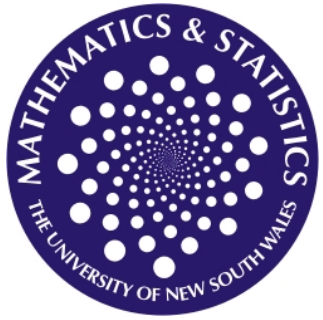
Statistics

MATH2089

UNSW
THE UNIVERSITY OF NEW SOUTH WALES



Semester 1, 2018 – Lecture 1

This lecture

1. Introduction

1.1 What is statistics?
1.2 The statistical process
1.3 Populations and samples
1.4 Random sampling

Additional reading: Section 1.1 in the textbook

Some quotes

"I am not much given to regret, so I puzzled over this one a while. Should have taken much more statistics in college, I think."

Max Levchin, Paypal Co-founder, Slide Founder, 2010

"I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding."

Hal Varian, Chief Economist at Google, 2009

1. Introduction

What is statistics?

- In order to learn about something, we must first collect observations, referred to as **data**

Definition

Statistics is the science of the (i) collection, (ii) processing, (iii) analysis, and (iv) interpretation of data

In short, statistics is learning from data and it allows us to gain new insights into the behaviour of many phenomena.

- Statistical concepts and methods are not only useful but indeed are often vital in understanding the world around us
- Further, it allows us to turn observational evidence into information for decision making, which is probably the most important aspect

What is statistics?

In engineering, this includes diversified tasks like

- calculating the average length of the downtimes of a computer
- predicting the reliability of a launch vehicle
- evaluating the effectiveness of commercial products
- studying the vibrations of airplane wings
- checking whether the level of lead in the water supply is within safety standards
- determining the strength of supports for generators at a power plant
- collecting and presenting data on the number of persons attending seminars on solar energy
- ...

What is statistics?

- Statistics is a discipline that makes use of mathematics, computer science and subject matter expertise
- Statistics considers the presence of randomness, uncertainty and variation, which are everywhere in real life

- If each computer had exactly the same length of downtime,
- If the level of lead was exactly identical everywhere and every time in the water supply,
- If each seminar attracted the same number of people,
- and if those values were known with absolute accuracy,

$\Rightarrow$ Then a single observation would reveal all desired information, we would not need statistics :-(

Statistics

Statistics allows us to describe, understand and control the variability insofar as possible and to **take this uncertainty into account** when making judgements and decisions.

Example 1: Does cloud seeding work?

Cloud seeding is the attempt to change the amount of precipitation that falls from clouds, by dispersing substances into the air that serve as cloud condensation.

The usual intent is to increase precipitation (rain or snow).

1. A natural question may be

   "Does cloud seeding using a given substance (say, silver nitrate) really work ?"

   $\Rightarrow$ research question

How can we answer this question ?

2. First, we should observe the amount of precipitation that falls from seeded clouds, as well as from unseeded clouds

   $\Rightarrow$ experiment, collection of data

Example 1: Does cloud seeding work?

For our experiment, we observe 52 clouds, 26 of which were chosen at random and seeded with silver nitrate.

The following rainfall (in acre-feet) are recorded:

Unseeded Clouds

```
1202.6 830.1 372.4 345.5 321.2 244.3 163.0 147.8
95.0 87.0 81.2 68.5 47.3 41.1 36.6 29.0 28.6 26.3
26.1 24.4 21.7 17.3 11.5 4.9 4.9 1.0
```

Seeded Clouds

```
2745.6 1697.8 1656.0 978.0 703.4 489.1 430.0 334.1
302.8 274.7 274.7 255.0 242.5 200.7 198.6 129.6
119.0 118.3 115.3 92.4 40.6 32.7 31.4 17.5 7.7 4.1
```

$\Rightarrow$ These values are our data.

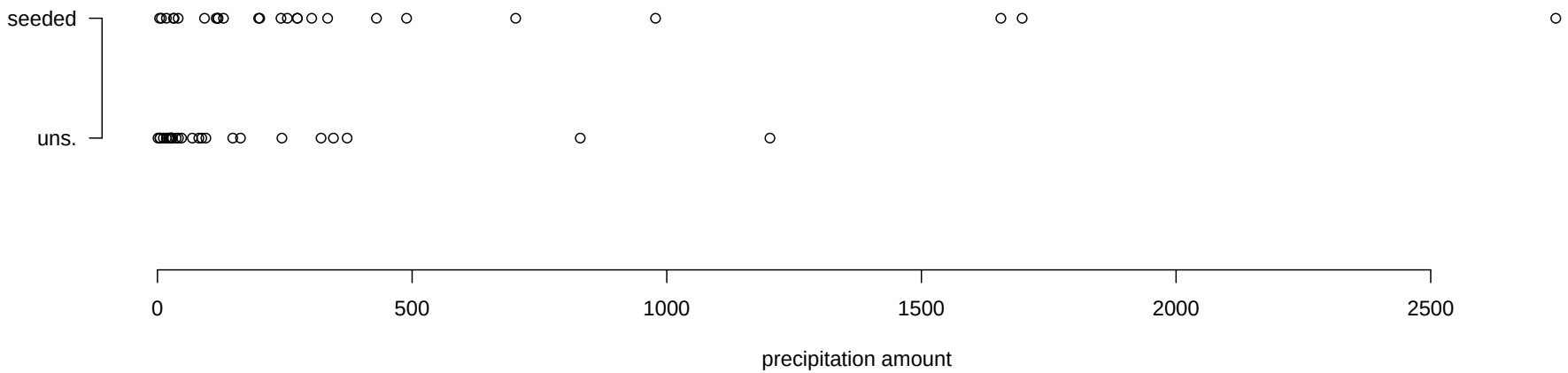
Of course, we observe variability: we could not expect each cloud (seeded or not) to give exactly the same amount of rain!

Example 1: Does cloud seeding work?

What can we do with those numbers?

3. We should present the data so that they are readily comprehensible

This includes graphical representations



as well as numerical summary measures

average prec. seeded $= 441.98$ average prec. unseeded $= 164.58$

$\Rightarrow$ The description and summarisation of data, is called **descriptive statistics** (Chapter 2)

Example 1: Does cloud seeding work?

At first sight, seeded clouds seem to give more precipitation than unseeded clouds.

Careful! - We must take into account the possibility of chance:

- Due to chance only, the 26 seeded clouds might be the clouds that would have given more rainfall anyway
- Due to chance only, the 26 unseeded clouds might be the clouds that would have given less rainfall anyway

$\Rightarrow$ Can we really conclude that the observed higher amount of rainfall for seeded clouds is due to seeding? Or is it possible that the seeding was not responsible for that but rather that the higher rainfall amount was just a **chance occurrence**?

- We have only observed 52 clouds. If we had observed 52 (or more) other clouds, would we observe different rainfall amounts?

$\Rightarrow$ Can we really generalise what we are seeing on a particular data set beyond that data set? How risky is it?

Example 1: Does cloud seeding work?

4. We should analyse and interpret the data bearing in mind that the observed features may be consequences of chance only

This part of statistics is called **inferential statistics** or **statistical inference** (Chapters 6-12).

- How far to go with generalising from an observed data set
- Are such generalisations reasonable or justifiable
- Do we need to collect more data

Some of the most important problems in inference concern the appraisal of the risks and consequences of making wrong decisions.

- Risks are often appraised by calculating **probabilities** of some events occurring
- We will discuss **probability theory** in more details in Chapters 3-5

Example 1: Does cloud seeding work?

5. Finally, we should draw conclusions from our investigations, that is, we should answer the question

   "Does cloud seeding using silver nitrate result in more rainfall than not cloud seeding using silver nitrate?"

The statistical process

The points (1) to (5) in the above example form the **typical procedure for statistical inference**:

1. Set clearly defined goals for the investigation; formulate the research question
2. Decide what data is required/appropriate and how to collect them; collect the data
3. Display, describe and summarise the data in an efficient way; check for any unusual data features
4. Choose and apply appropriate statistical methods to extract useful information from the data
5. Interpret the information, draw conclusions and communicate the results to others

Fact

Every step in this process requires understanding statistical principles and concepts as well as knowledge and skills in statistical methods.

Example 2: Hair colour and pain tolerance

An experiment conducted at the University of Melbourne suggests that there may be a difference in pain threshold for blonds and brunettes.

1. The research question is:

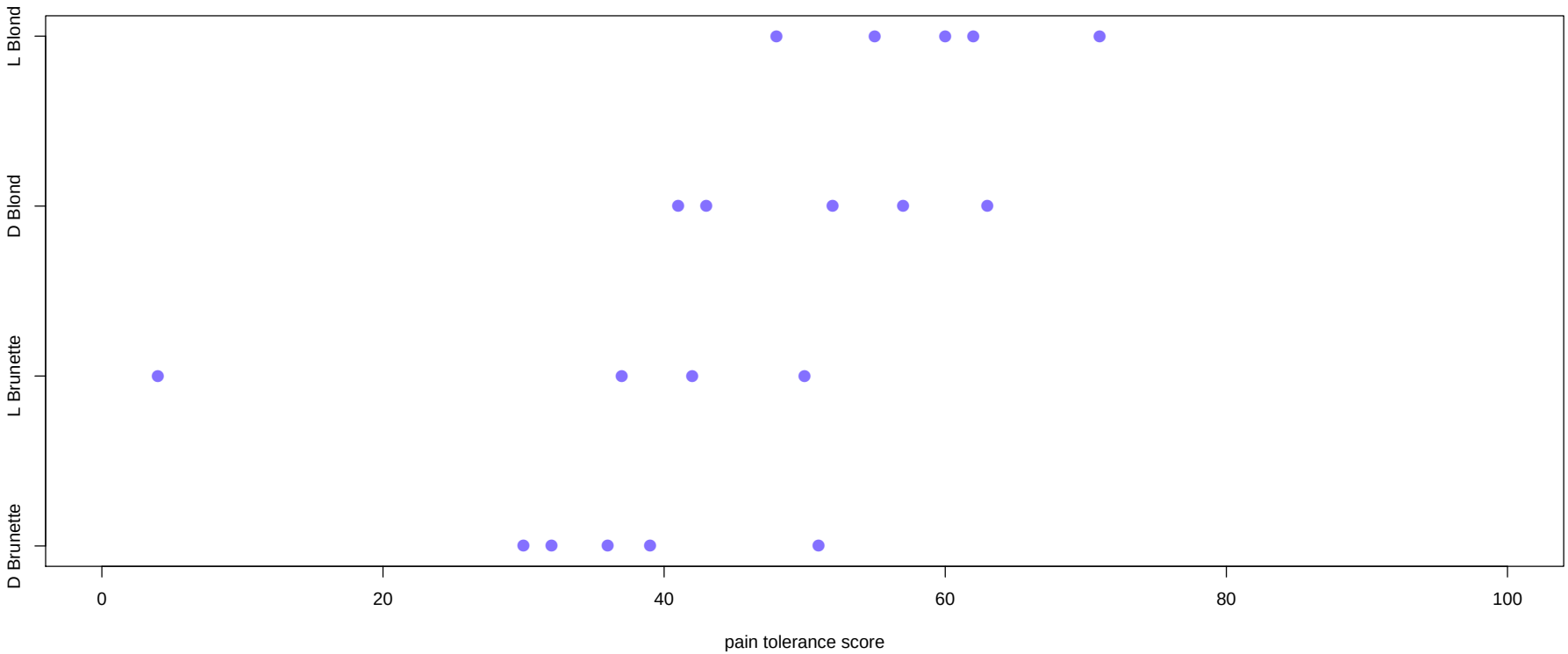
   "Is pain threshold related to hair colour?"

A group of 19 subjects was divided into light blond, dark blond, light brunette and dark brunette groups and a pain threshold score was measured for each subject

2. The data are ... (A higher score means a higher pain threshold)

3. (better seen in the next figure)

Example 2: Hair colour and pain tolerance



Example 2: Hair colour and pain tolerance

4. Pain threshold seems to increase with lighter hair colour, but is this effect real or just due to chance ? (the number of observations is quite small: we have 4 or 5 observations per hair colour)

$\Rightarrow$ We have to apply some inferential method to come to a conclusion

5. Depending on what we have observed, either

   It is clear from the data that pain tolerance is related to hair colour

   or

   The data do not allow us to conclude that pain tolerance is related to hair colour

Remark: In the latter case we won't say "pain tolerance is not related to hair colour". It might still be the case, but with such a small number of observations, we are not sure and it would be too risky to affirm it is.

Population

- Usually, we are interested in obtaining information about a total collection of elements, which is referred to as the **population**
- The elements are often called **individuals** (or units)
- Given the research question, we have observed some characteristic for each individual. This characteristic, which could be quantitative or qualitative, is called a **variable**

Example 1

In Example 1 (clouds seeding), the population consists of all the clouds of the sky. An individual is a cloud and the variable of interest is the amount of rainfall.

Example 2

In Example 2 (pain tolerance), the population consists of all blonds and brunettes of the world. An individual is one of those people and the variable of interest is the pain threshold score.

Sample

- It is often physically impossible or infeasible from a practical standpoint to obtain data on the whole population
- Think also of very expensive, or very time-consuming, or destructive experiments
- In most situations, we can only observe a subset of the population, that is, we must work with only partial information

The subset of the population which is effectively observed is called the **sample**.

The **data** are the measurements that are actually collected over the sample in the course of the investigation.

Note: sometimes, we may use "population" to designate the set of all potential measurements and "sample" to designate the subset of measurements actually observed (i.e., the data)

Sample

In Example 1, the sample consists of the 52 clouds whose rainfall amounts have been recorded.

(We might also consider that we have two samples: 26 seeded clouds and 26 unseeded clouds).

In Example 2, the sample consists of the 19 persons whose pain threshold scores have been measured.

Fact

The distinction between the data actually acquired (the sample) and the vast collection of all potential observations (the population) is a key to understanding statistics.

Sampling

- The process of selecting the sample is called sampling
- If the sample is to be informative about the total population, it must be representative of that population

$\Rightarrow$ Suppose you are interested in the average height of UNSW students, would you select the sample from the UNSW basketball team?

$\Rightarrow$ Suppose you are interested in the average age of UNSW students, would you select a sample made up of postgraduate students only?

- The quality of the data is paramount in a statistical study

  Your results are only as good as your data !

- Sampling must be carefully done, impartially and objectively

Random sampling

In practice, the only sampling scheme that guarantees the sample to be representative of the population is **random sampling**.

$\Rightarrow$ The individuals of the sample are selected in a totally random fashion, without any other prior consideration

Any non-random selection of a sample often results in one which is inherently biased toward some values as opposed to others.

$\Rightarrow$ We must not attempt to deliberately choose the sample according to some criteria

$\Rightarrow$ Instead, we should just leave it up to "chance" to obtain a sample which correctly covers the underlying population

The importance of random sampling

Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population, taking the randomness into account (using probabilities).

$\Rightarrow$ Not possible if the sample is not random!

Information drawn in a non-random sample cannot, as a rule, be generalised to larger populations.

Fact

The statistical procedures presented in this course may not be valid when applied to non-random samples.

$\Rightarrow$ Never unquestioningly accept samples without knowing how the data have been generated / collected / observed

Objectives

Now you should be able to:

- identify the role that statistics can play in the engineering problem-solving process ☐
- discuss how variability affects the data collected and used for making engineering decisions ☐
- discuss how probability theory is used in engineering and science ☐
- discuss the importance of random sampling ☐