Overview to first year statistics and data science courses

DATA1001/MATH1005/MAT1015/MATH1115

09 March 2018

Statistical education in the age of big data

Major in statistics in the age of big data

– Statistics is a quantitative science whose lifeblood is data. In the age of big data, statistics as a discipline is evolving and it is important that our statistical education evolves with it.

– Students need to understand that statistics is a discipline about solving problems and about generating new scientific knowledge through the deep understanding of data generated from other disciplines.

– A key aspect of statistics deals with the understanding, analysis, interpretation, presentation and organization of data, whether that data originates from the humanities, engineering, business, health, medicine or the physical and life sciences.

Changes in the statistics discipline in the last 10 years

Our environment has changed.

– **Data:** It is bigger and more complex. Our undergraduate program needs to reflect such changes.

– **Users:** Large data now guides many decisions and policy-making processes. This has generated an increased demand for non-data analytic people to acquire fundamental statistical thinking and intuition.

– **Technology:** The revolution in computing technology has provided the statistical discipline with better tools to understand data.

Together these changes have resulted in a dramatic change in the graduate attributes that are desirable for all University of Sydney graduates.

Our new curriculum is one step towards addressing such issues.

Key changes

We have looked at the changes in our environment and determined the topics and information that is important for a students. The construction of this new curriculum is a learning curves for ALL of us.

In particular we aim to:

– incorporate statistical thinking and intuition;

– develop statistical competence in handling real world data in the age of big data;

– give practical experience with the appropriate accompanying soft skills;

– remove unneccesary symbolism; and

– provide the foundation to link symbols with words.

What we didn't change: The formal concept of probability and distribution theory has always been introduced in STAT2011 (first semester of second year) and this continues to be the case.

Sydney Courses in Data Science & Statistics

Sydney Courses in Data Science & Statistics

- DATA1001 is part of a suite of foundational courses in Data Science and Statistics at the University of Sydney.
  - DATA1001 Foundations of Data Science (6cp)
  - MATH1005 Statistical Thinking with Data (3cp)
  - MATH1015 Biostatistics (3cp)
  - MATH1115 Interrogating Data (3cp)
  - OLE1631/2 Shark Bites and Other Data Stories (0/2cp)
- DATA1001 = MATH1005/1015 + MATH1115

Aim (What we tell the students)

Aim

- In all the courses, the students will learn how to **problem solve with data**, using **statistical thinking** and **computational skills.**
- They also develop essential soft skills of curiosity, collaboration and communication.

For tutors

Tutoring in the new first-year Data Science and Statistics suite of units will be different to what you are used to.

– Please read the teaching guide :

– Lab Preparation: Familiarise yourself with the lab worksheet & data that the students will be using in class;

– Understand the questions that they'll also be working on.

– Ask the lecturer if you have any questions before class (Ed Q&A for tutors)

– There are solutions at the end of the text book, please discuss with the first year teaching team if you don't understand or **disagree** with the solution.

– You will find there is no "right or wrong" way of doing the question.

– You will find there are many different R codes to achieve the same outcome.

Educational Design

Class	Statistical Thinking	Computational Skills	Soft Skills	Activities
Lectures	✔			Data Story, Statistical Concepts, Thinking Exercises
Labs	✔	✔		RGuide, RQuizzes, Data Stories
Projects	✔	✔	✔	Verbal and Written Reproducible Reports

Statistical Thinking

Here are some slides for the student in Lecture 1, you are encouraged to read through them.

6 mental habits

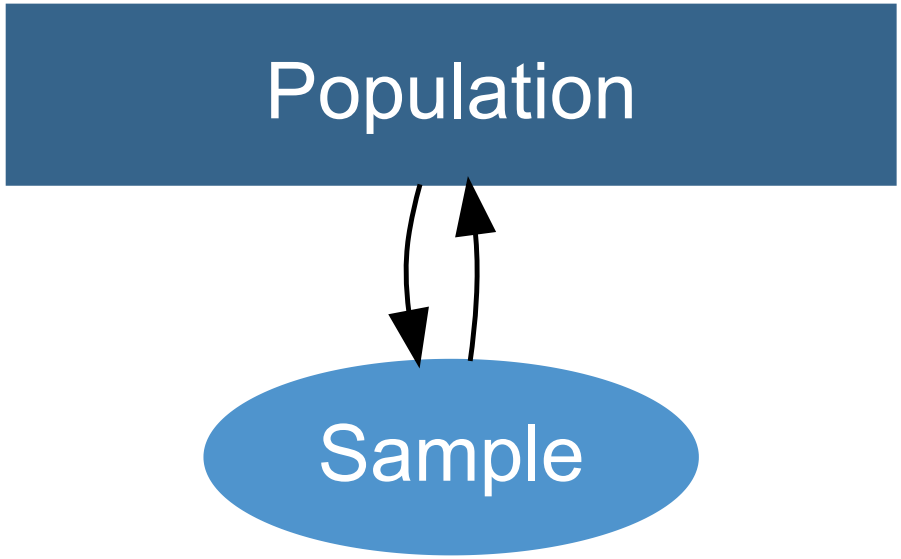
1. Understand the statistical process as a whole.
2. Always be skeptical.
3. Think about the variables involved.
4. Always relate the data to the context.
5. Understand (and believe) the relevance of statistics.
6. Think beyond the textbook.

Potential challenges for students

Topic 1: Experimental Design + Population and Samples

Population and Samples

To **problem solve with data**, we need to understand how a sample (usually the "data") relates to a population.



Experimental Design

– Week 1 we already covered Chapters 1 and 2.

– Design of Experiments looks simple but these two chapters have lots of depth and contain some of the hardest questions. The concept appears in many of the revision sections.

Example 1:

(Hypothetical.) In a clinical trial, data collection usually starts at "baseline," when the subjects are recruited into the trial but before they are assigned to treatment or control. Data collection continues until the end of followup. Two clinical trials on prevention of heart attacks report baseline dala on smoking, shown below. In one of these trials, the randomization did not work. Which one, and why?

Example 1 Solution:

In trial (i), something must have gone wrong with the randomization. The difference between 49.3% and 69.0% shows that the treatment group smoked less to begin with, which would bias any further comparisons. The difference cannot be due to the treatment, because baseline data say what the subjects were like before assignment to treatment or control. (More about this in chapter 27.)

Example 2:

The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.

1. Why did they study men and women and the different age groups separately?

2. The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly. (Page 25)

Example 2 Solution:

1. They were controlling for age and sex as possible confounders; this is discussed on p.13, with respect to a specific disease—lung cancer.

2. This is the wrong conclusion to draw. Ex-smokers are a self-selected group, and many people give up smoking because they are sick. So recent exsmokers include a lot of sick people. (Other epidemiological data show that if you quit smoking, you will live longer.)

Revision question from later on

Example 3 Question:

Census data show that in 1980, there were 227 milllion people in the US of whom 11.3% were 65 or older. In 2000, there were 281 million people, of who 12.3% were 65 or older. Is the difference in the percentages statistically significant ?

Example 3 Solution:

– You can try to do a two sample test, but the result is close to meaningless. Why?

– We have Census data on the whole population. There is no sampling variability to worry about. Census data are subject to many different errors, but it is not a chance model.

– The ageing of the population is real and this may makes a difference to the health care system.

Topic 2: Modelling Data

R.M.S.

– Talk about "root-mean-square"; the book focuses on the use of "words". The phrase "root-mean-square" says how to do the arithmetic, provided you remember to read it backwards:

- SQUARE all the entries, getting rid of the signs.
- Take the MEAN (average) of the squares.
- Take the square ROOT of the mean.

This can be expressed as an equation, with root-mean-square abbreviated to

$$\text{RMS} = \sqrt{\text{Mean of squared data}} \approx \text{Mean of absolute data}$$

Or

$$\text{RMS} = \sqrt{\frac{x_i^2}{n}} \approx \frac{|x_i|}{n}$$

22/57

SD vs SE

– deviation = entry - average

– R.M.S was introduce to describe SD being R.M.S deviation from average.

– The book very specifically uses SD and later on SE to keep the concept clear for the students. It is important we use such words and keep to this.

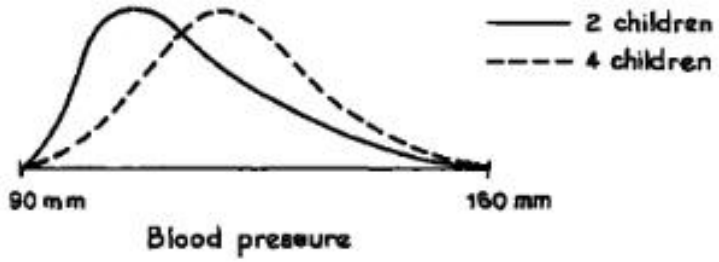
Note: In this course, we intentionally focus on statistical concepts in words. This is vital for collaborating with people from different fields. The mathematics is introduced in 2nd year - see the Maths Guide. However, here some simple mathematical notation is helpful.

For example: Later on we learn to describe

Individual measurement = exact value + chance error + bias

Example 4: Histogram

1. As a sideline, the Drug Study compared blood pressures for women having different numbers of children. Below are sketches of the histograms for women with 2 or 4 children. Which group has higher blood pressure? Does having children cause the blood pressures of the mothers to change? Or could the change be due to some other factor, whose effects are confounded with the effect of having children?



Note that this question links histogram to experimental design.

Example 4 Solution:

On the whole, the mothers with four children have higher blood pressures. Causality is not proved, there is the confounding factor of age. The mothers with four children are older. (After controlling the age, the Drug Study found there was no association left between number of children and blood pressure.)

Example 5:

7. Two histograms are sketched below. One shows the distribution of age at death from natural causes (heart disease, cancer, and so forth). The other shows age at death from trauma (accident, murder, suicide). Which is which, and why?

Example 5 Solution:

1. Natural causes. (ii) Trauma. Reason: Young people die of accidents, murder, etc. Old people die of heart disease, cancer, etc.

Normal approximation

– This is introduced very early before probability distributions.

– Here we are introducing a simple concept: If you have a curve how can you best describe this curve using a formula.

– We are also told that this curve have some very special properties .

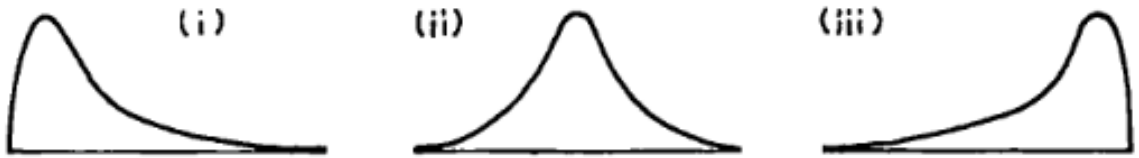
– We are giving student rule of thumbs … they are forced to think about the data.

– Please note that they will link this back to P(X < x) concept later in the book, so please be patient and do not introduce random variables in your tutorial at this stage. The concept of chance and random variables are introduced in Chapter 13 using the "Box model".

– You can see from the next two examples, students have a lot to think about without being introduced too many concepts at once.

Example 6:

11. One term, about 700 Statistics 2 students at the University of California, Berkeley, were asked how many college mathematics courses they had taken, other than Statistics 2. The average number of courses was about 1.1; the SD was about 1.5. Would the histogram for the data look like (i), (ii), or (iii)? Why?



Example 6 Solution:

Histogram (i) is right. With (ii), the average of 1.1 would be right in the middle, and a lot of people would be taking fewer than $1.1 - 1.5 = -0.4$ courses.

Histogram (iii) is even worse.

Example 7:

For women age 25-34 with full time jobs, the average income in 2004 was $32,000. The SD was $26,000, and 1/4 of 1% had incomes above $150,000. Was the percentage with incomes in the range from $32,000 to $150,000 about 40%,50%, or 60%? Choose one option and explain briefly.

Note: This is giving the student a chance to think about a histogram without seeing it.

Example 7 Solution:

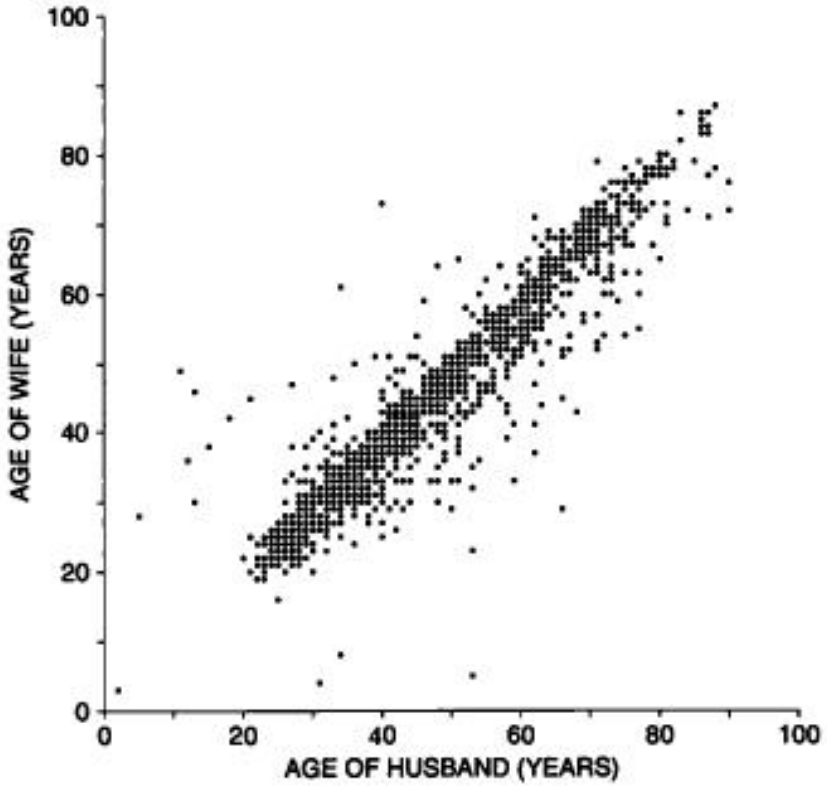
The histogram has a long right hand tail, the median is well below average. Therefore, a lot less than 50% are earning above $32,000. Choose 40%.

Topic 3: SD line and Correlation

Scatter plot

Example 8:

7. The scatter diagram below shows ages of husbands and wives in Ohio. Data were extracted from the March Current Population Survey. Or did something go wrong? Explain your answer.



SD Line

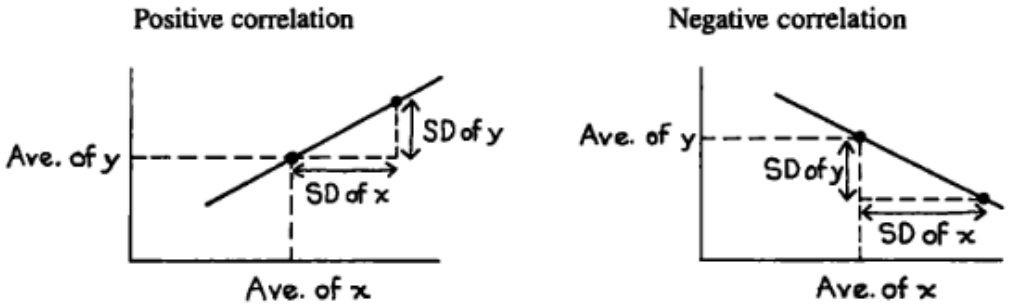
Figure 8 shows how to plot the SD line on a graph. The line goes through the point of averages, and climbs at the rate of one vertical SD for each horizontal SD. More technically, the slope is the ratio.

$$(\text{SD of } y)/(\text{SD of } x).$$

This is for positive correlations. When the correlation coefficient is negative, the SD line goes down; the slope is[7]

$$-(\text{SD of } y)/(\text{SD of } x).$$

Figure 8. Plotting the SD line.



Concept of correlation

The correlation coefficient is a measure of linear association, or clustering around a line. The relationship between two variables can be summarized by

· the average of the x-values, the SD of the x-values,

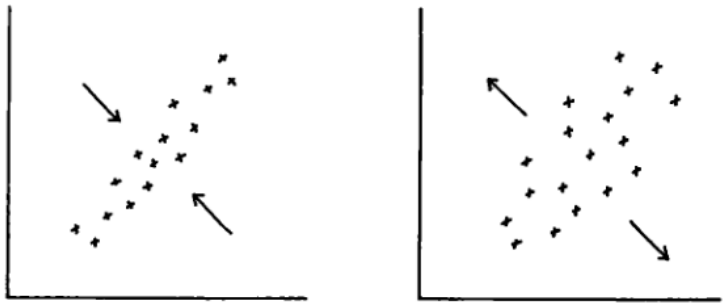
· the average of the y-values, the SD of the y-values,

· the correlation coefficient r.

Figure 5. Summarizing a scatter diagram. The correlation coefficient measures clustering around a line.

(a) Correlation near 1 means tight clustering.

(b) Correlation near 0 means loose clustering.

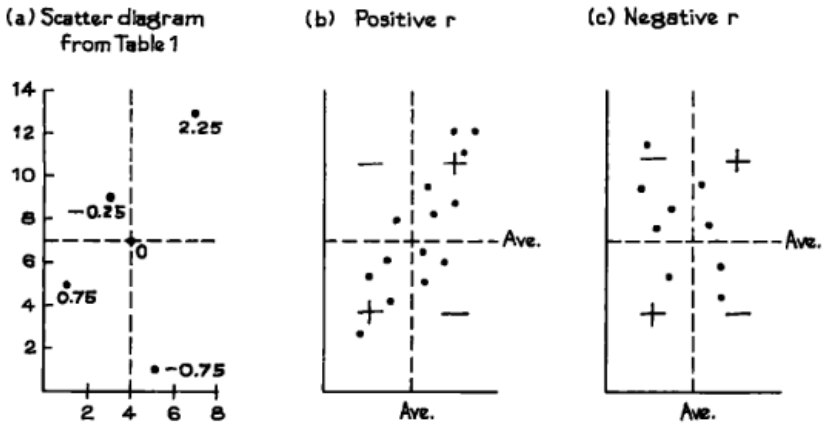


Correlation

- The text book teaches a very different way to calculate correlation.
- This is a conceptual idea
  - Step 1: if you move x into standard units and y into standard units
  - Step 2: Then the correlation is the average of the product

Why does r work as a measure of association? In figure 9a, the products are marked at the corresponding dots. Horizontal and vertical lines are drawn through the point of averages, dividing the scatter diagram into four quadrants. If a point is in the lower left quadrant, both variables are below average and are negative in

Figure 9. How the correlation coefficient works.



Correlation Formula

– This is given and discussed at the end.

Example 9

Suppose men always married women who were exactly 8% shorter. What would the correlation between their heights be?

Example 9: Solution

The correlation would be 1.00. All the points on the scatter diagram for height of wife vs. height of husband would lie on a straight line which slopes up. The slope of the line is 0.92, but correlation and slope are two different things.

– Read Chapter 9 carefully (a lot of depth in this chapter). We will potentially explore a lot more of this section in the advanced level.

Topic 4: Sample survey

Chapter 19

This section is completely new for first year.

– Read the text book for this section.

Example 10a:

A utility company serves 50,000 households, As part of a survey of customer attitudes, they take a simple random sample of 750 of these households. The average number of television sets in the sample households turns out to be 1.86 and the SD is 0.8. If possible, find a 95% confidence interval for the average number of television sets in all 50,000 households. If this isn't possible, explain why not. (Page 424)

Question 10b:

As part of the survey, all person aged 16 or over in the sample households are interviewed. This makes 1528 people, On the average, the sample people watched 5.2 hours of television the Sunday before this survey, and the SD was 4.5 hours. If possible, find a 95% confidence interval for the average number of hours spent watching television on the Sunday by all person age 16 and over in the 50,000 households. If this isn't possible, explain why not. (Page 425)

Note: Two very similar style questions, makes the student think before performing t-test or calculate confidence interval.

Topic 5: Box model

Box model

– The purpose of a **box model** is to analyze chance variability,

– The rest of the slides are created by Di for Lecture 22.

– I encourage you to read these slides when they are release, read the text book or simply attend the lecture.

The box model

The box model

- The **box model** is a simple way to describe the chance of generating a number
- We need to know:
  - the distinct numbers that go in the box ("tickets")
  - the number of each kind of tickets in the box
  - the number of draws from the box

The box model for gambling

The box model for gambling

- For a box model for gambling games,
  - the tickets represent the amount won (+) and lost (-) in each play
  - the chance of drawing a particular value, is the chance of winning that amount in 1 play
  - the number of draws is the number of plays
- The **net gain** is the sum of the draws from the box.

Examples

Example1 (Dice)

Throw a fair dice 25 times.

```
## Warning: package 'DiagrammeR' was built under R version 3.3.2
```

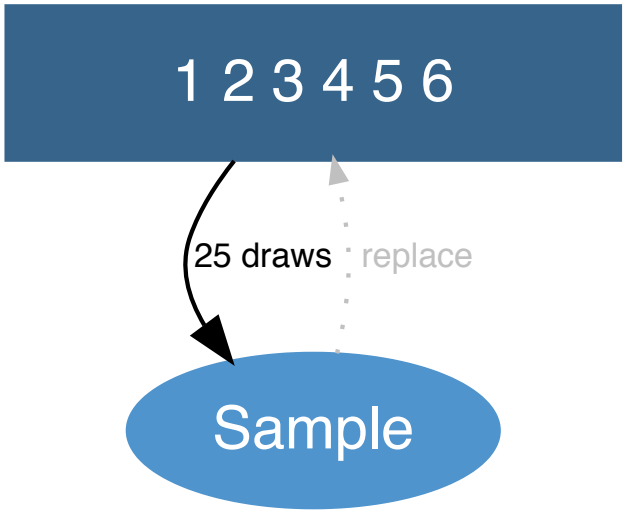


Illustration 1

Roll a dice 25 times

- the distinct numbers that go in the box ("tickets"): [1, 2, 3, 4, 5, 6]
- the number of each kind of tickets in the box: [one of each]
- the number of draws from the box: draw one ticket 25 times.

Simulation

```
set.seed(1)
```

```
## Warning in set.seed(1): '.Random.seed' is not an integer vector but of type
## 'NULL', so ignored
```

```
dietosses = sample(c(1:6), 25, repl = T)
dietosses
```

```
##  [1] 2 3 4 6 2 6 6 4 4 1 2 2 5 3 5 3 5 6 3 5 6 2 4 1 2
```

Example4 (Multiple Choice Quiz)

- A quiz has 25 multiple choice questions, with 5 answers each.
  - you get 4 points for a correct answer
  - you lose 1 point for an incorrect answer
- If you haven't studied and guess each answer, what is your expected score?

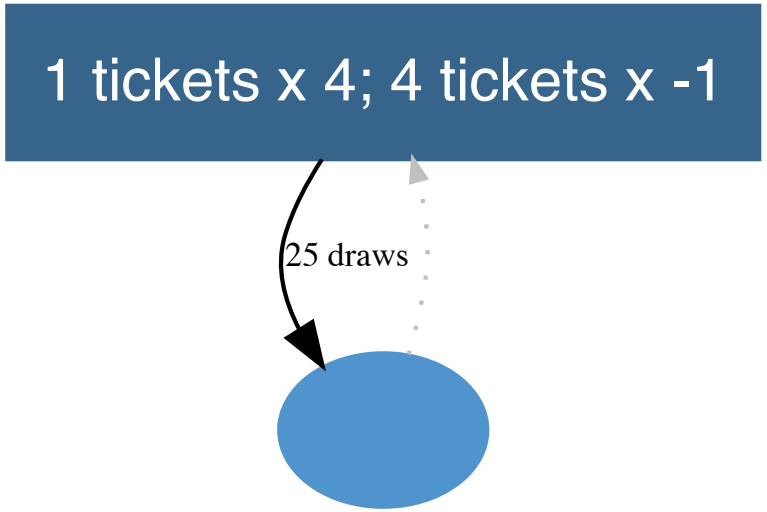
Illustration 2

Roll a dice 25 times

- the distinct numbers that go in the box ("tickets"): [+4, -1]
- the number of each kind of tickets in the box:

  [one ticket of +4 ; four tickets of -1]
- the number of draws from the box: draw one ticket 25 times.

```
## Warning: package 'DiagrammeR' was built under R version 3.3.2
```



Simulation

```
set.seed(1)
answer = sample(c(4, -1), 25, repl = T, prob = c(1/5, 4/5))
head(answer)
```

```
## [1] -1 -1 -1  4 -1  4
```

```
cumsum(answer)[25]
```

```
## [1] 0
```

```
table(answer)
```

```
## answer
## -1  4
## 20  5
```

Thank you