Faculty of Science

School of Mathematics and Statistics

MATH1041
Statistics for Life and Social Sciences

Lecture Notes

written by Pierre Lafaye de Micheaux, Jakub Stoklosa

Term 1, 2019

0 Organisation and Aims of MATH1041

- Eveything You Need to Know for a Good Start

1 One-Dimensional Exploratory Analysis

- Lecture 1: Introduction to Data Collection and Organisation
- Lecture 2: Variable Types
- Lecture 3: Numerical Summaries
- Lecture 4: RStudio and Graphical Summaries

0 Organisation and Aims of MATH1041

- Eveything You Need to Know for a Good Start

UNSW
SYDNEY

你好
goeiedag
nyob zoo
buongiorno
alapiale
bok
gamarjoba
привет
manao ahoana
namaskaram
vitaju
goddag
aloha
myttyn da
shalom
guten tag
bonjour
inda wadou
grüß gott
jó napot kívánok
namaste
labdien
ciao
godan dagin
selamat pagi
annyong haseyo
befa
adieu
こんにちは
kia ora
hoeien nuh'n't
guatertag
mingalaba
laba diena
mbolo
tere
bonan tagon
mba'eichapa
dana esteline
sabai di
ani sogomen
ndeewo
hallo
ola
houje
hello
ha
ina kwana
konnichi wa
hafa adai
dobar den
bon dia
idh-iyessbhène
hyvää päivää
salam
kem cho
gütte mòrje
hylo
ne yibeogo
demat
bonjou
ulaakkut
kia bo
namaskar
n'dja arè
moien
goede morgen
dia duit
mari-mari
azul
bondi
kalimera
ka tiana
hey
nliwale
ékooni
salâm
bonzour
mat ni kani
tchoumriepsou
buenos días
mwaramutse
dobardan
bonghjornu
yadra
god dag
osiyo
anissoroma
rojbas
egunon
mirë dita
barev
안녕하세요
mholo
ayo
γειά σου

RO DUC

Breaking the Ice

Please go to:

- http://j.mp/2F82Orq then Additions/Add Marker - Simple.
- https://pollev.com/pierrelafaye259 and take 1 minute to answer (anonymously) each question.

We can analyse such data using methods you will learn in MATH1041. And we can use that to improve the course!

The Agenda Slide

Overarching aim of the course: introducing statistics, the study of collecting, analysing and interpreting data (fundamental to any quantitative research)

First class of Week 1 (21/02/2019 - 2 hours)

Last time:

- Nothing :)

Today:

- Quick introduction to MATH1041 (slides 0.8–0.40)
- What data to collect and organising them in files (slides 1.4–1.16)
- Types of variables (slides 1.20–1.25)

Learning Outcomes for this Lecture

- Understand your responsibilities when undergoing the MATH1041 course, including participation in lectures, in class tutorials, online tutorials and out-of-class activities.
- Learn how to get face-to-face support for the MATH1041 course.
- Understand where you can find key materials for the course: Moodle + Lecture notes + Tutorial exercises + Software + Books.
- Remember that visual clues are used in the slides to help you digest them.
- Understand how your mastering of the course will be assessed.

Should You Be Here?

This course is primarily aimed at students intending to pursue a major in a field involving quantitative research but for which higher level mathematics or statistics is not essential.

Assumed knowledge: A level of knowledge equivalent to achieving a mark of at least 60 in HSC Mathematics is assumed; students who have taken HSC General Mathematics will not have achieved this level. Such students should have completed an appropriate Bridging Course before the commencement of Term 1. Otherwise, they are advised to seek the advice of the First Year Director (A/Prof Jonathan Kress, fy.MathsStats@unsw.edu.au).

Note: This course is **not** intended for students who propose to study a substantial amount of Mathematics beyond first year level.

Who Am I?

My name is **Pierre (P-Air) Lafaye de Micheaux**. I am a Senior Lecturer in the School of Mathematics and Statistics.

Previously, I was working in France and Canada. I have a background in Statistics and in Neuroscience.

As you may have guessed from my accent, my first language is French.

Contact

Pierre (P-Air) Lafaye de Micheaux

E-mail:	lafaye@unsw.edu.au
Telephone:	(00.[+612]) 9385 7029
Office:	2050 (*You are welcome during contact hours**)
Web-page:	https://web.maths.unsw.edu.au/˜lafaye

* Wednesday (Virtual consultation) + Thursday (Office), 4:30–5:30pm

My office is on the second floor of the Red Centre building. See H13 building on this map: http://www.facilities.unsw.edu.au/sites/all/files/KENC_Campus_Dec14.pdf (row H, column 14).

For administrative problems, contact the **Student Services Office** (Mrs Markie Lugton, RC-3072, Fy.mathsstats@unsw.edu.au).

⚠ Only use your official UNSW email address z1234567@unsw.edu.au and quote your student ID.

Who Are You?

How can we learn a bit more about you, a class of hundreds of people?

Answer: Statistics!

But first we need some **data** on the class . . .

We have put a short survey up on `UNSW Moodle` called "A survey about you!".

Thanks to everyone who has completed it! If you have not done the survey yet, please complete it **today or tomorrow**.

We will use these data to illustrate several statistical concepts.

Lectures

I will give 4 theatre lectures per week (total: 4×50 minutes).

Thursday	**14:00 – 16:00**	**Burrows Theatre**	Weeks:1-9
Friday	**14:00 – 16:00**	**Burrows Theatre**	Weeks:1-8,10

⚠ In Week 11, lectures will be given on Tuesday and Wednesday from 14:00 - 16:00.

In case you miss one of my lectures you can listen to the audio recording that will be made available each week on Moodle:

https://moodle.telt.unsw.edu.au/course/view.php?id=37650

Note: there will also be one hour of online lecture activity per week.

Classroom Tutorials

Classroom tutorials: (one per week, **compulsory attendance**)

See http://timetable.unsw.edu.au/2019/MATH1041.html for the room and time.

You should have already met your tutor at the first tutorial session.

Location may change. Check myUNSW before your first tutorial.

Online Tutorials

Online tutorials: one per week, **due each Sunday by 6pm**

The online tutorials will start in Week 2 within a web interface called Maple T.A., accessible (using your ZID/password) at:

https://mapletap.telt.unsw.edu.au:8443/mapleta/

These can be done either in the labs (in your lab time only; see http://timetable.unsw.edu.au/2019/MATH1041.html) or from whatever computer with Internet access. The computer labs (located in the basement, RC-G012, or on the Mezzanine level, RC-M020, of the School of Mathematics and Statistics in the Red Centre) will only be staffed for Weeks 2 and 3.

It is important that you attend your lab next week!

Assumptions About Student's Work

I will assume you have 3 courses per term (which should be a maximum). A standard assumed workload is 40 hours of work per week.

Consequently, I will assume that you can devote 13 hours per week studying for MATH1041. This time will be divided as follows:

- 4 hours of classroom lecture;
- 1 hour of online lecture activity;
- 1 hour of classroom tutorial;
- 1 hour of online tutorial;
- 6 hours of other out-of-class activities (review your lecture slides, attempt the exercises, practice for the exams, do your assignments, etc.).

If you can follow this schedule, I am highly confident that you will succeed this course!

Software to be Used

You will be using RStudio which is an interface to the freely available statistical language and data analysis software R[R Core Team, 2017].

RStudio consultants are available in the computer labs at times to be announced (on Moodle) to help you becoming familiar with RStudio: https://www.maths.unsw.edu.au/currentstudents/maple-lab-consultants

Install R (https://cran.r-project.org) first and then RStudio Desktop (http://www.rstudio.com/products/rstudio/download).

You can also try using your cell phone on one of these pages:

- https://rdrr.io/snippets/
- http://rextester.com/l/r_online_compiler
- https://jupyter.org/try

Books on R

To improve your skills in the **R** language, you can consult one of these ▹books◃ [Lafaye de Micheaux et al., 2013], in English, Chinese, French or Indonesian. This is not mandatory.

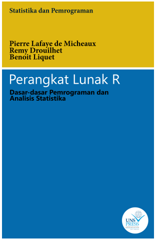
The English version can be downloaded for free here:

- `http://biostatisticien.eu/springeR/Rbook.pdf`

You can also order a (non-free) printed copy here:

- `http://www.springer.com/us/book/9781461490197`
- `https://link.springer.com/book/10.1007%2F978-1-4614-9020-3`

Statistics and Computing
Pierre Lafaye de Micheaux
Rémy Drouilhet
Benoit Liquet
The R Software
Fundamentals of Programming and Statistical Analysis
Springer


Statistika dan Pemrograman
Pierre Lafaye de Micheaux
Remy Drouilhet
Benoit Liquet
Perangkat Lunak R
Dasar-dasar Pemrograman dan Analisis Statistika

Lectures and Tutorials

Role of the course lecturer:

- ▹[McKeachie and Svinicki, 2014]◃ p. 71: “communicate the teacher’s enthusiasm about the subject” (I’ll do my best!);
- emphasise important points in the textbook;
- add supplementary material from other sources;
- raise questions and help students answer them;
- dwell on mathematical theoretical points;
- **teach students how to learn and think**.

Role of the tutors:

- explain the solution of selected exercises, answer questions;
- offer contact hours to answer your questions.

Materials: The Essentials

Course notes: PDF slides created using my **R** package `MATHxxxx` and made available each week on the Moodle website of the course.

Software: R/ `RStudio`.

Tutorial exercise booklet: can be downloaded from Moodle.

▹**Textbook:**◃ Moore D. S., McCabe G. P. and Craig B. A. (2017). *Introduction to the Practice of Statistics*, 9^{th} Ed. [Moore et al., 2017].

The textbook ($156.19) and the tutorial booklet can be purchased from the UNSW book shop `https://www.bookshop.unsw.edu.au`. *See map slide3: E15 bldg (row E, col. 14).*

Visual Clues: Icons

Week 1

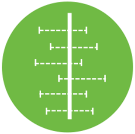
Week 2

Week 3

Week 4

Week 5

Week 6



Week 7

Week 8

Week 9

Week 10

Visual Clues: Icons

Abuse of language

Abuse of notation

Any question?

Will be assessed

Going back

Brainstorming time!

Well done!

Time for a break!

Brick of statistical knowledge

Use your calculator

Visual Clues: Icons

Have a close look

Very tricky!

Do it using your computer!

Give me some feedback

Time for group work

A hint

Learning Outcomes

Take notes

Consult Moodle

Time for a quiz

Read a few pages in the textbook

Visual Clues: Icons

Reflect Write down your reflections

Link theory and practice through research examples

Shiny by RStudio RShiny applet

Too bad, try again

MESSAGE Take home message

🎬 Watch this video

Wall of statistical knowledge (several bricks)

⚠ Warning

Your thoughts

Visual Clues: Colours

A question will be displayed in red.

Definition 0.0 (A definition)

Some content.

This is a **new term** defined directly in the text.

Exercise 0.0 (An exercise)

Some content.

And its Solution:

Example 0.0 (An example)

Some content.

Remark 0.0 (A remark)

Some content.

Visual Clues: Colours

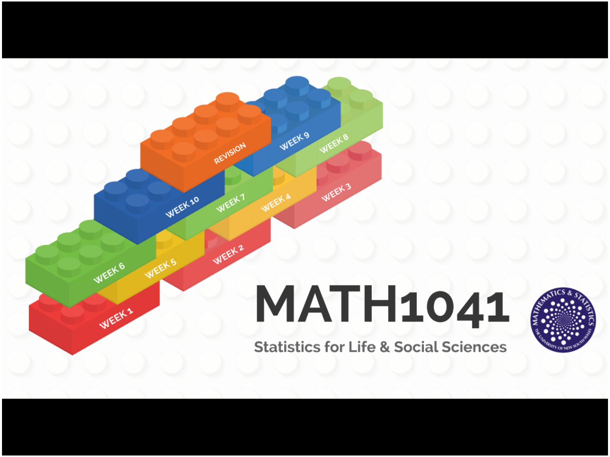
Assumptions are very important in Statistics!

Assumption A

Some content.

Concept Map

This is a diagram of key concepts and their relationships.



https://prezi.com/view/5Eet9eXPLVc6U1Q2yIQk/

Fill the Holes

I will successfully complete MATH1041 with a very good mark!

Text in green will only appear on the slides I will present in class, not on your version of the slides.

I recommend 'Xournal++' to comment PDF files on your laptop/tablet (see https://github.com/xournalpp/xournalpp/releases for Windows or Mac version; to download Linux binaries, see http://ppa.launchpad.net/andreasbutti/xournalpp-master/ubuntu/pool/main/x/xournalpp/). But this will remove interactive features (such as animations) that can only be activated with Acrobat Reader.

Assessments

Task	When Due	Weight	Duration
Online tutorials	Each Sunday, 6pm	10%	1 week
Mid-term test	Week 6	15%	45 minutes
Assignment (project)	Week 9	15%	2 weeks
Final examination	Between 6–18 May	60%	2 hours

Times, locations, etc. will be given closer to test dates.

You can drop courses via `myUNSW` up until the end of the teaching period, but there are implications for your enrolment status, academic record and/or fee or contribution liability, depending on when you drop (e.g., before the *census date*, which in Term 1 is 17 March 2019):

- https://student.unsw.edu.au/enrolment-drop-course

Assessment Details (I)

- The 10 **online tutorials** (10%) are designed to give immediate in-session feedback to students on their progress and mastery of the material. These tutorials are due at the end of each week (Weeks 2–11, Sunday 6 pm) and are done using the MapleTA system which can be accessed via Moodle.

- The **mid-term lab test** (15%) is designed to give students feedback on progress and mastery of the first parts of the course, under exam conditions and to evaluate progress towards the stated learning outcomes. It will be taken in the computer labs. Students will answer some questions on the computer and some others by writing in a booklet.

⚠ *Book a time for your mid-term test by the end of week 3.* (A booking page will be available on Moodle.)

Assessment Details (II)

- The **computing assignment** (15%) (a.k.a., *the Stats Project*) will be made available on Moodle at the end of Week 7, two weeks before it is due for submission. This activity will require the use of `RStudio`.

- The **final examination** (60%) will assess everything. A few practice exams and their solutions will be provided so that you should not be surprised by the format of this examination.

Special Consideration

- If you miss an assessment due to illness or misadventure you must apply for special consideration through `myUNSW` within 3 days.
- You must provide suitable documentation (e.g., a medical certificate) and have it verified at Student Central.
- Weekly Online Tutorials are already very flexible and special consideration for these is usually not granted.
- Additional assessment exams for term 1 2019 will be held during the period 22 May to 2 June.
- For details see
    - Course outline
    - `http://student.unsw.edu.au/special-consideration`
    - www.maths.unsw.edu.au/currentstudents/special-consideration-illness-misadventure

Getting Help

Seek help as soon as you need it!

Help with course work:

1. Ask me directly during the lecture (I will be be very happy to have questions!);
2. I will spend a few minutes outside the classroom after each class, answering specific questions;
3. Ask one of your classmates;
4. Use the Moodle MATH1041 Discussion Forum;
5. Come to my office on Thursdays, $16:30 - 17:30$;
6. Use my online office hour on Wednesdays, $16:30 - 17:30$;
7. Use the Statistics Walk-in Consulting Service: https://www.maths.unsw.edu.au/currentstudents/statistics-consultation-service

Some General Advice to Help you Study

- Practice a sport, or go to the gym, or at least walk
- Sleep enough (7 to 9 hours)
- Drink mainly water, a lot
- Eat well (fruits, vegetables), avoid sugar
- Work using a Pomodoro Technique (https://en.wikipedia.org/wiki/Pomodoro_Technique)
- Use proper revision techniques such as Mind Mapping, etc.: http://biostatisticien.eu/Revision-Techniques-booklet.pdf
- Manage your stress (e.g., meditation, yoga, cardiac coherence)
- First year of university has lots of great opportunities and is an exciting time, but it can be stressful for some students. If your studies are affected by difficulties in your personal life, consider seeking help from the UNSW Counselling Services.

Some Advice for MATH1041

- Make sure you read your student email regularly.
- Check the resources on Moodle
- Try to attend **all** lectures. Be on time.
- Bring the MATH1041 course notes each week (on *Moodle*).
- Take your own notes during the course, in the intended gaps.
- Work **regularly**, **from now on**. Learning is a **cumulative** process and relies **very strongly** on previous elementary bricks of knowledge: Plan – Prepare – Perform!
- Stay focused. Focus and motivation are key to success!
- Ask questions if you don't understand. The probability that "you are not the only one" is equal to 1. And I will be very happy to answer! Speak up and ask questions in your tutorial to get the most out of it.
- Try to form study groups and to work together.

Questions

Thank you to consult the **Course Outline** for more details.

Any question?

Course Aims

MATH1041 provides an introduction to **Statistics**: the science of collecting, organizing, analysing and interpreting numerical facts or qualitative description of objects, which we call **data**.

The goal of Statistics is to learn from data.

What do the data tell me?

Statistics plays a fundamental role in quantitative research (research involving data). Some examples of fields in which quantitative research plays a major role are: psychology, biology, physics, economics, . . .

You will learn statistics by doing. Our motto shall be **Practice!** as advocated by [College Report ASA Revision Committee, 2016].

What is Statistics?

- **A statistic**: a numerical summary of some data
- **Statistics**: the science of collecting, analysing and understanding of data measured with uncertainty.
- **Who needs to know about statistical methods?** Anyone who collects, analyses or wants to understand data in order to answer some research question. (And a lot of people do!)

An Example of a Research Question

Is the flu epidemic worse than before?

Note: This question is deliberately vague. We will get back to it.

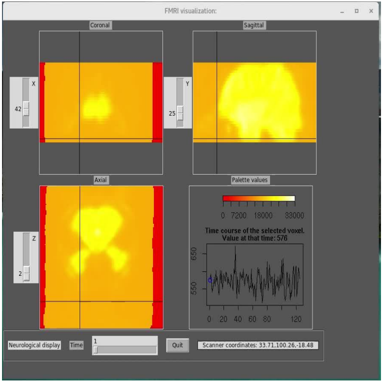
Skills To Be Developed in MATH1041

In this course, you will learn how to approach designing studies and analysing data to answer research questions like the previous one. In particular, at the end of this course, you will be able to:

1. Recognise which analysis procedure is appropriate for a given research problem involving one or two variables.
2. Understand principles of study design.
3. Apply probability theory to practical problems.
4. Interpret computer output for a statistical procedure.
5. Calculate confidence intervals and conduct hypothesis tests by hand for small datasets.
6. Understand the usefulness of Statistics in your professional area.
7. Apply statistical procedures on a computer using **R**/`RStudio`.

A Few Motivating Examples from my Research

An fMRI scanner can produce massive 4D data sets, recording the brain activation in 3D while the subject is performing a cognitive task.



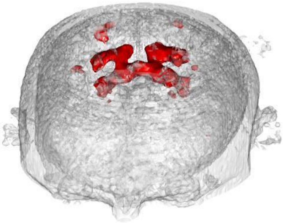
Time course of the selected voxel.
Value at that time: 576

Scanner coordinates: 33.71,100.26,-18.48

The R package `AnalyzeFMRI` can help you read and visualise such data sets.

Can we isolate the groups of brain cells (neurons) that are involved in the treatment of colour stimuli by the human brain?

This would help neurosurgeons to avoid damaging those regions when patients go under surgery for epilepsy so that they won't loose the ability to see colours.

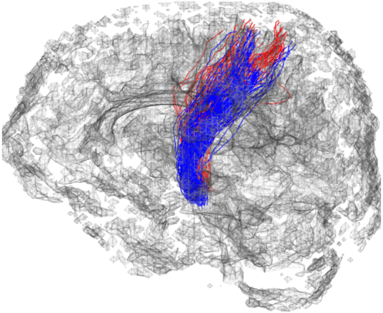


|< < ◁ ▷ > >| − ▸◂ +

What you will learn in MATH1041 will allow you to answer that!

Using Statistics, can we disentangle genetics from environmental factors for the geometry of motorcortex brain fibres?

Answering this question can help doctors to decide to treat a patient with Parkinson's disease using either gene therapy or by modifying his/her lifestyle (food, exercice, alcohool, smoking, etc.).



|◁ < ◁ ▷ > ▷| − ▶◀ +

We need advanced statistical theory that is currently being developed!

Let’s take a 5 minutes break!

1 One-Dimensional Exploratory Analysis

- Lecture 1: Introduction to Data Collection and Organisation
- Lecture 2: Variable Types
- Lecture 3: Numerical Summaries
- Lecture 4: RStudio and Graphical Summaries

Lecture 1: Data Collection and Organisation

1. Introduction to Data Collection and Organisation
2. Variable Types
3. Numerical Summaries
4. RStudio and Graphical Summaries

Relationship to Textbook [Moore et al., 2017]: **Section 1.1** “Data”, page 4–5.

Learning Outcomes for Lecture 1

- Data come with a context and a purpose
- Data sets, files and file format
- Population, cases, labels
- Variables, number of variables
- Sample size

Introduction

Data → Information

Roughly, data are just a bunch of numbers

In the pre-topic lesson, you have encountered several types of data sets. Let's open the '1041-old.csv' data set to realise that it's not easy for us, humans, to make sense of data! (Note: 'csv' stands for 'comma separated values'.)

A major goal of **Statistics** is to make data **informative**.

The Steps Involved in Statistics

- What data to collect?

Depends on what is the research question and who asks it.

- How to collect data in a clever way?

Design of experiments. Computer simulations. (Will be seen later.)

- How to organize your data?

In paper notebooks? files? data bases? even DNA for long term storage!

- How to describe your data?

File format and size. File content. Statistical descriptive summaries.

- How to analyse data?

Relationships, statistical inference. (Will be seen later.)

What Data to Collect: The Flu Example

3 mn

You are asked to study a flu epidemic. Define the population, the data that needs to be collected, the variables, etc.

Use the Think-Pair-Share technique [McKeachie and Svinicki, 2014]. Spend one minute thinking about the question, one minute writing down your thoughts (on the next slide) and one minute to compare what you wrote with your neighbour (ask his/her name first). List your top two variables in the chat thread (PollEveryWhere wordcloud).

Your Thoughts

Data Sets (I)

Pre-topic

A **data set** (or dataset) is a collection of **data** (i.e., numbers, qualifiers, pieces of information). Each value is known as a datum.

In Statistics, data sets usually come from actual **observations** collected on **cases**, obtained by sampling a **population** (of such cases).

Most commonly, a data set corresponds to the contents of a single database table, or a single statistical **rectangular table**. Each row in the table corresponds to the observations (values) of a few **variables** (such as height and weight) on one given element (case) of that population. Each column of the table represents a particular variable.

The data set may comprise data for one or more members, corresponding to the number of rows, and called the **sample size**.

After week 11, your lecturer will have a data set with all your marks. Can you state the population, the cases, the data, the variables and the sample size? (See slide Data Sets (III) for formal definitions.)

Data Sets (II)

Pre-topic

A data set also refers to a computer file having a record organization, namely a standard way that information is encoded for storage (called the **format** of the file).

Many different file formats exist, usually recognised by looking at the **file extension** (e.g., `.png` for images, `.txt` for stream of characters, `.html` for webpages, `.zip` for compressed files, `.xls` for Excel files, etc.).

Look into the Datasets folder in the ”Computing information” section on Moodle. Which formats do you recognise?

In MATH1041, we will mostly work with very simple rectangular data sets, stored in one of the following formats: `TXT`, `DAT`, `CSV`, `XLS`, `XLSX`.

Data Sets (III)

Pre-topic

You should **always** clearly identify the:

- **Population:** the collection of all individuals or items or objects under consideration in a statistical study, usually determined by what we want to know.
- **Cases:** the members, objects, units, subjects or individuals from the population, from which information (i.e., data) is obtained. Together they form the sample.
- **Labels:** the identification code of each individual.
- **Sample size** n: number of cases/observations in the sample.
- **Variable:** a characteristic of the cases that can be measured, collected, recorded or counted.
- **Number of variables** p: the total number of variables recorded, measured or collected.

And, most importantly, why do we have/need this specific data set? Do we hope to answer specific questions?

Class Survey Data

We are in the process of collecting some data on you via a student survey on Moodle. We did the same thing a few years ago on previous MATH1041 students. We asked 400 students a number of different questions, such as:

- their gender;
- their mode of transport to UNSW;
- satisfaction score of UNSW;
- amount of money spent on a hair cut;
- etc.

Let's open 'RStudio' and load the '1041-old.RData' file. Clicking on 'survey.df' in the Environment tab on the right enables us to display the content of this file.

We will get back to all kind of data collection mechanisms in Week 3.

File and Data Set Description

Describe the file containing the Class Survey Data (i.e., its format and size). Describe its content (population, cases, labels, variables, number of variables, sample size).

This question is related to concepts you have learned in the pre-topic lesson.

Answer

The RData format (usually with extension .RData, .rdata or .rda) is a format designed for use with **R**, for storing a complete **R** workspace or selected "objects" from a workspace in a form that can be loaded back by **R**. (source: https://www.loc.gov). The file ‘1041.RData‘ is stored in a **binary compressed format**. Its **size** is 321.4 KB.

If we click on ‘1041.RData‘ from the Files tab of RStudio, the ‘survey.df‘ data frame appears in the Global Environment. We can also see that there are "400 obs. of 20 variables".

The **population** is the set of all MATH1041 students that specific year. The **cases** are each individual student within this population. For privacy reasons, the **labels** (i.e., names of the students here) are not given, though we could have used codes for anonymization The **variables** are ‘gender‘, ‘hair.cost‘, ..., ‘labour.cost‘. There are $p = 20$ variables and the **sample size** is $n = 400$.

Class Survey Data

Remember that in the flu example, you were asked what kind of data you would need to collect. Here, it is the other way around: we assume you have access to a given data set. (Note there are tons of freely available data sets on the Internet now).

What kind of questions can we hope to answer with the Class Survey Data?

Use the Think-pair-share technique to brainstorm on this for 3 minutes with your neighbour. Write your thoughts on the next slide.

Your Thoughts

☕ Let’s take a 5 minutes break!

Lecture 2: Variable Types

1. Introduction to Data Collection and Organisation
2. Variable Types
3. Numerical Summaries
4. RStudio and Graphical Summaries

Relationship to Textbook [Moore et al., 2017]: **Section 1.1** “Data”, page 3–4.

Learning Outcomes for Lecture 2

- Categorical variables
- Quantitative variables
- Units of measurement

Quantitative or Categorical?

Definition 1.1 (Type of a variable)

The **type of a variable** is either categorical or quantitative. A **categorical** variable places an individual into one of several **categories**. A **quantitative** variable takes numerical **values**, measured on a scale.

Example 1.1

Variable ‘Age‘ is a quantitative variable whereas variable ‘Gender‘ is a categorical variable.

Note: for a quantitative variable, it is important to give the **units of measurements**. For example, the units of Age are years.

Knowing the type of a variable will help us to choose the right statistical technique.

Quantitative or Categorical?

Which of the following variables are quantitative, and which are categorical?

- satisfaction with UNSW (from 0 to 10)
- time travelling to UNSW
- method of travelling to UNSW

QUIZ Give your answer online: https://pollev.com/pierrelafaye259

Answer

- Satisfaction score with UNSW (from 0 to 10): a quantitative variable.
- Time spent travelling to UNSW: a quantitative variable.
- Method of travelling to UNSW: a categorical variable.

The Flu Example

3 mn

What are the types of the variables you chose to collect (in Lecture 1) to study the flu epidemic.

Use the Think-pair-share technique to brainstorm on this for 3 minutes.

Your Thoughts

Preparation for Our Next Lecture

1. Please complete the short introduction to `RStudio` online module.
2. If you can, bring your laptop to class (not mandatory though).
3. Read textbook pages xix–xxi (6 mn) and 2–7 (9 mn).

That's it for today's lectures! See you tomorrow :)

The Agenda Slide

Overarching aim of the course: introducing statistics, the study of collecting, analysing and interpreting data (fundamental to any quantitative research)

Second class of Week 1 (22/02/2019 - 2 hours)

Last time:

- Data collection and organisation
- Variable types (categorical, quantitative)

Today:

- Numerical summaries
- Graphical summaries

For next time:

- Read appropriates pages in the textbook
- Do your pre-topic lesson (on Moodle)

Lecture 3: Numerical Summaries

This lecture, we will meet common types of numerical summaries of data – ways of summarising the key properties of data using a few numbers.

1. Introduction to Data Collection and Organisation
2. Variable Types
3. Numerical Summaries
4. RStudio and Graphical Summaries

Relationship to Textbook [Moore et al., 2017]: **Section 1.3** “Describing Distributions with Numbers”, pages 27–40.

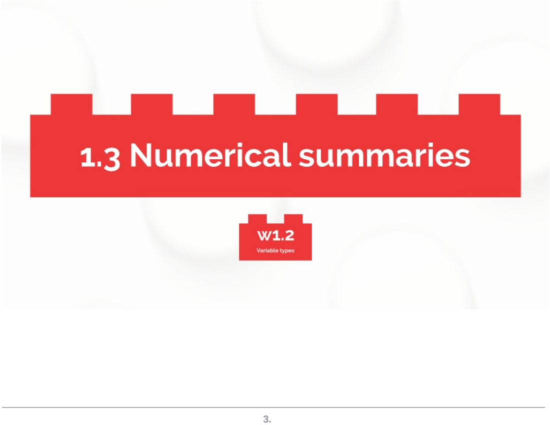
Learning Outcomes for Lecture 3

- Describe the main features of a set of data: statistical distribution of a variable, counts, percent/proportion
- Measures of the center of a distribution: mean, median
- Measures of spread: variability, standard deviation
- Outliers, percentiles and five number summaries

Data analysis for one or two variables

	one variable		**two variables**		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary			
useful graphs:					

Concept Map



1.3 Numerical summaries

W1.2

Variable types

3.

https://prezi.com/view/5Eet9eXPLVc6U1Q2yIQk/

Exploratory Analysis

Until now, we have described characteristics of a dataset and its main constituents. Let's move to a description of the data themselves!

Exploratory analysis consists of **describing the main features** of the data in a dataset. This description can be done by providing numerical summaries of the **distribution** of the variables involved, such as:

- **proportions** or **percentages**
- **mean** or **average**
- **median**
- **interquartile range (IQR)**
- **standard deviation**

The numbers are aids to understanding, not "the answer" per se.

Exploratory Analysis

Which numerical summary to use depends on:

- Whether you are summarising **one variable** or looking at the relationship between **two variables**.
- Whether the variables are **quantitative** or **categorical** (qualitative).

In today's class we will focus on numerical summaries when we have **only one variable**.

Statistical Distribution of a Variable

Pre-topic

Definition 1.2 (Statistical distribution)

The statistical distribution of a variable is :

1. the set of **possible** **values** of that variable;
2. **together with** their **counts** or, for quantitative variables, the counts of the values falling in some predefined ranges.

The distribution of a categorical variable is a **table of frequencies**:

1. list of the possible categories (even those not observed);
2. together with the count, percent or proportion of cases in each category. **Note:** % = proportion × 100

What is the statistical distribution of the variable ‘gender‘ in the ‘1041-old.csv‘ file (on Moodle)?

We will discuss the distribution of a quantitative variable later.

Statistical Distribution of a Variable

1 mn

Why do we need to compute the statistical distribution of a variable? When we replace some data by their statistical distribution, what do we loose? Is it problematic?

Use the ”minute paper” strategy, which consists in spending 1 minute to **write** *your* answer to this question. This is a very valuable tool since ”active thinking is vital for effective learning” [McKeachie and Svinicki, 2014] p. 70.

Distribution of a Variable: Your thoughts

We loose the information about the labels of the cases (e.g., who they are), or even the individual values for quantitative variables. But we end up with a very concise and informative summary of our data!

Note that if we shuffle the data, the distribution computed with these new data will remain exactly the same. Also all numerical summaries computed thereafter will remain the same.

Distribution of Mode of Transport

Mode	frequency	percentage
train/bus	146	36.50
bus	94	23.50
walking	61	15.25
car/train/bus	46	11.50
car	39	9.75
other	8	2.00
motorbike	4	1.00
bicycle	2	0.50

Australian Open Women Tennis Final

Students were asked if they watched the Australian Open Women Tennis Final (yes/no).

Table 1.5: Distribution of `Did you watch?`

Did you watch?	frequency	percentage
Yes	49	12.25
No	351	87.75

Numerical Summary of Gender

Consider the data from the class survey:

- Is gender a quantitative or categorical variable?
- What type of numerical summary would you use for the gender of MATH1041 students?

Give your answer online: https://pollev.com/pierrelafaye259

Answer

- Is gender a quantitative or categorical variable?

A categorical variable

- What type of numerical summary would you use for the gender of MATH1041 students?

A table of frequencies, which is computed in **R** with its `table()` function.

```
load("Data/1041.RData")
table(gender)
```

```
#> gender
#>   1   2
#> 160 240
```

Recommended Numerical Summary

If you want to (fully) summarise **one categorical variable**, use a:

table of frequencies or percentages

Summarising Quantitative Variables

1 mn

Is it a good idea to compute a table of frequencies or percentages for a **quantitative** variable? Why? Why not?

Take one minute to discuss that question with your neighbour.

Answer

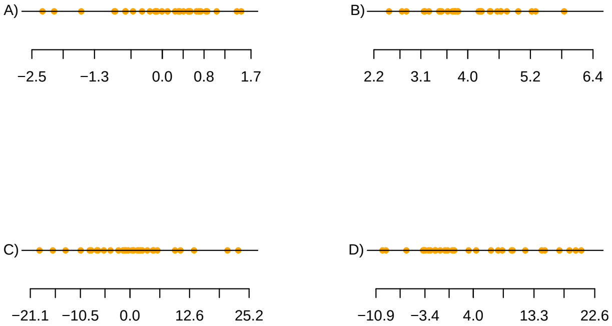
A quantitative variable may take an infinite number of different possible values. It is not possible to list all these values in a table!

So what should we do?

Summarising Quantitative Variables

2 mn

What do you observe? Try to describe / summarise / compare these four different data sets. (**Pay attention to the x-axis values.**)



Take two minutes to write your description.

Your Summary of These Four Data Sets

Not easy!

A): Most points seem close to 0, with points on each side. Roughly, points are between -2 and 2.

B): Most points seem close to 4, with points on each side. Roughly, points are between 2 and 6.

C): Most points seem close to 0, with points on each side. Roughly, points are between -21 and 25.

D): Most points seem close to 4, with points on each side. Roughly, points are between -11 and 22.

We need some consistent way to describe such data sets!

Measures of Location

Given measurements of a quantitative variable, an obvious question is

How large (or small) are the values?

Measures of **location** tell us how large (or small) the typical value is.

A useful measure for location could be the **mean**. For example, the mean satisfaction rating with UNSW was:

$$7.84$$

How was this value calculated?

The Mean

Definition 1.3 (Mean)

To find the **mean** of a set of observations, add their values and divide by the number of observations.

Mathematically, let $\boldsymbol{x} = \{x_1, x_2, x_3, \ldots, x_n\}$ be the n observations (the data points), the mean is calculated as:

$$\mathrm{mean}(\boldsymbol{x}) = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Textbook notation: the textbook refers to the mean as $\bar{x}$.

The mean is just another name for what is commonly called the **average** value of a set of numbers (in Statistics, the set of these numbers is the data set).

Example – Average Income

3 mn

A small company employs **four** young engineers, who each earn \$70,000, **two** senior engineers, who each earn \$88,000 and the owner (also an engineer), who gets \$160,000. The latter claims that on average, the company pays \$88,000 to its engineers and, hence, is a good place to work.

Do you agree with this claim? If no, why?

Use the Think-Pair-Share technique to brainstorm for 3 minutes.

Your Thoughts

The Median – An Alternative to the Mean

The mean of the seven salaries is indeed

$$\frac{4 \times 70,000 + 2 \times 88,000 + 160,000}{7} = \$88,000$$

but this hardly describes the situation.

The mean can be heavily influenced by outliers.

On the other hand, the **median** is the **middle** observation in the data set. This value is $\$70,000$, a much better representation of what an employed engineer earns with the firm.

How do we compute a median?

Definition of the Median

Definition 1.4 (Median)

The **median** is the “middle value”.

For n values **sorted** as $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, the median is computed as:

- $\text{median}(\boldsymbol{x}) = x_{(n+1)/2}$ if n is odd; and
- $\text{median}(\boldsymbol{x})$ equals the average of $x_{(n/2)}$ and $x_{(\frac{n}{2}+1)}$ if n is even.

Textbook notation: the textbook refers to the median as M.

We try to find a value M such that 50% of the observations are below M while 50% are above M.

M is **not** equal to $\frac{n+1}{2}$.

Computing the Median for UNSW Satisfaction

A **subset** of the full data values of UNSW satisfaction ratings led to the following 26 values (already sorted).

5, 5, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8,
8, 8, 8, 9, 9, 9, 9

What is the median of these data?

Give your answer online: https://pollev.com/pierrelafaye259

Answer

First, notice that the number of observations is even ($n = 26$).

We need to use the average of $x_{(n/2)}$ and $x_{(n/2+1)}$.

We compute ($n/2 = 26/2 = 13$) and ($n/2 + 1 = 26/2 + 1 = 14$).

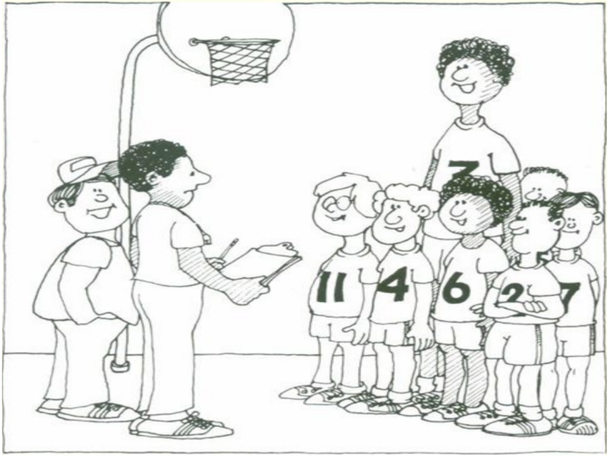
So the Median is the average of the 13th and 14th observations.

Since $x_{(13)} = 7$ and $x_{(14)} = 8$, we have $M = (7 + 8)/2 = 7.5$

Using R/RStudio

```r
subset <- c(5, 5, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9)
median(subset)
```

```
#> [1] 7.5
```



Should we scare the opposition by announcing our mean height or lull them by announcing our median height?

Mean Versus Median

Outliers are unusually large (or small) data values that tend to be quite far away from where the bulk of the data is contained.

The **mean** can be grossly affected by outliers, whereas the **median** is **robust** (resistent) to outliers.

We will come back to dealing with outliers later.

Mean Versus Median

	mean	median
UNSW satisf. (full data set)	7.84	8
Travel time	51.99 min	50 min
Hairdresser labour cost	$842,655.9	$50

Using R/RStudio

```r
load("Data/1041.RData")
c(mean(unsw.satisf), median(unsw.satisf))
```

```
#> [1] 7.84 8.00
```

```r
c(mean(travel.time), median(travel.time))
```

```
#> [1] 51.99 50.00
```

```r
c(mean(labour.cost), median(labour.cost))
```

```
#> [1] 842655.9     50.0
```

RShiny Applet

`https://math1041.teaching.unsw.edu.au`

Use the above web applet to investigate the effect (on the mean and the median) of moving one observation far away from the others. Note that you can drag a point with your mouse.

The Quartiles Q_1 and Q_3

We can further examine our data by looking at the medians of the top and bottom halfs of the data.

These measures are known as the **first** and **third quartiles**.

Definition 1.5 (Q_1 and Q_3)

- The first quartile Q_1 is the median of the observations whose position in the ordered data are to the left of location of the overall median.
- The third quartile Q_3 is the median of the observations whose position in the ordered data are to the right of location of the overall median.

Note: The median can be thought of as the second quartile, $M = Q_2$.

Find the median M. Then use the same procedure once for the set of numbers to the left of M, and once for those on the right of M.

Computing Q_1 and Q_3 for UNSW satisfaction

Again, recall the UNSW satisfaction subset from a few slides ago:

`5, 5, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9`

What are Q_1 and Q_3 for these data?

Give your answer online: https://pollev.com/pierrelafaye259

Answer

As there are 26 observations in this subset, then half the data will consists of 13 observations. So for Q_1 we use the first 13 data observations, and for Q_3 we use the last 13 data observations.

For one half of the data, our new n will be equal to 13, since this is an odd number we use:

$$Q_1 = x_{((13+1)/2)} = x_{(7)} = 7$$

and

$$Q_3 = x_{(13+(13+1)/2)} = x_{(20)} = 8.$$

Notice that for Q_3 we have 13 + (13+1)/2, the 13 being added because we need to include the second half of the data.

Using R/RStudio

```r
subset <- c(5, 5, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9)
quantile(subset, c(0.25, 0.75))
```

```
#> 25% 75%
#>   7   8
```

Note: if n is odd, e.g., $n = 11$, then the first half of the data will contain 6 obs. and the second half of the data also 6 obs. The sixth obs. (which is the median) is used twice to compute both Q_1 and Q_3.

Several rules exist for calculating quartiles. Just report the values that your software gives.

Measures of Spread

A measure of location (such as the mean or median) alone can be misleading.

For example, two countries with the same median family income are very different if one has extremes of wealth and poverty, and the other has little variation among families.

So in addition to reporting the location of our data we should also report the **spread** of our data.

IQR – A Measure of Spread.

A simple measure of spread is the **interquartile range**.

Definition 1.6 (Interquartile range)

$$Q_3 - Q_1 = \text{interquartile range} = \text{IQR}$$

Example – IQR for UNSW Satisfaction

For the 26 UNSW satisfaction scores, recall that $Q_1 = 7$ and $Q_3 = 8$.

This means that the interquartile range is:

$$\text{IQR} = Q_3 - Q_1 = 8 - 7 = 1.$$

With **R**:

```r
IQR(subset)
```

```
#> [1] 1
```

Standard Deviation – Another Measure of Spread

Another measure of spread is the **standard deviation**.

Definition 1.7 (Standard deviation)

The standard deviation $\mathrm{sd}(\boldsymbol{x})$ is computed using the formula:

$$\sqrt{\frac{(x_1 - \text{mean}(x))^2 + (x_2 - \text{mean}(x))^2 + \cdots + (x_n - \text{mean}(x))^2}{n-1}}$$

Note: The **variance** is another useful measure of spread. **To find the variance we square the standard deviation**. We will come back to the variance later on in the course.

Textbook notation: the textbook refers to the standard deviation as s and the variance as s^2.

Calculating standard deviations (and variances) using the above formula can be very tedious by hand, especially if we have lots of data. In MATH1041 we will use `RStudio` to find standard deviations and variances.

For small data sets, it is possible (but a bit tedious) to calculate the standard deviation using the "statistics mode" on your calculator. There is a standard deviation button: usually $\boxed{\sigma_{n-1}}$ or $\boxed{s_x}$.

Standard Deviation for UNSW Satisfaction

Recall our subset on satisfaction with UNSW:

$$n = 26$$

$$x_1 = 5,\ x_2 = 5,\ x_3 = 6,\ x_4 = 7, \ldots, x_{26} = 9.$$

You could use your calculator to show that for this dataset,

$$\mathrm{mean}(\boldsymbol{x}) \simeq 7.46$$

and that the standard deviation for this data set is 1.067 (to 3 decimal places):

$$\mathrm{sd}(\boldsymbol{x}) = \sqrt{\frac{(5-7.46)^2 + (5-7.46)^2 + \cdots + (9-7.46)^2}{25}} \simeq 1.067.$$

```
c(mean(subset), sd(subset))
```

```
#> [1] 7.461538 1.066987
```

Note: this will be explored in tutorials.

IQR versus Standard Deviation

	IQR	Standard Deviation
UNSW satisf. (full data set)	2	1.3
Travel time	60 min	37.81 min
Labour cost	$63.75	$7,632,285

Standard Deviation is Influenced by Outliers

For the labour cost example we get the following standard deviation:

$$\$7,632,285$$

But if outliers are removed from the data (say, values greater than $\$1,000$) then we get:

$$\$94.7563 \approx 95$$

which is about $80,000$ times smaller than the above!

```
sd(labour.cost)
```

```
#> [1] 7632285
```

```
sd(labour.cost[labour.cost < 1000])
```

```
#> [1] 94.7563
```

IQR is Hardly Affected by Outliers

For the labour cost data we get an IQR of:

$$\text{IQR} = \$63.75$$

With the outliers (values greater than $\$1,000$) omitted we get an IQR of:

$$\text{IQR} = \$45$$

The difference between these IQR values is much smaller compared to the standard deviation (see previous slide).

```
IQR(labour.cost)
```

```
#> [1] 63.75
```

```
IQR(labour.cost[labour.cost < 1000])
```

```
#> [1] 45
```

Shiny by RStudio https://math1041.teaching.unsw.edu.au

Recommended Numerical Summaries

If you want to summarise **one quantitative variable**, use:

Measures of:	location	spread
Commonly used:	**mean**	**standard deviation**
Robust to outliers:	**median** (M)	**interquartile range** (IQR)

Five-Number Summaries

We can obtain a nice summary of the data by compiling some of the measures that we encountered.

Textbook advocates the **five-number summary** as:

Min.	Q_1	M	Q_3	Max.

where **Min.** and **Max.** are the smallest and largest values in the data set.

Example – Five-Number Summary

For the subset of 26 UNSW satisfaction scores, the five-number summary is:

Min.	Q_1	M	Q_3	Max.
5	7	7.5	8	9

Think about how long these values took us to calculate. Now let's use RStudio to find a **six-number summary**:

```
summary(subset)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
#>   5.000   7.000   7.500   7.462   8.000   9.000 
```

R even gives you the mean!

Five-Number Summary for Travel Times (in mn)

We can easily do this for another variable:

Min.	Q_1	M	Q_3	Max.
0	20	50	80	300

```r
load("Data/1041.RData")
summary(travel.time)
```

```
#>     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     0.00   20.00   50.00   51.99   80.00  300.00
```

What Did We Learn?

- Summarise a set of data by a small number of numerical values.
- Type of a variable determines the choice of numerical summary to use
- Statistical distribution of a variable as a concise but almost complete summary of data
- Numerical exploratory analysis (a.k.a., numerical summary) on one variable
- Calculating the location and spread of data using numerical summaries
- Limitations of some of these statistical tools

The Keywords Slide

- Location and spread
- Mean and median
- Q_1 and Q_3
- IQR and standard deviation
- Five-number summary

Let's take a 10 minutes break!

Lecture 4: RStudio and Graphical Summaries

This lecture, we will introduce `RStudio` and meet common graphs used for visualising data.

1. Introduction to Data Collection and Organisation
2. Variable Types
3. Numerical Summaries
4. RStudio and Graphical Summaries

Relationship to Textbook [Moore et al., 2017]: **Section 1.2** "Describing Distributions with Graphs", pages 8–23.

Learning Outcomes for Lecture 4

- `RStudio`
- Recommended graphical tools: histograms and boxplots.
- Discuss: tails, extreme values, overall pattern, shape.
- Mention other plots: bar graph, pie chart, stemplot, time plots.

RStudio

Pre-topic

Most statistical procedures are most easily implemented using a computer, and a **statistical package** specially developed for data analysis.

To conduct all of our analysis in MATH1041, we will use a very well-known and popular Statistics program called `RStudio`.

To learn more about **R** and `RStudio`, you can watch several videos (see next slide) from "MarinStatsLectures":

`https://www.youtube.com/user/marinstatlectures/featured`

RStudio

Pre-topic

Watch these videos. Install **R** and ‘RStudio‘ and play with them.

- *What is RStudio* (5 mn): https://www.youtube.com/watch?v=riONFzJdXcs
- *Installing R and RStudio* (5 mn): https://www.youtube.com/watch?v=cX532N_XLIs
- *Getting started with* **R**, *Part I* (8 mn): https://www.youtube.com/watch?v=UYclmg1_KLk
- *Importing data from Excel into* **R** (7 mn): https://www.youtube.com/watch?v=qPk0YEKhqB8

Other Statistics Packages

Pre-topic

Some other programs/software used in Statistics:

- SAS
- SPSS (PASW)
- Excel
- Minitab
- S+ (S-PLUS)

These are non-free and will not be used in MATH1041. But you might encounter one of them further in your studies or in your professional life.

The Role of Graphs

Graphing data is a key step in analyses – it is important to use the appropriate graph(s) for your situation!

Recall that

Data → Information

where data are just a bunch of numbers.

A major goal of **Statistics** is to make them **informative**.

It is easy with a graph to describe / convey the main features of a set of data.

Tools for Making Data Informative

- Summary measures (previous lecture).
- **Graphical tools** (this lecture).

Why Should We Use Graphs?

```
#>   [1] -3.626454 -2.816357 -3.835629 -1.404719 -2.670492 -3.820468
#>   [7] -2.512571 -2.261675 -2.424219 -3.305388 -1.488219 -2.610157
#>  [13] -3.621241 -5.214700 -1.875069 -3.044934 -3.016190 -2.056164
#>  [19] -2.178779 -2.406099 -2.081023 -2.217864 -2.925435 -4.989352
#>  [25] -2.380174 -3.056129 -3.155796 -4.470752 -3.478150 -2.582058
#>  [31] -1.641320 -3.102788 -2.612328 -3.053805 -4.377060 -3.414995
#>  [37] -3.394290 -3.059313 -1.899975 -2.236824 -3.164524 -3.253362
#>  [43] -2.303037 -2.443337 -3.688756 -3.707495 -2.635418 -2.231467
#>  [49] -3.112346 -2.118892  3.398106  2.387974  3.341120  1.870637
#>  [55]  4.433024  4.980400  2.632779  1.955865  3.569720  2.864945
#>  [61]  5.401618  2.960760  3.689739  3.028002  2.256727  3.188792
#>  [67]  1.195041  4.465555  3.153253  5.172612  3.475510  2.290054
#>  [73]  3.610726  2.065902  1.746367  3.291446  2.556708  3.001105
#>  [79]  3.074341  2.410479  2.431331  2.864821  4.178087  1.476433
#>  [85]  3.593946  3.332950  4.063100  2.695816  3.370019  3.267099
#>  [91]  2.457480  4.207868  4.160403  3.700214  4.586833  3.558486
#>  [97]  1.723408  2.426735  1.775387  2.526599
```

Using RStudio, compute the median of these data. First, you will download the data set artificialdata.csv from Moodle. Next, to read the data, you can use this instruction:

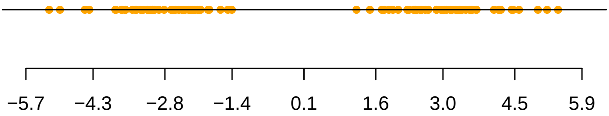
```
x <- read.csv(file = "artificialdata.csv")
```

Is the median a good summary / representative of this data set? If not, why? (Hint: Look at the sorted data: **R** function sort().)

Why Should We Use Graphs?

You should realise that you summarised all the data with just one number, which you expect to be the best representative of this data set. How many cases have a value close to the median? None!

Let's plot these data.



Take home message: always plot your data!

Note: a simple **R** instruction to use: `plot(x, rep(1, length(x)))`

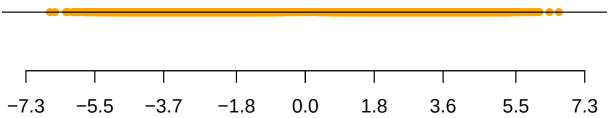
Why Should We Use Graphs?

Using RStudio, compute the median of the 10,000 data contained in the object x, that is created by the following instruction:

```
set.seed(1) ; n <- 5000 ; x <- c(rnorm(n, -3), rnorm(n,
3)) ; x
```

We will get back to the meaning of this instruction later.

Here is the plot of these data:



Do you think the median is a good summary / representative of this data set? If not, why? (Hint: Look at the sorted data.)

Why Should We Use Graphs?

Difficult to say!

We are going to see a new type of graph, called a histogram, that will help us answer that question!

Recommended Graphical Tools

Pre-topic

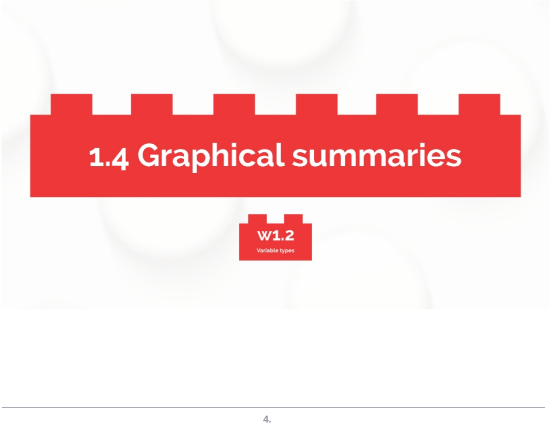
- If you want to graphically summarise **one variable**:
  - and it is **categorical**: a **bar chart**.
  - and it is **quantitative**: a **histogram** or a **boxplot**.
- If you want to explore the relationship between **two variables**:
  - this will be done next week

Watch these videos:

- `https://www.youtube.com/watch?v=Eph_Y0BmHU0` (5 mn)
- `https://www.youtube.com/watch?v=Hj1pgap4U0Y` (5 mn)
- `https://www.youtube.com/watch?v=U64yNvlhv9I` (4 mn)

Consider the variable `labour.cost` in the `1041-old` data set on Moodle. What is the type of this variable? Create, using RStudio, an appropriate graph for this variable.

Concept Map



https://prezi.com/view/5Eet9eXPLVc6U1Q2yIQk/

Data analysis for one or two variables

	one variable		**two variables**		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary			
useful graphs:	barchart	boxplot or histogram			

Recommended Graphical Tools

Which type of graph to use depends on:

- Whether you are summarising **one variable** or looking at the relationship between **two variables**.
- Whether the variables are **quantitative** or **categorical** (qualitative).

Which graphs do you know? When should they be used?

Give your answer online: https://pollev.com/pierrelafaye259

Class Survey Data

Recall the “Class Survey Data” that was collected on MATH1041 students via a student survey a few years ago. We asked 400 students a number of different questions, such as:

- their gender;
- their mode of transport to UNSW;
- satisfaction score of UNSW;
- amount of money spent on a hair cut;
- etc.

Class Survey Data

Identify the types of the variables involved in the following questions (that is, whether they are quantitative or categorical), and what sort of graph you would use to summarise them:

- gender of MATH1041 students?
- satisfaction with UNSW (from 0 to 10)?
- time spent travelling to UNSW?
- method of travelling to UNSW?

QUIZ

Give your answers online: https://pollev.com/pierrelafaye259

Answer

- Gender is a categorical variable. We could use a bar chart.
- Satisfaction score is quantitative variable. We could use a histogram or a boxplot.
- Time spent travelling to UNSW is quantitative variable. We could use a histogram or a boxplot.
- Method of travelling to UNSW is a categorical variable. We could use a bar chart.

Let's Practice

Consult Moodle to access the Class Survey data set.

Use `RStudio` to create the graphs to summarise these variables. Write down the codes you used.

We are going to produce these graphs together.

If you did not bring your laptop or did not install RStudio, that's okay, you will do these in labs. You can also try to use this website:

- `https://jupyter.org/try`

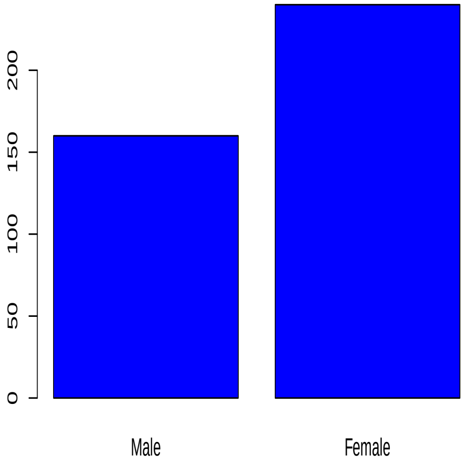
Here is a video that explains how to use it on your phone:

- `https://www.youtube.com/watch?v=1Vej5_O0zIY`

Gender of MATH1041 students (a **bar chart**)

```r
load("Data/1041.RData")
barplot(table(factor(gender, levels = 1:2,
labels = c("Male", "Female"))),
col = "blue")
```

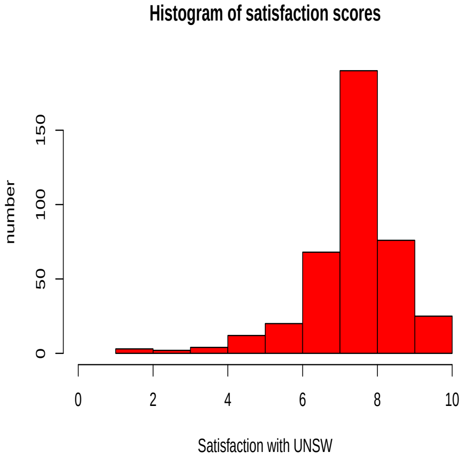
Gender of MATH1041 students (a **bar chart**)



Satisfaction with UNSW (a **histogram**)

```r
hist(unsw.satisf, col = "red",
main = "Histogram of satisfaction scores",
xlim = c(0, 10),
xlab = "Satisfaction with UNSW",
ylab = "number")
```

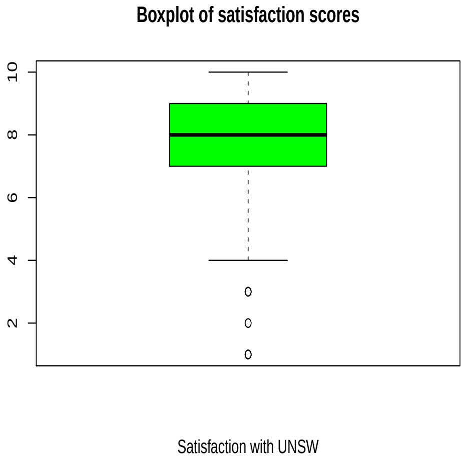
Satisfaction with UNSW (a **histogram**)



Satisfaction with UNSW (a **boxplot**)

```r
boxplot(unsw.satisf, col = "green",
main = "Boxplot of satisfaction scores",
xlab = "Satisfaction with UNSW")
```

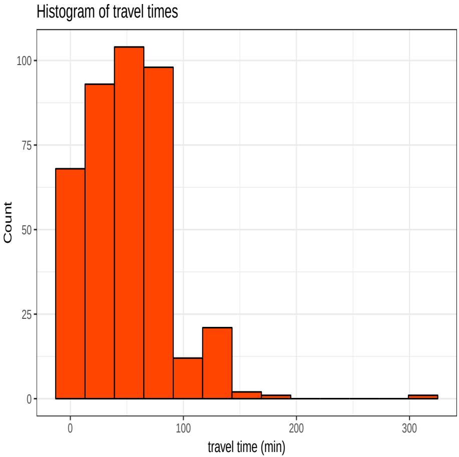
Satisfaction with UNSW (a **boxplot**)



Time travel to UNSW (a **histogram**)

```r
library("ggplot2")
ggdata <- data.frame(travel.time = travel.time)
ggplot(ggdata, aes(x = travel.time)) +
geom_histogram(binwidth = 26,
fill = "orangered", color = "black") +
theme_bw() +
labs(title = "Histogram of travel times",
x = "travel time (min)", y = "Count")
```

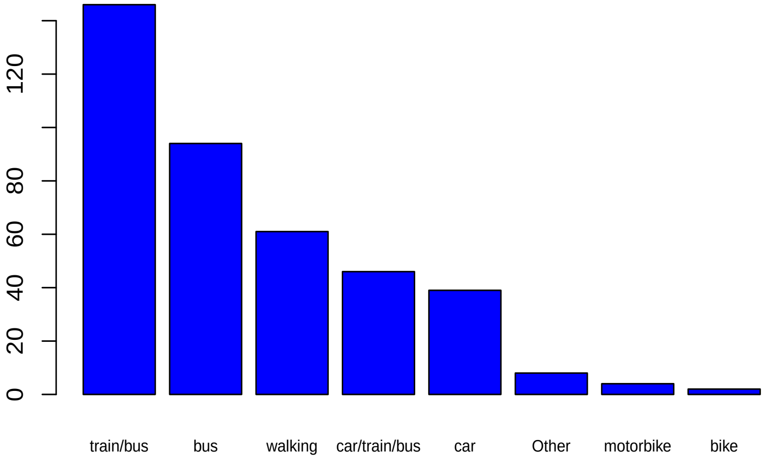
Time travel to UNSW (a **histogram**)



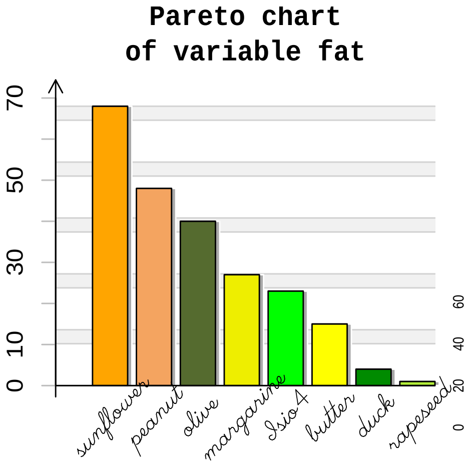
Method of travel to UNSW (a **bar chart**)

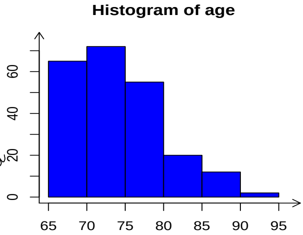
```r
transport <- factor(transport,
levels = c(1:7, "Other:"),
labels = c("walking", "bike", "motorbike",
"car", "bus", "train/bus",
"car/train/bus", "Other"))
barplot(sort(table(transport), decreasing = TRUE),
col = "blue", cex.names = 0.7)
```

Method of travel to UNSW (a **bar chart**)



Find the differences between a histogram and a bar chart.





Access this quiz online: https://pollev.com/pierrelafaye259

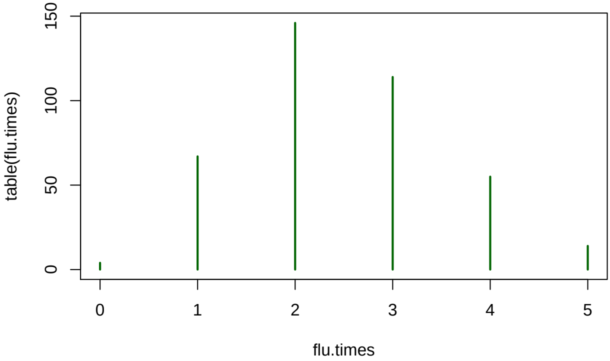
Bar Plot for a Discrete Variable

When a variable is **discrete** rather than continuous (*i.e.*, it can only take a small number of integer values), then it is better to create a **bar graph** (bar plot) where bars have a very narrow width. Why?

Consider the variable `flu.times` which counts how many times a MATH1041 student got the flu until now.

Bar Plot for a Discrete Variable

Since we do not have all the data from the survey yet, this plot was created using artificial data.



We will revisit these plots and talk more about what a discrete variable is in Week 4.

Construction of a Histogram

Shiny by RStudio https://math1041.teaching.unsw.edu.au

Let's spend some time to explain the construction of a histogram.

Example – Travel Times to UNSW

Now that we can plot our data, we can also **comment** on **location** and **spread** using these graphs.

Always comment your graphs!

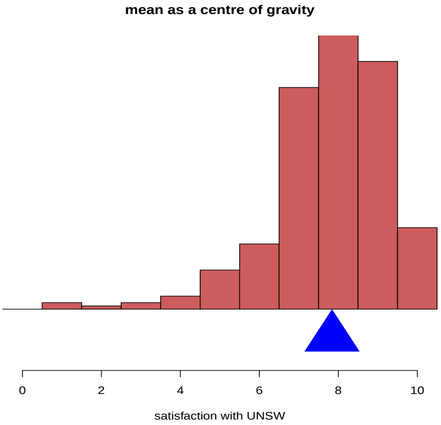
For example, recall that the **mean** satisfaction rating with UNSW was:

$$7.84.$$

This can be seen on a histogram.

In last lecture, we introduced the mean. It has a physical interpretation as the **centre of gravity** of the distribution of data.

- `https://www.youtube.com/watch?v=bZgcUaHPEYc` (2 mn)
- `https://www.youtube.com/watch?v=R8wKV0UQtlo` (2 mn)

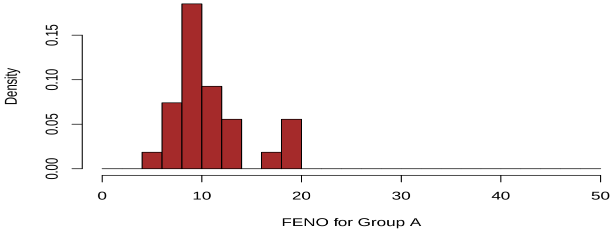


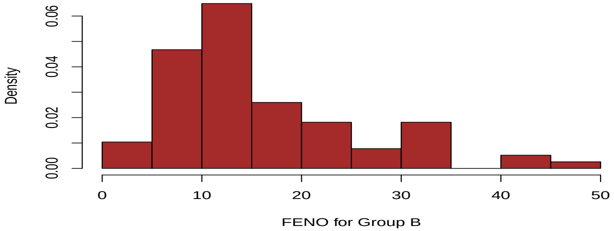
Measures of Spread – The Asthma Example

A few years ago a colleague was involved in a study that explored possible genetic differences between asthmatics and non-asthmatics.

He recorded observations for the quantative variable `FENO`= Fraction of Expired Nitric Oxygen (“biomarker” for asthma), for two groups A and B with genetic differences.

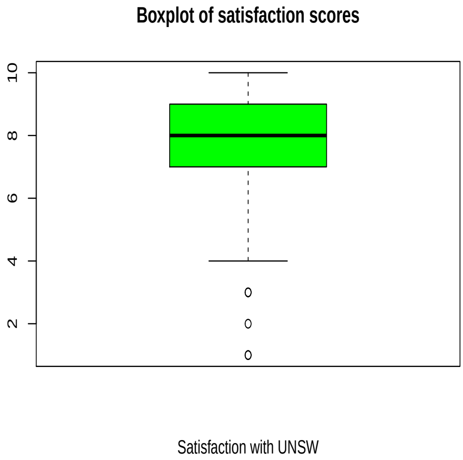
Let’s have a look at the **location** and **spread** for both groups using histograms.





Time for a break!

Medians and Boxplots



Median corresponds to the bold horizontal bar in the **boxplot**!

Medians and Boxplots

How about the edges of the box?

These correspond to the medians of the lower and upper half of the data. Recall from the last lecture that these are called:

$$Q_1 = \text{first quartile}$$

and

$$Q_3 = \text{third quartile}$$

How about the stems/whiskers?

They are drawn at the :

- smallest observation between $Q_1 - 1.5 \times \text{IQR}$ and Q_1
- largest observation between Q_2 and $Q_2 + 1.5 \times \text{IQR}$

Five-Number Summary and Boxplot

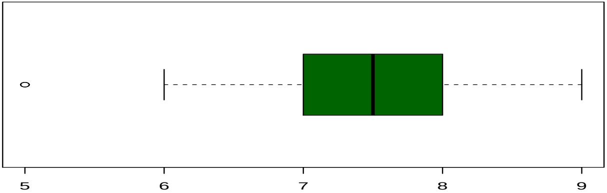
In fact, the **five-number summary** can be used to construct a boxplot.

We use the reduced UNSW scores example here.

```
subset <- c(5, 5, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9)
summary(subset)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   5.000   7.000   7.500   7.462   8.000   9.000
```

```
boxplot(subset, horizontal = TRUE, col = "darkgreen")
```

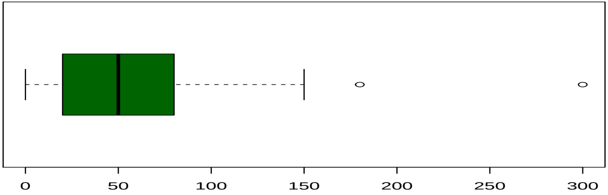


With all data now.

```
load("Data/1041.RData")
summary(travel.time)
```

```
#>     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
#>     0.00    20.00    50.00    51.99    80.00   300.00
```

```
boxplot(travel.time, horizontal = TRUE, col = "darkgreen")
```



Boxplot Terminology

[Moore et al., 2017] use the terms **boxplot** and **modified boxplot**.

Boxplot	Stems/whiskers go from the box to the minimum and maximum. (visual representation of five-number summary)
Modified boxplot	Stems use $1.5 \times$ IQR rule (all boxplots given in our slides are of this variety)

In this course, as in **R**, we will call the latter a "boxplot" (without the "modified"): [Moore et al., 2017] is the exception rather than the rule.

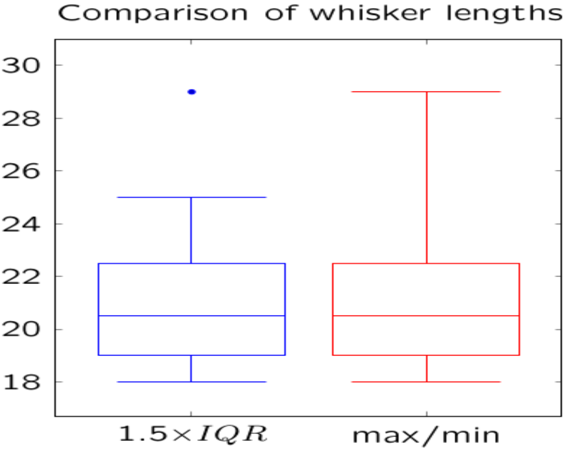


Figure 1.3: Boxplot Terminology

Note: The width of the box is chosen for aesthetic reasons only.

Outlier Identification

How do we decide if an observation is an outlier?

There is no clear-cut answer, but [Moore et al., 2017] recommend the

$1.5\times$ IQR Criterion for Outliers:

Observation is a **suspected outlier**

$$\Updownarrow$$

More than $1.5\times$IQR lower than Q_1; or

More than $1.5\times$IQR higher than Q_3.

A suspected outlier is an observation outside the whiskers on the boxplot.

Example – The $1.5 \times$ IQR Criterion

Using the five-number summary below, apply the suspected outlier rule to the following ‘age‘ data collected on 20 randomly chosen students.

18 18 19 19 19 19 19 19 20 20 21 21 21 21 22 23 24 24 25 29

Min.	Q_1	M	Q_3	Max.
18	19	20.5	22.25	29

Give your answer online: https://pollev.com/pierrelafaye259

Answer

$$\text{IQR} = Q_3 - Q_1 = 22.5 - 19 = 3.5$$

$$1.5 \times \text{IQR} = 1.5 \times 3.5 = 5.25$$

$$Q_1 - 1.5 \times \text{IQR} = Q_1 - 5.25 \quad = \quad 19 - 5.25 = 13.75$$

the lower suspected outliers are → **no lower suspected outliers**

$$Q_3 + 1.5 \times \text{IQR} = Q_3 + 5.25 \quad = \quad 22.5 + 5.25 = 27.75$$

the upper suspected outliers are → **29 is a suspected outlier**

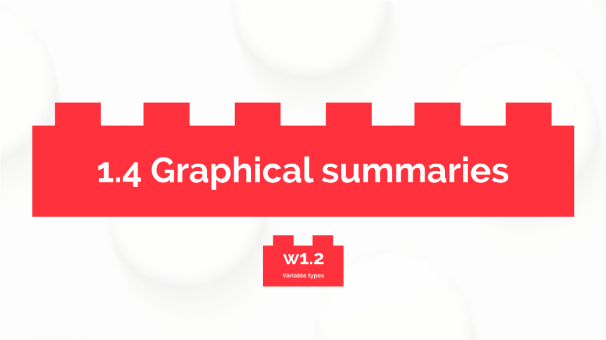
Example – Boxplot

Using the five-number summary and the 1.5 x IQR value calculated from the previous slide, construct a boxplot for the ‘age‘ data collected on the 20 randomly chosen students.

Do it by hand and using `RStudio`.

We will only be using ‘RStudio‘ for that today.

Concept Map



What to Look for in a Graph

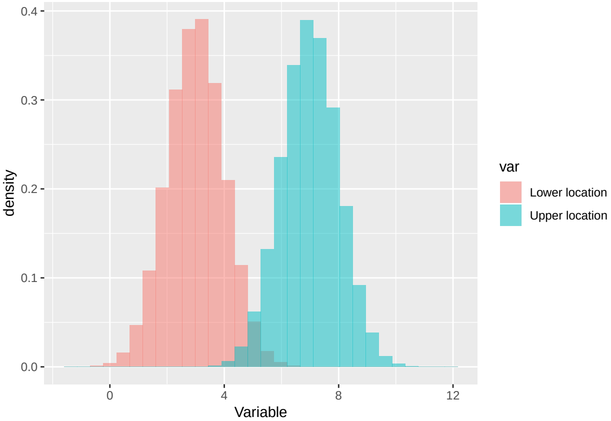
When **commenting** on a graph of a **quantitative** variable, consider describing the **overall pattern** of the data in terms of:

- the **location** (where most of the data are) and **spread** (or variability) of the data;
- the **shape** of the data (symmetric, left-skewed or right-skewed).

And also, indicate:

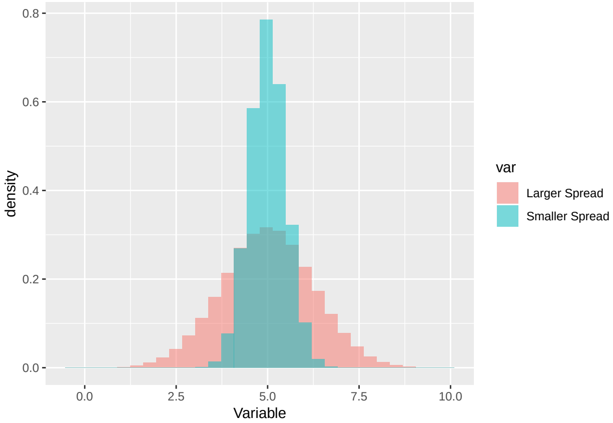
- if there are any unusual observations (called **suspected outliers**).

Location



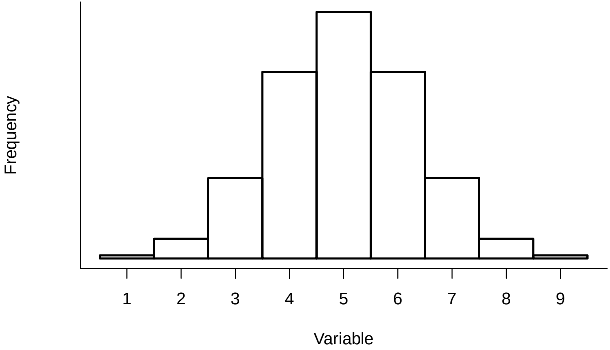
Two histogams with a different location.

Spread

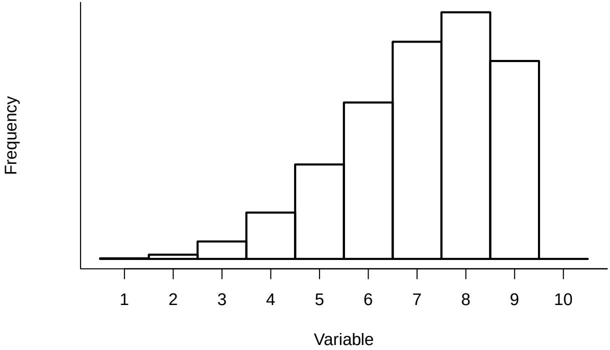


It is easier to make a comment on the spread when we compare two histograms, or when we have an idea of the usual spread of a given variable.

Typical Shapes: Symmetric

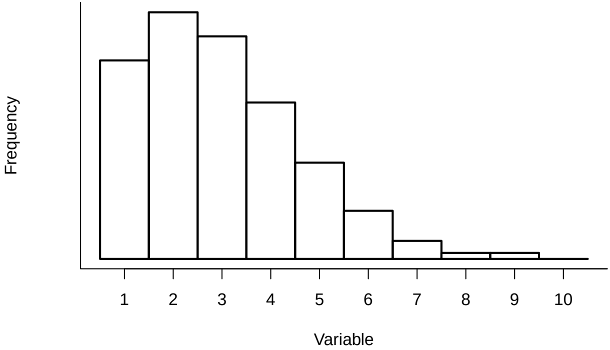


Typical Shapes: Skewed Towards the Left



Long **tail** on the left. Observing **extreme values** on the left is not uncommon.

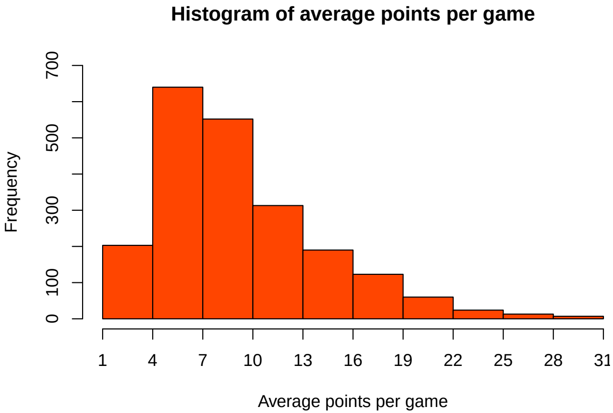
Typical Shapes: Skewed Towards the Right



Long **tail** on the right. Observing **extreme values** on the right is not uncommon.

NBA Average Points per Game

The following histogram depicts the scoring **average** of players from the National Basketball Association (NBA) up to the 2008 season.



Comment on the location, spread and shape of the histogram.

Answer

- **Location:** Most observations appear to be between 0 and 9.
- **Spread:** There appears to be a fairly large spread.
- **Shape:** The distribution is clearly skewed to the right.
- There does not appear to be any unusual observations.

You make (several) graphs to gain a better understanding of your data. You choose the graph to present in your report that best illustrates the message you want to convey.

Use `RStudio` to create a box plot for the scoring **average** of NBA players.

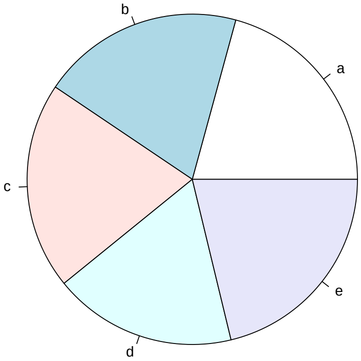
Consult Moodle to access the data set `NBA.txt`.

Some Other Graphical Tools

- **Stem-and-leaf plots** (e.g., page 13 in [Moore et al., 2017])
- **Pie charts** (**controversial** – better to use a bar chart.)
- **Time plots** (e.g., page 24 in [Moore et al., 2017]). Suitable for time ordered data. Common in financial pages of newspapers.
- **Dot plots** "Poor person's" histogram.

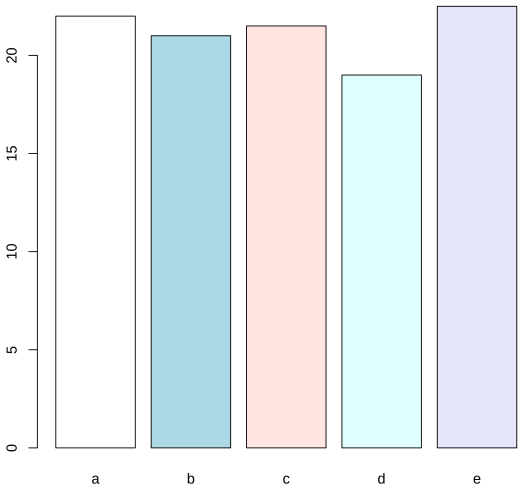
Pie Chart Problem

Rank categories by increasing surface.



Pie Chart Problem

Rank categories by increasing surface.



Fancy Graphs

We have covered some fundamental graphical tools.

But new tools are constantly being developed and modified.

Depending on the problem at hand, there is nothing to stop you devising your own graphical display!

A good example of an improvised graphical display is the moving bubble plot used by Prof Hans Rosling in:

http://www.youtube.com/watch?v=jbkSRLYSojo

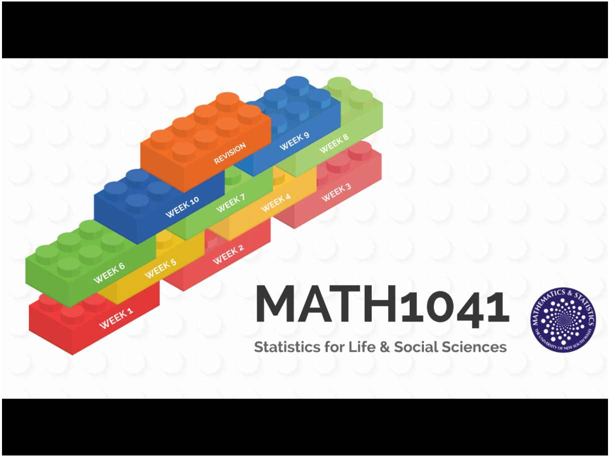
What Did We Learn?

- Introduced the statistical software `RStudio`
- Graphical Exploratory Analysis (a.k.a. graphical summary) on one variable
- Describe the main features of a set of data using a graph

The Keywords Slide

- Bar chart, bar plot, bar graph
- Histogram and boxplot
- Location/shape/spread
- (Suspected) Outlier

The Wall of Knowledge Slide



https://prezi.com/view/5Eet9eXPLVc6U1Q2yIQk/

The Post-Mastery Quiz Slide

Give your answer online: https://pollev.com/pierrelafaye259

What Did **you** Learn?

Reflect

Exercise 1.1

Reflect on what **you** have learned in this chapter and try to relate it to your prior knowledge. What did you already know? What is new? (This could be a theoretical result, a way of thinking, some software skills, etc.). Think about a real life context where you could apply this (new) knowledge.

Write Down Your Reflections

Reflect

The Feedback Slide

Take two minutes to write your reactions to the first day (anonymously on Moodle).

Preparation for Our Next Lecture

1. Try to find on internet or scan in any newspaper a graph different to the ones presented in this lecture. Post-it on Moodle and comment on its advantages/disadvantages. We will vote for the best finding!
2. Consider reading the textbook pages related to what we have covered so far.

That's it for today's lecture, enjoy the weekend!

College Report ASA Revision Committee (2016).
GAISE (Guidelines for Assessment and Instruction in Statistics), Education College Report.
The American Statistical Association.

Lafaye de Micheaux, P., Drouilhet, R., and Liquet, B. (2013).
The R Software: Fundamentals of Programming and Statistical Analysis.
Statistics and Computing. Springer New York.

McKeachie, W. J. and Svinicki, M. (2014).
McKeachie's Teaching Tips.
Wadsworth Publishing, Cengage Learning, 14th ed edition.

Moore, D. S., McCabe, G. P., and Craig, B. A. (2017).
Introduction to the Practice of Statistics.
W.H. Freeman, 9th ed edition.

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