Task 3 STA4020

Nzambuli Daniel

2024-06-14

TASK: Optimizing Corn Yield with Simulated RCBD Data

The Agricultural Research Institute (ARI) is planning an experiment to investigate the impact of corn variety selection on yield, considering the influence of soil fertility. They need your help to simulate data and analyze it beforehand. The experiment involves three promising corn varieties (V1, V2, V3) planted within six distinct blocks representing different soil fertility levels. Each block will contain a randomized plot for each variety.

QUESTIONS:

a) Simulate the yield data with some random variation between varieties and blocks.

Maize production in Africa ranges from `0.9 tonnes` per hectare to `12 tonnes` per hectare

```
varieties = c("V1", "V2", "V3")
blocks = c("Block1", "Block2", "Block3", "Block4", "Block5", "Block6")
trt = length(varieties)
reps = length(blocks)

cat("There are", trt, "of maize on\n\t ", reps, "blocks")
```

```
## There are 3 of maize on
##    6 blocks
```

Generate the data

```
min_havrest = 0.9
max_havrest = 12

# generate random data
country_data = sample(seq(min_havrest, max_havrest), 9)
mn_yield = mean(country_data)
sd_yield = sd(country_data)

cat("A sample of values from the minimum yield of each country to the max yield in Africa:\n",
country_data, "\nwas used to generate the mean:\n", mn_yield, "\nand standard deviation:", sd_y
ield)
```

```
## A sample of values from the minimum yield of each country to the max yield in Africa:
##  7.9 8.9 11.9 0.9 1.9 2.9 4.9 10.9 6.9
## was used to generate the mean:
##  6.344444
## and standard deviation: 3.94053
```

Sample base yields

```
# seed twice remember math modelling when the data kept changing on each run
set.seed(123)
a = sample(seq(min_havrest, max_havrest), 9)
set.seed(123)
yield = rnorm(3, mean(a), sd(a)) # is also a random function
yield
```

```
## [1]  4.721113  5.939690 12.539469
```

Fill the yield table

```
default_yield = c(V1 = yield[1], V2 = yield[2], V3 = yield[3])
effect = rnorm(reps, mn_yield, sd_yield)
names(effect) = blocks
data = data.frame(Block = character(), Variety = character(), Yield = numeric())



# no need to seed so that each crop has a different yield
for (b in blocks) {
  for (v in varieties) {
    avg = default_yield[[v]] + effect[[b]]
    acc_yld = rnorm(1, avg, 5)
    data <- rbind(data, data.frame(Block = b, Variety = v, Yield = acc_yld))
  }
}

data
```

```
##     Block Variety     Yield
## 1  Block1      V1  9.115088
## 2  Block1      V2 18.682383
## 3  Block1      V3 20.960823
## 4  Block2      V1 13.578877
## 5  Block2      V2 13.347010
## 6  Block2      V3 16.614170
## 7  Block3      V1 26.758388
## 8  Block3      V2 21.531652
## 9  Block3      V3 15.809093
## 10 Block4      V1 16.388591
## 11 Block4      V2 11.736431
## 12 Block4      V3 15.361049
## 13 Block5      V1  4.990671
## 14 Block5      V2  2.169100
## 15 Block5      V3 10.254445
## 16 Block6      V1  5.233796
## 17 Block6      V2  1.144103
## 18 Block6      V3 20.366284
```

Pivot wider for better viewing

```
data_long =tidyr::pivot_wider( data = data,
  names_from = Block,
  values_from = Yield
)
data_long
```

```
## # A tibble: 3 × 7
##   Variety Block1 Block2 Block3 Block4 Block5 Block6
##   <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 V1        9.12   13.6   26.8   16.4   4.99   5.23
## 2 V2       18.7    13.3   21.5   11.7   2.17   1.14
## 3 V3       21.0    16.6   15.8   15.4  10.3   20.4
```

b) State the null and alternative hypotheses for the effect of variety (or blocking) on yield.

H_0 there is no statistically significant mean in the yield of the different maize varieties in the blocks

$$\mu_{v1} = \mu_{v2} = \mu_{v3}\ in\ all\ blocks$$

H_1 there is at least one mean yield that is statistically and significantly different from the other means in the blocks.

c) Perform an ANOVA to assess the effect of variety on yield while accounting for the block effect as a random factor.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
me errors
```

```
wrk_data = data %>% mutate(
  Variety = as.factor(Variety),
  Block = as.factor(Block)
)
wrk_data
```

```
##     Block Variety     Yield
## 1  Block1      V1  9.115088
## 2  Block1      V2 18.682383
## 3  Block1      V3 20.960823
## 4  Block2      V1 13.578877
## 5  Block2      V2 13.347010
## 6  Block2      V3 16.614170
## 7  Block3      V1 26.758388
## 8  Block3      V2 21.531652
## 9  Block3      V3 15.809093
## 10 Block4      V1 16.388591
## 11 Block4      V2 11.736431
## 12 Block4      V3 15.361049
## 13 Block5      V1  4.990671
## 14 Block5      V2  2.169100
## 15 Block5      V3 10.254445
## 16 Block6      V1  5.233796
## 17 Block6      V2  1.144103
## 18 Block6      V3 20.366284
```

```
anv = aov(Yield~Variety+Block, data = wrk_data)

summary(anv)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Variety      2   85.8   42.90   1.381 0.2953
## Block        5  455.1   91.02   2.931 0.0696 .
## Residuals   10  310.5   31.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Based on the p-value, can you conclude a significant difference in yield between the varieties at a 5% significance level? Explain your reasoning.

Check p-value

```
check_p_val = function(pval, sig_level, element){
  cat("for the element", element, "\n\n")
  if(pval < sig_level){
    cat("At a significance level of:\n", sig_level,"\nwe reject H_0.\n\tConclude:\n\t\tthere at least one mean yield of the maize varieties that is statistically and significantly different from the other yields")   }else{
      cat("At a significance level of:\n", sig_level,"\nwe fail to reject H_0.\n\tConclude:\n\t\tno statistically significant mean yield of the maize varieties")
    }
  }
```

Extract the p-values

```
pvals = summary(anv)[[1]][,"Pr(>F)"]
pvals
```

```
## [1] 0.29530157 0.06955086         NA
```

```
variety_pval = pvals[1]
block_pval = pvals[2]
```

Check for variety

```
check_p_val(variety_pval, 0.05, "variety")
```

```
## for the element variety
##
## At a significance level of:
##  0.05
## we fail to reject H_0.
##  Conclude:
##      no statistically significant mean yield of the maize varieties
```

Check for block

```
check_p_val(block_pval, 0.05, "block")
```

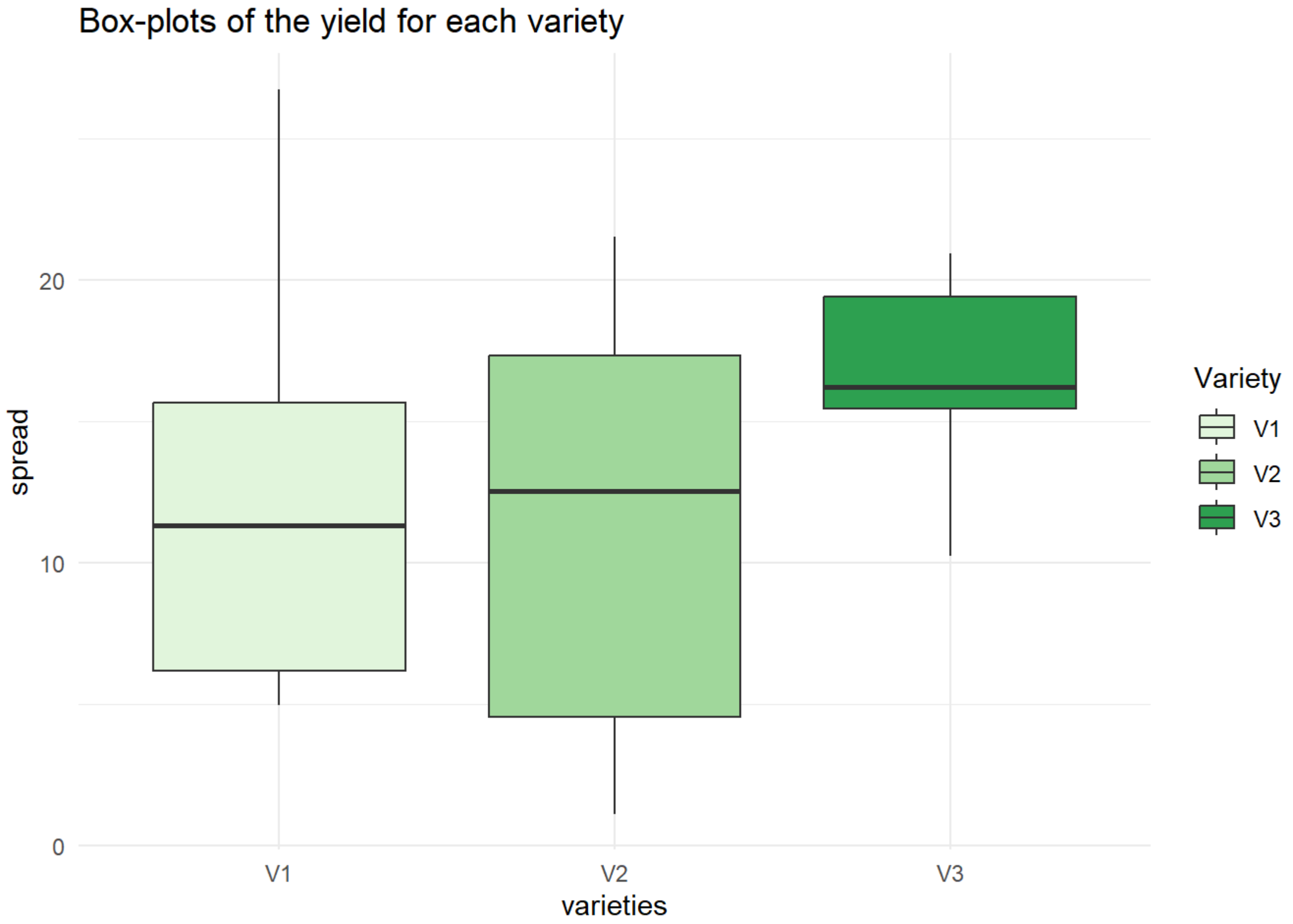
```
## for the element block
##
## At a significance level of:
##  0.05
## we fail to reject H_0.
##  Conclude:
##      no statistically significant mean yield of the maize varieties
```

e) Create box-plots to visually compare the distribution of yield across the different corn varieties.

```
library(ggplot2)

ggplot(wrk_data, aes(x = Variety, y = Yield, fill = Variety))+
  geom_boxplot()+
  labs(
    title = "Box-plots of the yield for each variety",
    x = "varieties",
    y = "spread"
  )+
  theme_minimal()+
  scale_fill_brewer(palette = "Set4")
```

```
## Warning: Unknown palette: "Set4"
```



f) Based on your analysis of the simulated data, summarize your key findings regarding the effect of corn variety on yield.

From the p-value there is a statistically significant difference in at least one of the means in the corn varieties.

I will perform an LSD post-hoc analysis to identify the mean and how they are different from each other

```
library(agricolae)
lsd_res = LSD.test(aov(Yield~Variety, data = wrk_data), "Variety", p.adj = "none")

lsd_res$groups
```

```
##        Yield groups
## V3 16.56098      a
## V1 12.67757      a
## V2 11.43511      a
```

Observations

Box-plots

1. For the **Median** variety `V1` and `V2` have values between 10 and 20 tonnes.
2. Variety `V1` has the largest inter quantile range (IQ)
3. The tails of `V2` and `V3` are the longest. For V2 the highest yields are larger than the majority of the rest of the data. This changes to the lowest yields for V3 which are lower than the other datsets.
4. V3 appears to be significantly different from V1 and V2 with higher yields

Post-hoc analysis

1. V1 and V3 are the most significantly different varieties with V2 not being easily distinguishable from the other two varieties

Conclusion

- **V1** shows a somewhat wider IQR than V2, suggesting more variability in yield within V1.
- There appear to be no outliers in any of the varieties, as all data points fall within the range of the whiskers.

g) How can the results from the actual experiment, if they confirm the simulated findings, be used to inform future corn breeding programs and selection strategies for maximizing yield?

- **V3** tends to have a higher median and mean yield compared to V1 and V2, which might suggest it is a more productive variety under the conditions tested.
- **V2** displays the least variability in yield with an average yield higher than the recorded average for Africa, which could be advantageous if consistent output is desired.
- **V1** shows more variability, which might suggest it is more sensitive to environmental or experimental conditions.

Based on this findings a cross-breed of V2 and V3 which takes advantage of V2 consistent production and V3 high yields may result in higher and consistent yield. This will meet the maximizing yield goal.

If maximum yield is the target V3 can be planted alone and if consistency is the target V2 will do well. The farmers should be careful not to overgrow just one crop and exhaust the production ability of the land.

V1 can also be bread with V2 to try and remove the variation in its breed production. It can also be breed with V3 to improve its maximum production. This can help reduce its sensitivity to the conditions in the blocks and maybe develop a better breed

References

Global Yield Gap Atlas. (2020, August 12). *ssa-maize - Global Yield Gap Atlas*. Www.yieldgap.org. https://www.yieldgap.org/ssa-maize

Knoema. (2020, January 1). *Maize Yield by country, 2020 - knoema.com*. Knoema. https://knoema.com/atlas/topics/Agriculture/Crops-Production-Yield/Maize-yield