Introduction to R

May 8 - 9, 2017

Yusri Yusup, PhD
Environmental Technology
School of Industrial Technology
Universiti Sains Malaysia

Why use R?

- R is free and available on many operating systems (OSX, Windows, Linux).
- R has many statistical tools (10000+ packages)
- Statistical analysis using R is reproducible

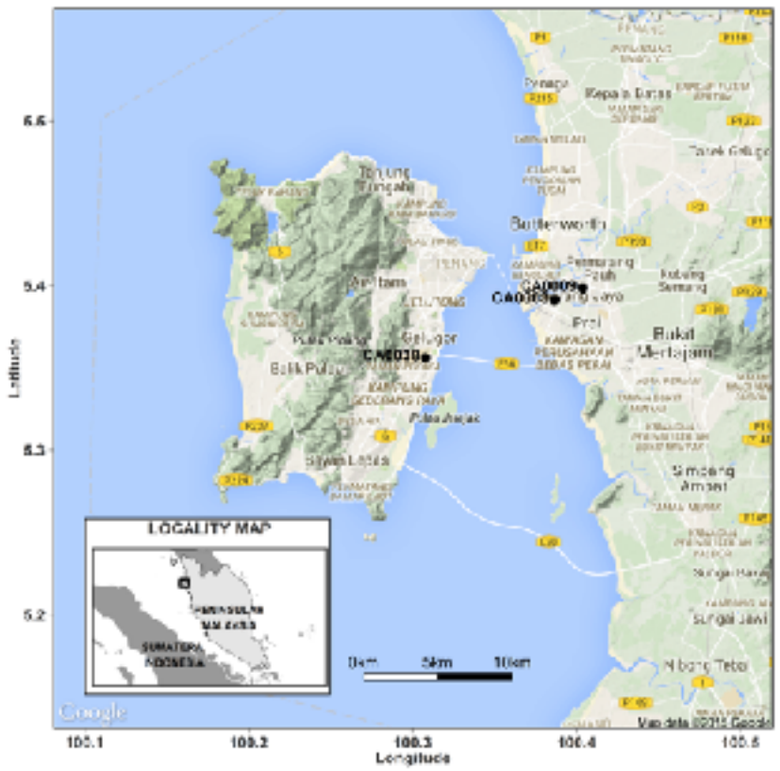
R compared to other commercial softwares

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

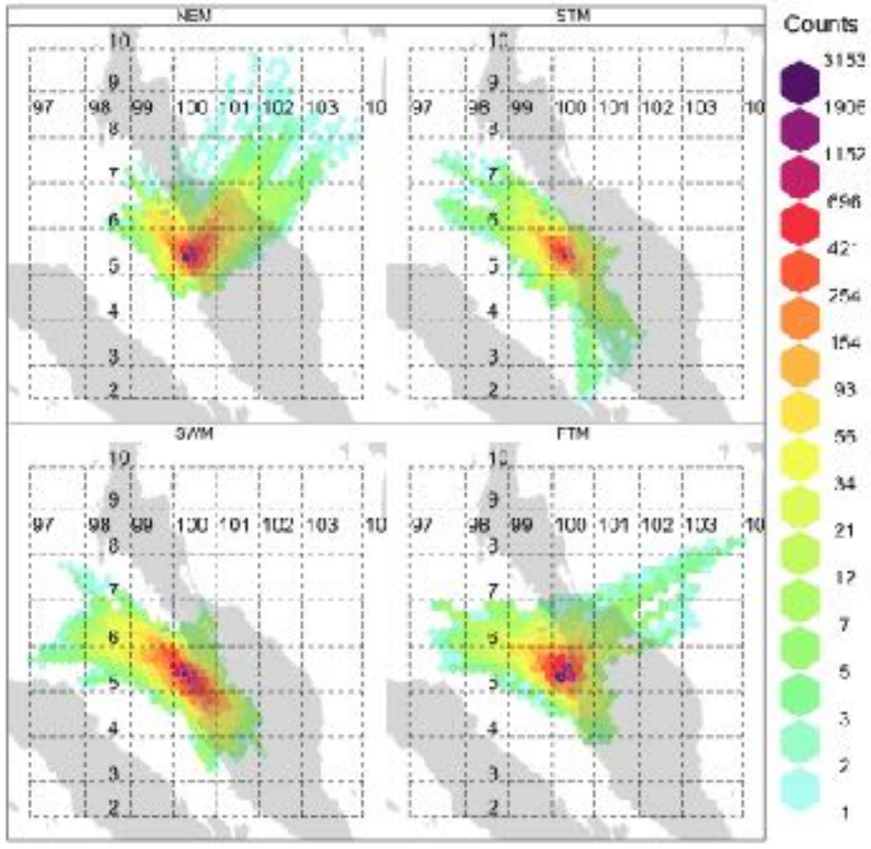
Notable R applications

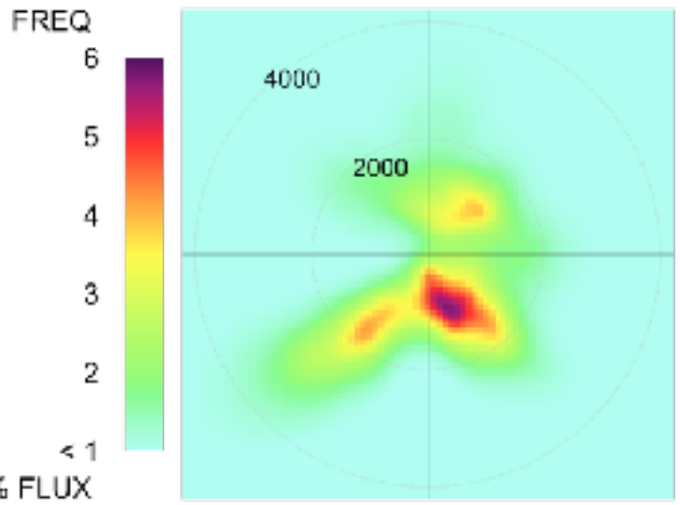
- Genomics - study the structure of genes
  - VERY large sets of data (millions upon millions)
  - exploratory-based research, discovering trends and relationships in large datasets
  - sometimes freely available on the web waiting for somebody to make discoveries
- Large physical systems - e.g., Earth's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

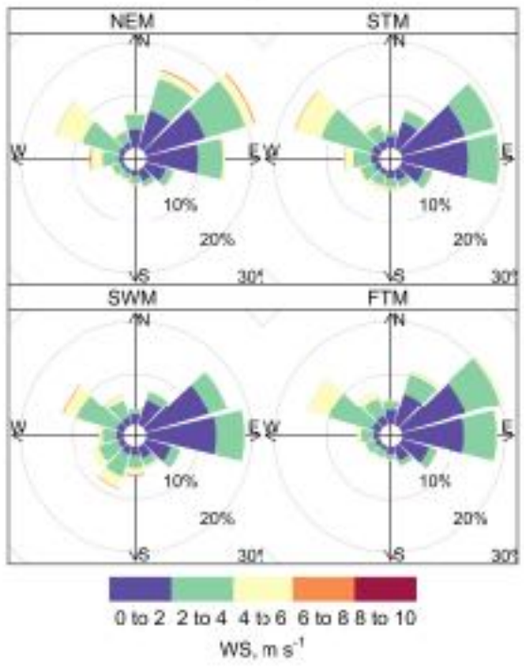
Notable R applications

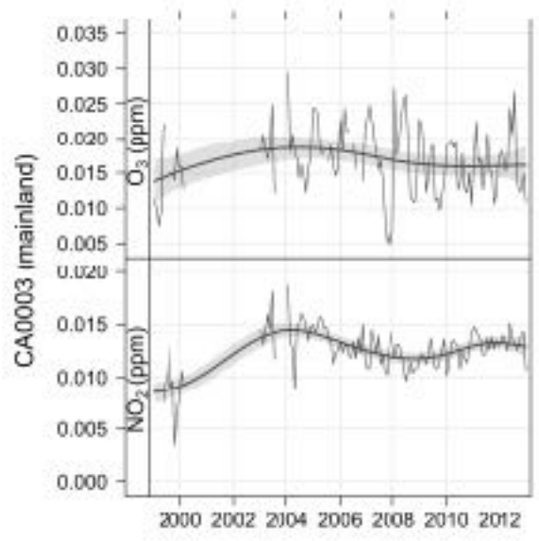


WMKB (mainland)









Course objectives

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

Course topics

1. Overview of R
2. Installing and navigating R
3. Data management (with some plotting)
4. Descriptive statistics
5. Correlation and regression
6. ANOVA

Topic 1: Overview of R

Learning outcome

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

- explain what R is about.

Topic 1: Overview of R

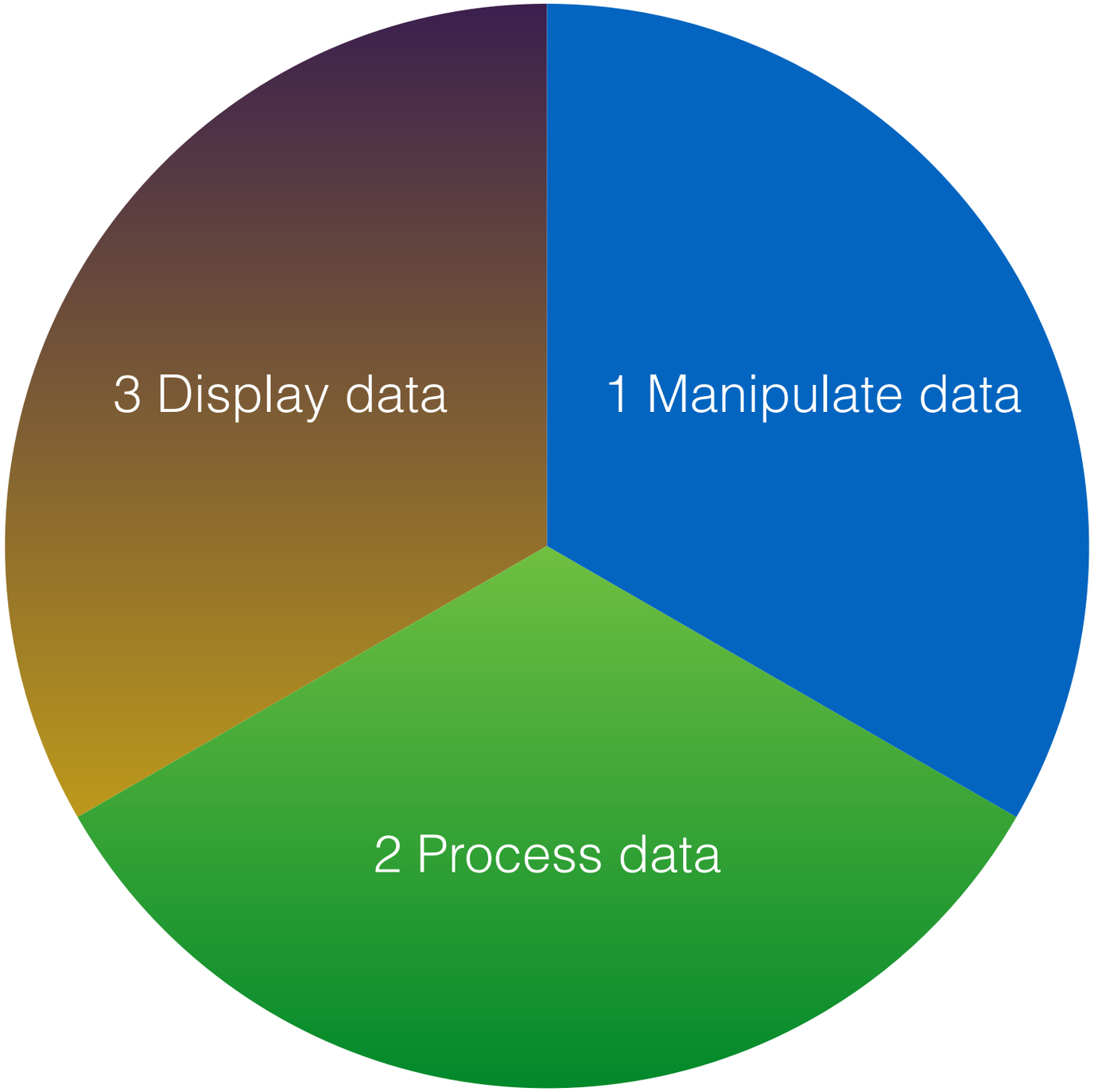
- R is a robust data analysis tool
- R is open source (FREE!!!)
- R is popular

  “R is the most popular language for data scientists — and it's been around for almost 20 years — and so by sheer force of numbers and time, R has more extensions than any other data science software. R is the primary tool used for statistical research: when new methods are developed, they're not just published as a paper — they're also published as an R package. That means R is always at the cutting edge of new methodologies.” - Revolutions, Microsoft (2017)

- 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)

Topic 1: Overview of R

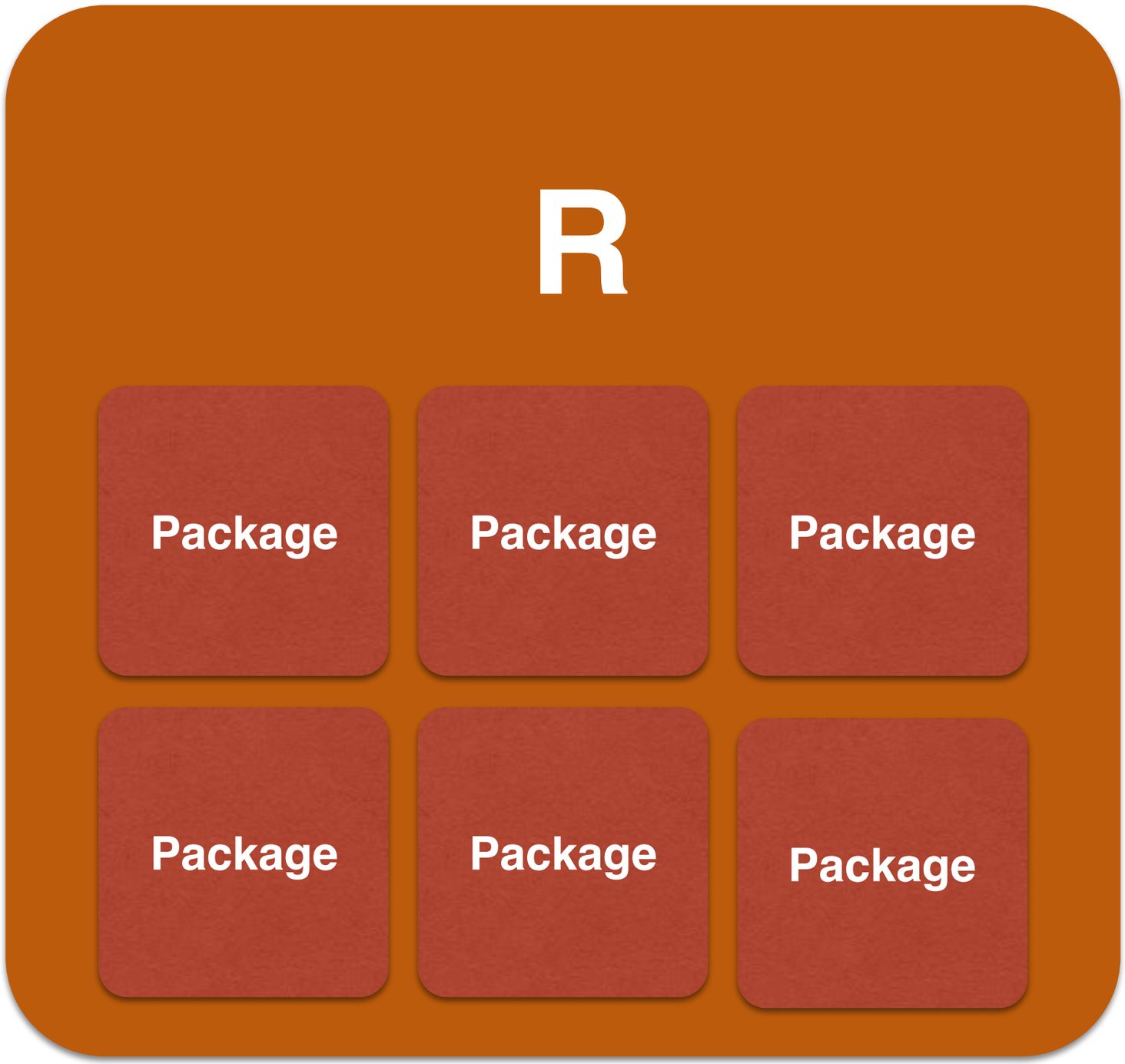
What can you do in R?



Topic 1: Overview of R

- R consists of about 25 standard/base packages

- Other packages (> 10000) available to download within R or RStudio



Resources

- You can download R's notes at: http://cran.r-project.org/doc/manuals/R-intro.pdf

- You can download a brief intro of R at: http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Citing R in Research Papers

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

  R Development Core Team (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

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

Install R

Download R ver. 3.4.0 from
http://cran.r-project.org/bin/windows/base/

Install RStudio

Download RStudio ver. 1.0.143 from
https://www.rstudio.com/products/rstudio/download2/

Run R

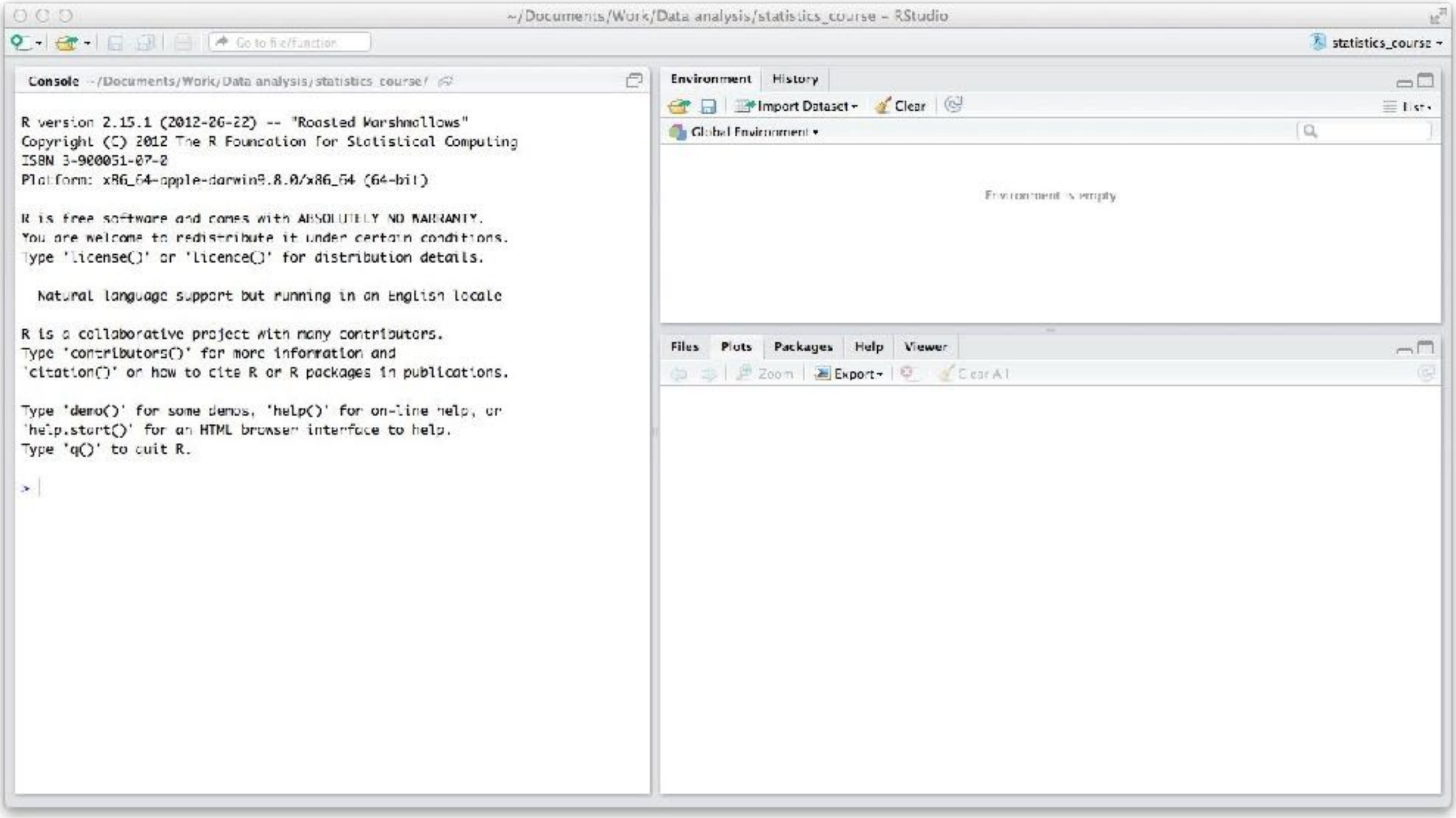
R Console

~/Documents/Work/Data analysis

Help Search

```
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.52 (6188) x86_64-apple-darwin9.8.0]

[Workspace restored from /Users/Yusri/Documents/Work/Data analysis/.RData]
[History restored from /Users/Yusri/Documents/Work/Data analysis/.Rapp.history]

>
```

Run RStudio


~/Documents/Work/Data analysis/statistics_course – RStudio
Go to file/function
statistics_course
Console ~/Documents/Work/Data analysis/statistics_course/
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' or how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Import Dataset
Clear
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Zoom
Export
Clear All

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).

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

Topic 3: Data management (and some plotting)

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.

Project Management

- 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)

- 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)

- 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)

- 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.

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

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
```

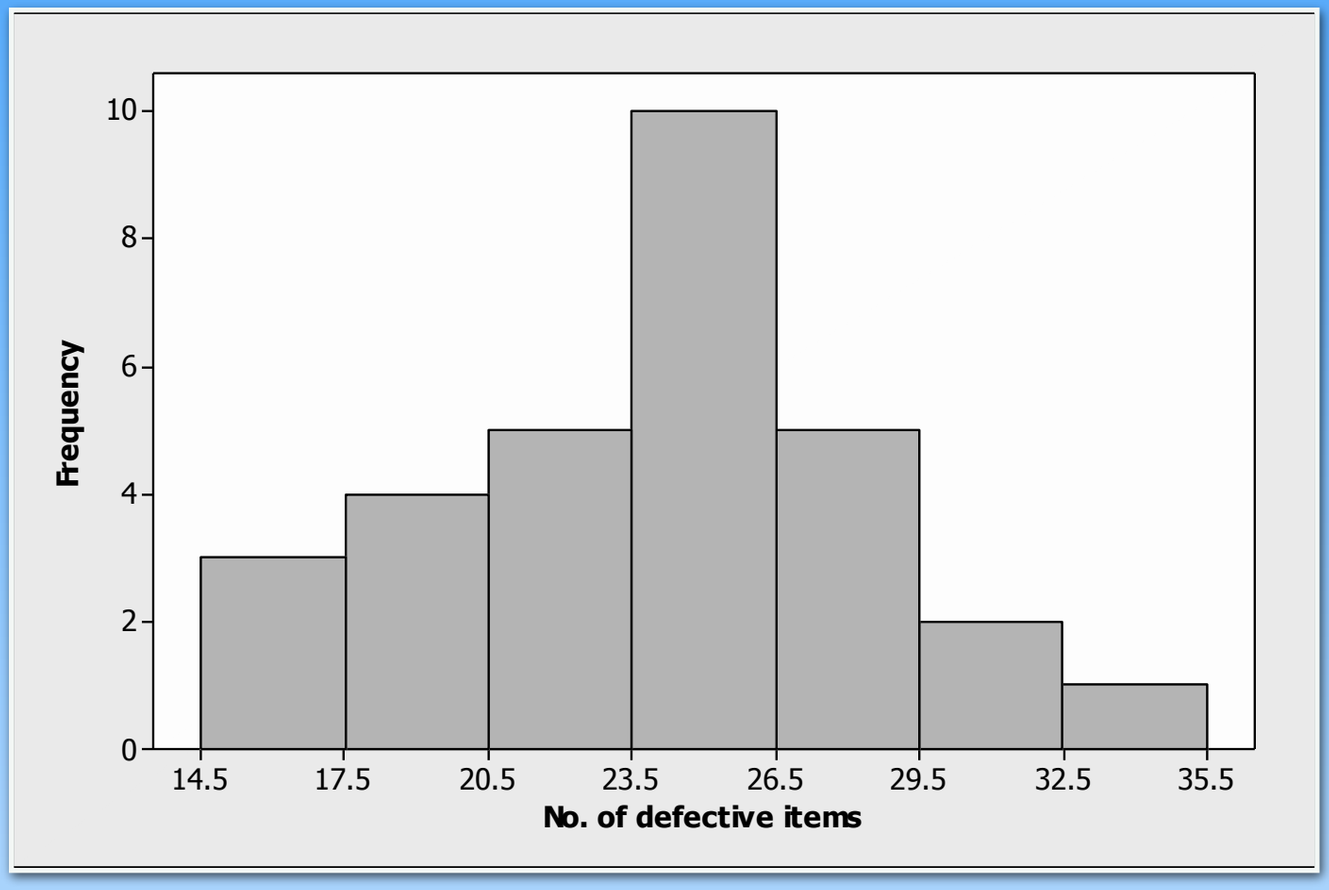
Popular plots in R

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

Graphs for grouped data in a frequency distribution

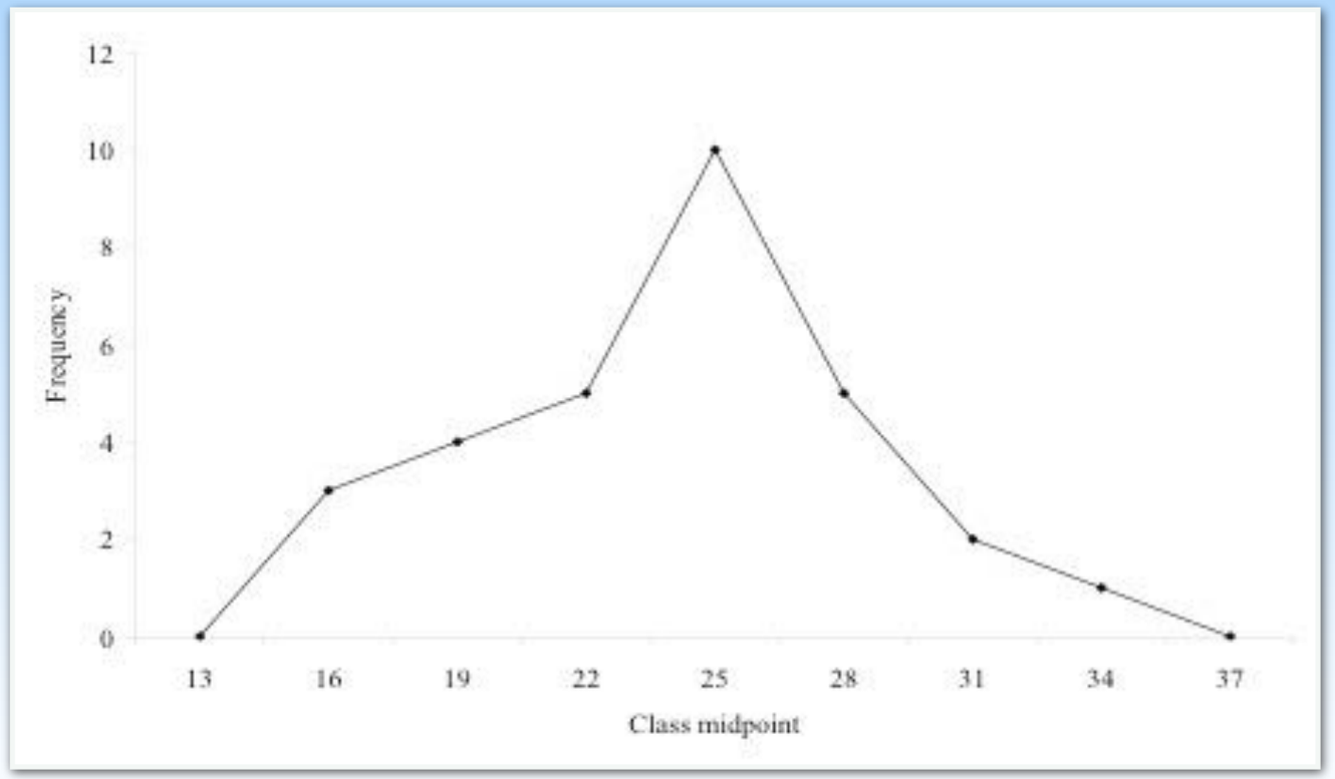
- A **histogram** is defined as a bar graph that displays the data found in a frequency distribution using the class width to represent the base of the bar and the frequency to represent the height of the bar.

- 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.



Histogram

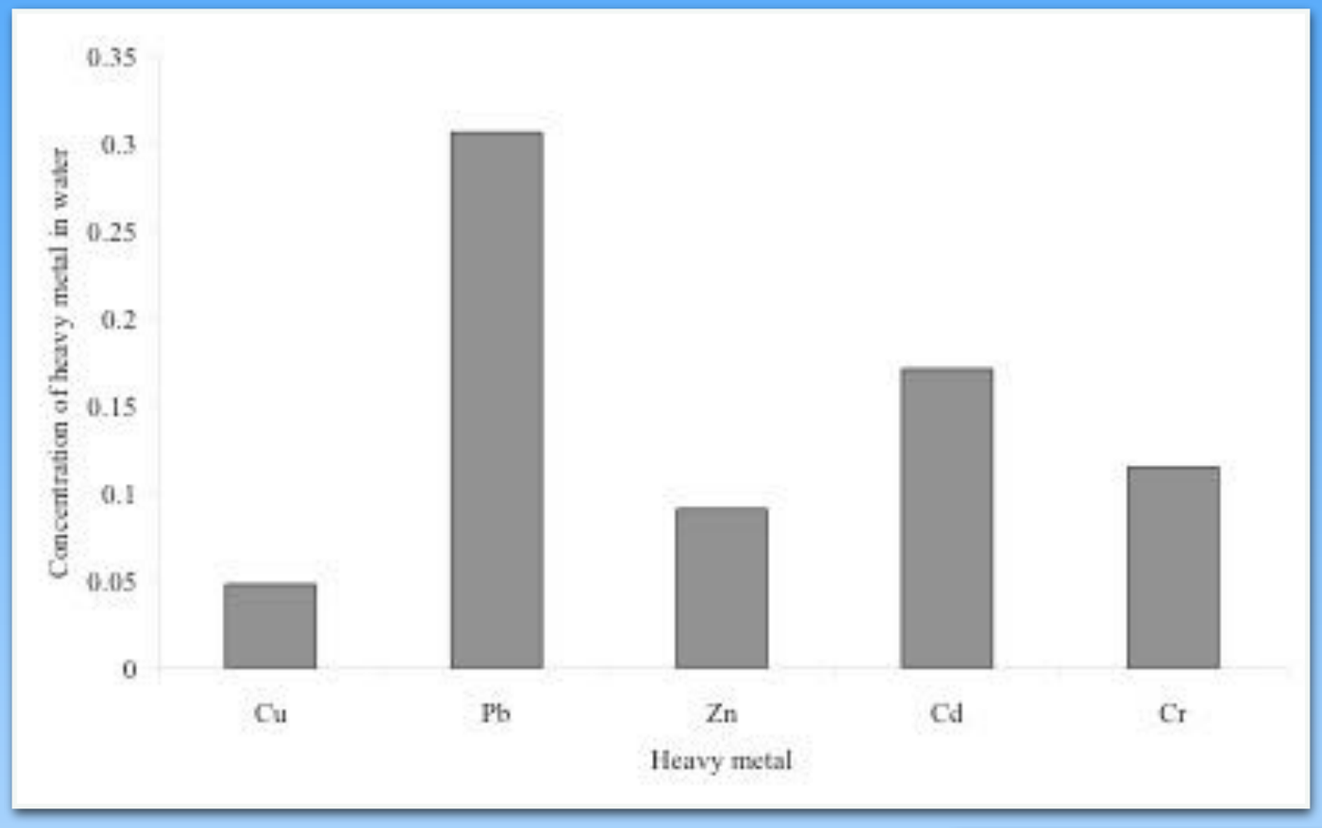
Frequency polygon



Graphs for ungrouped data

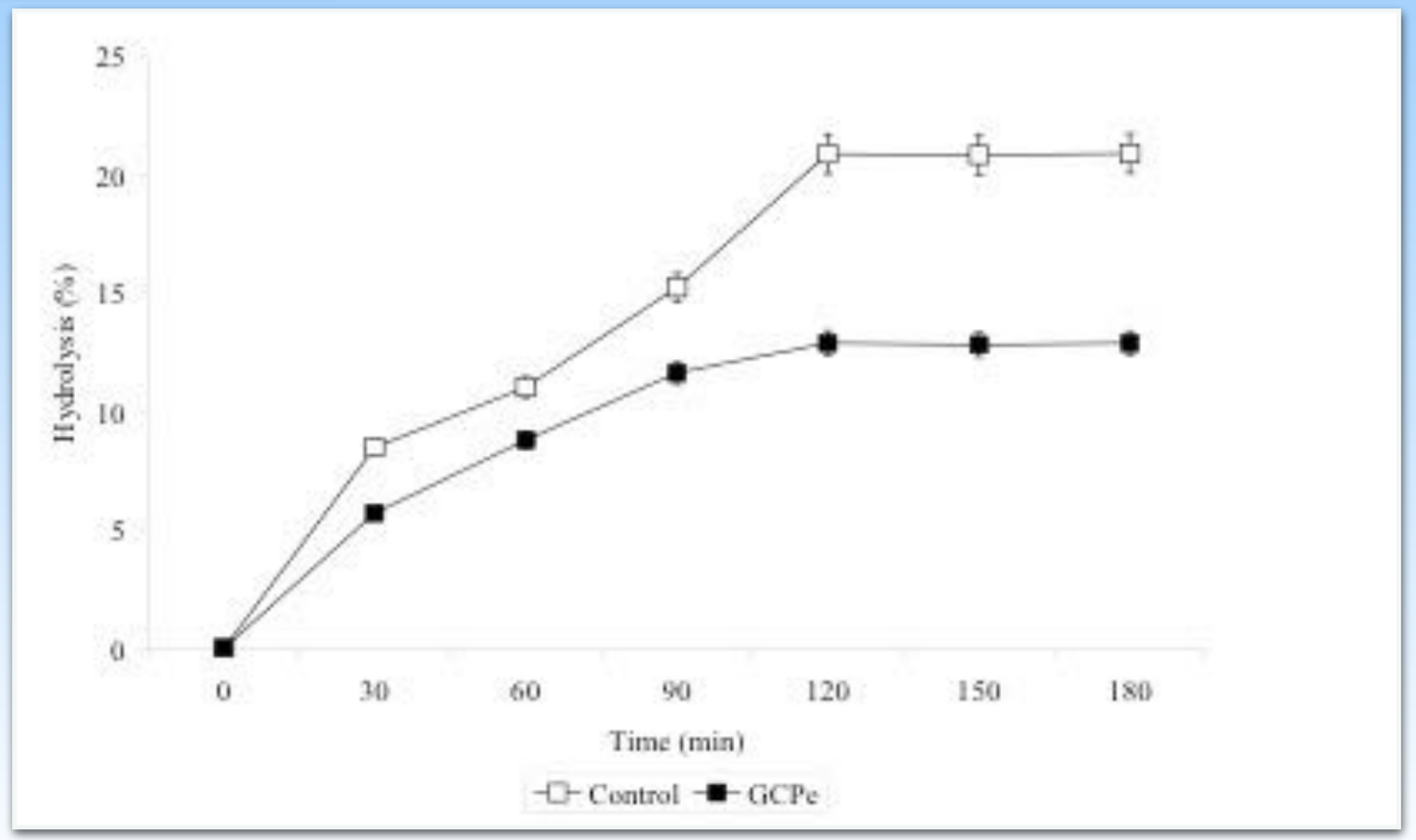
- **<u>Pareto chart</u>** is defined as a bar graph used to present categorical data. This chart consists of bars where each bar represents a category. The base of each bar represents the category and the heights represent the frequencies or relative frequencies.

- **<u>Time series graph</u>** 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.



Pareto chart

Time series plot



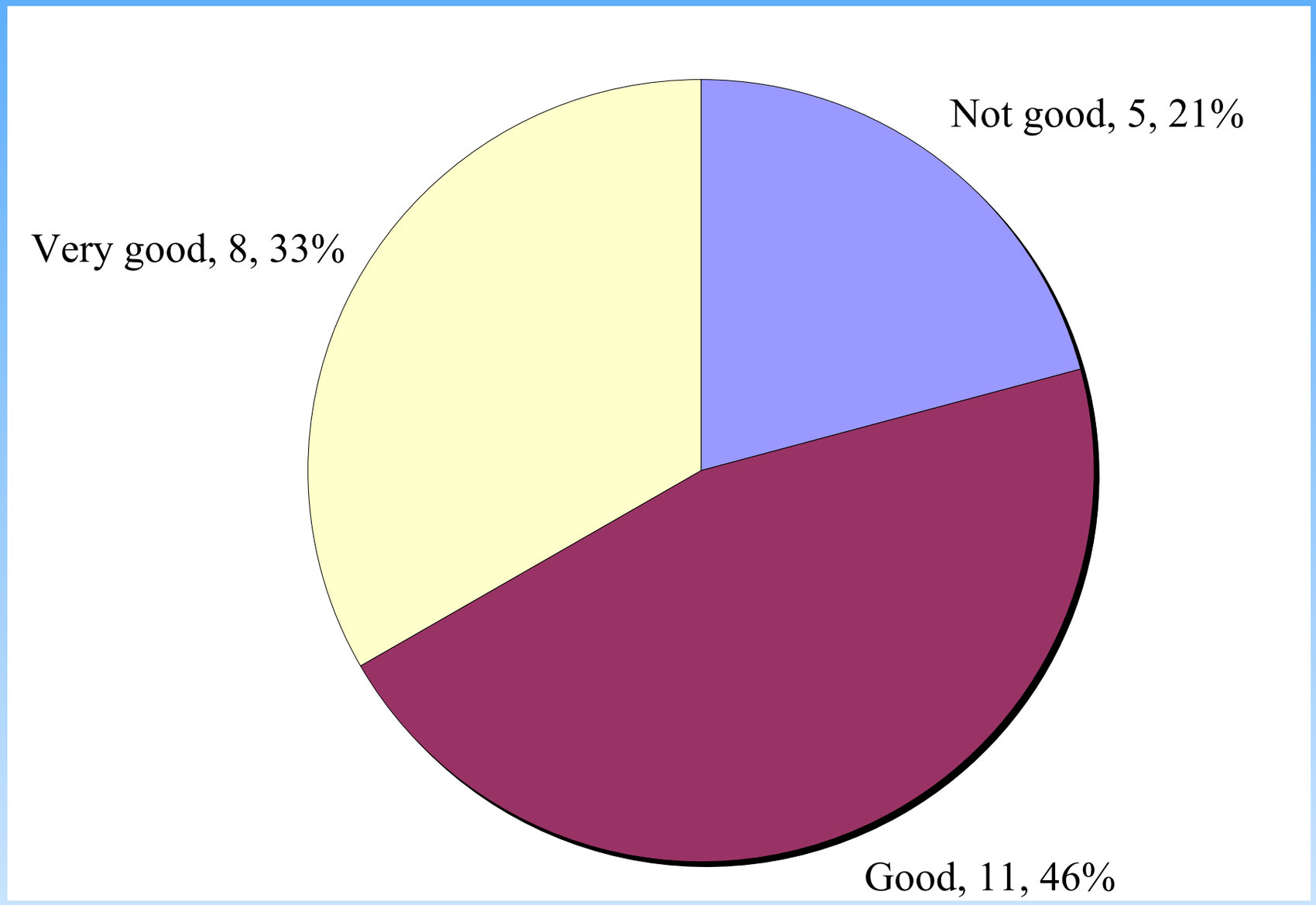
- Pie chart is used to present categorical data. It is a circle that is divided into sectors according to the percentage of frequencies of each category.

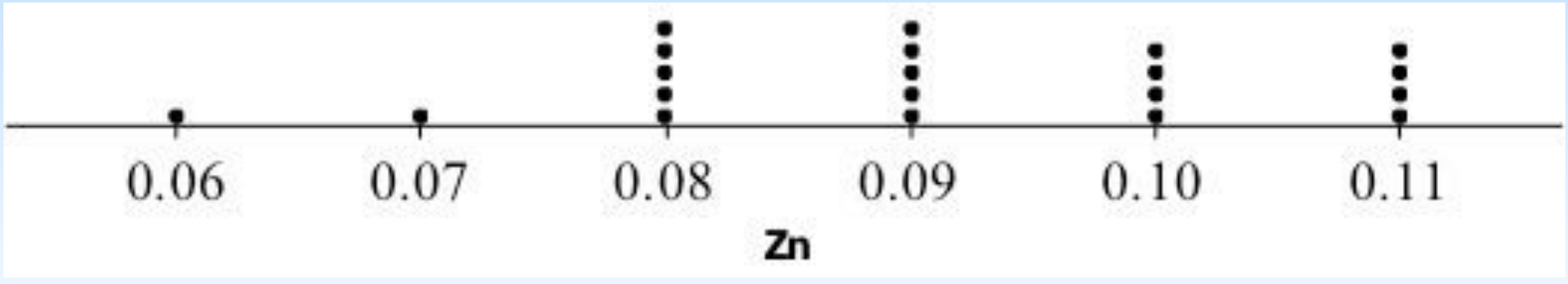
- The area of each sector can be calculated by using the following formula:

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

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

Pie chart





Dot plot

Hands-on 1: Distribution Analysis

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

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

2. Plot the distribution

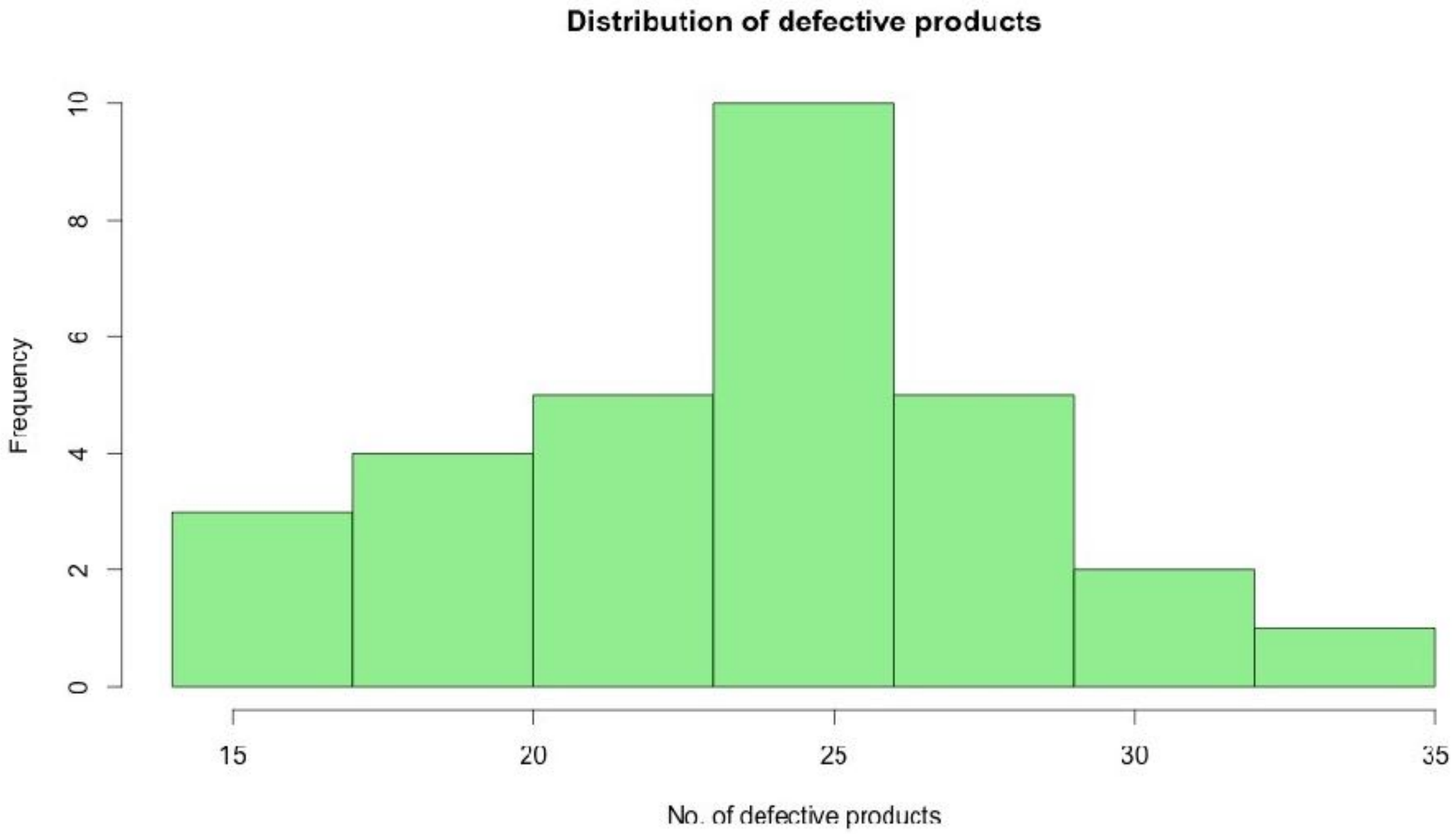
```
> hist_info<-
hist(ho1_data$no_defect,breaks=seq(14,35,by=3),xla
b="No. of defective products",main="Distribution
of defective products",col='lightgreen')
```

```
> 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

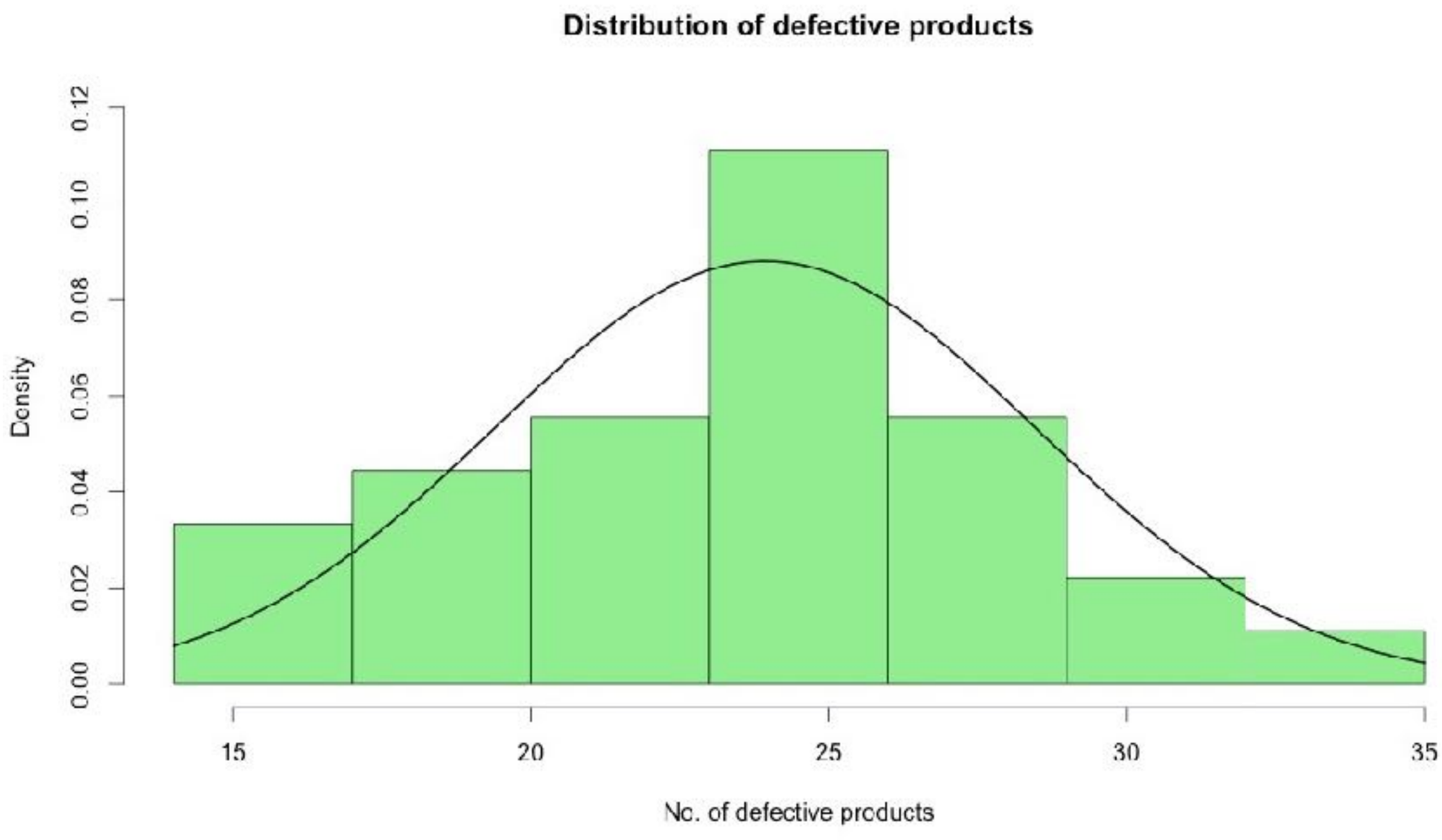


Hands-on 1 Solution: Distribution Analysis

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="No. of defective
products",main="Distribution of defective
products”,col='lightgreen',freq=FALSE,ylim=
c(0,0.12))

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

Hands-on 1 Solution: Distribution Analysis

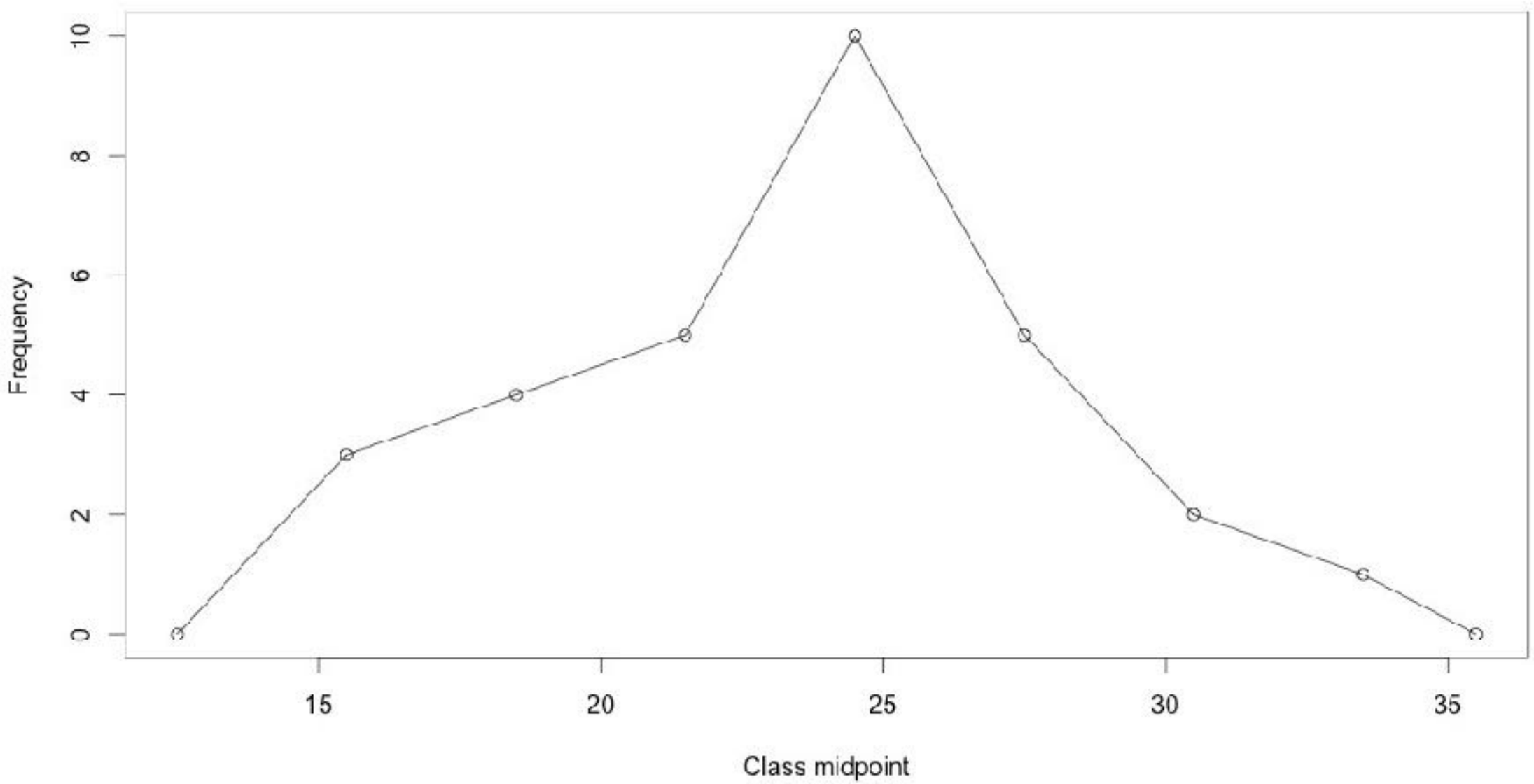


Hands-on 1 Solution: Distribution Analysis

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



Hands-on 2: Barplot

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

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)
```

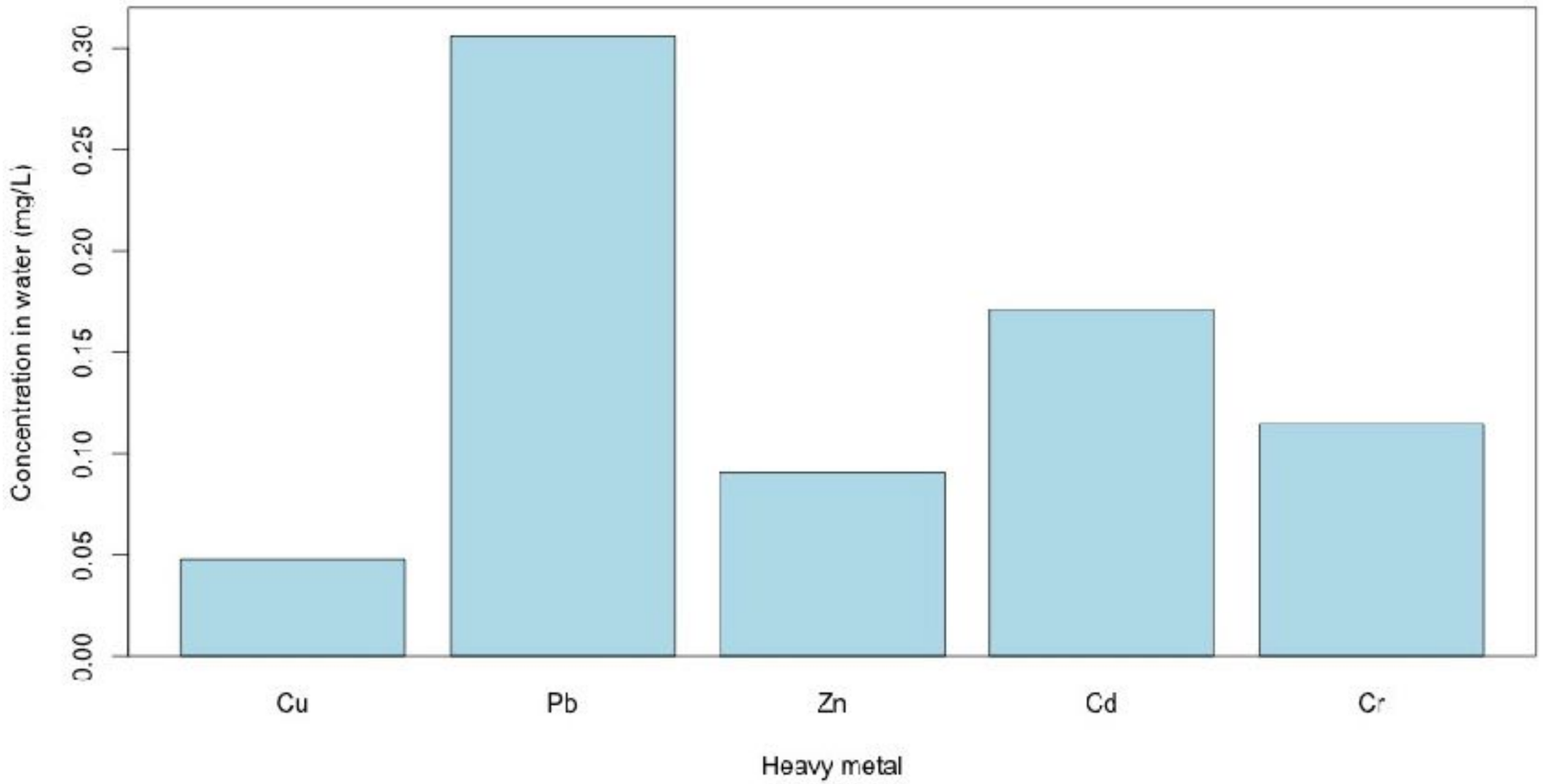
You can rename the headers of the data frame by using the following commands:

```
> 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



Hands-on 3:
Time Series

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

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.*

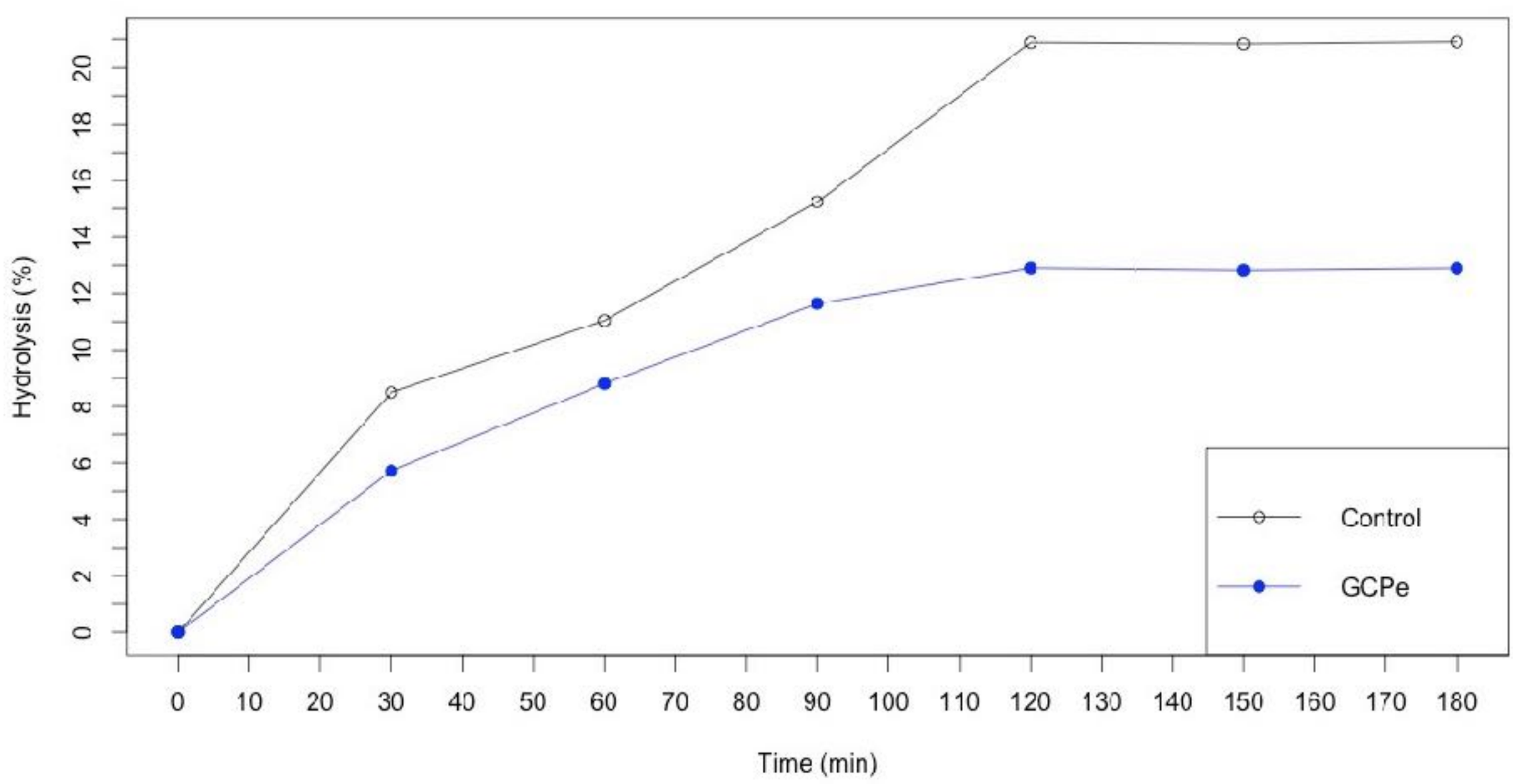
```
> legend("bottomright",c("Control","GCPe"),lty=c(1,1),pch=c(1,19),col=c('black','blue'))
```

```
> axis(1,at=seq(0,200,by=10))
```

```
> axis(2,at=seq(0,30,by=1))
```

```
> box()
```

Hands-on 3 Solution: Time Series



Hands-on 4:
Pie Charts

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

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()`.

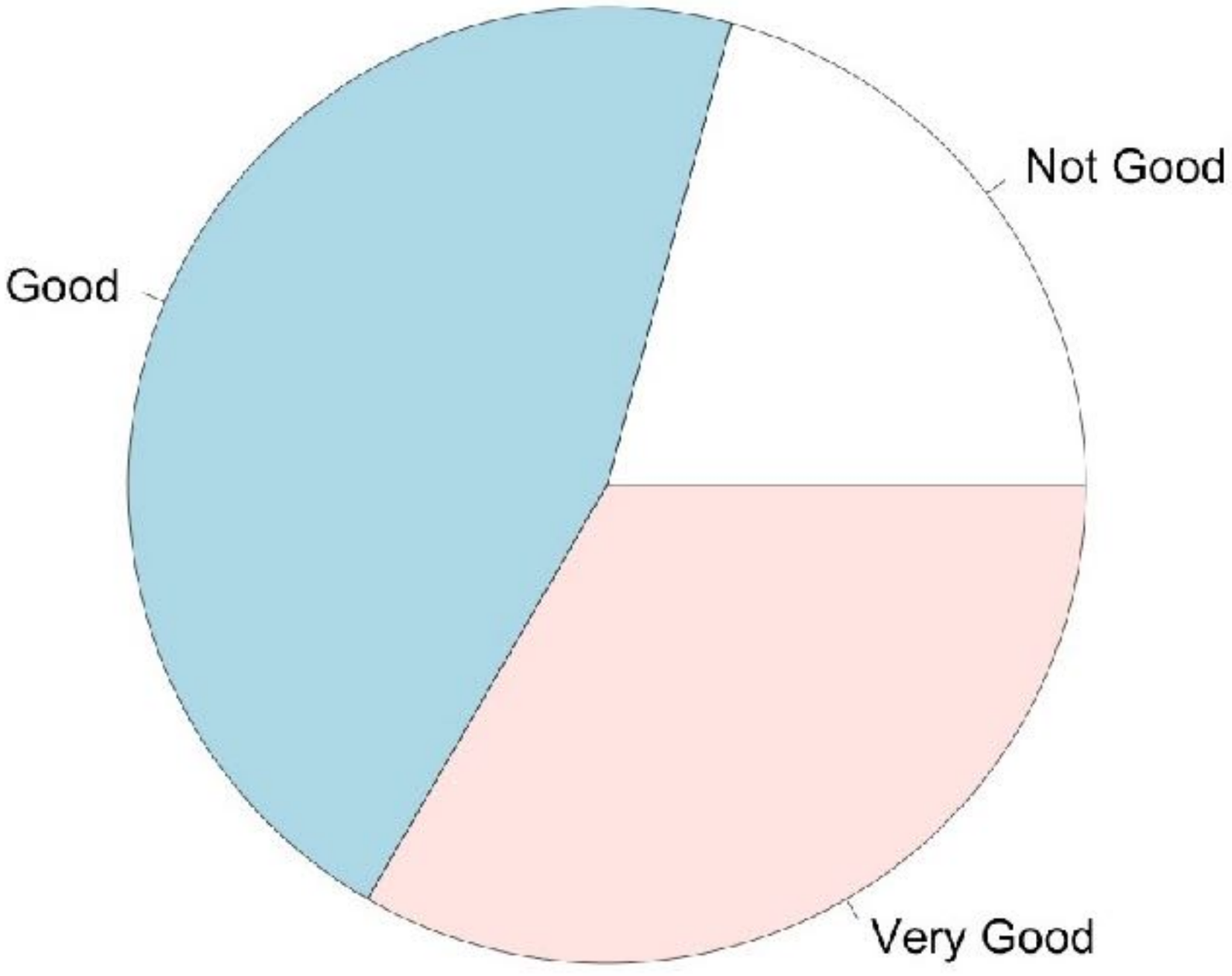
2. Plot the pie chart

```
> 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



Hands-on 5:
Zn in Water - Dot Plot

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

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

Solution:

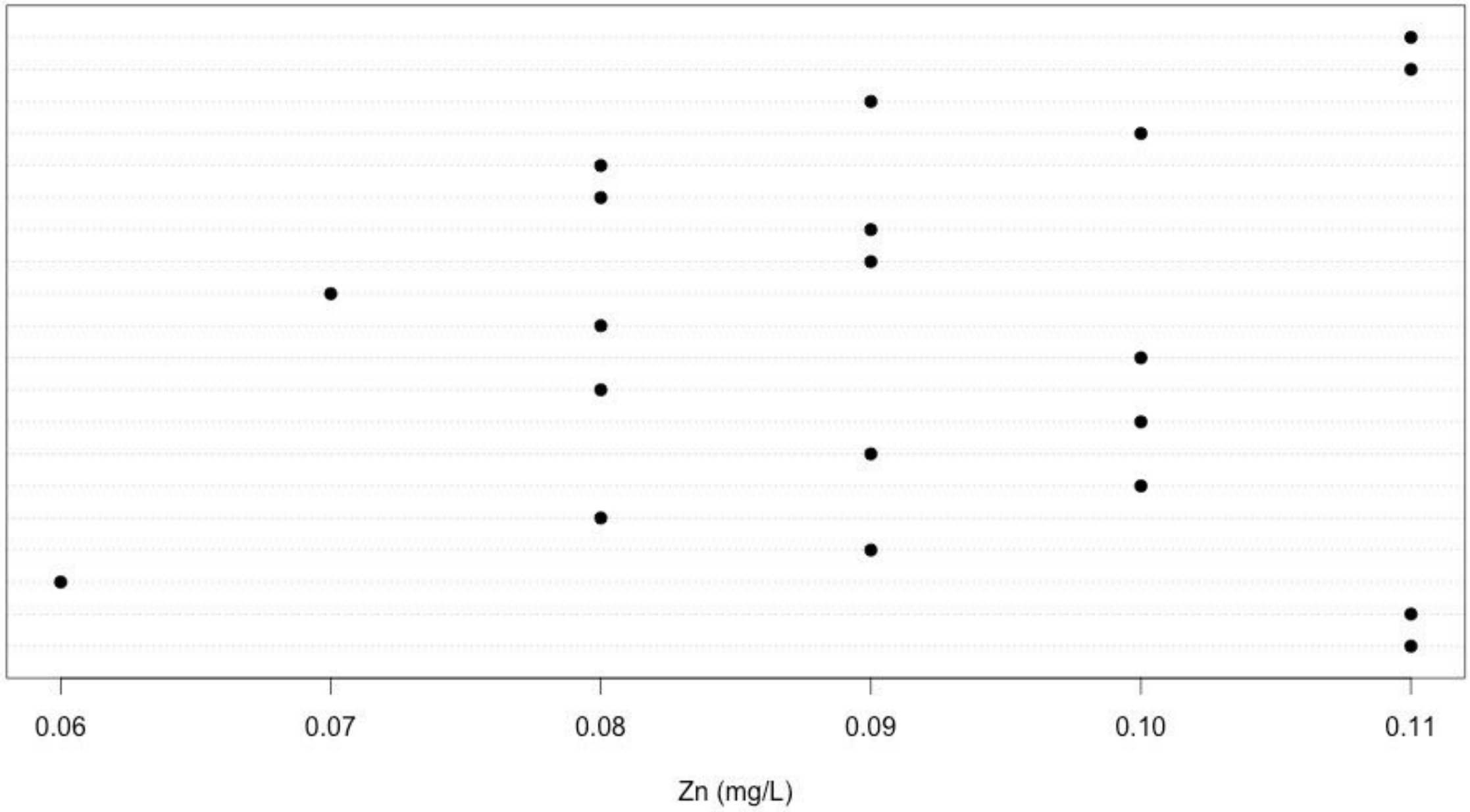
1. Import the data

```
> 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)
```

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



Topic 4: Descriptive statistics

Learning outcome

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

- describe data using descriptive statistics in R.

The arithmetic mean

- 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:

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

is the variable used to represent the data values

The median

- 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.

- 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.

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.

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).

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:

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

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

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:

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

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

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()`

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

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

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

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()`

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()`

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

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

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

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

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: 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.

Correlation

- 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

- 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:

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

- 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

- 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.

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, β_0 represents the Y intercept of the regression equation, β_1 is the slope of the regression equation, and $\boldsymbol{\varepsilon}$ is the error term.

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 Ŷ.

Calculating formula

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

- 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

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:

$$R^2 = \frac{Explained\ \ Variation}{Total\ \ Variation}$$

$$1 \leq R^2 \leq 0$$

Multiple Regression

- 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:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \quad . \; . \; . \quad + b_k X_k$$

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

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

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')
```

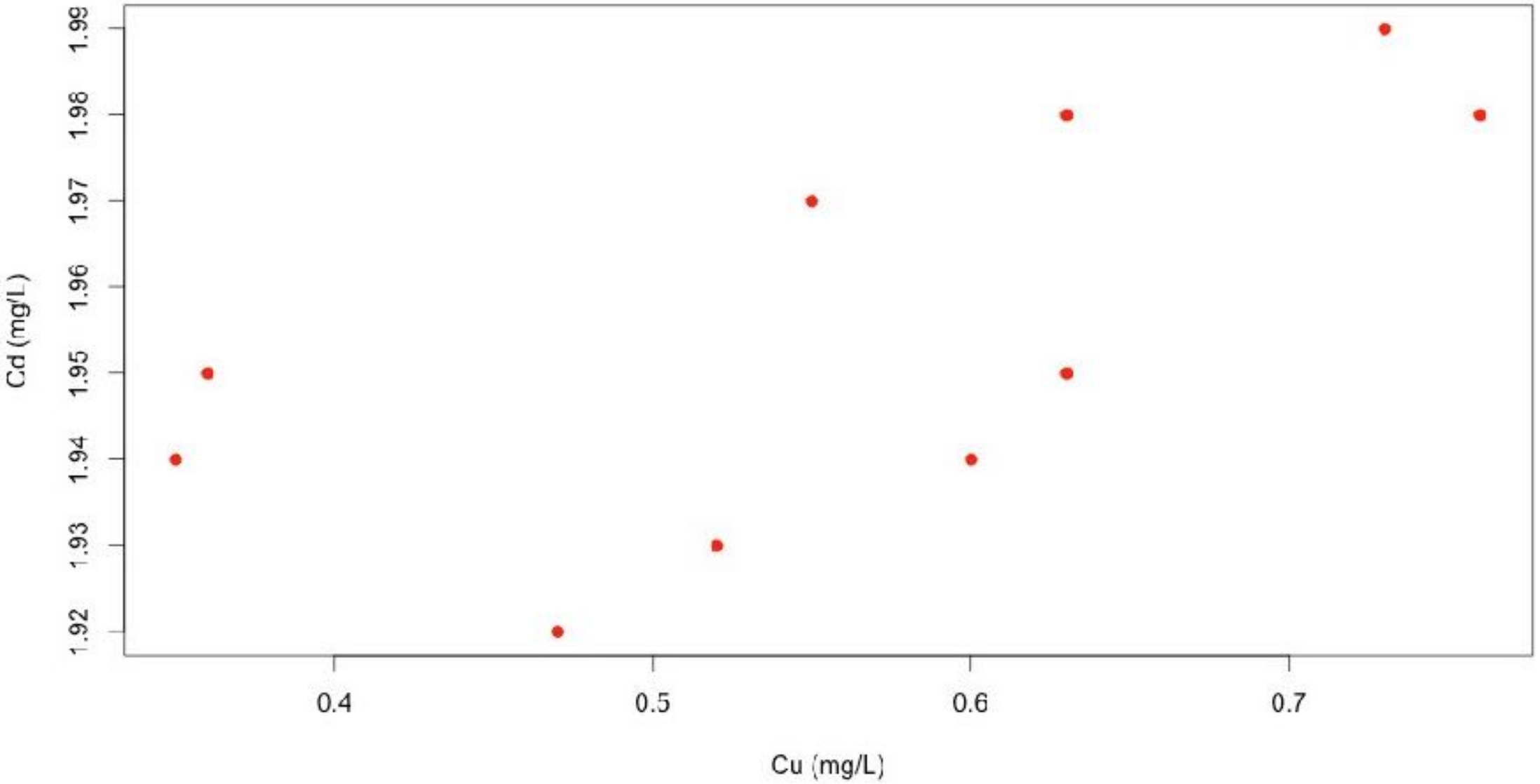
The plot shows a linear relationship between copper and cadmium concentrations.

3. Calculate the correlation coefficient, r

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

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

Hands-on 19 Solution: Correlation coefficient



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

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

Solution:

1. Import the data

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

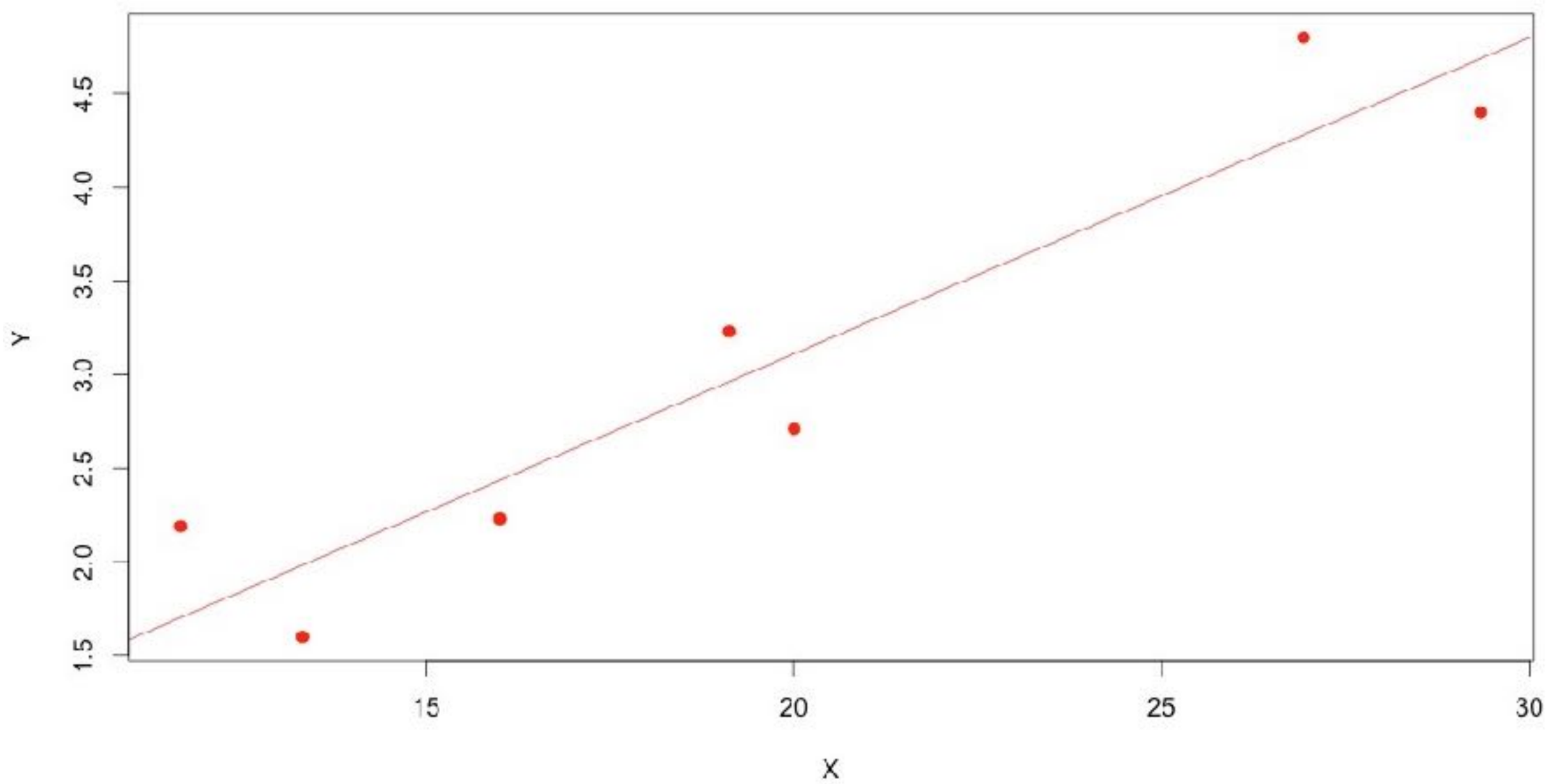
2. Calculate the linear regression coefficient

```
> 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

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



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

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)

Topic 6: ANOVA

Learning outcome

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

- conduct ANOVA.

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.

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.

One-way analysis of variance

- 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

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

- 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		

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 α = 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

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)
```

Hands-on 22 Solution: ANOVA

	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 α = 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

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

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

Interpretation of results:

We always look for significant interactions between factors first. An interaction occurs when the effect of one factor changes for different levels of the other factor.

Since the p-value (< 2e-16) of the interaction between **ratio:ripe** indicates that it is significant, then it cannot be determined if there is a significant difference between means of the individual **ratio** and **ripe** factors. The significant difference between **ratio** and **ripe** factors cannot be analyzed because they are affected by both factors simultaneously. However, from the F value (the highest value), we can make the preliminary assessment that **ripe** is more significantly different (variable) than **ratio**.