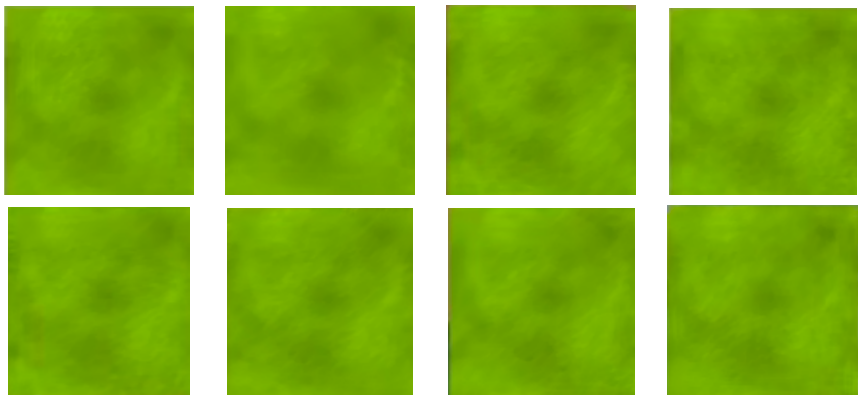
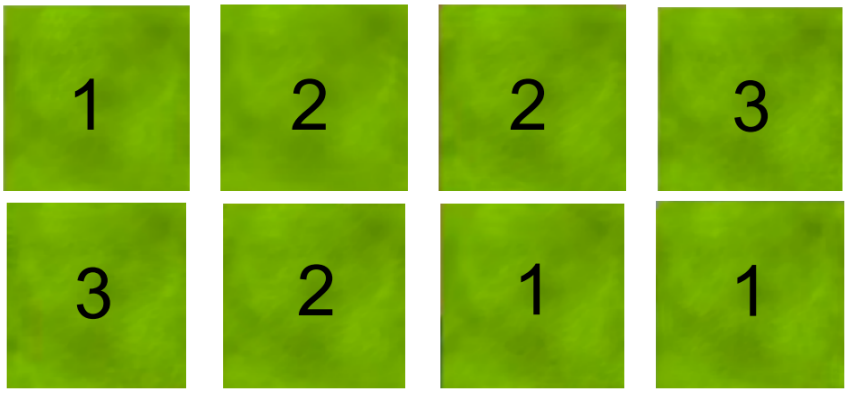
One Way ANOVA - An Agricultural Example

Ananda Biswas

Suppose we have 8 plots and 3 varieties of a crop. We want to study the yield of the different varieties of the crop.

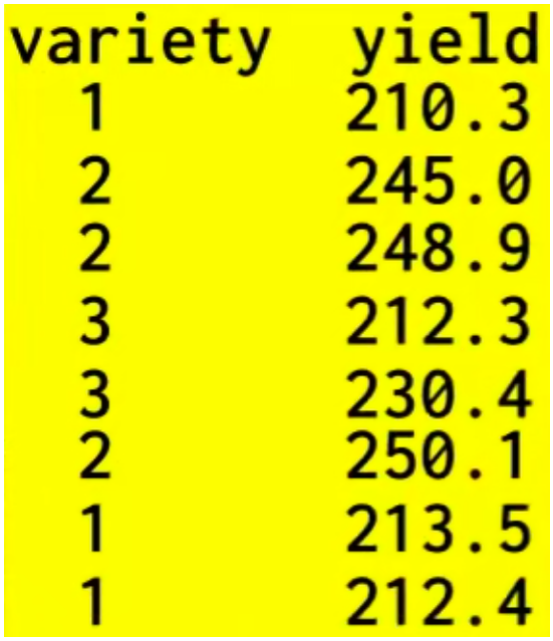


First we randomly select 3 plots for variety **1**, then another 3 plots for variety **2** and the rest 2 plots are used for variety **3**.



Here our blackbox diagram is

Our data set is



variety	yield
1	210.3
2	245.0
2	248.9
3	212.3
3	230.4
2	250.1
1	213.5
1	212.4

We can have a linear model as follows :

$$y_{ij} = \alpha_i + \epsilon_{ij}$$

where y_{ij} is the observed yield in the jth plot of the ith variety of crop ;

α_i is the true yield of the ith variety of crop;

ϵ_{ij} is the random error in the jth plot of the ith variety of crop.

Here, $i = 1, 2, 3$. For $i = 1, 2$, $j = 1, 2, 3$ and for $i = 3$, $j = 1, 2$.

• We may be interested in seeing which variety of the crop gives more yield. Then we can come up with another linear model :

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where μ is on an average, the true yield of the crop for all varieties;

α_i is the additional yield of the ith variety of the crop.

See that μ is unidentifiable from the above model.

We shall put a restriction on α_is that $\sum_{i=1}^{3} \alpha_i = 0$. This also boosts the interpretation of the model - μ is the average yield of the crop; α_i is additional yield of ith variety. So when we average the true yields of all the varieties, we get μ. $\left[\frac{1}{3}\sum_{i=1}^{3}(\mu + \alpha_i) = \mu\right]$

• Suppose variety **1** of the crop is widely used, 2 new varieties of the crop has been discovered and we want to study their effectiveness.

In such a scenario, yet another linear model can be :

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where μ is the benchmark yield of the crop when variety **1** is used;

$\alpha_1 = 0$, and α_2, α_3 are additional yields of the variety **2** and variety **3** respectively.

```
getwd()
```

```
## [1] "D:/Programming Languages/R/Linear Statistical Models - Arnab Chakraborty/004"
```

```
agri = read.csv('variety_and_yield_dataset.csv')

agri
```

```
##   variety yield
## 1       1 210.3
## 2       2 245.0
## 3       2 248.9
## 4       3 212.3
## 5       3 230.4
## 6       2 250.1
## 7       1 213.5
## 8       1 212.4
```

```
dim(agri)
```

```
## [1] 8 2
```

```
names(agri)
```

```
## [1] "variety" "yield"
```

The variable *variety* is a factor.

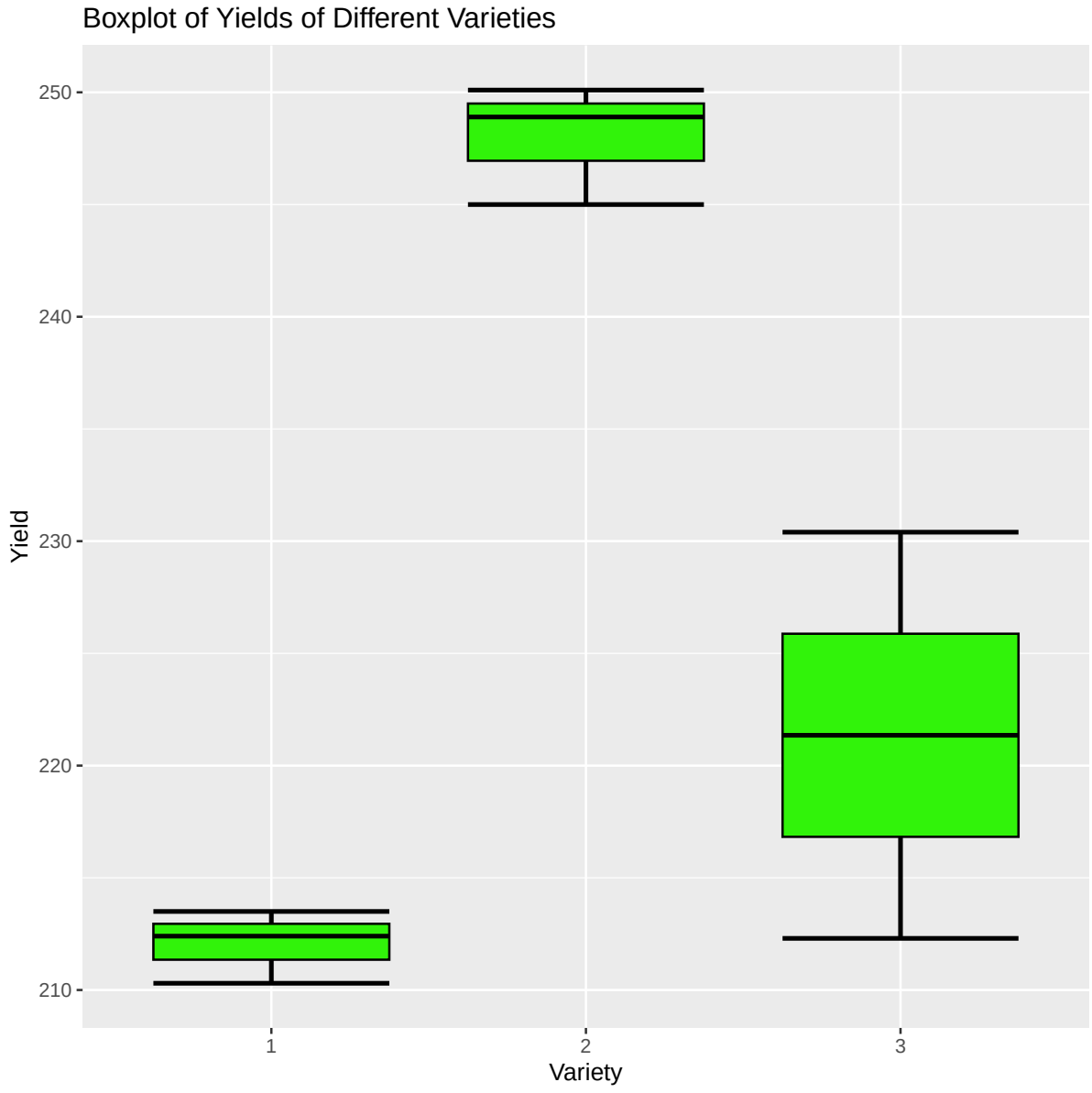
```
agri$variety = factor(agri$variety)
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
## Warning: package 'ggplot2' was built under R version 4.2.2
## Warning: package 'tibble' was built under R version 4.2.3
## Warning: package 'tidyr' was built under R version 4.2.3
## Warning: package 'readr' was built under R version 4.2.2
## Warning: package 'purrr' was built under R version 4.2.3
## Warning: package 'dplyr' was built under R version 4.2.3
## Warning: package 'stringr' was built under R version 4.2.3
## Warning: package 'forcats' was built under R version 4.2.2
## Warning: package 'lubridate' was built under R version 4.2.2
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0
--
```

```
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts()
--
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
agri %>%
  ggplot(aes(x = variety, y = yield)) +
  stat_boxplot(geom = "errorbar", linewidth = 1) +
  geom_boxplot(fill = "#31F30A", color = "black") +
  labs(x = "Variety", y = "Yield", title = "Boxplot of Yields of Different Varieties")
```



We shall fit the model :

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where μ is on an average, the true yield of the crop for all varieties;

α_i is the additional yield of the ith variety of the crop.

```
fit1 = lm(yield ~ variety, agri)
```

```
fit1

##
## Call:
## lm(formula = yield ~ variety, data = agri)
##
## Coefficients:
## (Intercept)     variety2     variety3
##     212.067       35.933        9.283
```

Surprisingly, we do not get any estimate for **variety1**. Let's see the model matrix.

```
model.matrix(fit1)

##   (Intercept) variety2 variety3
## 1           1        0        0
## 2           1        1        0
## 3           1        1        0
## 4           1        0        1
## 5           1        0        1
## 6           1        1        0
## 7           1        0        0
## 8           1        0        0
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$variety
## [1] "contr.treatment"
```

R never constructs a model matrix that is not full column rank. If the design matrix does not become full column rank, R will throw away columns to make it full column rank.

Here α_1 is forced to be 0.

Let us fit the linaer model :

$$y_{ij} = \alpha_i + \epsilon_{ij}$$

where y_{ij} is the observed yield in the jth plot of the ith variety of crop ;

α_i is the true yield of the ith variety of crop.

```
fit2 = lm(yield ~ variety - 1, agri)
```

```
fit2

##
## Call:
## lm(formula = yield ~ variety - 1, data = agri)
##
## Coefficients:
## variety1  variety2  variety3
##    212.1     248.0     221.3
```

```
model.matrix(fit2)

##   variety1 variety2 variety3
## 1        1        0        0
## 2        0        1        0
## 3        0        1        0
## 4        0        0        1
## 5        0        0        1
## 6        0        1        0
## 7        1        0        0
## 8        1        0        0
## attr(,"assign")
## [1] 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$variety
## [1] "contr.treatment"
```

See that previously α_1 was forced to 0. We can interpret that variety **1** is the benchmark variety and the intercept is benchmark yield.

```
fit1$coefficients

## (Intercept)    variety2    variety3
##  212.066667   35.933333    9.283333
```

```
fit2$coefficients

## variety1 variety2 variety3
## 212.0667 248.0000 221.3500
```

In **fit1**, R has reported the benchmark yield and the *additional yields from variety **1*** of variety **2** and **3**.

In **fit2**, R has reported the estimated yields of all the varieties.

We can easily observe that, the two reports are equivalent. If we add the intercept(the benchmark yield) and add additional yield of variety2, variety3 in **fit1**, we can get the estimated yield of variety **2** and **3** as given by fit2 respectively.

We can also estimate σ^2.

```
summary(fit1)

##
## Call:
## lm(formula = yield ~ variety, data = agri)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -1.7667 -3.0000  0.9000 -9.0500  9.0500  2.1000  1.4333  0.3333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  212.067      3.496  60.663  2.3e-08 ***
## variety2      35.933      4.944   7.268 0.000771 ***
## variety3       9.283      5.527   1.680 0.153887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.055 on 5 degrees of freedom
## Multiple R-squared:  0.9179,Adjusted R-squared:  0.8851
## F-statistic: 27.96 on 2 and 5 DF,  p-value: 0.00193
```

The **Residual standard error** is the estimated σ^2.