Basic R training

Anil Adhikari and Dev Paudel

3/13/2022

Data input in R

There are three different ways of data input into R.

- Use RStudio features - Method 1.
  File > Open file > [select the file you want to input]
- Use RStudio features - Method 2.
  File > Import Dataset > From CSV / From Excel > specify input features
- Using R commands.
  Naviagte to the directory with data files and use appropriate commands.

```
# check your working directory
getwd()
```

```
## [1] "/Users/aniladhikari/Desktop/RLab"
```

```
# set your working directory to the folder with data files
setwd("/Users/aniladhikari/Desktop/RLab")

# View the file names
list.files()
```

```
## [1] "NAPA_R_tutorial.html" "NAPA_R_tutorial.pdf"  "NAPA_R_tutorial.Rmd"
## [4] "Wheat_91.txt"         "wheatc.csv"
```

```
# Use correct file name to run input commands
dat1 <- read.csv("Wheatc.csv")

# For text files
dat2 <- read.table("Wheat_91.txt", sep = "\t", header = TRUE)
```

Data exploratory analysis

1. Check if the data is in correct format. For example, check for typos, alphabets in place of numbers, missing row column names, missing data etc.
2. Check the distribution of data for outliers. There can be two types of outliers, true outliers, outliers due to inaccuracy in data collection, input, unit differences etc. We should try to see if there are any outliers of the second kind and rectify them.

```
# checking data for correct data type and format
head(dat1)
```

```
##   variety yield
```

```
## 1       A  22.2
## 2       D  23.9
## 3       B  24.1
## 4       D  21.7
## 5       C  25.9
## 6       C  18.4
```

```
# check data types in each columns
str(dat1)
```

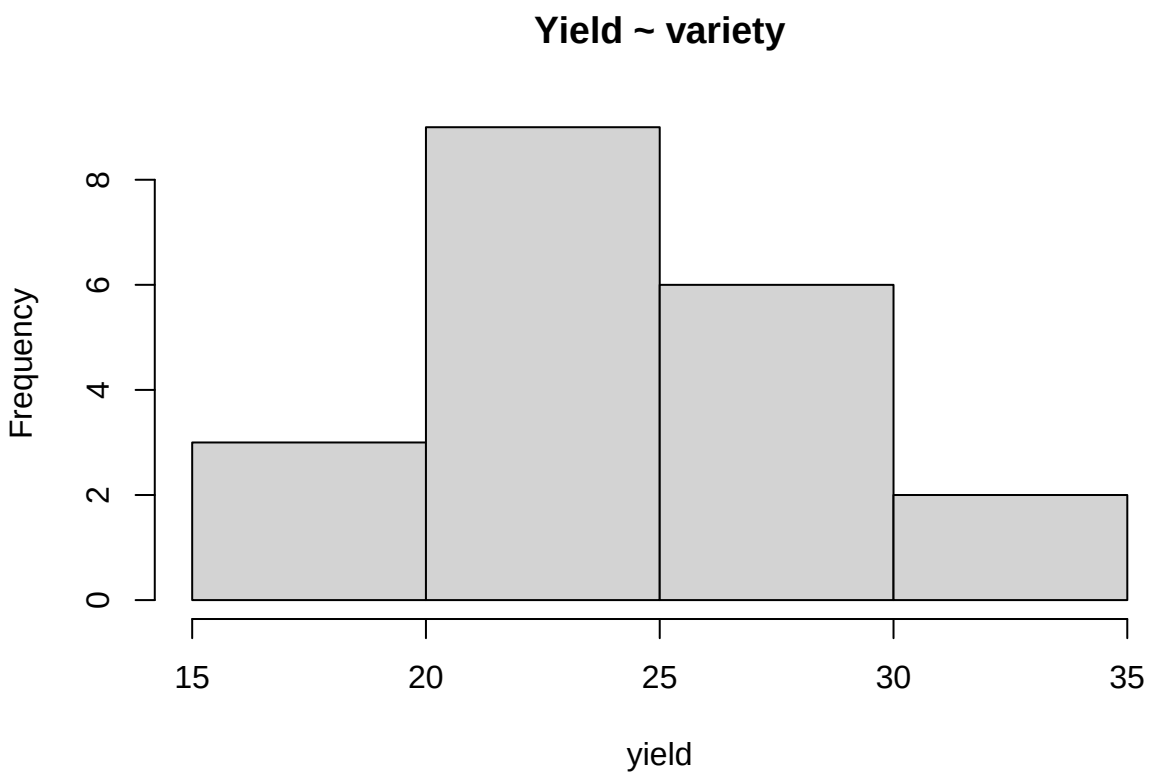
```
## 'data.frame':    20 obs. of  2 variables:
##  $ variety: chr  "A" "D" "B" "D" ...
##  $ yield  : num  22.2 23.9 24.1 21.7 25.9 18.4 24.8 28.2 17.3 26.4 ...
```

```
# check for missing data
table(is.na(dat1))
```

```
##
## FALSE
##    40
```

```
# checking for data accuracy

# basic histogram for distribution
hist(dat1$yield, main = "Yield ~ variety", xlab = "yield")
```



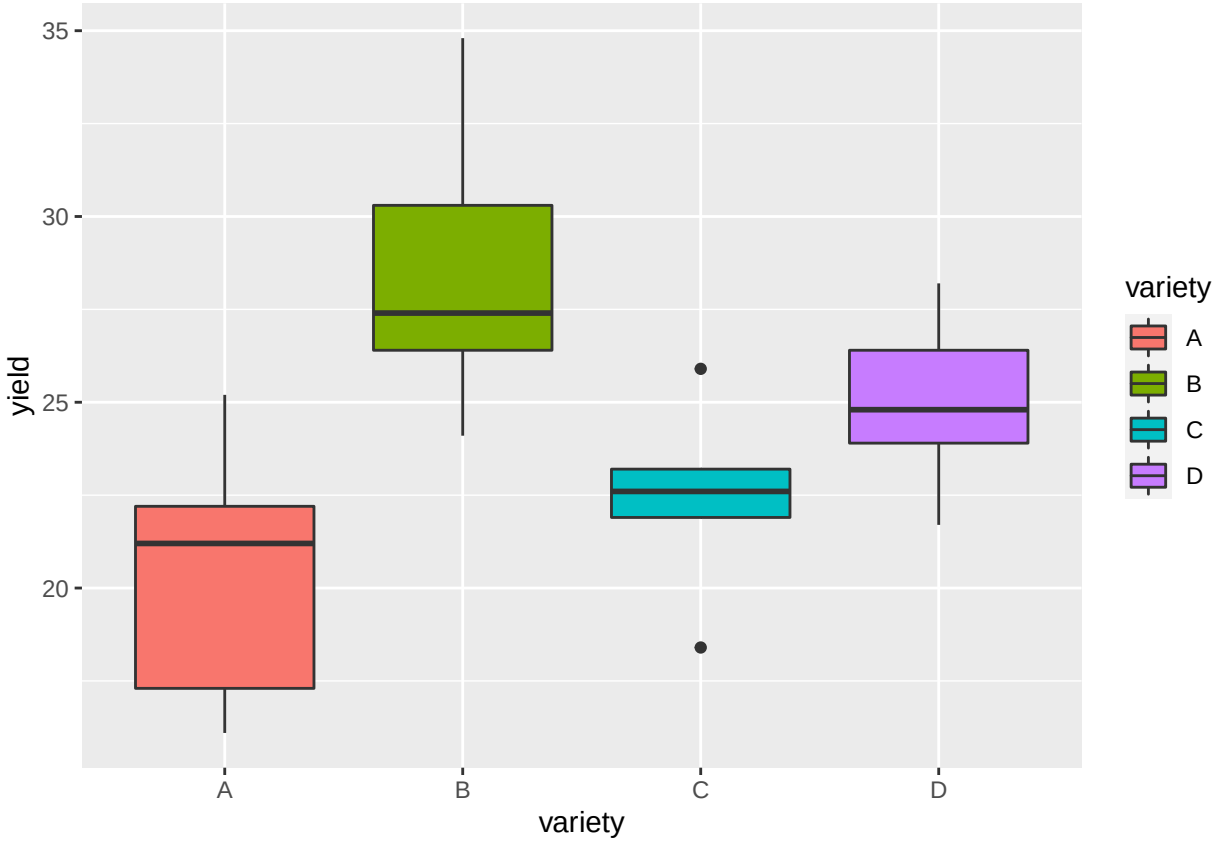

```
# create a boxplot
library(ggplot2)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'tibble'
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
```

```
## 'rlang::last_warnings' when loading 'pillar'
```

```
ggplot(dat1, aes(x=variety, y=yield, fill = variety)) +
  geom_boxplot()
```



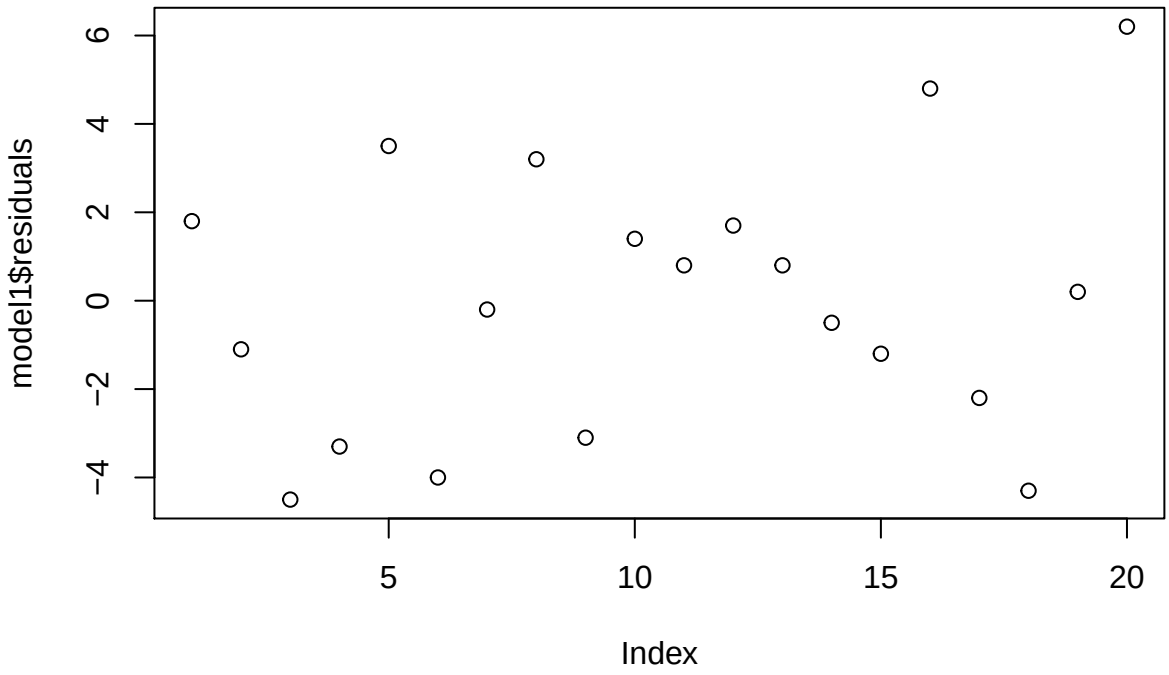
Analysis of variance

```
model1 <- lm(yield ~ variety, data = dat1)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## variety    3  188.2  62.733  5.6901 0.00756 **
## Residuals 16  176.4  11.025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

check model fit

```
plot(model1$residuals)
```



Post-hoc mean comparison

```
library(agricolae)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'
```

```
# Tukey's HSD
HSD <- HSD.test(model1,"variety", alpha = 0.05, group = TRUE)
HSD
```

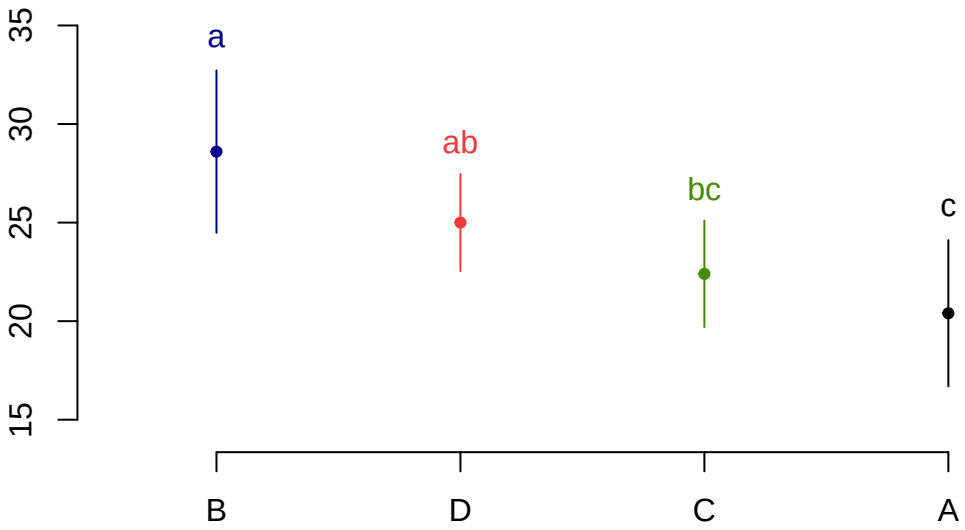
```
## $statistics
##   MSerror Df Mean       CV      MSD
##    11.025 16 24.1 13.77756 6.008142
##
## $parameters
##    test  name.t ntr StudentizedRange alpha
##   Tukey variety   4         4.046093  0.05
##
## $means
##   yield      std r  Min  Max  Q25  Q50  Q75
## A  20.4 3.708773 5 16.1 25.2 17.3 21.2 22.2
## B  28.6 4.118859 5 24.1 34.8 26.4 27.4 30.3
## C  22.4 2.700926 5 18.4 25.9 21.9 22.6 23.2
## D  25.0 2.466779 5 21.7 28.2 23.9 24.8 26.4
##
## $comparison
## NULL
##
## $groups
##   yield groups
## B  28.6      a
```

```
## D  25.0     ab
## C  22.4      b
## A  20.4      b
##
## attr(,"class")
## [1] "group"
```

```
# LSD test
out<-LSD.test(model1,"variety")
out
```

```
## $statistics
##   MSerror Df Mean       CV  t.value      LSD
##    11.025 16 24.1 13.77756 2.119905 4.451801
##
## $parameters
##         test p.ajusted  name.t ntr alpha
##   Fisher-LSD      none variety   4  0.05
##
## $means
##   yield      std r     LCL     UCL  Min  Max  Q25  Q50  Q75
## A  20.4 3.708773 5 17.2521 23.5479 16.1 25.2 17.3 21.2 22.2
## B  28.6 4.118859 5 25.4521 31.7479 24.1 34.8 26.4 27.4 30.3
## C  22.4 2.700926 5 19.2521 25.5479 18.4 25.9 21.9 22.6 23.2
## D  25.0 2.466779 5 21.8521 28.1479 21.7 28.2 23.9 24.8 26.4
##
## $comparison
## NULL
##
## $groups
##   yield groups
## B  28.6      a
## D  25.0     ab
## C  22.4     bc
## A  20.4      c
##
## attr(,"class")
## [1] "group"
```

```
# graphical representation
plot(out,variation="SD") # variation standard deviation
```



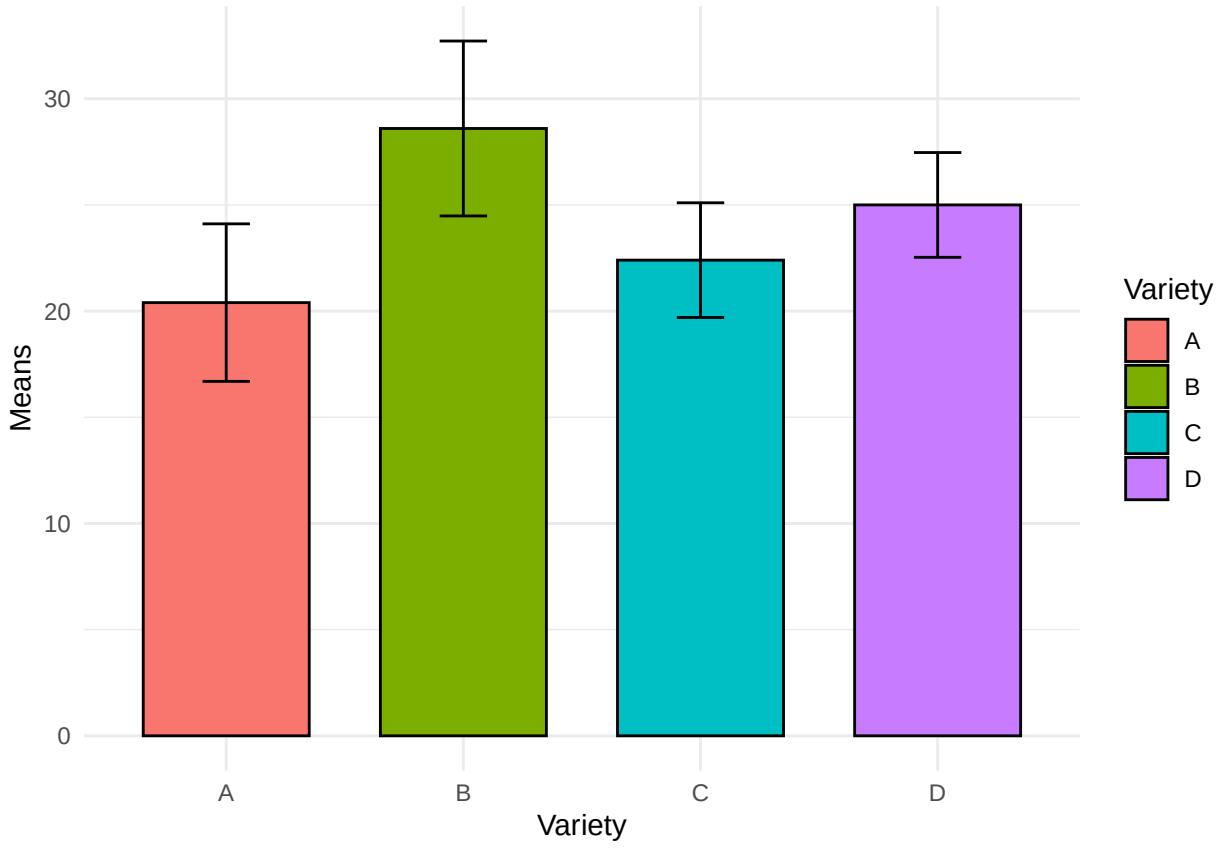

```
# bar graphs with SD
# extract means and SD
Means <-out$means$yield
SD <- out$means$std

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
tbl <- tibble(Variety = rownames(out$means), Means = Means, SD = SD)

ggplot(tbl, aes(x = Variety, y=Means, fill = Variety)) +
  geom_bar(stat = "identity", width = 0.7, color = "black") +
  geom_errorbar(aes(ymin = Means - SD, ymax = Means + SD), width = 0.2) +
  theme_minimal()
```


30
20
10
0
Means
A
B
C
D
Variety
Variety
A
B
C
D