R Basics and Exploratory Statistics

Hakunawadi Pswarayi

University of Nottingham, MAPS Project

Course Outline

- Some Common errors with the R software
- Objects and object names
- Data types and structures in R
- Handling data in R
- Graphics in R
- Exploratory data analysis

The skills to take home

- Writing and debugging R scripts
- Generating and manipulating R data
- Making R graphics
- Carrying out exploratory data analysis

Common Errors With R and R Scripts

- Script missing things or has more things than necessary:
    - **a bracket, comma, full stop, etc,.**
- Package not installed or loaded:
    - Not installed: **”there is no package called ‘abc’”**
    - Not loaded: **”could not find function ”mutate”**
- **Working directory not set**:
    - Most common with beginners
    - Error message: **”cannot open the connection”**
- **To resolve errors, read and understand the error message**

The Mother of Errors!

- Error message difficult to understand and trace.
- Source of **great frustration!**
- Watch out for **differences** in variable names between scripts and data files
    - especially when a function or script is used for different variables

Objects and object names

- We manipulate **objects** and **object names** with R scripts/commands
- Objects are **data**, and object names, **variables**
- Objects are assigned names
- e.g., mass, 200:
  - **200** is the data object,
  - and **mass** the object name
- To manipulate the object (200), we manipulate the object name (mass).

Data Types

- Simplest data type is a **scalar**
  - which is a single value of some variable
  - e.g. the value 22, or the name ”Bert”
- **Numeric** data
- **Character** data
- **Logical** data

Data Structures

- Vectors
- Factors
- Matrices
- Data Frames

To handle data, first set a working directory

- **Essential** when running data and scripts from your PC drive:
- Three ways of setting a working directory
- 1. Using the "**setwd()**" function:
    - you need to know the working directory path
    - which I find complicated
- 2. Using the "**Session**" Tool on the Tool menu
    - Click "Session" - "Set working directory" - "Choose directory"
    - Navigate to the folder you want as working directory
    - Then click "**Open**" on the dialogue box

The file browser

- 3. Using the ”**File Browser**” tab
    - Navigate to the folder/file you want as working directory
    - Then ”click” on the ”**More**” tool (with cog/wheel next to it)
    - Then ”click” on ”**Set as Working Directory**”

Reading Simple Data Files into R

- Different data **formats** are read in by different **functions**
- Data formats:
  - **.txt files**
  - **.dat files**
  - **.csv files** (comma-separated values)
  - **.excel files**
- The different functions:
  - **read.delim()**, reading .txt files
  - **read.table()**, reading .dat files
  - **read.csv()**, reading .comma-separated values files
  - **readxl()**, reading excel files

Function and arguments

- The typical arguments:
  - File name
  - header
  - stringsAsFactors
- Example of function and arguments:
- read.delim("vari.txt", header = T, stringsAsFactors = T)

Writing data out of R

- Data can be written out in.dat .txt, and .csv formats
- Data is **directly** written to the working directory
- Sloppiness **deletes** files, and give **errors**
    - Data file with identical name and format with one being written out of R is automatically deleted/overwritten.
    - An **open** data file with an identical name and format to that being written will generate an error

Subsetting and Cleaning Data

- Subsetting is extracting subsets of R objects from:
  - **data frames**
  - **lists**
  - **matrices**
  - **vectors**
  - **factors**
- Subsetting **operators**:
  - They are: [, [[, $
- Cleaning data is removing **NA** values

Histograms

- Created with **single** variables.
  - to **visualize** data distributions.
- Using the function: **hist()**
- The arguments, e.g.,:
  - label: **xlab**
  - title: **main**
  - colour
  - and many others waiting to be **explored!**

Multivariate graphs

- Created with **multiple variables.**
  - to visualize how they **relate** to each other.
- Using the function: **plot()**
- The arguments, e.g.,:
  - axes and title labels: **xlab, ylab, main**
  - axes limits: **xlim, ylim**
  - symbols: **(pch)**, symbol fill: **(bg)**, symbol size: **(cex)**, symbol colour: **(red, blue, etc.)**
  - type: l, b, c, o,h, s and **n**
    - type **n**: plots **differentiating sources of a variable**
    - e.g., yield from different regions
  - **legend** and its **position**

Functions and arguments

- The **par()** function
- The arguments:
  - **mfrow**: number of rows and columns into which our device should be split
  - **mar**: to adjust the margins for each individual graph
- The **layout()** function
- For the **finer** control of the layout of our graphics
- The main argument:
  - **matrix** that specifies the locations for each graphic
  - e.g., rbind(1, 2:4)

Graph formats

- The formats: **pdf, png and jpeg**
  - The functions: **pdf(), png() and jpeg()**
- To print a graph, first **create** the device:
  - pdf(”name.pdf”)
  - png(”name.pgn”)
  - jpeg(”name.jpeg”)
- Then **close** the device after: **dev.off()**
  - otherwise all your graphics onwards will be pdf etc. documents

Exploratory Data Analysis

- The initial data investigations with:
- **summary statistics**
  - mean, median, skewness coefficients, number of outliers
- and **graphics:**
  - Histograms, Q-Q Plots, Boxplots
- To check for data conformity to **assumptions**
  - e.g., conformity to normal distribution
- To check for data anomalies (**e.g., outliers**)
  - that might affect statistics

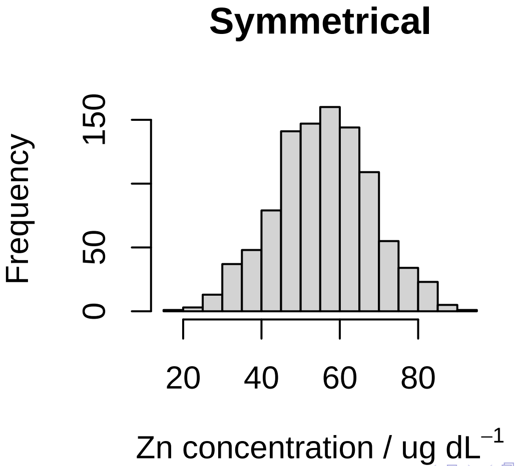
Distributions and the Assumptions of Statistical Inference

- Data distribution should match assumptions of statistical inference
- Otherwise **validity** of interpretation of results is affected
- We are mostly interested with assumptions of **normally** distributed data
- Therefore **skewed** distributions violate the normal distribution assumptions

Symmetrical Distribution

- One type of a normal distribution
- Most frequent values are around the mid-point of the range of values in the data set.
- Have similar sized right and left tails.
- The distribution of interest for our **assumptions** of statistical inference.
  - Because it gives **efficient/precise** estimates of the mean and variances

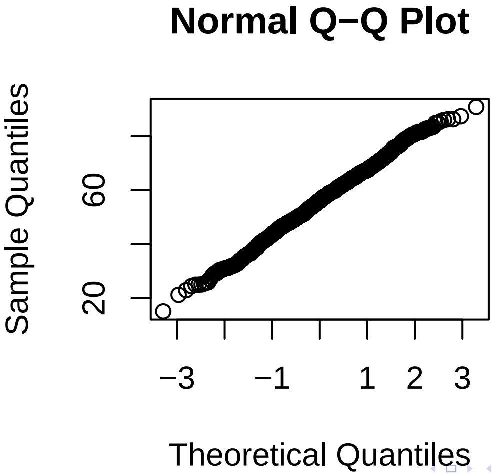
Histogram of a Symmetrical Distribution



Q-Q Plots

- A plot of the ordered values of a data set against the equivalent quantiles of a standard version of a specified distribution.
- Used to assess if the sample data came from some theoretical distribution (e.g., Normal distribution).
- The plot is expected to lie on **straight line** when assessing for normal distribution, and the sample data are from a normal random variable.

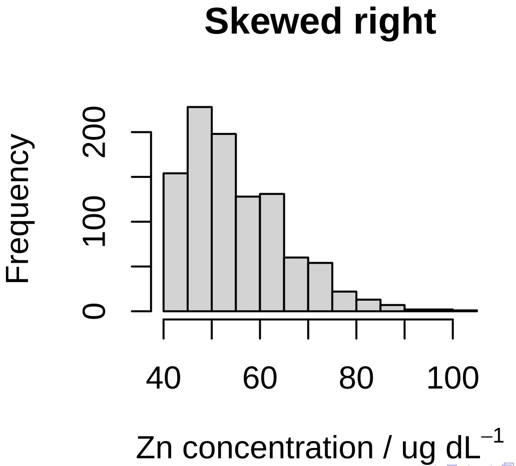
Q-Q Plot of a Symmetrical Distribution



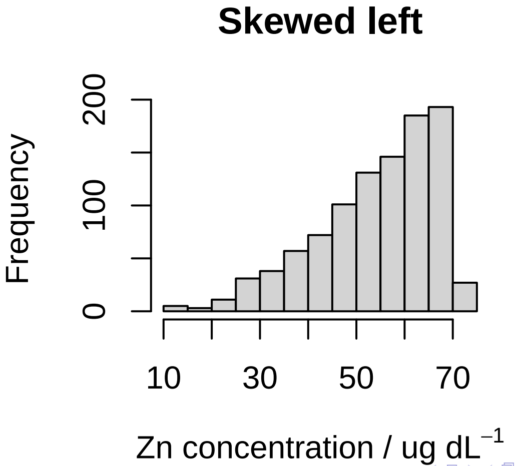
Skewed distributions

- Positive (Right) or negative (left) skewdness:
  - **Positive**: Most frequent values found below the mid-point, with an upper tail of large values:
  - **Negative**: Most frequent values found above the mid-point, with a lower tail of small values

Histogram of Right Skewed Distribution



Histogram of Left Skewed Distribution



Importance of Skewed Distributions

- Skewed distributions give **inefficient** estimates of the mean and variances
  - Inefficient estimates have **large** errors
  - Hence, **less** precise estimates
- Skewness affects **validity** of inferences
  - e.g., comparison of 2018 and 2019 annual household incomes in the **next** slide

Importance of Skewed Distributions

2018	2019
6000	5000
6500	6000
7000	6500
7500	7000
8000	7500
8500	8200
9000	23100
9500	30000
10100	40000
8011	**14811**

Importance of Skewed Distributions

- Although the mean **increased** in 2019, households became **poorer**
- The mean is generally **strongly** affected by a few wealthier households
- Hence, the mean can **mislead** when data are **skewed**.

The Pearson coefficient

- Coefficient depends on the **mean cube** of the difference for the data from the **mean**
- Hence, is very susceptible to **outliers**
- When coefficients are outside range [-1, 1]:
  - The validity of inference is affected
  - Positive coefficient means right-skewed distribution
  - Negative coefficient means left-skewed distribution
  - **Investigate outliers**, or **transform** distribution

The Octile coefficient

- Based on whether the octiles are symmetrical about the **median**
- Hence, **immune** to effects of outliers compared to Pearson's
- When coefficients are outside the range [-0.2, 0.2]:
  - The validity of inference is affected
  - Transform the distribution if possible

Location of the values: the mean and median

- Data for variables (e.g., serum zinc for WRA) **vary**
- This variation gives data distributions
- We **summarize** data distributions by the mean, median and mode

The Mean

- A **typical/representative** value of a data distribution
- The central tendency in a data distribution
- A simple arithmetic average of a sample:

$$\mu = \frac{1}{N}\Sigma_{i=1}^{N} x_i \tag{1}$$

- e.g., $(61 + 62 + 63 + 64 + 65) \div 5 = 63g/dL$
- The most widely used statistic
- However, a **non-robust** statistic
- Because it is easily influenced by single value changes in a data set
- E.g., $(61 + 62 + 63 + 64 + \mathbf{90}) \div 5 = 68$

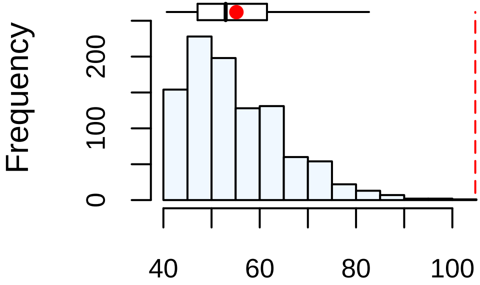
The Median

- A value at the **middle** of a numerically ordered data set
  - e.g, $61, 62, \mathbf{63}, 64, 65$
- Or the average value of **two middle** values in a numerically ordered data set
  - e.g., $61, \mathbf{62}, \mathbf{63}, 64 : (62 + 63)/2 = 62.5$
- Half data set values are less than the median, the other half larger
- A rough estimate of the centre of a distribution

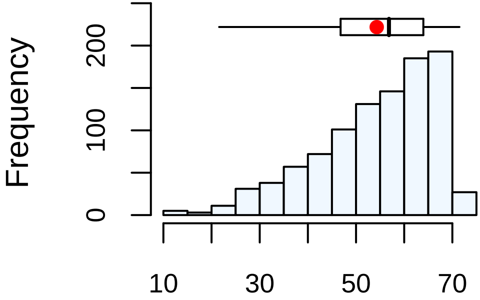
The Median

- **Robust** statistic
- Because it is not easily influenced by single value changes
- e.g., $(61, 62, \mathbf{63}, 64, 65)$ to this $(61, 62, \mathbf{63}, 64, \mathbf{90})$
- Although the last value changed from 65 to 90, the median **remained 63**

The mean is larger than median - right skewed



The mean is smaller than median - left skewed



Is About Outliers

- The **Untypical** values in datasets that **may or may not be erroneous**
- The **Extreme** observations lying an abnormal distance from other values
  - Very large values e.g., 21, 23, 20, 22, 25, **45**
  - Very small values e.g., **9**, 21, 23, 20, 22, 25, 23

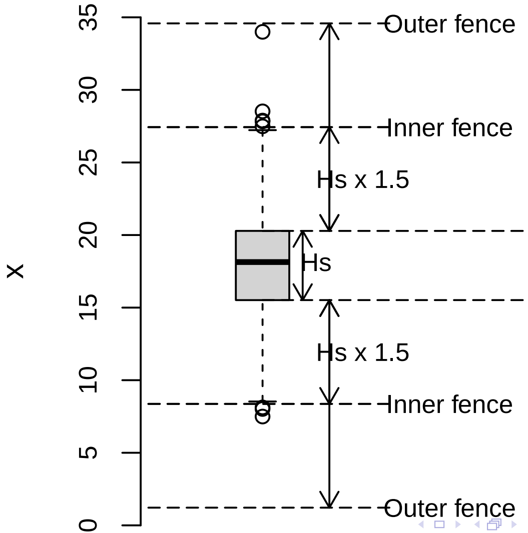
Importance Of Outliers

- They may induce skewness in data distributions
- The mean and standard deviation **highly sensitive** to outliers
  - e.g., $21, 23, 20, 22, 25, 23$ Mean: **22.3**, Stdev: **1.8**
  - e.g., $21, 23, 20, 22, 25,$ **45**,: Mean: **26**, Stdev: **9.5**.
  - Hence, can unduly influence statistics calculated.
  - Which leads to **incorrect inferences** about data

Tukey Criterion

- Outliers are values **beyond** 1.5 or 3 times the **interquartile range**
- The Interquartile Range/Hspread (HS):
  - A measure of the variability of a variable in a data set
  - Defined as the absolute difference between Quartile 3 (75th percentile) and Quartile 1 (25th percentile).
- For **MAPS**, outliers are values beyond 3 times the interquartile range

Boxplot, Interquartile Range, and Outliers



Types and Origin

- **Impossible** data: e.g., negative Zinc values due to:
  - Erroneous data entry
  - Erroneous sample analyses
  - **Unusual** data values, e.g., large values above outer fence due to:
    - Contamination
    - Soils with high nutrient content
    - True genetic values

Dealing With Outliers

- Impossible data values
  - **Edit** erroneous data entries
  - **Re-analyse** erroneously analysed samples if spares are available
  - Unsual data values
    - Use **robust** estimators
    - **Remove** outliers (some purposes)
    - **Retain** outliers (other purposes)

Based on Residuals, not Raw Data

- Transformations are carried out when skewness coefficients are **outside** the range
- Investigate skewness on distribution of **residual**, not raw data.
- If you base on raw data distributions:
    - you are on the highway to **wrong conclusions**
    - because complexities in raw data (**e.g., subpopulations**) may lead to skewed distributions.

Based on Residuals, not Raw Data

- Data complexities may be due to:
  - Differences in the **means** of subpopulations
  - e.g., rural and urban sub-populations
- Raw data: original observed values
- Residuals: the difference between the **observed** value and the **fitted** value (e.g., a subgroup mean) from some proposed conceptual model is for data.

Based on Residuals, not Raw Data

- Fit the proposed model first.
- Then investigate residuals for skewness
  - Investigate outliers, or transform distribution when the Pearson coefficient is outside the range $[-1, 1]$
  - Transform distribution when the octile coefficient is outside the range $[-0.2, 0.2]$

The Methods

- Two methods, Log and BoxCox.
- Method of choice depends on the **type** and skewness **severity**
- Or the outcomes of transformation
  - Better method gives the **smallest** coefficient of skewness
- Log method:
  - Cannot be used with **negative** types of skewness
  - May not be adequate with **severely** skewed distributions
- The BoxCox method
  - For severely skewed distributions
  - For **both** negative and positive skewed distributions