COMP828 Summary Statistics and Plots

Nuttanan Wichitaksorn

Department of Mathematical Sciences
Auckland University of Technology

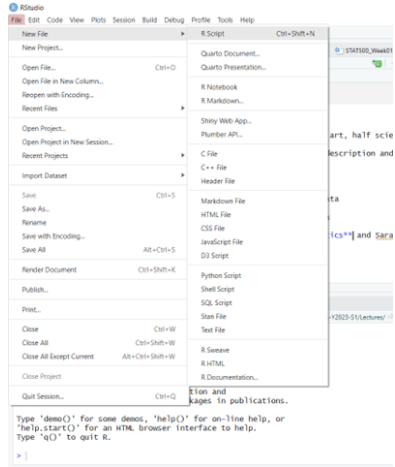
R Scripts[1] [2]

- Working with R and RStudio through R scripts is the most efficient way and you are encouraged to use them for coding/statistical programming.
- In addition, it is useful for reproducible work so that R users can learn from each other.
- R scripts allow you to write the complex command lines and save them for later use.
- You can also make some comments using # in the R scripts and that won't be run/executed.

[1] http://mercury.webster.edu/aleshunas/R_learning_infrastructure/R%20scripts.html
[2] https://bookdown.org/ndphillips/YaRrr/writing-r-scripts-in-an-editor.html

R Scripts (cont.)

To open a new R script, go to “File” > “New File” > “R Script” or press “Ctrl+Shift+N”.



R Scripts (cont.)

Sample Code:

```
# COMP824 Week 2 R Code

x <- rnorm(50) # Generate 50 normal random numbers
y <- rnorm(x)  # Generate another 50 normal random numbers
plot(x,y)      # Make a plot of x and y

mean(x)        # Calculate mean of x
w <- 1:20      # Make a sequence from 1 to 20
z <- 1 + sqrt(w)/2  # Calculate z
```

Summary Statistics

- Statistics is the art of learning from data. (Probably half art, half science)
- It is concerned with the collection of data, it subsequent description and analysis, which often leads to the drawing of conclusions.
- There are two main branches of statistics:

  Descriptive Statistics: describing and summarizing of data

  Inferential Statistics: drawing of conclusions from data
- In the first few weeks, we will mainly discuss **descriptive statistics**.

Summary Statistics (cont.)

- It is often useful to summarize a dataset using a single number that represents all the observations.
- There are many ways to do this but the three most common are *mean*, *median* and *mode*, which are the measures of central tendency.[3]
- The other group of descriptive statistics, which are frequently used, are *standard deviation*, *variance*, *range*, *quartiles*, and *interquartile range*.[4]
- Other useful statistics are *proportion*, *minimum*, *maximum*, etc.

[3] https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-central-tendency

[4] https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-spread

Summary Statistics (cont.)

Mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

Example: Motor Trend Car Road Test (an `R` built-in dataset) `help(mtcars)`

```
mean(mtcars$mpg)
# or
mpg <- mtcars$mpg
mean(mpg)
```

The average fuel consumption is ? miles per gallon.

The mean cannot be calculated for categorical data, as the values cannot be summed.

Summary Statistics (cont.)

Mode is the most commonly occurring value in a distribution.

Interestingly, there is no `R` function that can find the mode.

```
mode(mtcars$gear)
```

```
[1] "numeric"
```

Is this the mode we want? Find out what the `mode` function in `R` means.

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

However, some datasets may have no mode at all.

Summary Statistics (cont.)

Median is the middle value in distribution when the values are arranged in ascending or descending order.

```
median(mtcars$wt)
# see also
sort(mtcars$wt)
```

The median is less affected by outliers and skewed data than the mean and is usually the preferred measure of central tendency when the distribution is not symmetrical.

The median cannot be identified for categorical nominal data as it cannot be logically ordered.

Summary Statistics (cont.)

Variance and **standard deviation** are measures of the spread of the data around the mean. They summarise how close each observed data value is to the mean value.

The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero.

```
sd(mtcars$hp)
```

```
[1] 68.56287
```

```
var(mtcars$hp)
```

```
[1] 4700.867
```

What about `sd(mtcars$hp)^2`?

Summary Statistics (cont.)

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.

The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.

The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.

Summary Statistics (cont.)

Interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest.

The IQR is often seen as a better measure of spread than the range as it is not affected by outliers.

The best way to find quartiles and interquartile range in `R` is to use boxplot.

A useful function in `R` to find some frequently used statistics is `summary()`.

Plots

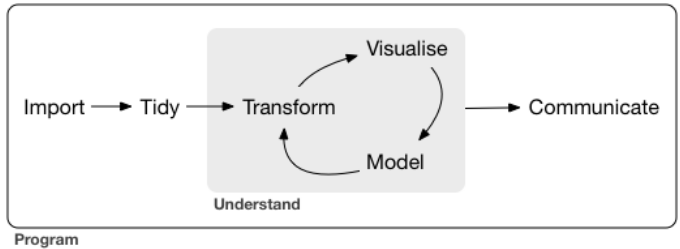


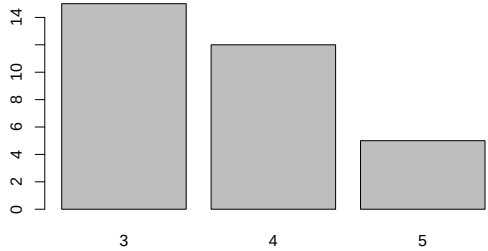
Figure 1: https://r4ds.had.co.nz/explore-intro.html

Plots can be used to visualize the data and/or communicate with the audience.

Plots (cont.)[5]

Barplots can only be done on qualitative variables. A barplot is a tool to visualize the distribution of a qualitative variable.

```
barplot(table(mtcars$gear))
```



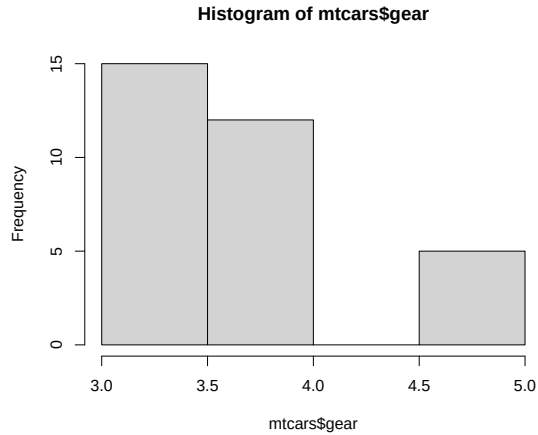
Notice the `table()` function.

[5]https://statsandr.com/blog/descriptive-statistics-in-r/

Plots (cont.)

Histogram is the better option to see the distribution.

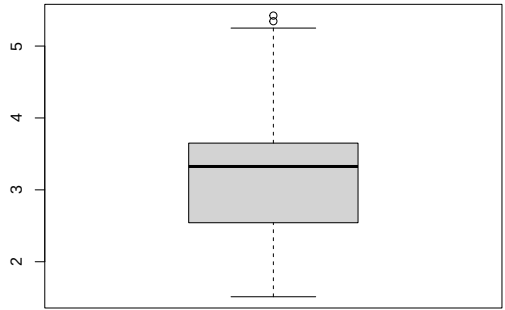
```
hist(mtcars$gear)
```



Plots (cont.)

Boxplot summarizes the data and provides some meaningful statistics.

```
boxplot(mtcars$wt)
```



Plots (cont.)

Scatter plot is useful to visualize the relationship of two quantitative variables.

```
plot(mtcars$mpg,mtcars$wt)
```