

Case Study : Analysis of a seriously unbalanced design

Prof L. Gentzbittel *Digital Agriculture Laboratory - Skoltech

Prof C. Ben †Digital Agriculture Laboratory - Skoltech

April, 5th 2021 - Skoltech

CASE STUDY PRESENTATION

Ten genotypes are assessed at two different locations.

For different reasons, different numbers of measurements or replicates are performed at each site

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. Installation 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

### 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())
# use of the constraint 'set-to-zero' for ANOVAs ## will see later in this script
options(contrasts=c('contr.treatment','contr.poly'))
#can also use 'contr.sum' for a 'sum-to-zero' constraint
```

LOADING REQUIRED METHODS FOR ANALYSIS

```
## Loading of the R packages needed for the analysis.
library(car)          # Levene's test and types of SS
library(agricolae)    # Newman-Keuls & Tukeys tests
library(ggplot2)
library(dplyr)
library(openxlsx)     # to load excel files
```

STARTING THE ANALYSIS

```
## loading data file
StemLength <- read.xlsx("08_UnbalancedDesign.xlsx", sheet = 1)
str(StemLength)
```

*l.gentzbittel@skoltech.ru

†c.ben@skoltech.ru

```
## 'data.frame': 12 obs. of 3 variables:
## $ StemLength: num 30 50 30 35 55 100 102 20 22 75 ...
## $ Genotype : chr "G1" "G1" "G1" "G1" ...
## $ Site : chr "S1" "S1" "S1" "S1" ...

## check the design
with(StemLength,
     table(Genotype, Site)) ## seriously unbalanced

##          Site
## Genotype S1 S2
##          G1 5 2
##          G2 2 3

## ANOVA tables of unbalanced design depends on order of factor fitted in the model !
## fit ANOVA2 mmodels, with interaction:
ana1 <- aov( StemLength ~ 1 + Genotype * Site, data = StemLength)
summary(ana1)

##          Df Sum Sq Mean Sq F value    Pr(>F)
## Genotype    1      3      3    0.041    0.845
## Site        1   9490   9490  125.699 3.59e-06 ***
## Genotype:Site 1      3      3    0.035    0.857
## Residuals    8     604      76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ana2 <- aov( StemLength ~ 1 + Site * Genotype, data = StemLength)
summary(ana2)

##          Df Sum Sq Mean Sq F value    Pr(>F)
## Site        1   8451   8451  111.935 5.56e-06 ***
## Genotype    1   1042   1042   13.805 0.00591 **
## Site:Genotype 1      3      3    0.035 0.85716
## Residuals    8     604      75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## clear case !

#####
## Taking unbalanceness into account using different types of sum-of-squares
#####

# create the 'most complete' model
model <- lm( StemLength ~ 1 + Site * Genotype, data = StemLength )

anova(model) ## type I analysis, the default method for R ## OK

## Analysis of Variance Table
##
## Response: StemLength
##          Df Sum Sq Mean Sq F value    Pr(>F)
## Site        1 8451.1  8451.1  111.9349 5.564e-06 ***
## Genotype    1 1042.3  1042.3   13.8054 0.005909 **
## Site:Genotype 1    2.6     2.6   0.0346 0.857164
## Residuals    8  604.0    75.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova(model, type = "II") ## interesting

## Anova Table (Type II tests)
```

```
##
## Response: StemLength
##           Sum Sq Df  F value    Pr(>F)
## Site      9490.3  1 125.6994 3.592e-06 ***
## Genotype   1042.3  1  13.8054 0.005909 **
## Site:Genotype  2.6  1   0.0346 0.857164
## Residuals    604.0  8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model, type = "III") ## usual in agronomy
```

```
## Anova Table (Type III tests)
##
## Response: StemLength
##           Sum Sq Df  F value    Pr(>F)
## (Intercept) 8000.0  1 105.9603 6.836e-06 ***
## Site      5315.7  1  70.4068 3.093e-05 ***
## Genotype    515.7  1   6.8307 0.03096 *
## Site:Genotype  2.6  1   0.0346 0.85716
## Residuals    604.0  8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```