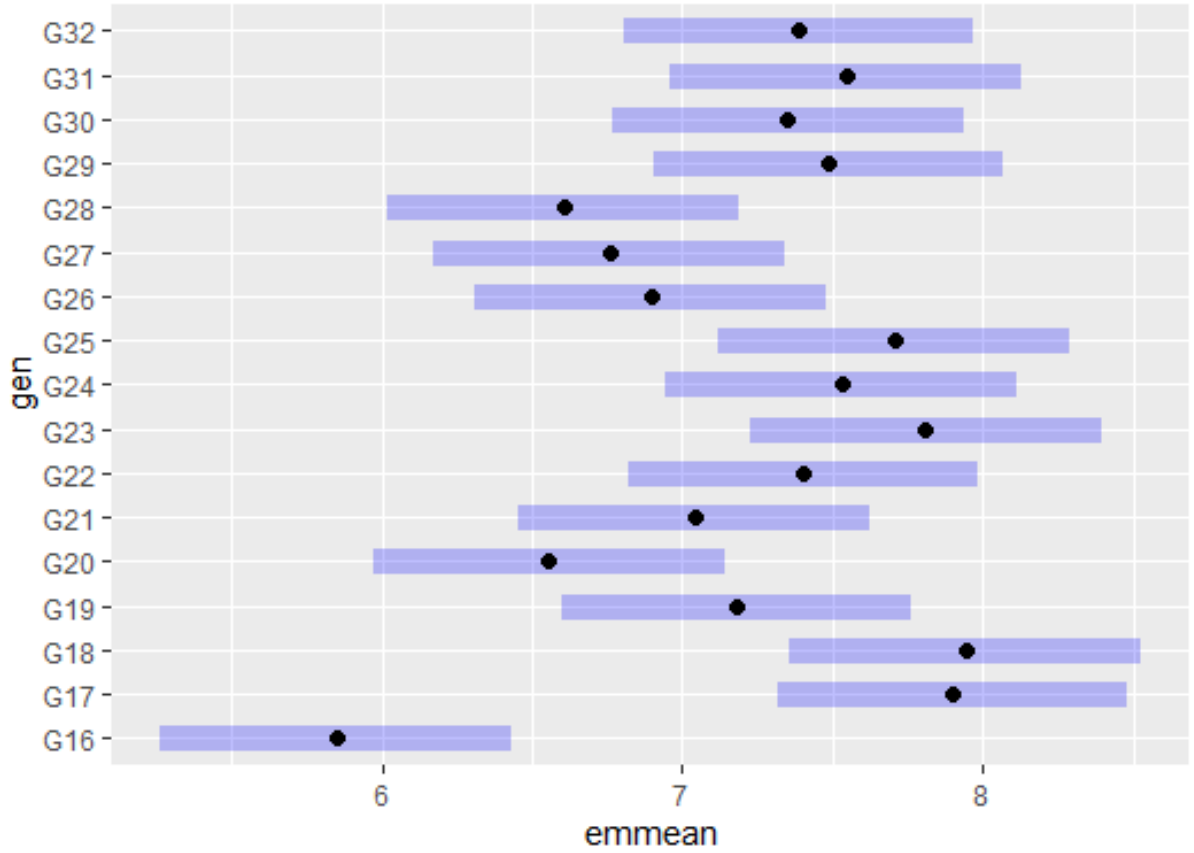
```
#Anova
anova(fm)
```

```
> anova(fm)
Analysis of Variance Table

Response: yield
          Df  Sum Sq Mean Sq F value    Pr(>F)
gen       16 14.8680 0.92925  3.7696 0.0005615 ***
Residuals 34  8.3814 0.24651
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let´s Practice!

- Calculate the estimated means for the different levels of the variable

```
# Obtaining estimated marginal means:
library(emmeans)
(fm_means <- emmeans(fm, "gen"))
plot(fm_means)
```



Let´s Practice!

- Pairwise comparisons with Tukey adjustment

```
#Pairwise comparisons
# Pairwise comparisons with Tukey adjustment:
pairs(fm_means)
```

```
> # Pairwise comparisons with Tukey adjustment:
> pairs(fm_means)
 contrast  estimate    SE df t.ratio p.value
 G16 - G17  -2.0473 0.405 34  -5.050  0.0015
 G16 - G18  -2.0913 0.405 34  -5.159  0.0011
 G16 - G19  -1.3300 0.405 34  -3.281  0.1372
 G16 - G20  -0.7083 0.405 34  -1.747  0.9312
 G16 - G21  -1.1917 0.405 34  -2.940  0.2657
 G16 - G22  -1.5513 0.405 34  -3.827  0.0395
 G16 - G23  -1.9607 0.405 34  -4.837  0.0027
 G16 - G24  -1.6793 0.405 34  -4.143  0.0178
 G16 - G25  -1.8563 0.405 34  -4.579  0.0056
 G16 - G26  -1.0453 0.405 34  -2.579  0.4693
 G16 - G27  -0.9090 0.405 34  -2.242  0.6904
 G16 - G28  -0.7557 0.405 34  -1.864  0.8913
 G16 - G29  -1.6380 0.405 34  -4.041  0.0232
 G16 - G30  -1.5007 0.405 34  -3.702  0.0535
 G16 - G31  -1.6943 0.405 34  -4.180  0.0162
 G16 - G32  -1.5370 0.405 34  -3.791  0.0431
 G17 - G18  -0.0440 0.405 34  -0.109  1.0000
 G17 - G19   0.7173 0.405 34   1.769  0.9245
 G17 - G20   1.3390 0.405 34   3.303  0.1310
 G17 - G21   0.8557 0.405 34   2.111  0.7706
 G17 - G22   0.4960 0.405 34   1.224  0.9974
 G17 - G23   0.0867 0.405 34   0.214  1.0000
 G17 - G24   0.3680 0.405 34   0.908  0.9999
 G17 - G25   0.1910 0.405 34   0.471  1.0000
```

Let´s Practice!

- Pairwise comparisons with Tukey adjustment

```
# Pairwise comparisons with Bonferroni correction:
pairs(fm_means, adjust = "bonferroni")
```

```
> pairs(fm_means, adjust = "bonferroni")
 contrast  estimate    SE df t.ratio p.value
 G16 - G17  -2.0473 0.405 34  -5.050  0.0020
 G16 - G18  -2.0913 0.405 34  -5.159  0.0015
 G16 - G19  -1.3300 0.405 34  -3.281  0.3259
 G16 - G20  -0.7083 0.405 34  -1.747  1.0000
 G16 - G21  -1.1917 0.405 34  -2.940  0.7986
 G16 - G22  -1.5513 0.405 34  -3.827  0.0721
 G16 - G23  -1.9607 0.405 34  -4.837  0.0038
 G16 - G24  -1.6793 0.405 34  -4.143  0.0292
 G16 - G25  -1.8563 0.405 34  -4.579  0.0082
 G16 - G26  -1.0453 0.405 34  -2.579  1.0000
 G16 - G27  -0.9090 0.405 34  -2.242  1.0000
 G16 - G28  -0.7557 0.405 34  -1.864  1.0000
 G16 - G29  -1.6380 0.405 34  -4.041  0.0392
 G16 - G30  -1.5007 0.405 34  -3.702  0.1026
 G16 - G31  -1.6943 0.405 34  -4.180  0.0262
 G16 - G32  -1.5370 0.405 34  -3.791  0.0797
 G17 - G18  -0.0440 0.405 34  -0.109  1.0000
 G17 - G19   0.7173 0.405 34   1.769  1.0000
 G17 - G20   1.3390 0.405 34   3.303  0.3071
 G17 - G21   0.8557 0.405 34   2.111  1.0000
 G17 - G22   0.4960 0.405 34   1.224  1.0000
```

Let´s Practice!

- Fit the model with random effects

```
library(nlme)
fme <- lme(yield ~ 1, random = list(gen = ~ 1), data = dados)
fme
```

```
Linear mixed-effects model fit by REML
  Data: dados
  Log-restricted-likelihood: -48.51987
  Fixed: yield ~ 1
(Intercept)
   7.230471

Random effects:
 Formula: ~1 | gen
        (Intercept)  Residual
StdDev:   0.4770553 0.4964977

Number of Observations: 51
Number of Groups: 17
```

References

- Chapter 7 - Analysis of Variance I: The One-Way Classification[1] (for a more classical view)

- Chapter 2 - Completely Randomized Designs[2]

1. Steel, R. G. & Torrie, J. H. Principles and Procedures of Statistics: A Biometrical Approach. 2nd Edition. (1980).

2. Casella, G. Statistical Design. (2008).