# Statistical Analyses
## Using R

**Date:** August 25 - 26, 2015
**Venue:** Computer Lab, School of Ind. Tech., USM

USAINS GROUP OF COMPANIES
Wholly owned by Universiti Sains Malaysia
(473883-H)

UNIVERSITI SAINS MALAYSIA

*Instructor:*
*Yusri Yusup, PhD*
*Environmental Technology*
*School of Industrial Technology*
*Universiti Sains Malaysia*
*Penang.*

# Course Schedule

## Day 1: Overview of R

| | |
|---|---|
| 8:30 A.M.: | Registration |
| 9:00 A.M.: | Introduction to R |
| 10:30 A.M.: | *Tea break* |
| 10:45 A.M.: | Installing and navigating R (hands-on) |
| 1:00 P.M.: | *Lunch* |
| 2.00 P.M.: | Data management (hands-on) |
| 3:00 P.M.: | *Tea break* |
| 3:15 P.M.: | Descriptive statistics (hands-on) |
| 5:00 P.M.: | End |

## Day 2: Statistical analyses using R

| | |
|---|---|
| 9:00 A.M.: | Hypothesis testing (hands-on) |
| 10:30 A.M.: | *Tea break* |
| 10:45 A.M.: | ANOVA (hands-on) |
| 1:00 P.M.: | *Lunch* |
| 2:00 P.M.: | Correlation and regression (hands-on) |
| 3:00 P.M.: | *Tea break* |
| 3:15 P.M.: | Non-parametric tests (hands-on) |
| 5:00 P.M.: | End |

# Statistical Analyses using R

Yusri Yusup

1

# Why use R?

- R is free and available on many operating systems (OSX, Windows, Linux).
- R has many statistical tools (4000+ packages)
- Statistical analysis using R is reproducible

2

# R compared to other commercial softwares

| | |
|---|---|
| **R**<br>**Price:** Free<br>**OS:** Available on all OS<br>**Interface:** Command-based<br>**Analyses:** reproducible<br>**Update:** User-dependent and frequent<br>**Customizable:** High | **Minitab (5 users)**<br>**Price:** RM6000 (RM3500 to update)<br>**OS:** Windows<br>**Interface:** Point-and-click<br>**Analyses:** reproducible<br>**Update:** Developer-dependent<br>**Customizable:** Low |
| **SPSS (standard)**<br>**Price:** RM23000 per year<br>**OS:** Windows, Mac OS, Linux<br>**Interface:** Point-and-click<br>**Analyses:** reproducible<br>**Update:** Developer-dependent<br>**Customizable:** Low | **SAS**<br>**Price:** RM36400 per year (commercial)<br>**OS:** Windows, Linux, Unix<br>**Interface:** Point-and-click and command-based<br>**Analyses:** reproducible<br>**Update:** Developer-dependent<br>**Customizable:** Low |

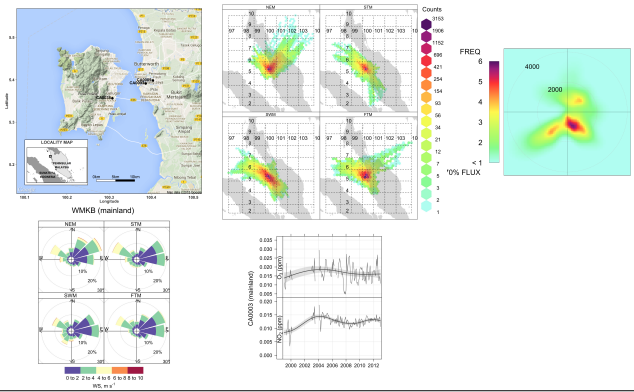
3

# Notable R applications

- Genomics - study the structure of genes
  - VERY large sets of data (millions upon millions)
  - exploratory-based research, discovering trends and relationships in large datasets
  - sometimes freely available on the web waiting for somebody to make discoveries
- Large physical systems - e.g., Earth's atmosphere
  - VERY large data sets available online

4

# Notable R applications




5

# Course objectives

- To instruct participants on how to use R to manage and analyze data
- To teach participants how to plot figures using R

6

# Course topics

1. Overview of R
2. Installing and navigating R
3. Data management (and some plotting) in R
4. Mean, median, mode, variance, and standard deviation

7

# Course topics

6. Hypothesis testing and confidence interval
7. Correlation and regression
8. ANOVA
9. Non-parametric tests

8

# Topic 1: Overview of R

Learning outcome

At the end of this topic, the participant will be able to:

- explain what R is about.

9

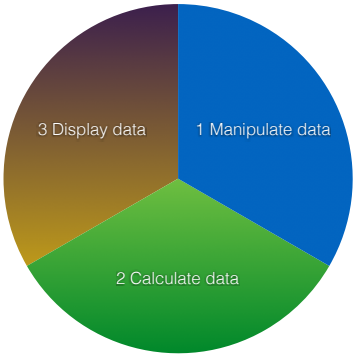
# Topic 1: Overview of R

- R is a robust data analysis tool
- R is open source (FREE!!!)
- R is popular
- Command-based interface makes it easy to document the data analysis method
- Large online user community (stackoverflow.com and you can use Google to search)

10

# Topic 1: Overview of R

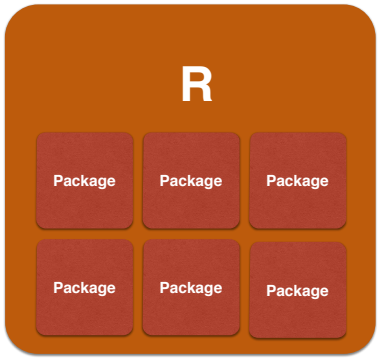
## What can you do in R?




11

# Topic 1: Overview of R

- R consists of about 25 standard/base packages
- Other packages available to download within R or RStudio




12

13

# Resources

- You can download R's notes at: http://cran.r-project.org/doc/manuals/R-intro.pdf
- You can also download a PDF on the many packages (maybe not up-to-date) at: http://www.lsw.uni-heidelberg.de/users/christlieb/teaching/UKStaSS10/R-refman.pdf
- You can download a brief intro of R at: http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

14

# Citing R in Research Papers

- The programmers of R ask that any analysis done using R be cited as:

  *R Development Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.*

15

# Topic 2: Installing and Navigating R

Learning outcome

At the end of this topic, the participant will be able to:

- install and navigate RStudio.

# Topic 2: Installing and Navigating RStudio

16

**Install R**

Download R ver. 3.2.1 from
http://cran.r-project.org/bin/windows/base/

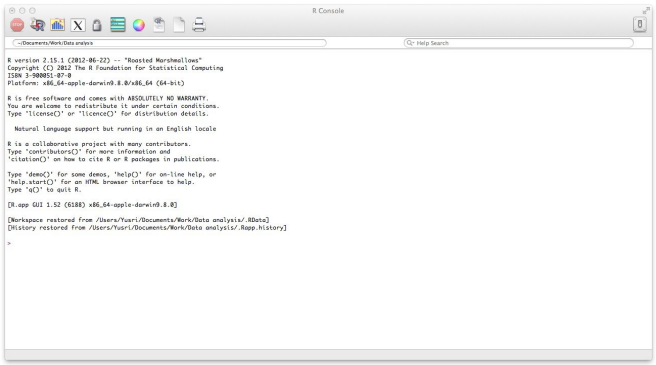
**Install RStudio**

Download RStudio ver. 0.99.473 from
http://cran.r-project.org/bin/windows/base/

# Run R

17

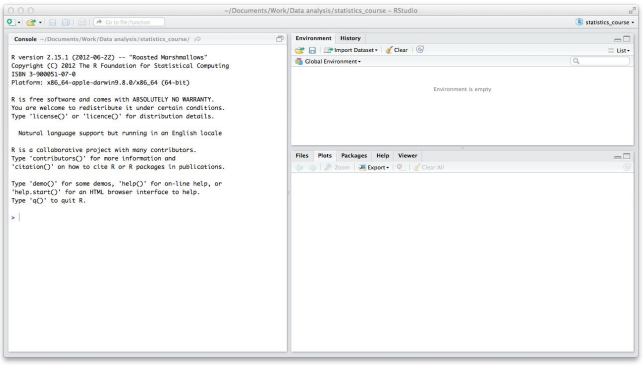
18



19

# Run RStudio

20




21

# Topic 2: Installing and Navigating RStudio

- You can determine the R version by looking at the console.
- You can determine the RStudio version by clicking on About RStudio
- Create new **R scripts** by clicking on **File > New File > R Script**. You can create custom analyses using scripts.
- The other popular way to communicate your R analysis is by using **R Markdown**. (another topic for another time). Example: http://rpubs.com/yusriy/wd2015_06_27

22

# Topic 2: Installing and Navigating RStudio

- If your console gets messy, you can clear it by pressing '**ctrl + L**'.
- You can import data by point-and-click by clicking on **Tools > Import Dataset**.
- You can export your plots at the lower left panel and clicking on the button '**Export**'.
- You can view help on functions by typing '**?**' in front of the function in the **console**. Example: ?mean

23

# Topic 3: Data management (and some plotting) in R

Learning outcome

At the end of this topic, the participant will be able to:

- input or generate data into RStudio.
- import data.
- differentiate the types of data classes used in R.
- use 'functions'.
- plot histogram, frequency polygon, barplot, time series, pie chart, dot plots.
- modify plots.

24

# Project Management in R

- My recommendations on how to start a **data analysis project**:
  - Create a **main folder**
  - Create **subfolders**:
    - **data** - to house all your data
    - **R** - to store all your scripts
    - **figs** - to store all your generated figures
    - **docs** - to keep any relevant documents
- Next level: **version control** - e.g. GitHub (after you are familiar with R), example: http://yusriy.github.io/R_stat_analysis/

# Topic 3: Data management (and some plotting) in R

25

- R uses command-based user interface. Command prompt is ">"
- Insert values into variables by using the "arrow" operator (<-) or equal operator"="
- Variables are case-sensitive
- Try,

```
> x <- 3
> y = 4
> data <- c(1,2,3,4)
> list_of_data <- 1:20 #Create a sequence from -1 to 20 with interval of 1
> data2 <- seq(from=0,to=5,by=0.5) #Create a sequence with interval of 0.5
```

# Topic 3: Data management (and some plotting) in R

26

- There are many different "functions" in R, some of them only available in installed "packages"
- Create a 2 by 2 with element 1, 2, 3, 4

```
> matrix_A <- matrix(c(1,2,3,4),2,2))
```

# Topic 3: Data management (and some plotting) in R

27

- Find out more about the function by using the symbol '?' like ?matrix
- Functions can be used by inserting "arguments" into "()". There could more than 1 argument and sometimes return a value.
- In the case of the *matrix* function, the value return is the matrix itself.

28

# Function: `ls()`

- Type `ls()`.
- This function would list all the variables in the workspace
- There are different data types
  - logical: TRUE, FALSE
  - character/string: a, b, c, computer, statistics, research
  - **numeric: 0.2, -1.0, 101325.2 (default setting)**
  - integer: -1, 0, 3, 4, -1201
  - atomic (same as vector, matrix): [2, 3], [1,4], [1, 2; 3, 4]
- `class()` can be used to determine the class of the the variable

29

# Function: `rm()`

- `rm()` can be used to remove variables from the workspace
- Example,

```
> rm(a, matrix_a)
> rm(list=ls()) #delete all variables in the workspace
```

30

# Popular plots in R

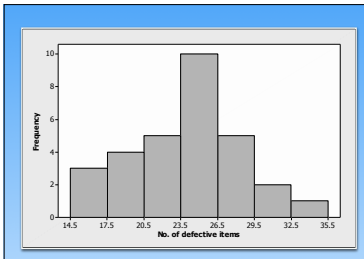
- Histogram and frequency polygon plots
- Barplot
- Time series plot
- Pie charts
- Dot plots

31

# Graphs for grouped data in a frequency distribution

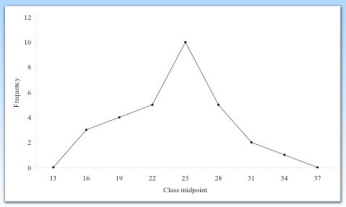
- A **histogram** is defined as a bar graph that displays the data found in a frequency distribution using the class width to represent the base of the bar and the frequency to represent the height of the bar.
- A **frequency** polygon is defined as a graph that displays the data found in a frequency distribution using straight lines to connect the points that are placed at the class midpoint. The height of each point represents the frequency of the class.

32




Histogram

Frequency polygon

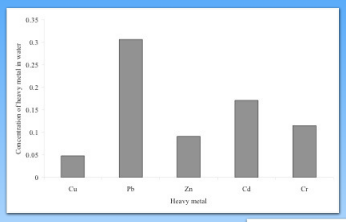



33

# Graphs for ungrouped data

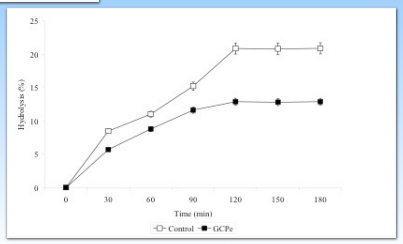
- **Pareto chart** is defined as a bar graph used to present categorical data. This chart consists of bars where each bar represents a category. The base of each bar represents the category and the heights represent the frequencies or relative frequencies.
- **Time series graph** is used when the data are collected over a period of time. It shows the changes which occur in the phenomena over that period of time.

34




Pareto chart

Time series plot




35

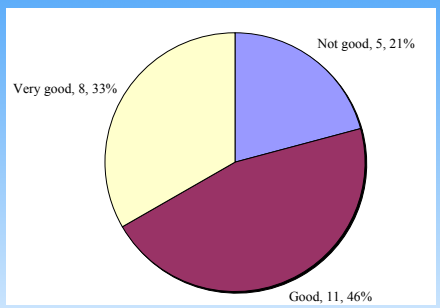
- Pie chart is used to present categorical data. It is a circle that is divided into sectors according to the percentage of frequencies of each category.
- The area of each sector can be calculated by using the following formula:

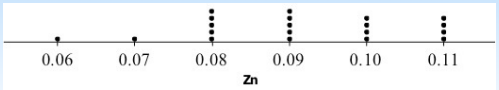
$$Degrees = \frac{Frequency\ of\ each\ category}{Total\ of\ all\ frequencies} \times 360^0$$

- A dot plot is defined as a graph in which each value is plotted as a point (dot) along a horizontal line.

36

Pie chart




Dot plot

# Hands-on 1: Distribution Analysis

37

**Defective products distribution analysis**
For thirty days, a factory concerned about the quality of its products collected data on the number of defective products returned by customers. Using R, construct a frequency distribution chart and analyze its distribution.

Data: ho1_data.txt

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 24 | 28 | 27 | 23 | 22 | 26 | 21 | 29 | 19 |
| 33 | 27 | 19 | 26 | 25 | 18 | 25 | 26 | 24 | 25 |
| 21 | 15 | 30 | 31 | 25 | 16 | 25 | 19 | 23 | 29 |

# Hands-on 1 Solution: Distribution Analysis

38

**Solution:**

1. Import data from text file

Ensure your working directory is the same as the data file.

```
> getwd()
```

```
> ho1_data <- read.table("ho1_data.txt",header = TRUE)
```

*Note: You can import data from directories other than the working directory by using the argument in read.table "file". "Header" is an argument to tell read.table that the data file contains headers (default is FALSE). Try leaving "header" as FALSE, what would happen?*

```
> ho1_data <- read.table(file="/Desktop/ho1_data.txt",header = TRUE)
```

*OR to open a window where you can choose the file:*

```
> ho1_data <- read.table(file.choose(),header=TRUE)
```

# Hands-on 1 Solution: Distribution Analysis

39

2. Plot the distribution

```
> hist_info<-hist(ho1_data
$no_defect,breaks=seq(14,35,by=3),xlab="No. of
defective products",main="Distribution of
defective products",col='lightgreen')
```

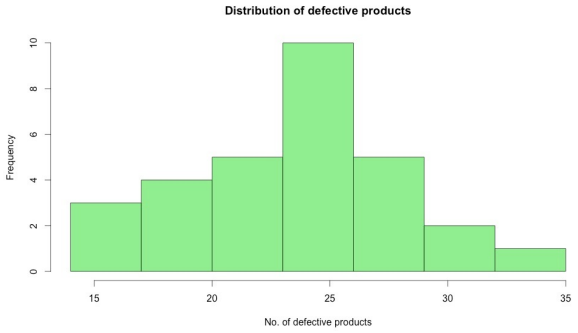
```
> hist_info
```

You can see classes, counts, etc. by typing `hist_info` because you assigned the info of the histogram into it by using the operator `<-`.

In this case you have specified the classes by looking at the data first and setting `breaks` and using the function `seq()`.

# Hands-on 1 Solution: Distribution Analysis

40




# Hands-on 1 Solution: Distribution Analysis

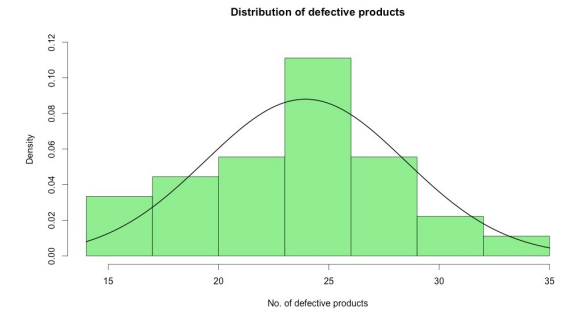
41

You can plot a density distribution where the total area under the graph is 1 by setting `freq` to `FALSE`.

```
> hist_info<-hist(ho1_data
$no_defect,breaks=seq(14,35,by=3),xlab="N
o. of defective
products",main="Distribution of defective
products",col='lightgreen',freq=FALSE,yli
m=c(0,0.12))

> curve(dnorm(x,mean=mean(ho1_data
$no_defect),sd=sd(ho1_data
$no_defect)),add=TRUE,col="black",lwd=2)
```

# Hands-on 1 Solution: Distribution Analysis

42



# Hands-on 1 Solution: Distribution Analysis

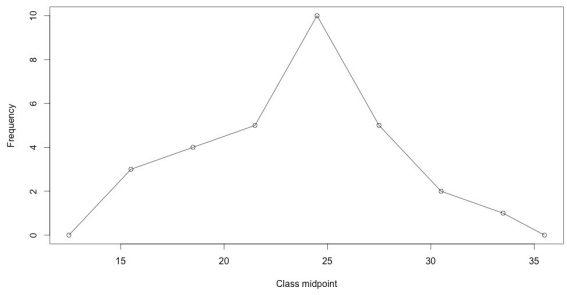
43

To draw a frequency polygon plot where frequency starts and end at zero, you need to add the before and after class midpoints and zeros for the frequency class.

```
> x <- c(12.5,hist_info$mids,35.5)

> y <- c(0,hist_info$counts,0)

> plot(x,y,type='o',xlab='Class
midpoint',ylab='Frequency')
```

# Hands-on 1 Solution: Distribution Analysis

44




# Hands-on 2: Barplot

45

**Heavy metals in water**
A study was conducted to determine the average concentration of heavy metals (in mg/L) in water. Construct a bar plot based on the following data.

Data: ho2_data.xlsx

| Cu | Pb | Zn | Cd | Cr |
|---|---|---|---|---|
| 0.048 | 0.306 | 0.091 | 0.171 | 0.115 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 2 Solution: Barplot

46

**Solution:**

1. Import data

Since the data is in Excel format, you need to download, install, and load the package "XLConnect".

```
> install.packages("XLConnect")
> library(XLConnect)
```

Then import the data.

```
> ho2_data <- readworksheet(loadworkbook("ho2_data.xlsx"),sheet=1)
```

You can rename the headers of the data frame by using the following commands:

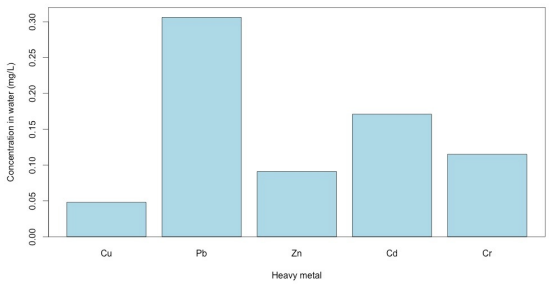
```
> names(ho2_data)[1] <- "metal"
> names(ho2_data)[2] <- "concentration"
```

2. Draw the bar plot

```
> barplot(ho2_data$concentration,xlab="Heavy metal",ylab="Concentration in water
(mg/L)",names.arg=ho2_data$metal,ylim=c(0,0.32),col='lightblue')
```

# Hands-on 2 Solution: Barplot

47




# Hands-on 3: Time Series

48

**Starch hydrolysis of noodles**
An experiment was run to study the changes in hydrolysis of starch (%) in noodles over a period of 3 hours. Two types of noodles were used: one as a control (without banana flour) and the other with Cavendish peel flour (i.e., banana flour) or GCPe. Construct a time series plot.

Data: ho3_data.xlsx

| Time (min) | 0 | 30 | 60 | 90 | 120 | 150 | 180 |
|---|---|---|---|---|---|---|---|
| Control | 0 | 8.495 | 11.038 | 15.239 | 20.887 | 20.839 | 20.909 |
| GCPe | 0 | 5.720 | 8.803 | 11.640 | 12.902 | 12.826 | 12.896 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 3 Solution: Time Series

49

**Solution:**

1. Import the data

```
> ho3_data<-readWorksheet(loadWorkbook('ho3_data.xlsx'),sheet=1)
```

2. Plot the time series

```
> plot(ho3_data$Time..min.,ho3_data$Control,type='o',xlab='Time (min)',ylab='Hydrolysis (%)',axes=FALSE)
> lines(ho3_data$Time..min.,ho3_data$GCPe,type='o',pch=19,col='blue')
```

*Note #1:* `type='o'` *is to set the type of lines plotted onto the figure,* `'o'` *is 'overplotted', type* `?plot` *in the R console to see other types of lines.*

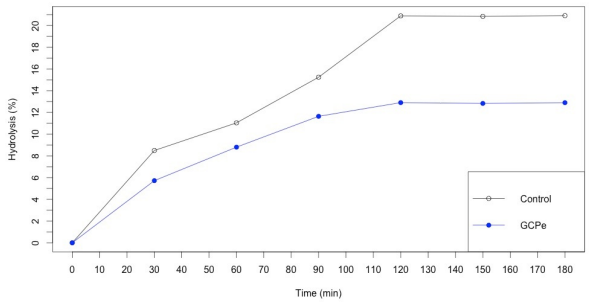
*Note #2:* `axes=FALSE` *is used so that you can control how R plots the axes.*

*Note #3: You can overlap plots by plotting the initial data and then plotting over this original plot with the next set of data using* `lines()` *or* `points()`*. Caution that the default axes follows the initial data plotted.*

```
> legend("bottomright",c("Control","GCPe"),lty=c(1,1),pch=c(1,19),col=c('black','blue'))
> axis(1,at=seq(0,200,by=10))
> axis(2,at=seq(0,30,by=1))
> box()
```

# Hands-on 3 Solution: Time Series

50




# Hands-on 4: Pie Charts

51

**Pie Charts in R**

Construct a pie chart from the following data:

Data:

| Class | Frequency |
|---|---|
| Not good | 5 |
| Good | 11 |
| Very good | 8 |
| **Total** | **24** |

# Hands-on 4 Solution: Pie Charts

52

**Solution:**

1. Input data into R

```
> ho4_data<-data.frame(c('Not Good', 'Good','Very Good'),c(5,11,8))
> names(ho4_data)<-c('Class','Frequency')
```

*Note: You can name the headers of the data frame by using the function* `names()`.

2. Plot the pie chart

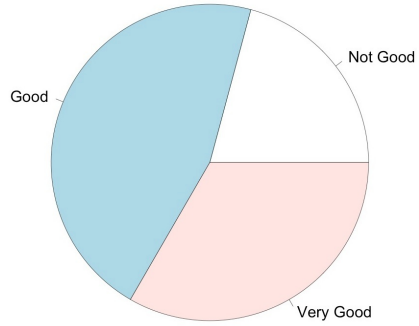
```
> par(mar=c(0.01,0.01,0.01,0.01))
> pie(ho4_data$Frequency,ho4_data$Class,cex=2)
```

*Note: You can change the margins of the plots using the function* `par()` *and the argument mar. The first numeric item is the bottom margin, left, top, and right margin sequentially.*

*Note:* `cex = 2` *because to increase the labels so that it is clearer if you want to increase the size of the pie chart.*

# Hands-on 4 Solution: Pie Charts

53




# Hands-on 5: Zn in Water - Dot Plot

54

**Dot plot of Zn concentrations in water**
Use dot plot to represent zinc (Zn in mg/L) concentrations collected from 20 sampling points.

Data: ho5_data.xlsx

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.11 | 0.11 | 0.06 | 0.09 | 0.08 | 0.10 | 0.09 | 0.10 | 0.08 | 0.10 |
| 0.08 | 0.07 | 0.09 | 0.09 | 0.08 | 0.08 | 0.10 | 0.09 | 0.11 | 0.11 |

Data courtesy of School of Industrial Technology, USM

55

# Hands-on 5 Solution: Zn in Water - Dot Plot

**Solution:**

1. Import the data

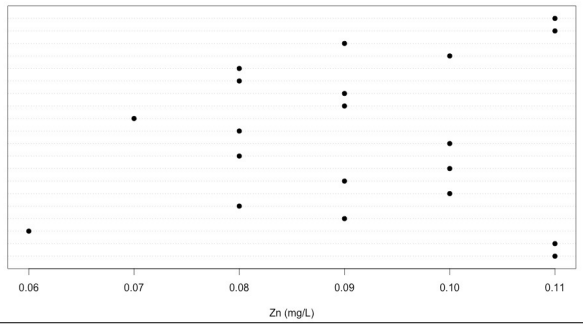
```
> ho5_data <-
readWorksheet(loadWorkbook('ho5_data.xl
sx'),sheet=1)
```

2. Plot the dot plot

```
> dotchart(ho5_data$Zn..mg.L.,xlab='Zn
(mg/L)',pch=19)
```

56

# Hands-on 5 Solution: Zn in Water - Dot Plot




57

## Topic 4: Mean, median, mode, variance, and standard deviation

Learning outcome

At the end of this topic, the participant will be able to:

- describe data using descriptive statistics in R.

# The arithmetic mean

58

- The arithmetic mean is defined as the sum of all data values divided by the total number of values. It is also known as the average or mean. The sample mean for ungrouped data is:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is the variable used to represent the data values

58

# The median

59

- The median is the middle value of the data array after arranging the data in ascending or decreasing order. The symbol for the median is *MD*. The median is found based on the number of values in the data set as follows:
- If the number of data values is odd then the median is $MD = X_n$

where $X_n$ is the middle value after arranging the data in increasing or decreasing order.

59

60

- If the number of data values is even then the median is the average of the two middle values after arranging the data in increasing or decreasing order,

$$MD = \frac{X_n + X_{n+1}}{2}$$

where $X_n$ and $X_{n+1}$ are the two middle values.

60

# Mode

- The mode is defined as the value that occurs most frequently in the data set. The symbol for the mode is *M*.
- Note:
  - If only one value occurs with the greatest frequency, then the data set is said to be unimodal.
  - If two values occur with the greatest frequency, then the data set is said to be bi-modal.
  - If more than two values occur with the greatest frequency, then the data set is said to be multi-modal.
  - If no value is repeated, then the data set has no mode.

61

61

# Range

- Range is defined as the difference between the highest and lowest value. The symbol for range is *R*.

  *R = highest value - lowest value*

- Note:
  - The range is considered as the simplest measure of variance.
  - The range depends only on two values to measure the dispersion and as a result this measure is very sensitive to extreme values (large or low).

62

62

# Variance and standard deviation

- Variance is defined as the average of the squared deviations of the values from the mean. The symbol for the sample variance is and is calculated using the following formulas:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$S^2 = \frac{n\ (\sum_{i=}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2}{n\ (n-1)}$$

63

63

64

Standard deviation is defined as the square root of the variance. The symbol for the sample standard deviation is and is calculated by using the following formula:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X - \bar{X})^2}{(n-1)}}$$

$$S = \sqrt{\frac{n\ (\sum_{i=1}^{n} X_i^2) - (\sum_{i=}^{n} X_i)^2}{n\ (n-1)}}$$

64

65

# Topic 4: Mean, median, mode, variance, and standard deviation

- You can use R to conduct any statistical analyses you require.
- We will start with the basics: mean, median, mode, variance, and standard deviation.
- Functions you will use are:
  - `mean()`
  - `median()`
  - `table()`, `names()`, and `max()`
  - `var()`
  - `sd()`

66

# Hands-on 6: Mean pH

**Mean pH of green banana pulp**
The pH of green banana pulp is an important physico-chemical parameter. The pH data obtained from 12 different samples are listed below. Find the mean pH of the 12 samples.

Data:

| | | | | | |
|---|---|---|---|---|---|
| 4.49 | 4.37 | 4.75 | 5.64 | 4.73 | 4.59 |
| 4.54 | 5.37 | 5.58 | 5.65 | 5.53 | 5.47 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 6 Solution: Mean pH

67

**Solution:**

1. Input data

```
> ho6_data <-
c(4.49,4.37,4.75,5.64,4.73,4.59,4.54,5.37,5.58
,5.65,5.53,5.47)
```

2. Calculate the mean

```
> x <- mean(ho6_data)
```

*Note: If you want to use the mean in further calculations, then you can assign the value to another variable such as* `x` *above.*

# Hands-on 7: Median

68

**Determining median from a dataset**
The following dataset has 9 data points (odd number of data points). Determine the median of this dataset using R.

Data:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.20 | 0.50 | 0.51 | 0.53 | 0.67 | 0.70 | 0.78 | 0.78 | 0.81 |

# Hands-on 7 Solution: Median

69

**Solution:**

1. Input the data

```
> ho7_data<-
c(0.20,0.50,0.51,0.53,0.67,0.70,0.78,0.78,0.81
)
```

2. Calculate the median

```
> MD <- median(ho7_data)
```

*Note: If the number of data points is even, then R would average the two middle numbers to obtain the median.*

# Hands-on 8: Mode and `table()`

70

**Mode of particulate matter (PM1) at a palm oil mill**
The following data represent the amount of PM1 (μg/m$^3$) in air at a Penang palm oil mill. Find the mode.

Data:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 58.19 | 67.31 | 120.28 | 67.36 | 80.25 | 108.76 | 74.08 | 40.61 | 30.96 | 108.64 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 8 Solution: Mode and `table()`

71

**Solution:**

1. Import the data

```
> ho8_data <- read.csv('ho8_data.csv')
```

*Note: csv stands for 'comma-separated values' a quite common format.*

2. Find the mode

```
> table(ho8_data$pm1)
```

*Note #1:* `table()` *would group the data according to number of instances. You can determine the mode from the group with the highest number of instances.*

*Note #2: In this example, since there is no group with the highest number of instances, then there is no mode for this dataset. Trying another case where there is a mode, let's add another value where we will create a dataset with a mode.*

```
> test_data <- c(ho8_data$pm1,108.64)
> test_data <- table(test_data)
> mode <- names(test_data)[test_data==max(test_data)]
```

*You can see that the class "108.64" has "2" has its number of instances and thus this class is the mode. We use* `names()` *here because the mode class is the header of the highest number of instances. We use* `max()` *to find the maximum number of instances.*

# Hands-on 9: Range

72

**Range of Mg concentration in water**
The following data are the Mg concentration in water. Find the range of Mg concentration in water.

Data: ho9_data.csv

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10.53 | 37.4 | 16.8 | 37.785 | 20.37 | 30.95 | 15.135 | 32.28 | 42.46 | 8.255 |
| 17.145 | 13.895 | 4.35 | 16.125 | 9.35 | 25.26 | 15.45 | 4.08 | 7.86 | 9.745 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 9 Solution: Range

73

**Solution:**

1. Import the data

```
> ho9_data <- read.csv('ho9_data.csv')
```

2. Calculate the range

```
> temp_data <- range(ho9_data$Mg)
```

```
> mg_range <- temp_data[2] - temp_data[1]
```

*Note:* `range()` *will create a variable with two values: the lowest and the highest values of the dataset. To obtain the difference between these two values, i.e., the range, you have to minus them such as shown above.*

# Hands-on 10: Variance and Standard Deviation

74

**Variance and standard deviation of Mg in water**
Using the Mg data before, calculate the variance and standard deviation of this dataset.

Data: ho9_data.csv

# Hands-on 10 Solution: Variance and standard deviation

75

**Solution:**

1. Calculate variance

```
> Mg_var <- var(ho9_data$Mg)
```

2. Calculate standard deviation

```
> Mg_std <- sd(ho9_data$Mg)
```

*Note: There are a number of options when calculating variance and standard deviations, the same goes for other functions as well. You can take a look at these options by typing* `?var` *or* `?sd`.

# Topic 5: Hypothesis testing and confidence interval

76

Learning outcome

At the end of this topic, the participant will be able to:

- conduct hypothesis testing.
- determine confidence interval.
- plot probability distributions.

## Hypothesis testing

77

- Hypothesis testing is a common method of drawing inferences about a population based on statistical evidence from a sample. Hypothesis testing is one of the most important tools of application of statistics to real life problems. Most often, decisions are required to be made concerning populations on the basis of sample information.

## Definitions

78

- Statistical hypothesis is a claim about a population parameter. The claim may or may not be true.
- Null hypothesis is a statistical hypothesis which states that there is no difference between a population parameter (such as mean, proportion, …) and a specific value. In other words, it states that a population parameter is equal to some claimed value. The symbol for null hypothesis is $H_0$.

79

- Alternative hypothesis is a statistical hypothesis which states that there is a difference between a population parameter and a specific value, or the parameter has a value that differs from the null hypothesis. The symbol for alternative hypothesis is $H_1$.

80

# Definitions

- Test statistic is a value calculated from the sample data and used to make a decision whether the null hypothesis should be rejected or not.
- Critical region (rejection region) is the range of values of the test statistic which indicates that the null hypothesis should be rejected.

81

- Non-critical region (non-rejection region) is the range of values of the test statistic which indicates that the null hypothesis should not be rejected.
- Critical value is any value that separates the critical region (where the null hypothesis should be rejected) from non-critical region.
- Significance level is the probability that the test statistic will lie in the critical region when the null hypothesis is true. The symbol for significance level is $\alpha$. Researchers usually choose $\alpha$ to be 0.05, 0.01, and 0.10.

# Errors in hypothesis testing

82

- Type I error is defined as the event of rejecting the null hypothesis when the null hypothesis is true. The probability of type I error is called the significance level .
- Type II error is defined as the event of failing to reject the null hypothesis when the null hypothesis is false. The probability of **Type II error is**.

# Z test for testing claims about a mean

83

- The Z-test is a statistical test used to test the mean when the population is normally distributed and σ is known.
- Z-test is used when the sample size is large. The formula for z test is:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

# t distribution

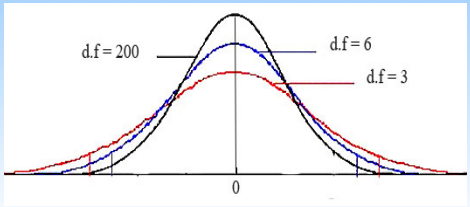
84

The t distribution is a family of curves based on the concept of degrees of freedom (d.f) (number of values that are free to vary after a test statistic has been calculated) which is related to the sample size. Some important properties of the t distribution are given below:

- It has bell shape as the standard normal distribution, where wider shape reflects greater variability.
- The mean of t distribution is equal to 0.

85

- The standard deviation is greater than 1.
- As the sample size increases the t distribution gets closer to the standard normal distribution.
- t distribution is symmetric about the mean.




86

# t test

The t test is a statistical test used to test the mean when the population is normally or approximately normally distributed and σ is unknown. Furthermore, t-test is used when the sample size is small n < 30. The formula for t-test is:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$ with degrees of n -1 freedom

87

## Hands-on 11: Comparing reported mean and sampling mean

**Testing a claim**

A manager of a confectionary company claims that the average number of cakes sold daily is more than 1750. A random sample of 36 days was selected to test the manager's claim. The sample data showed that the average is 1765 cakes. The standard deviation of the population is 100 cakes. Is there enough evidence to support the claim? You can assume that the population is normally distributed.

## Hands-on 11: Comparing reported mean and sampling mean

88

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis, $H_0$: $\mu \leq 1750$ cakes

Alternative hypothesis, $H_1$: $\mu > 1750$ cakes (manager's claim)

2. Input the data

```
> mean = 1765
> sd = 100
> n = 36
> manager_claim = 1750
```

## Hands-on 11: Comparing reported mean and sampling mean

89

2. Calculate the test value *z*

```
> z = (mean - manager_claim) / (sd / sqrt(n))
> alpha = 0.05
> z.alpha = qnorm(1 - alpha)
> result <- alpha > z.alpha
```

*Note #1: We calculated z using the formula,* $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1765 - 1750}{100/\sqrt{36}} = 0.9$

The critical value for a right-tailed at $\alpha = 0.05$ (alpha) is 1.64, since `z > z.alpha` is false then the null hypothesis cannot be rejected and thus the manager's claim is false.

*Note #2:* `qnorm()` *gives the Z-score at the interested probability, in this case 95%*

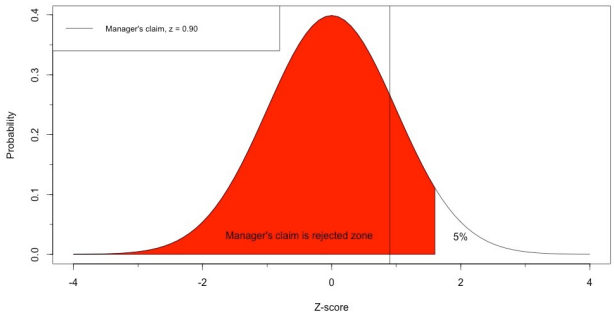
## Hands-on 11: Comparing reported mean and sampling mean

90

3. Visualize the Z-distribution

```
> x <- seq(-4,4,by=0.1)
> y <- dnorm(x)
> plot(x,y,type='l',xlab='Z-score',ylab='Probability')
> minor.tick()
> i <- x[which(x < z.alpha)]
> i <- length(i)
> polygon(c(-4,x[i],1.60),c(0,y[i],0),col='red')
> lines(c(z,z),c(-0.1,0.5))
> legend("topleft",c("Manager's claim, z = 0.90"),lty=1,col=c("black"),cex=0.8)
> text(c(2,-0.5),c(0.03,0.03),c("5%","Manager's claim is rejected zone"))
```

# Hands-on 11: Comparing reported mean and sampling mean




91

# Hands-on 12: Testing validity of reported mean

**Average weight of dried fruits packets**

The label on a dried fruits packet exhibit a weight of 275 grams (g). A sample of size 10 packets was selected and checked. The mean and standard deviation were 277.25 g and 2.725 g respectively. Does it appear that the mean weight is 275 g. Assume that the distribution is normally distributed, $\alpha$ = 0.05.

92

# Hands-on 12 Solution: Testing validity of reported mean

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis, $H_0$: $\mu$ = 275 g (manufacturer's claim)

Alternative hypothesis, $H_1$: $\mu \neq$ 275 g

2. Input data and determine the critical values for a two-tailed test

```
> sample_mean = 277.25
> mean = 275
> sd = 2.725
> n = 10
```

At $\alpha$ = 0.05 and degree of freedom (DOF) = 9 the critical values are:

```
> q <- qt(c(0.025,0.975),df=9)
> q
[1] -2.262157 2.262157
```

3. Calculate the test value t

```
> t <- (sample_mean - mean) / (sd / sqrt(n))
```

93

# Hands-on 12 Solution: Testing validity of reported mean

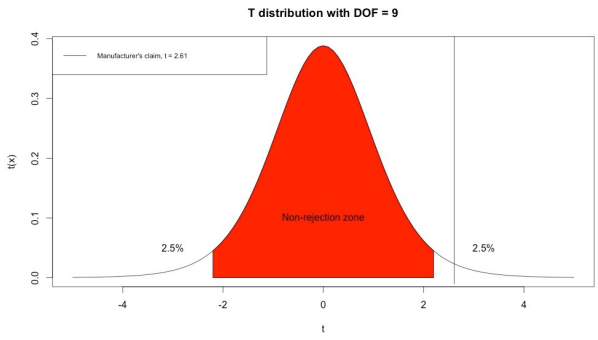
94

4. Make a decision

Since the test value t = 2.611055 is in the non-rejection zone, then the null hypothesis cannot be rejected. The manufacturer's claim is not valid. There are more dried fruits in the packets than what is labelled.

# Hands-on 12 Solution: Testing validity of reported mean

95




# Hands-on 12 Solution: Testing validity of reported mean

96

Optional: Visualizing the distribution in R

```
> x <- seq(-5,5,by=0.1)
> y <- dt(x,df=9)
> plot(x,y,xlab='t',ylab='t(x)',type='l',main="T distribution with DOF = 9")
> i <- x[which(x < q[1])]
> i <- length(i) + 1
> k <- x[which(x < q[2])]
> k <- length(k)
> polygon(c(-2.2,x[i:k],2.2),c(0,y[i:k],0),col='red')
> lines(rep(q[2],2),c(-0.01,0.5))
> legend("topleft","Manufacturer's claim, t = 2.61",lty=1,cex=0.7)
> text(c(-3,3.2),c(0.05,0.05),c("2.5%","2.5%"))
> text(0,0.1,"Non-rejection zone")
```

# Hands-on 13: Hypothesis testing

97

**Pesticide residue in wells**
The water wells that supply drinking water to nearby residents of two locations were sampled for pesticide residue. The results of the analysis are given below. Test the hypothesis that more than 20% of location 1 wells test positive for pesticide residue. Use $\alpha = 0.05$.

Data:

| Location | Positive | Negative |
|---|---|---|
| 1 | 51 | 123 |
| 2 | 21 | 356 |

# Hands-on 13 Solution: Hypothesis testing

98

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis, $H_0$: $\mu \leq 20\%$

Alternative hypothesis, $H_1$: $\mu > 20\%$

2. Determine the critical value of one right-tailed test

At $\alpha = 0.05$, critical value is,

```
> z.alpha <- qnorm(1 - 0.05)
```

# Hands-on 13 Solution: Hypothesis testing

99

3. Input the data

```
> positive <- 51
> negative <- 123
> total1 <- positive + negative
> p_hat <- positive / total1
> p = 0.20
> q = 1 - p
```

4. Calculate Z value

```
> z = (p_hat - p) / sqrt(p * q / total1)
```

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.2931 - 0.20}{\sqrt{(0.20)(0.80)/174}} = 3.0703$$

5. Compare

Since `z > z.alpha` (3.0703 > 1.64), thus the null hypothesis can be rejected and thus more than 20% of location 1 wells are contaminated with pesticide residue at 95% confidence interval.

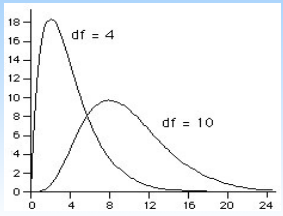
Optional:

You can try to visualize the distribution using what you have learned previously.

# Chi-square distribution

100

The chi-square distribution is a family of curves based on the concept of degrees of freedom (df) which is similar to t distribution. The symbol for chi-square is $\chi^2$ (read Ki). A chi-square variable cannot be negative and the distribution is positively skewed.




## One-sample hypothesis tests about a standard deviation or variance

101

The $\chi 2$ test for a single sample is:

$$\chi^2 = \frac{(n-1)\ S^2}{\sigma^2}$$ with n-1 degrees of freedom

n = sample size,

$s^2$ = is the sample variance, and

$\sigma^2$ = is the population variance.

# Hands-on 14: Chi-square

102

**pH variance of banana**
A nutritionist claims that the variance of pH in bananas is 0.644. A sample of 20 bananas has a standard deviation of 0.8. Is there enough evidence to reject the nutritionist's claim? Use $\alpha$ = 0.05.

# Hands-on 14 Solution: Chi-square

103

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis, $H_0$: $\sigma^2 = 0.644$

Alternative hypothesis, $H_1$: $\sigma^2 \neq 0.644$

2. Determine the critical value

At $\alpha = 0.05$ and DOF = 19, critical value is,

```
> qchisq(0.025,df=19)
[1] 8.906516
> qchisq(0.975,df=19)
[1] 32.85233
```

# Hands-on 14 Solution: Chi-square

104

3. Calculate the chi-square test values $\chi^2 = \frac{(n-1)\ S^2}{\sigma^2} = \frac{(20-1)\ (0.8)^2}{0.644} = 18.88$

```
> n = 20
> sample_sd = 0.8
> sd = 0.644
> chi_test <- (n - 1) * (sample_sd^2) / sd
> chi_test
[1] 18.88199
```

4. Compare with critical values

Since 18.88199 is between 8.906516 and 32.85233, then the null hypothesis cannot be rejected and thus the nutritionist's claim is valid.

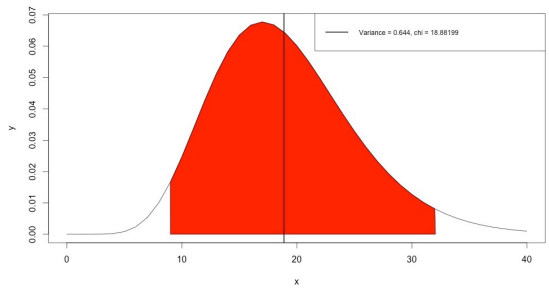
# Hands-on 14 Solution: Chi-square

105

Optional: Visualizing the chi-square distribution

```
> x <- seq(0,40,by=1)
> y <- dchisq(x,df=19)
> plot(x,y,type='l')
> lb <- qchisq(0.025,df=19)
> ub <- qchisq(0.975,df=19)
> i <- x >= lb & x < ub
> polygon(c(lb+0.1,x[i],ub-0.8),c(0,y[i],0),col='red')
> lines(c(chi_test,chi_test),c(-0.1,0.08),lwd=2)
> legend('topright',"Variance = 0.644, chi =
18.88199",lty=1,lwd=2,cex=0.75)
```

# Hands-on 14 Solution: Chi-square




106

# Confidence interval and sample size

- Estimating the value of a parameter from data obtained from a sample is one aspect of inferential statistics which is called estimation.
- A point estimate is of limited usefulness because it does not reveal the uncertainty associated with the estimate.
- Confidence intervals provide more information than point estimates, since they provide a range of plausible values for the unknown parameter.

107

# Definition

- Point estimate is a single value used to estimate a population parameter. For instance, the sample mean $\bar{x}$ is the best point estimate of the population mean, $\mu$.
- Interval estimate is a range of values used to estimate a population parameter. This interval may or may not contain the value of the parameter being estimated.

108

109

- Confidence level is the probability 1 - $\alpha$ associated with a confidence interval, or, is the probability that the interval estimate will contain the parameter, assuming that the estimation process is repeated a large number of times. Confidence level is also called confidence coefficient or degree of confidence.
- Confidence interval (CI) is an interval estimate of a parameter determined by using data obtained from a sample based on specified confidence level.

110

## Estimating a population mean: σ known or n ≥ 30

The confidence interval for the mean when the sample size is large or σ is known for a specific $\alpha$ is given by

$$\bar{X} - Z_{\alpha/2}\ [\frac{\sigma}{\sqrt{n}}] < \mu < \bar{X} + Z_{\alpha/2}\ [\frac{\sigma}{\sqrt{n}}]$$

$\bar{x}$ = sample mean,

σ = population standard deviation,

$Z_{\alpha/2}$ = critical Z value based on the desired confidence level, and *n* is the sample size.

111

## Confidence interval (CI) for the population variance and standard deviation

The confidence interval for the variance is:

$$\frac{(n-1)\ S^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)\ S^2}{\chi^2_L}$$ with d.f = *(n - 1)*

and for standard deviation is:

$$\sqrt{\frac{(n-1)\ S^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)\ S^2}{\chi^2_L}}$$ with d.f = *(n - 1)*

# Hands-on 15: Confidence interval

112

**Finding confidence interval**

A hot sauce company rates its sauce on a scale of spiciness from 1 to 20. A sample of 75 bottles of hot sauce is taste-tested. The mean and standard deviation of the sample are 13 and 2.5 respectively. Find the 95% confidence interval for the spiciness of the hot sauce produced by this company.

# Hands-on 15 Solution: Confidence interval

113

**Solution:**

1. Determine the critical value

At $\alpha$ = 0.05/2 = 0.025 (two-tailed), critical value is,

```
> qnorm(1 - 0.025)
[1] 1.959964
```

2. Input data and apply formulas $\bar{X} - Z_{\alpha/2}\left[\frac{S}{\sqrt{n}}\right] < \mu < \bar{X} + Z_{\alpha/2}\left[\frac{S}{\sqrt{n}}\right]$

```
> alpha <- qnorm(1-0.025)
> mean = 13
> sd = 2.5
> n = 75
> low_int <- mean - alpha * (sd/sqrt(n))
> high_int <- mean + alpha * (sd/sqrt(n))
```

The 95% confidence interval of the spiciness of the hot sauce is 12.434 < μ < 13.566

# Hands-on 16: t Confidence interval

114

**99% confidence interval**

A student measuring the boiling temperature of a certain liquid obtained the average reading of 7 different samples to be 101°C. If the student knows that the standard deviation of this procedure is 1.1°C, What is the 99% confidence interval of the population mean?

# Hands-on 16 Solution: t Confidence interval

115

**Solution:**

1. Determine the critical value

At $\alpha$ = 0.01/2 = 0.005 and DOF = 6, critical value is,

```
> alpha <- qt(1-0.005,df=6)
> alpha
> [1] 3.707428
```

2. Input data and apply formulas $\bar{X} - t_{\alpha/2}\left[\frac{S}{\sqrt{n}}\right] < \mu < \bar{X} + t_{\alpha/2}\left[\frac{S}{\sqrt{n}}\right]$

```
> mean = 101
> sd = 1.1
> n = 7
> low_int <- mean - alpha * (sd/sqrt(n))
> high_int <- mean + alpha * (sd/sqrt(n))
```

The 99% confidence interval of the boiling temperature of the liquid is 99.46°C ≤ μ ≤ 102.54°C

# Hands-on 17: True proportions at a 90% confidence interval

116

**90% confidence interval**

The water wells that supply drinking water to a residential area were tested. The results show that from the 500 samples collected, 12% of the samples were tested positive for pesticide residue. Find the the true proportion of samples containing pesticide residue at a 90% confidence interval.

# Hands-on 17 Solution: True proportions at a 90% confidence interval

117

**Solution:**

1. Determine the critical value

At $\alpha$ = 0.01/2 = 0.005 and DOF = 6, critical value is,

```
> alpha <- qnorm(1 - 0.05)
> alpha
> [1] 1.644854
```

2. Input data and apply formulas $\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

```
> p_hat = 0.12
> q = 1 - p_hat
> n = 500
> low_int <- p_hat - alpha * sqrt(p_hat * q / n)
> high_int <- p_hat + alpha * sqrt(p_hat * q / n)
```

The 90% confidence interval of the true proportions of percentage of wells tested positive for pesticide residue is 9.60% < μ < 14.4%

# Hands-on 18: Confidence interval of variance

**Confidence interval of mercury concentration variance**

An environmentalist is curious on what is the 95% confidence interval of the variance and standard deviation of mercury concentrations in water. Twenty water samples were analyzed and she discovered that the standard deviation of this sample set is 2 mg/L. What is the 95% confidence interval of the variance of mercury concentrations in water?

118

# Hands-on 18 Solution: Confidence interval of variance

**Solution:**

1. Determine the critical value

At $\alpha$ = 0.05/2 = 0.025 and DOF = 19, critical values are,

```
> alpha = 0
> alpha[1] <- qchisq(0.025,df=19)
> alpha[2] <- qchisq(1 - 0.025,df=19)
> alpha
> [1] 8.906516 32.852327
```

2. Input data and apply formulas $\frac{(n-1)\ S^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)\ S^2}{\chi^2_L}$

```
> n = 20
> sd = 2
> int = 0
> int[1] <- (n - 1) * sd^2 / alpha[1]
> int[2] <- (n - 1) * sd^2 / alpha[2]
```

The 95% confidence interval of the variance of mercury concentrations in wells is 2.31 mg/L < μ < 8.53 mg/L

119

# Topic 6: Correlation and regression

Learning outcome

At the end of this topic, the participant will be able to:

- calculate correlation coefficient.
- conduct regression analyses.
- conduct multi-variable linear regression.

120

# Correlation

121

- Correlation is defined as a statistical method used to determine whether a relationship between two or more variables exists. Simple correlation refers to the relationship between only two variables, while multiple correlation refers to the relationship between more than two variables.
- Scatter diagram is a graph of paired X-Y data values used to study the behavior of two variables. Scatter diagram consists of a horizontal X-axis to represent the range of one variable and a vertical Y-axis to represent the range of the second variable.

# Simple correlation coefficient

122

- Correlation coefficient is used to measure the strength and direction of the linear relationship between two quantitative variables X and Y. Correlation coefficient is called Pearson product moment correlation coefficient. The symbol for the sample correlation coefficient is *r* and for the population correlation coefficient is ρ.
- Correlation coefficient is calculated by using the following formula:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2]\,[n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}}$$

123

- Interpretation of the linear correlation coefficient. The range of *r* is from -1 to +1. If the value of correlation coefficient is close to +1, this means a strong positive linear relationship between the two variables, while the value of close to -1 means a strong negative linear relationship between the two variables. Furthermore, if the value of *r* is close to zero, this indicates that there is no significant linear relationship between the two variables.

# Simple regression

124

- In simple regression there are only two variables: independent variable, also called explanatory variable, or predictor variable and another variable called dependent variable also called a response variable whilst in multiple regression there are two or more independent variables and one dependent variable. Independent variables are used to predict the dependent variable in both simple and multiple regression.

125

The regression equation that describes the relationship between two variables (the relationship between a dependent variable Y and one independent variable X) is given below:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where, $\beta_0$ represents the Y intercept of the regression equation, $\beta_1$ is the slope of the regression equation, and $\boldsymbol{\varepsilon}$ is the error term.

126

Predicted, estimated, or fitted model and can be written as follows:

$$\hat{Y} = b_0 + b_1 X$$

Where b0 and b1 are estimation for β0 and β1 and $\hat{Y}$.

# Calculating formula

127

The formula for calculating $b_0$ and $b_1$ using the least squares method are:

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - [\sum X]^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

# Interpretation of regression equation

128

- A positive sign for the regression coefficient in the fitted model indicates that the ability of the independent variable to increase the response, whilst a negative sign indicates that the ability of the independent variable to decrease the response.

# Coefficient of determination

129

The coefficient of determination is defined as the ratio of the explained variation to the total variation. The symbol for coefficient of determination is $R^2$. It is calculated as:

$$R^2 = \frac{Explained\ \ Variation}{Total\ \ Variation} \qquad 1 \le R^2 \le 0$$

# Multiple Regression

130

- The regression equation can be used to describe the relationship between dependent variable Y and more than one independent variable ($X_1$, $X_2$, . . ., $X_k$). In the case of several independent variables the regression is called Multiple Regression and used to study the relationship between one dependent variable and several independent variables. The general form of the estimated multiple regression equation is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ . \ . \ . \ + b_kX_k$$

where, $\hat{Y}$ represents the predicted value of the dependent variable, $b_0$, $b_1$,…, $b_k$ are parameters to be estimated, and $X_1$, $X_2$,…, $X_k$ are the independent variables.

# Hands-on 19: Correlation coefficient

131

**Relationship between copper and cadmium in sediment**

The concentration of copper and cadmium in sediment is shown below. Construct a scatter diagram between the two metal concentrations and calculate the correlation coefficient.

Data: ho19_data.xlsx

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Cu (mg/L)** | 0.63 | 0.73 | 0.35 | 0.76 | 0.6 | 0.36 | 0.63 | 0.52 | 0.55 | 0.47 |
| **Cd (mg/L)** | 1.95 | 1.99 | 1.94 | 1.98 | 1.94 | 1.95 | 1.98 | 1.93 | 1.97 | 1.92 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 19 Solution: Correlation coefficient

132

**Solution:**

1. Import data

```
> library(XLConnect)
> ho19_data <- readWorksheet(loadWorkbook('ho19_data.xlsx'),sheet=1)
> names(ho19_data)<-c('Cu','Cd')
```

2. Plot cadmium versus copper concentration

```
> plot(ho19_data$Cu,ho19_data$Cd,xlab='Cu (mg/L)',ylab='Cd (mg/
L)',pch=19,col='red')
```

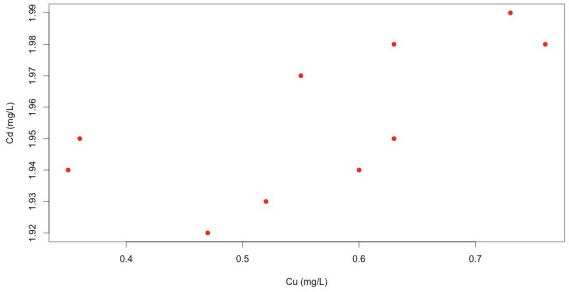
The plot shows a linear relationship between copper and cadmium concentrations.

3. Calculate the correlation coefficient, r

```
> r_cu_cd <- cor(ho19_data$Cu,ho19_data$Cd)
> r_cu_cd
[1] 0.6709394
```

$$r = \frac{n\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2]\ [n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}}$$

# Hands-on 19 Solution: Correlation coefficient




133

# Hands-on 20: Effect of one variable to another

**Effect of FRAP and total phenolic content of date palm fruits**

An experiment was carried out to analyze the edible parts of date palm fruits for their antioxidant activities using a ferric reducing/antioxidant method (FRAP). The objective of the study is to determine the effect of FRAP on total phenolic content (TPC). The results are given below.

Data: ho20_data.xlsx

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| FRAP (X) | 20.00 | 26.93 | 16.00 | 13.32 | 29.34 | 11.66 | 19.12 |
| TPC (Y) | 2.71 | 4.8 | 2.23 | 1.6 | 4.4 | 2.19 | 3.23 |

Data courtesy of School of Industrial Technology, USM

134

# Hands-on 20 Solution: Effect of one variable to another

**Solution:**

1. Import the data

```
> ho20_data <- readworksheet(loadworkbook('ho20_data.xlsx'),sheet=1)
```

2. Calculate the linear regression coefficient

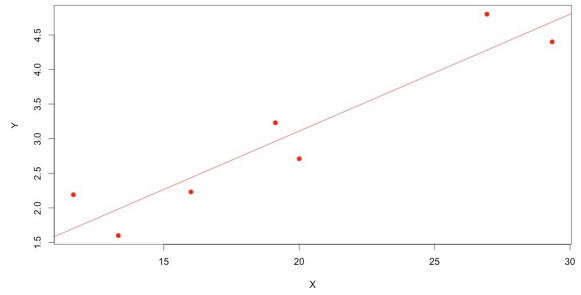
```
> lm_Y_X <- lm(Y ~ X, data = ho20_data)
> lm_Y_X
Call:
lm(formula = Y ~ X, data = ho20_data)


Coefficients:
(Intercept)            X
    -0.2655       0.1688
```

Using normal notations, $b_0$ = -0.2655 and $b_1$ = 0.1688 from the general equation $Y = b_0 + b_1 X$. The linear regression equation is Y = -0.2655 + 0.1688 X.

The constant $b_1$ is positive, which indicates that Y (FRAP) affects X (TPC) positively, if FRAP increases by 1 unit, TPC would also increase by 0.1688 units

135

# Hands-on 20 Solution: Effect of one variable to another




136

# Hands-on 21: Multi-variable linear regression

**Indoor air quality and physical properties**
The relationship between multi-factor scores and toluene metabolite concentrations (TDI) was studied to understand the behavior of indoor air components at different polyurethane factories. The data for 2 independent variables: relative humidity (RH, %) and dry bulb temperature ($T_d$, °C) and 1 dependent variable, TDI (μg/m$^3$), were collected.

Data: ho21_data.csv

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **TDI (Y)** | 81 | 79 | 78 | 76 | 75 | 59 | 58 | 57 | 55 | 53 |
| **RH (X1)** | 50 | 50 | 51 | 51 | 53 | 40 | 40 | 40 | 41 | 43 |
| **$T_d$ (X2)** | 35 | 35 | 33 | 33 | 33 | 30 | 28 | 28 | 27 | 27 |

Data courtesy of School of Industrial Technology, USM

137

# Hands-on 21: Multi-variable linear regression

**Solution:**

1. Import the data

```
> ho21_data <- read.csv('ho21_data.csv')
```

2. Perform multi-variable linear regression

```
> lmTDI <- lm(TDI ~ RH + TD,data=ho21_data)
> lmTDI


Call:
lm(formula = TDI ~ RH + TD, data = ho21_data)


Coefficients:
(Intercept)           RH           TD
   -40.2601       0.5842       2.6066
```

The multi-variable linear regression equation is TDI = -40.2601 + 0.5842 RH + 2.6066 TD (or Y = -402601 + 0.5842 X1 + 2.6066 X2)

138

# Topic 7: ANOVA

Learning outcome

At the end of this topic, the participant will be able to:

- conduct ANOVA.

139

## Analysis of variance (ANOVA)

- Analysis of variance is an important technique for analyzing and exploring the variation of a continuous response variable (dependent variable) measured at different levels of one or more independent variables.
- Analysis of variance is defined as a method of testing hypotheses about the equality of three or more population means by analyzing the sample variance.
- An ANOVA decomposes the observed variance in a continuous response into components due to different sources.

140

## Assumptions of ANOVA

The following assumptions must be satisfied in order to carry out an ANOVA:

- Normality - The samples must be obtained from populations which are normally or approximately normally distributed.
- Independence - The samples must be independent.
- Homogeneity - The variances of the populations must be equal.

141

# One-way analysis of variance

142

- In one-way ANOVA there is only one independent variable (X) (called factor) at different levels (groups) and the objective is to study the effect of different levels (groups) on a continuous response (Y) measured at different levels of X.

# One-way ANOVA Table

143

| Source of variation | Degrees of freedom d.f | Sum of squares SS | Mean squares MS | F |
|---|---|---|---|---|
| Between | $k - 1$ | $SS_B$ | $MS_B = \frac{SS_B}{k-1}$ | $F = \frac{MS_B}{MS_E}$ |
| Within (error) | $N - k$ | $SS_E$ | $MS_E = \frac{SS_E}{N-k}$ | |
| Total | $N - 1$ | $SS_T$ | | |

# Two-way analysis of variance

144

- The idea of one-way analysis of variance can be extended to study the effect of two factors (each factor has at least two levels) on response variable.
- The technique for analyzing the effect of two independent variables is called two-way analysis of variance.

# ANOVA Table

| F | MS | SS | d.f | S.O.V |
|---|---|---|---|---|
| A | p - 1 | $SS_A$ | $MS_A = \frac{SS_A}{p-1}$ | $F = \frac{MS_A}{MS_E}$ |
| B | q - 1 | $SS_B$ | $MS_B = \frac{SS_B}{q-1}$ | $F = \frac{MS_B}{MS_E}$ |
| Interaction | (p - 1)(q - 1) | $SS_{AB}$ | $MS_{AB} = \frac{SS_{AB}}{(p-1)(q-1)}$ | $F = \frac{MS_{AB}}{MS_E}$ |
| Error | pq(n - 1) | $SS_E$ | $MS_E = \frac{SS_E}{N-k}$ | |
| Total | pqn -1 | $SS_T$ | | |

145

# Hands-on 22: ANOVA

**Difference between means using ANOVA**

A researcher wants to study the effect of time on the ribose-induced Millard reaction by measuring pH. Four replicates were used with the data listed below. Test the hypothesis that there is no difference among the pH means at different temperatures, i.e., there is no effect of time on pH. Use $\alpha = 0.05$.

Data: ho22_data.csv

| Time | pH | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 15 | 5.37 | 5.37 | 5.38 | 5.37 |
| 30 | 5.24 | 5.23 | 5.25 | 5.25 |
| 45 | 5.17 | 5.18 | 5.18 | 5.19 |
| 60 | 5.07 | 5.08 | 5.09 | 5.07 |

Data courtesy of School of Industrial Technology, USM

146

# Hands-on 22 Solution: ANOVA

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis: $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hypothesis: $H_1$: At least one mean is different from the other means

2. Input and manipulate the data

```
> ho22_data<-read.csv('ho22_data.csv')

> ho22_data<-t(ho22_data) # This line is to transpose the data so that the columns
become rows

> ho22_data<-ho22_data[-1,] # This line to remove the first row

> colnames(ho22_data)<-c('t15','t30','t45','t60') # To rename the columns to reflect the
different factors

> ho22_data <- as.data.frame(ho22_data) # To change the data from matrix to data frame

> anova_data <- stack(ho22_data) # This line is to combine the factors and pH to become
a two-column data frame for ANOVA.
```

3. Conduct ANOVA

```
> result <- aov(values ~ ind, data = anova_data)

> summary(result)
```

147

# Hands-on 22 Solution: ANOVA

148

| | Df | Sum sq | Mean sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| ind | 3 | 0.1826 | 0.060873 | 885.424 | | |
| Residuals | 12 | 0.0000825 | 0.00006875 | | | |
| --- | | | | | | |
| Signif. codes: | 0 '***' | 0.001 '**' | 0.01 '*' | 0.05 '.' | 0.1 ' ' | 1 |

4. Compare

The critical F value at $\alpha$ = 0.05 with DOF 3 and 12 is

```
> qf(0.95,df1=3,df2=12)
[1] 3.490295
```

Since computed F value = 885.424 > 3.490295, then the null hypothesis can be rejected and thus the means are not the same at 95% confidence interval.

# Hands-on 23: ANOVA 2

149

**Multi-factor ANOVA**
A study was conducted to study the effect of composition ratio (%) and stage of ripeness on final viscosity. Six different compositions (factors = 1, 2, 3, 4, 5, and 6) and two stages of ripeness (ripe, r, and unripe, u) were chosen. Each combination was replicated 3 times. The results are given below.

Data: ho23_data.csv

| Composition ratio | Stage of ripeness | | | | | |
|---|---|---|---|---|---|---|
| | Ripe (r) | | | Unripe (u) | | |
| 1 | 190.83 | 190.67 | 190 | 272.63 | 276.5 | 276.17 |
| 2 | 205.08 | 205.33 | 205.83 | 331.54 | 331.13 | 329.84 |
| 3 | 218.67 | 221.08 | 219.33 | 333.75 | 334.54 | 335.13 |
| 4 | 222.75 | 222.67 | 224.58 | 327.42 | 327.8 | 326.88 |
| 5 | 206.67 | 210.75 | 209.08 | 288.75 | 288.38 | 288.29 |
| 6 | 255.75 | 254.75 | 257.75 | 255.13 | 254.29 | 254.67 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 23 Solution: ANOVA 2

150

**Solution:**

1. Import and manipulate the data

```
> ho23_data <- read.csv('ho23_data.csv')
> viscosity <- c(ho23_data$r,ho23_data$r.1,ho23_data$r.2,ho23_data
$u,ho23_data$u.1,ho23_data$u.2)
> ratio <- as.factor(rep(seq(1,6),times=6))
> ripe <- rep(c('r','u'),each=18)
> ho23_df <- data.frame(viscosity,ratio,ripe)
```

2. Conduct ANOVA and display result

```
> result<-aov(viscosity ~ ratio + ripe + ratio*ripe, data = ho23_df)
> summary(result)
```

# Hands-on 23 Solution: ANOVA 2

| | Df | Sum sq | Mean sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| ratio | 5 | 8893 | 1779 | 1341 | <2e-16 | *** |
| ripe | 1 | 64285 | 64285 | 48462 | <2e-16 | *** |
| ratio:ripe | 5 | 15560 | 3112 | 2346 | <2e-16 | *** |
| Residual | 24 | 32 | 1 | | | |
| --- | | | | | | |
| Signif. codes: | 0 ‘***’ | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’ | 0.1 ‘’ | 1 |

Explanation of results by Assoc Prof Dr Abbas…

151

# Topic 8: Non-parametric tests

Learning outcome

At the end of this topic, the participant will be able to:

- conduct sign test.
- Wilcoxon signed-rank and rank sum tests.

152

# Non-parametric tests (distribution free)

- In many cases the data do not follow normal distribution. In such cases it is important to use statistical tests that do not require the data to follow a particular distribution. Such tests are called non-parametric tests.

153

## Uses of non-parametric methods

154

Non-parametric tests should be employed in any of the following cases:

- When the sample is so small and that is difficult to
  - verify the normality assumption
  - with nominal data
  - with ordinal (rank) data

## One sample sign test

155

- Sign test for a single sample is the simplest non-parametric test which is used to test the hypothesis whether the median (proportion) of a set of data equal to a specific value.

## Wilcoxon signed-rank test (Matched pairs)

156

- Wilcoxon signed-rank test is used for paired samples, or before and after which takes into account the actual magnitude of the differences.
- Wilcoxon test can be used in place of t-test for dependent samples. This test does not require the condition of normality.

# Hands-on 24: Sign test

157

**Sign test**

An environmentalist claimed that the concentration of iron (Fe) in the Juru River sediment is 38 mg/L. He collected and analyzed 20 samples at different sampling locations. The results of his analysis are shown below. Use $\alpha$ = 0.05 to test the environmentalist's claim.

Data: ho24_data.csv

| | | | | |
|---|---|---|---|---|
| 40.33 | 37.79 | 40.03 | 36.41 | 36.05 |
| 38.00 | 36.53 | 37.84 | 35.06 | 37.98 |
| 37.21 | 36.7 | 39.71 | 37.47 | 38.00 |
| 36.77 | 35.73 | 36.31 | 38.74 | 37.76 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 24 Solution: Sign test

158

**Solution:**

1. Import the data

```
> ho24_data <- read.csv('ho24_data.csv',header=FALSE)
> names(ho24_data) <- 'Fe'
```

2. State the null and alternative hypothesis

Null hypothesis, $H_0$: concentration = 38 mg/L (claim)

Alternative hypothesis, $H_1$: concentration ≠ 38 mg/L

3. Compare each value with the claim and determine whether it is above the mode (+) or below (-)

```
> claim = 38
> sign_pos <- ho24_data$Fe > claim
> sign_neg <- ho24_data$Fe < claim
> sign_no <- ho24_data$Fe == claim
```

# Hands-on 24 Solution: Sign test

159

```
> no_pos <- sum(sign_pos) # count number of + and -
> no_neg <- sum(sign_neg)
> no_no <- sum(sign_no)
```

4. Conduct the test (the default is 95% confidence interval)

```
> binom.test(no_pos,no_pos + no_neg)

     Exact binomial test

data:  no_pos and no_pos + no_neg
number of successes = 4, number of trials = 18, p-value = 0.03088
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.06409205 0.47637277
sample estimates:
probability of success
             0.2222222
```

# Hands-on 24 Solution: Sign test

160

Since 0.03088 < 0.05, so the null hypothesis can be rejected and thus the environmentalist's claim is incorrect.

# Hands-on 25: Sign test 2

161

**Sales, before and after advertising**
A new cake was introduced to Malaysia in 2013. After a year of poor sales, the company started an advertising campaign. The sales before and after the advertising campaign were recorded in thousands of RM for a period of one month in 10 markets. The sales data are given in the following table. Can it be concluded that the sales have increased after the advertising campaign at $\alpha = 0.05$?

Data: ho25_data.csv

| Market | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before (RM) | 3.5 | 3.3 | 2.8 | 2.7 | 2.5 | 3.5 | 1.7 | 2.9 | 2.8 | 2.4 |
| After (RM) | 4.1 | 4.5 | 4.3 | 5.2 | 4.6 | 4.9 | 3.5 | 4.5 | 4.4 | 3.9 |

# Hands-on 25 Solution: Sign test 2

162

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis, $H_0$:

There is no difference in sales before and after advertisement, $md_i - md_0 = 0$.

Alternative hypothesis, $H_1$:

There is a difference in sales before and after advertisement, $md_i - md_0 \neq 0$.

2. Input data

```
> ho25_data<-read.csv('ho25_data.csv')
```

3. Calculate the difference between before and after

```
> diff <- ho25_data$after - ho25_data$before
> pos <- sum(diff > 0)
> neg <- sum(diff < 0)
> zero <- sum(diff == 0)
```

# Hands-on 25 Solution: Sign test 2

163

4. Conduct the test (the default is 95% confidence interval)

```
> binom.test(pos,pos + neg)


     Exact binomial test


data:  pos and pos + neg
number of successes = 7, number of trials = 9, p-value = 0.1797
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3999064 0.9718550
sample estimates:
probability of success
             0.7777778
```

Since 0.1797 > 0.05, so the null hypothesis cannot be rejected and thus there was no increase in sales after the advertisement campaign. Note: There has to be at least 8 positives before there is a 95% CI increase in sales.

# Hands-on 26: Wilcoxon signed-rank test

164

**Wilcoxon signed-rank test**

A new type of noodle was introduced for people who are diabetic. The researcher claims that this noodle will not increase a diabetic's glycemic index and blood sugar level. The blood sugar level was measured before and after eating the noodle and the results (in %) are given below. At $\alpha$ = 0.05, is there sufficient evidence to support the researcher's claim?

Data: ho26_data.csv

| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Before | 57 | 59 | 61 | 58 | 64 | 62 | 59 |
| After | 58 | 59 | 62 | 58 | 65 | 62 | 61 |

Data courtesy of School of Industrial Technology, USM

# Hands-on 26 Solution: Wilcoxon signed-rank test

165

**Solution:**

1. State the null and alternative hypothesis

Null hypothesis, $H_0$: There is no difference in blood sugar level before and after eating the noodle.

Alternative hypothesis, $H_1$: There is a difference in blood sugar level before and eating the noodle.

2. Input the data

```
> ho26_data<-read.csv('ho26_data.csv')
```

3. Conduct the Wilcoxon signed-rank test

```
> wilcox.test(ho26_data$before,ho26_data$after,paired=TRUE)

     Wilcoxon signed rank test with continuity correction

data:  ho26_data$before and ho26_data$after
V = 0, p-value = 0.08897
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(ho26_data$before, ho26_data$after, paired = TRUE) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(ho26_data$before, ho26_data$after, paired = TRUE) :
  cannot compute exact p-value with zeroes
```

4 Interpret the results

Since p-value = 0.08897 > 0.05, we cannot reject the null hypothesis, and thus there is no difference in blood sugar level before and after eating the noodle.

166

## Wilcoxon rank-sum test (Two independent samples)

- Wilcoxon rank-sum test is a non-parametric alternative to the two independent sample t-tests. The formula and the procedure will be given in the next example.

167

## Hands-on 27: Wilcoxon rank sum test

**Two independent samples**
A researcher wants to test whether there is a difference in ash contact (%) between green banana peel and ripe banana peel. The ash content was measured on a dry basis for varieties with the data shown below.

Data: ho27_data.csv

| Green peel (G) | Ripe peel (R) |
|---|---|
| 17.134 | 4.763 |
| 17.130 | 4.766 |
| 17.120 | 4.771 |
| 17.128 | 4.767 |
| 17.132 | 4.768 |
| 17.122 | 4.773 |
| 17.128 | 4.765 |
| 17.128 | 4.769 |
| 17.133 | 4.789 |
| 17.128 | 4.763 |

Data courtesy of School of Industrial Technology, USM

168

## Hands-on 27 Solution: Wilcoxon rank sum test

**Solution:**
1. State the null and alternative hypothesis

Null hypothesis, $H_0$: There is no difference in ash content between green and ripe banana peel.

Alternative hypothesis, $H_1$: There is a difference in ash content between green and ripe banana peel.

2. Input the data

```
> ho26_data <- read.csv('ho26_data.csv')
```

3. Conduct the Wilcoxon rank sum test

```
> wilcox.test(ho27_data$G,ho27_data$R,paired=FALSE)

     Wilcoxon rank sum test with continuity correction

data:  ho27_data$G and ho27_data$R
W = 100, p-value = 0.0001717
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(ho27_data$G, ho27_data$R, paired = FALSE) :
  cannot compute exact p-value with ties
```

4. Interpret the results
Since p-value = 0.0001717 < 0.05, we can reject the null hypothesis and thus there is a difference in ash content between green and ripe banana peel.