Design of experiments

Mid-semester practical test

Pranav Gopalkrishna, 1940223

2021-10-26

# Data set

This data set will be used for all the questions. It contains information about crop yields grown in the same time period, differing in

* Plots
* Fertilizer applied
* Soil density
* Irrigation

We will be primarily focused on the effect of fertilizer types on crop yield (with soil density as the second factor, if needed).

setwd("~/Documents/Study/computerScience/programming/r/data/")  
myData = read.csv("cropYield.csv")  
head(myData)

## X yield block irrigation density fertilizer  
## 1 1 90 A control low N  
## 2 2 95 A control low P  
## 3 3 107 A control low NP  
## 4 4 92 A control medium N  
## 5 5 89 A control medium P  
## 6 6 92 A control medium NP

summary(myData)

## X yield block irrigation density fertilizer  
## Min. : 1.00 Min. : 60.00 A:18 control :36 high :24 N :24   
## 1st Qu.:18.75 1st Qu.: 86.00 B:18 irrigated:36 low :24 NP:24   
## Median :36.50 Median : 95.00 C:18 medium:24 P :24   
## Mean :36.50 Mean : 99.72 D:18   
## 3rd Qu.:54.25 3rd Qu.:114.00   
## Max. :72.00 Max. :136.00

# Question 1

## Aim

Choose the data, which has 2 independent (factors) variables each with at least 5 replications and apply suitable the technique(s) of design of experiment to check for the significance of different components.

## Variables

x1 = myData$fertilizer  
x2 = myData$density  
y = myData$yield  
data = data.frame(x1, x2, y)  
head(data)

## x1 x2 y  
## 1 N low 90  
## 2 P low 95  
## 3 NP low 107  
## 4 N medium 92  
## 5 P medium 89  
## 6 NP medium 92

## HYPOTHESES

### Difference in fertilizer means

H0: There is no difference in the means of fertilisers

H1: Means are not equal with respect to fertiliser

### Difference in soil density means

H0: There is no difference in the means of densities

H1: Means are not equal with respect to density

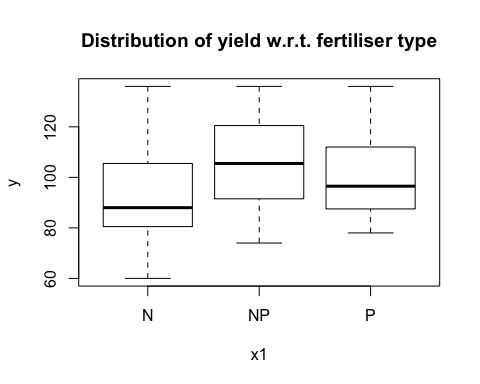
### Difference in interactions of fertilizer and soil density

H0: There is no difference in the means of the interactions between density and fertiliser

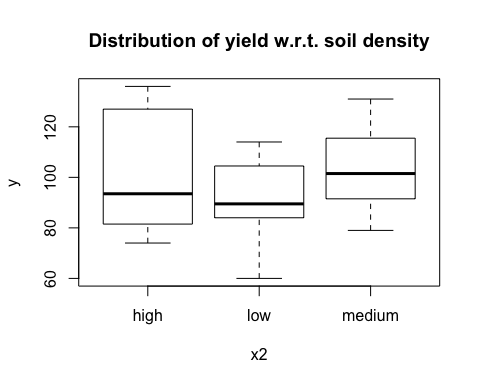
H1: There is difference in the means of the interactions between density and fertiliser

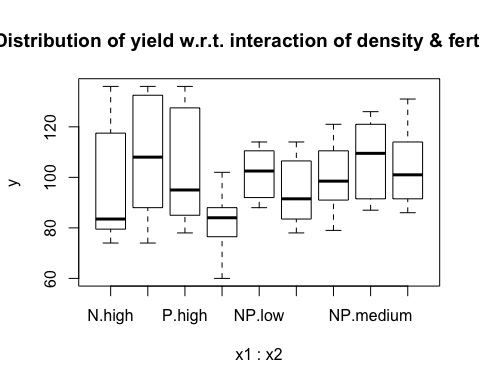
***Significance level = 0.05***

## VISUALISING DISTRIBUTION OF YIELD W.R.T. FACTORS boxplot(y~x1, main = "Distribution of yield w.r.t. fertiliser type")



boxplot(y~x2, main = "Distribution of yield w.r.t. soil density")



boxplot(y~x1:x2, main = "Distribution of yield w.r.t. interaction of density & fertiliser")  


As we can see in the above plots, means of both factors' levels as well as their variances differ. Their interactions also seem to be having some effect on the mean yields and the variation of yields. However, this difference may not be significant, which is what we aim to check. Also, the presence of large variation in interaction based-yields and soil density-based yields indicates that they may not be significant factors.

## ANOVA

model = aov(y~x1+x2+x1:x2)  
# x1:x2 denotes all the interactions of the levels of x1 and x2.  
summary(model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 2 1977 988.7 3.159 0.0493 \*  
## x2 2 1758 879.2 2.809 0.0678 .  
## x1:x2 4 305 76.2 0.244 0.9125   
## Residuals 63 19716 312.9   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As we can see, only x1 i.e. fertilizer type has significantly different mean yields within its levels, given a 0.05 significance level. Hence,

**we reject H0 for fertilizer type alone**

will perform post-hoc analysis for fertilizer types, to see which levels' mean yields are significantly different.

## POST-HOC ANALYSIS

library(multcompView)  
model = aov(y~x1)  
t = TukeyHSD(model)  
t$x[,'p adj']

## NP-N P-N P-NP   
## 0.04018125 0.29966031 0.58490135

As we can see, only the fertilizer types **NP and P** may be said to produce significantly different mean yields.

# Question 2

## Aim

State the assumptions of ANOVA using suitable dataset.

## Variables

x = myData$fertilizer  
y = myData$yield  
data = data.frame(x, y)  
head(data)

## x y  
## 1 N 90  
## 2 P 95  
## 3 NP 107  
## 4 N 92  
## 5 P 89  
## 6 NP 92

## ASSUMPTION 1: Response if normally distributed in the population

### Hypotheses H0: Sample of responses is drawn from a normal population H1: Sample of responses is not drawn from a normal population

If p-value is greater than the level of significance, we accept the null hypothesis. For our purposes, let’s fix the level of significance at 0.05 or 5% (meaning that if the sample values are from the 5% lowest probability region of the estimated normal distribution i.e. the region under the distribution curve with 5% of the values with the lowest probabilities of occuring, then the sample values are considered to be out of the normal distribution i.e. not normally distributed).

### **Test**

shapiro.test(y)

##   
## Shapiro-Wilk normality test  
##   
## data: y  
## W = 0.9587, p-value = 0.01866

# p-value = 0.3311 > 0.05,

Hence, we may conclude that the sample is taken from a normally distributed population (normally distributed with respect to yield).

## ASSUMPTION 2:Variances of responses of each treatment level are equal in the population

(This test must be used when we may conclude that the population is normally distributed (with respect to the given dependent variable)

### Hypotheses

H0: Variances within each group is equal to the others  
H1: Variances within each group differ for at least two groups

In our case, a group contains the outcomes of a particular fertilizer.

### **Test**

bartlett.test(y ~ x, myData)

##   
## Bartlett test of homogeneity of variances  
##   
## data: y by x  
## Bartlett's K-squared = 0.038857, df = 2, p-value = 0.9808

p-value = 0.6517 > 0.05. Hence, we may conclude there is close to equal variation within each group, when measured for the population. Of course, we cannot measure for the population (yet), so this is an estimate, given a 5% significance level. For our case, this means the variations in yield for each type of fertilizer’s application are estimated to be close to equal in the population.

### ASSUMPTION 3: Experimental error is normally distributed

Experimental error in Statistics is the difference between the (generally estimated) true value and the measured value of a characteristic.

In our case, the experimental error would be the difference between the measured recovery time for a drug type and the estimated true or mean value of the recovery time for the drug type. It is akin to variation of outcomes for the same treatment, or in this case, for the same variety.

To test if the experimental error is normally distributed, we will first obtain a set of residuals (i.e. differences between measured and estimated true values). Then, we will run them through the Shapiro-Wilk test.

### Test

e = residuals(aov(y ~ x, data = myData))  
# "residuals" is a generic function which extracts model residuals from objects returned by modeling functions.  
# Note that .anova is an attribute of a list such as circumferences, and .residual is an attribute of .anova.  
# We will run the data set "circumferences.anova.residual" through the Shapiro-Wilk test.  
shapiro.test(e)

##   
## Shapiro-Wilk normality test  
##   
## data: e  
## W = 0.96047, p-value = 0.02349

p-value = 0.1044 > 0.05. Hence, we may conclude that the experimental errors follow a normal distribution.

# Question 3

## Variables

## y = myData$yield x = myData$fertilizer # Treatment x\_levels = length(unique(x)) total\_observations = length(x) FUNCTIONS USED

# Correction factor  
cf = sum(y)^2 / length(y)  
# Treament sum of squares  
lms = function(level, regressor, response)  
# (Level mean square)  
{  
 sum = 0  
 n = 0  
 for(i in c(1:(length(regressor))))  
 {  
 if(regressor[i] == level)  
 {  
 sum = sum + response[i]  
 n = n + 1  
 }  
 }  
 return(sum^2 / n)  
}  
rss = function(regressor, response)  
# (Regression sum of squares)  
{  
 sum = 0  
 levels = unique(regressor)  
 for(level in levels)  
 {  
 sum = sum + lms(level, regressor, response)  
 }  
 return(sum - cf)  
}  
# Total sum of squares  
tss = 0  
for(i in y)  
{  
 tss = tss + i^2  
}  
tss = tss - cf

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## RBD

bf = myData$density # Blocking factor  
bf\_levels = length(unique(bf))

Soil density size is chosen as the blocking factor since it may be a significant factor in affecting yields (apart from fertilizer type).

data = data.frame(x, bf, y)  
head(data)

## x bf y  
## 1 N low 90  
## 2 P low 95  
## 3 NP low 107  
## 4 N medium 92  
## 5 P medium 89  
## 6 NP medium 92

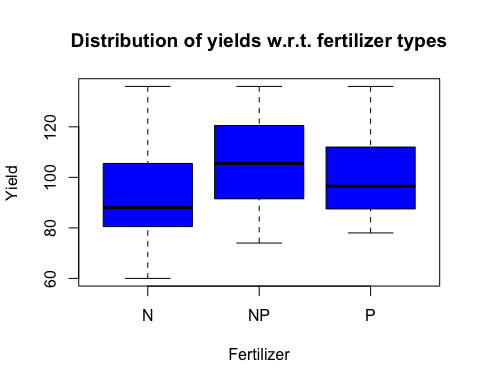
summary(data)

## x bf y   
## N :24 high :24 Min. : 60.00   
## NP:24 low :24 1st Qu.: 86.00   
## P :24 medium:24 Median : 95.00   
## Mean : 99.72   
## 3rd Qu.:114.00   
## Max. :136.00

### Visualizing effect of treatment and blocking factor

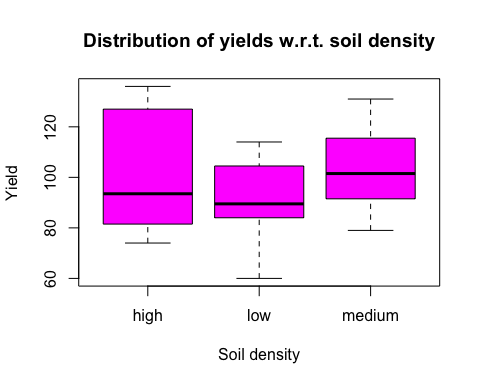
#### Visualizing distribution of yields w.r.t. treatment i.e. fertilizer types…

boxplot(y~x, main = "Distribution of yields w.r.t. fertilizer types", xlab = "Fertilizer", ylab = "Yield", col = "blue")



#### Visualising distribution of yields w.r.t. blocks i.e. soil density...

boxplot(y~bf, main = "Distribution of yields w.r.t. soil density", xlab = "Soil density", ylab = "Yield", col = "magenta")



### LINEAR REGRESSION MODEL

# Multiple linear regression model for two regressors...  
model = lm(y~x + bf)  
summary(model)

##   
## Call:  
## lm(formula = y ~ x + bf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.736 -14.153 -1.944 13.931 40.014   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.9861 4.5553 21.071 <2e-16 \*\*\*  
## xNP 12.7500 4.9901 2.555 0.0129 \*   
## xP 7.6667 4.9901 1.536 0.1292   
## bflow -10.0417 4.9901 -2.012 0.0482 \*   
## bfmedium 0.8333 4.9901 0.167 0.8679   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.29 on 67 degrees of freedom  
## Multiple R-squared: 0.1573, Adjusted R-squared: 0.1069   
## F-statistic: 3.126 on 4 and 67 DF, p-value: 0.02032

### # Here, we see error DF as 67. error\_df = 67 ERROR MEAN SQUARE

# Error sum of squares...  
ess = tss - rss(x, y) - rss(bf, y)  
# Error mean square...  
ems\_rbd = ess / error\_df

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

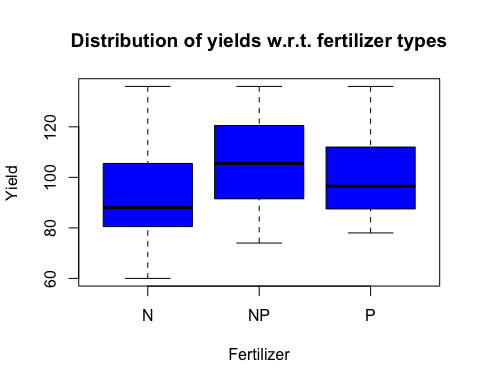
## CRD

myData = data.frame(x, y)  
summary(myData)

## x y   
## N :24 Min. : 60.00   
## NP:24 1st Qu.: 86.00   
## P :24 Median : 95.00   
## Mean : 99.72   
## 3rd Qu.:114.00   
## Max. :136.00

### Visualising distribution of yields w.r.t. treatment i.e. fertilizer types…

boxplot(y~x, main = "Distribution of yields w.r.t. fertilizer types", xlab = "Fertilizer", ylab = "Yield", col = "blue")



### LINEAR REGRESSION MODEL

# Multiple linear regression model for two regressors...  
model = lm(y~x)  
summary(model)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.917 -13.083 -2.917 13.354 43.083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92.917 3.627 25.622 <2e-16 \*\*\*  
## xNP 12.750 5.129 2.486 0.0153 \*   
## xP 7.667 5.129 1.495 0.1395   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.77 on 69 degrees of freedom  
## Multiple R-squared: 0.08324, Adjusted R-squared: 0.05667   
## F-statistic: 3.132 on 2 and 69 DF, p-value: 0.04987

### # Here, we see error DF as 69. error\_df = 69 ERROR MEAN SQUARE

# Error sum of squares...  
ess = tss - rss(x, y)  
# Error mean square...  
ems\_crd = ess / error\_df

# RELATIVE EFFICIENCY  
# EMS of CRD  
ems\_crd

## [1] 315.6377

# EMS of RBD  
ems\_rbd

## [1] 298.8155

# Relative efficiency of RBD w.r.t. CRD  
re = 100\* (1/ems\_rbd) / (1/ems\_crd)  
re

## [1] 105.6296

Hence, we see that in this case, RBD has proved to be around 5.6296% more efficient than CRD.