Topic 1: Introduction

ENVX2001 – Applied Statistical Methods

Januar Harianto

Introduction

- Coordinator - Floris van Ogtrop; floris.vanogtrop@sydney.edu.au
- **Januar Harianto** - Lecturer for Topics 1, 2 & 3
  - januar.harianto@sydney.edu.au
  - thomas.bishop@sydney.edu.au
- Additional contact details on Canvas

About me

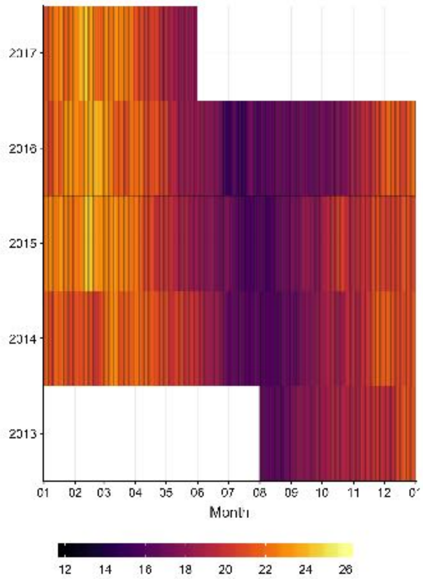
Ecophysiologist and climate scientist

One Tree Island

Marine invertebrates - echinoderms

Sea urchin

Driven to R by necessity (but I love it!)



Schedule

Lectures

- Check Canvas for final locations (e.g. Bosch is currently **CLOSED** due to rain damage)
- Tuesdays, 8:00 AM
- Wednesdays, 9:00 AM

Practicals (Labs) at Australian Technology Park

- 3 hours - with computers - see personal timetable
- Thu, Fri: **10am-1pm**
- Fri: **2pm-5pm**

Australia Technology Park (ATP)

Address:

Biomedical Building (C81)

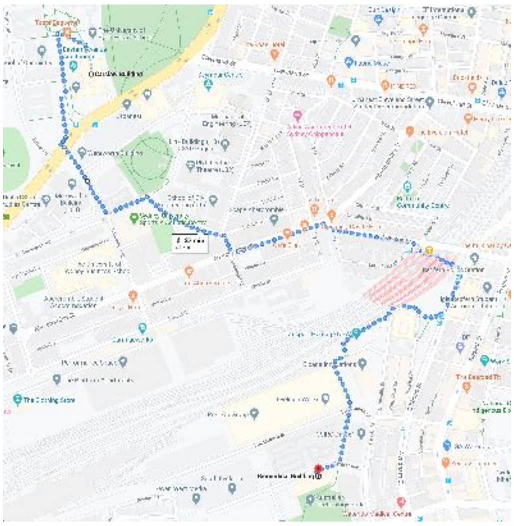
1 Central Avenue

Australian Technology Park,

Eveleigh, NSW 2015

! Important !
The ATP is a 30-minute walk from Carslaw Building.

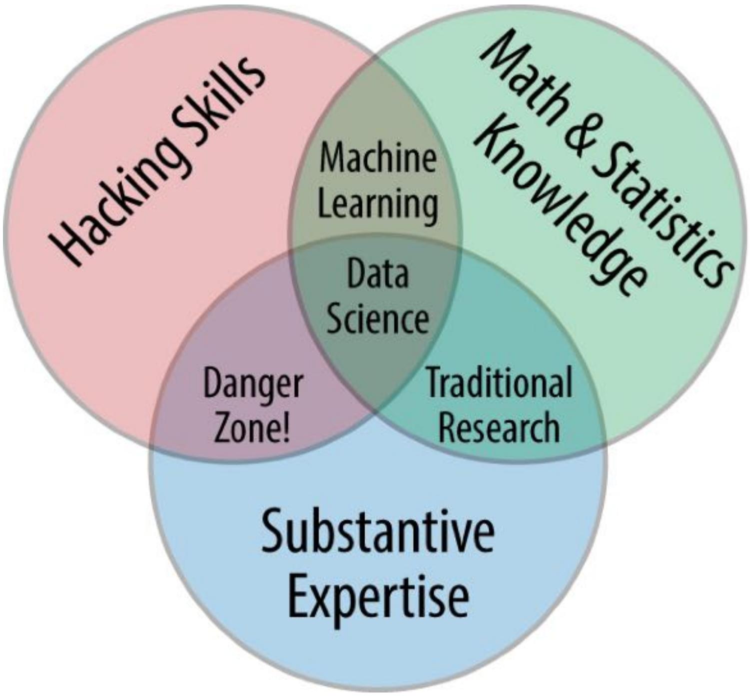
Wakling instructions are here.



Outline

Week	Lecturer	Lecture Topic	Assessment	Lab
Part 1: Designed Studies				
1	Januar Harianto	Introduction: Surveys		Lab 1
2	Januar Harianto	Surveys: Sampling designs		Lab 2
3	Januar Harianto	ANOVA I: One-way Analysis of Variance (ANOVA)		Lab 3
4	Aaron Greenville	ANOVA II: Introduction to experimental design	Report 1	No Labs
5	Aaron Greenville	ANOVA III: ANOVA with blocking		Lab 4
6	Aaron Greenville	ANOVA IV: ANOVA with 2 or more factors		Lab 5
Part 2: Finding Patterns in Data				
7	Liana Pozza	Regression I: Multiple linear regression	Report 2	Lab 6
8	Liana Pozza	Regression II: Variable selection		Lab 7
9	Liana Pozza	Regression III: Predictive modelling		Lab 8
10	Mathew Crowther	Multivariate analysis I: Principal component analysis (PCA)		Lab 9
11	Mathew Crowther	Multivariate analysis II: Clustering		Lab 10
12	Mathew Crowther	Multivariate analysis III: MDS and MANOVA		Lab 11
Part 3: Revision				
13	TBA	Revision	Presentation	Lab 12

Data Science

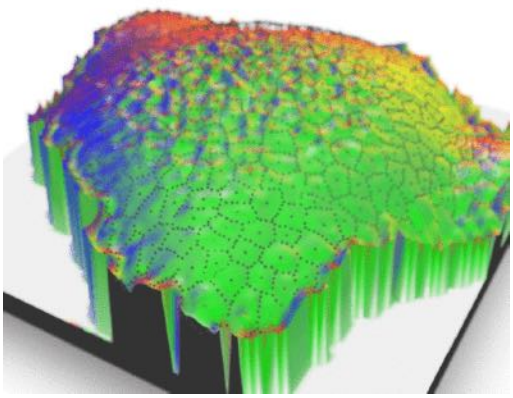


The Data Science Venn diagram by Drew Conway. Source

Big Data

Mechanical stress on plant tissue

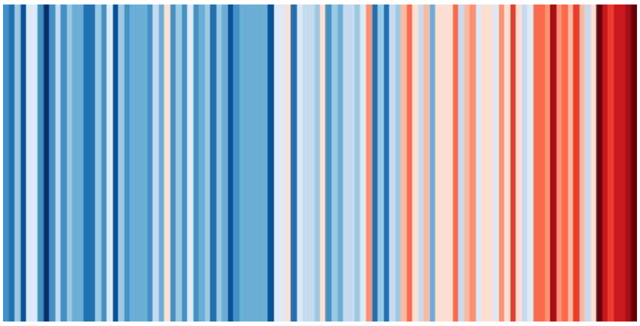
Paper source



R Code available here.

Warming in Australia, 1910-2019

Data Source



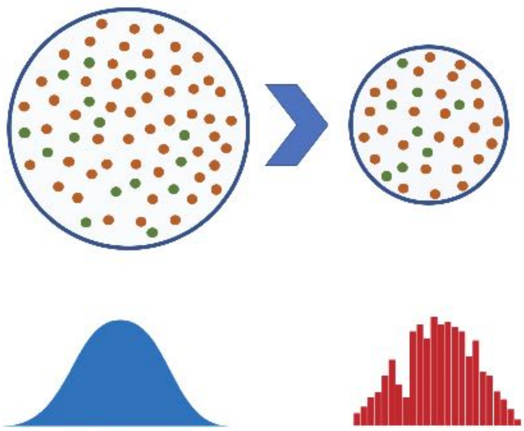
Each stripe is one year. Highest temperature is 1.5 °C above the baseline 1961-90 average.

Instructions for R here.

Learning outcomes

- Demonstrate proficiency in designing sample schemes and analysing data from them using using R;
- Describe and identify the basic features of an experimental design; replicate, treatment structure and blocking structure;
- Demonstrate proficiency in the use or the statistical programming language R to apply an ANOVA and fit regression models to experimental data;
- Demonstrate proficiency in the use or the statistical programming language R to use multivariate methods to find patterns in data
- Interpret the output and understand conceptually how its derived of a regression, ANOVA and multivariate analysis that have been calculated by R;
- Write statistical and modelling results as part of a scientific report;
- Appraise the validity of statistical analyses used in publications.

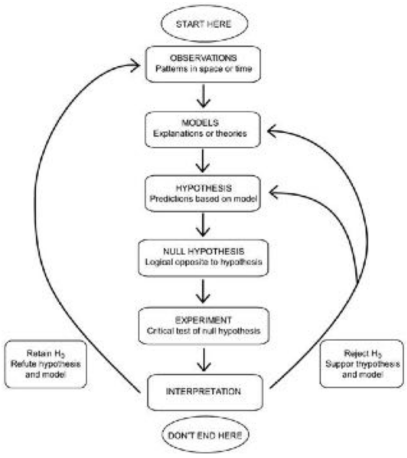
Learning outcomes



Sample and analyse data properly

Image source: Instance Selection: The myth behind Data Sampling

Learning outcomes



Design robust experiments

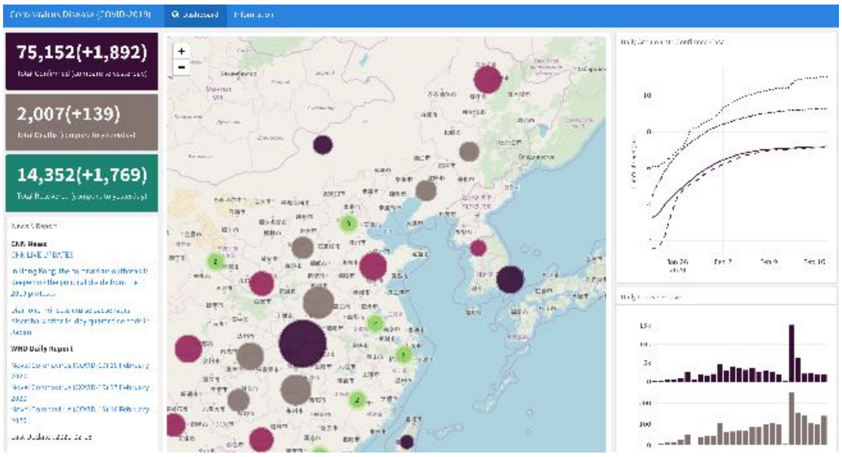
Image source: Underwood (1996) Experiments in Ecology, Cambridge University Press

Learning outcomes

Do in **R**: hypothesis testing, multivariate analysis, regression modelling

Image source: CRAN

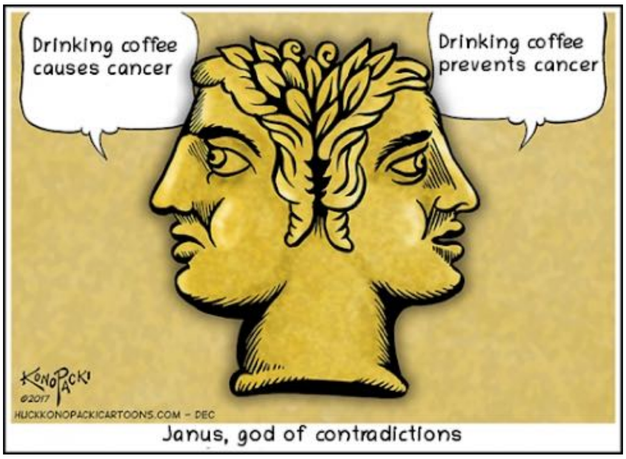
Learning outcomes



Present data as reproducible reports

Image source: COVID-19 Report Dashboard

Learning outcomes



Evaluate the validity of existing research

i.e. recognise good or bad research practices

Image source: Why cognitive bias is a researcher's worst enemy

Assessment

The latest information about assessments are found in the Unit outline.

Description	Date	Weight
Online quizzes (10)	Multiple weeks	5%
Report 1	Week 4	10%
Report 2	Week 7	10%
Presentation	Week 13	20%
Exam	TBA	55%

- **Reports**: for Topics 1 - 6
- **Semi-weekly quizzes**: online multiple choice questions based on the lectures and practicals
- **Exam**: allowed 1 A4 double-sided page of notes + provided equation sheet
- **Presentation**: See Canvas for details

Reference material

1. Lectures Slides
2. Practical Exercises
3. Books
    - Mead R, Curnow RN, Hasted AM (2002) Statistical methods in agriculture and experimental biology.
    - Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press: Cambridge, UK.

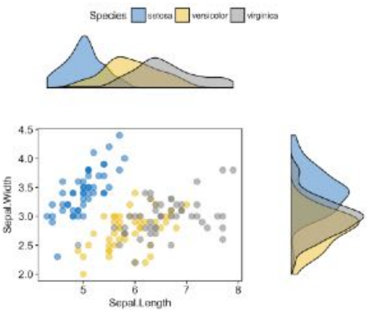
Extra help!

- edstem.org - access through Canvas
- Practicals and Tutorials - see Canvas
- Drop-in session: Mondays 11 am - 1 pm, from Week 3
- Appointments with lecturer
- Online documentation
  - https://stats.stackexchange.com/ - questions about stats
  - https://rseek.org/ - questions about R (also stats in R)
  - Google

The past, present and future

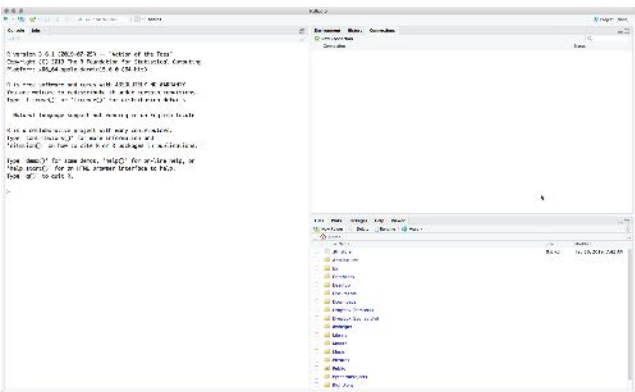
- Core for most of you:
  - 1st year: Introduction to statistical methods (ENVX1002). Code shared with DATA1001
  - **2nd year: Applied statistical methods (ENVX2001)**
- Elective
  - 2nd or 3rd year: Statistics in the natural sciences (ENVX3002)

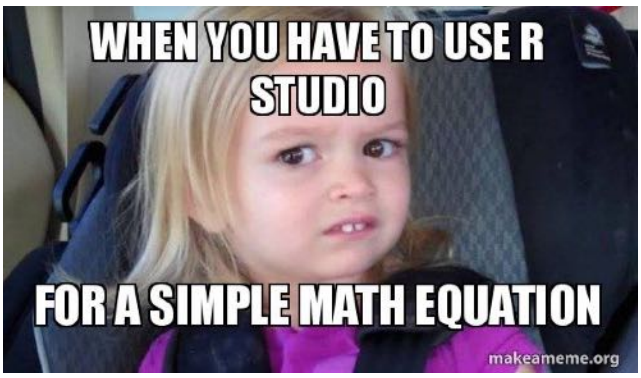
- Free and open source
- Download from CRAN



- More than 15,000 packages and counting

- Download from RStudio website
- Integrated Development Environment





Learning R

- RGuide by the School of Mathematics and Statistics - *short and sweet*
- R Module (under development) - *new to R*
- Lab 1 Session will re-introduce you to R, but try R before coming to the Lab!

Source: Adventure Time Season 1 Ep 25

Outline: Topic 1

Designs

- Sample vs. experimental designs

Revision

- mean, variance, standard error
- central limit theorem
- confidence intervals

Outline

Designs

- Sample vs. experimental designs

Revision

- mean, variance, standard error
- central limit theorem
- confidence intervals

Designs: Learning Outcomes

At the end of this topic students should be able to:

- Explain differences between
    - samples & populations
    - standard error & standard deviation;
- Describe key features of their data using
    - summary statistics,
    - graphical summaries and
    - confidence intervals;
- Demonstrate proficiency in the use of R for calculating summary statistics and generating graphical summaries and performing 1-sample t-tests.

Designs: Why do we care?

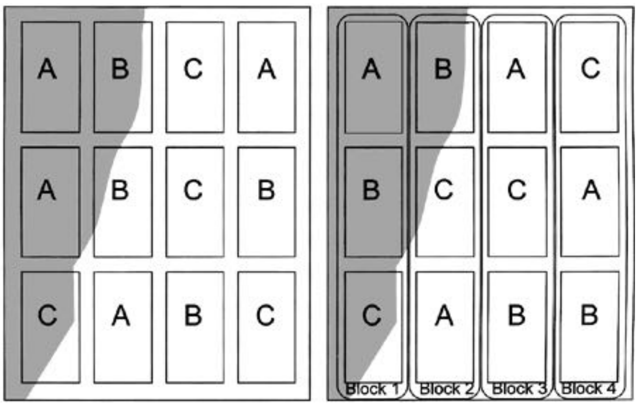
> *"To call in a statistician after the experiment has been done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."*
>
> Ronald Fisher

Also:

> (On a badly designed experiment).
>
> *"That's not an experiment you have there, that's an experience."*
>
> Ronald Fisher

More about R. Fisher.

Designs: Example



- **Completely random design** vs. **randomised block design**.
- Treatments (A, B, and C) are randomised on the left.
- On the right, treatments are replicated and *blocked* - each block contains one plot of each treatment.
- Shade could represent *anything* - patterns, shades, environmental factors.

Designing Research and Demonstration Tests for farmers' fields. Source.

Designs: What is an experiment?

> *"...a procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact."*
>
> Oxford Dictionary

Two types of experiments:

Controlled.experiments	**Observational.studies**
Comparative	Absolute
Manipulative	Mensurative

Designs: Different types of science

- Controlled experiments
- Observational studies
- Modelling
- Model development
- Methodology development

Designs: A video

Revision

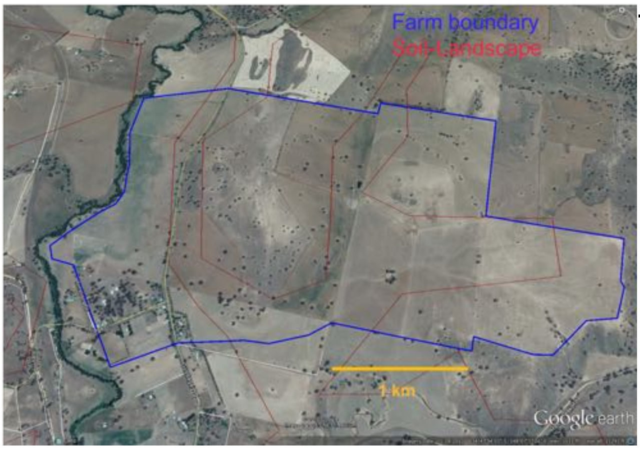
Designs

- Sample vs. experimental designs

Revision

- mean, variance, standard error
- central limit theorem
- confidence intervals

Data Story



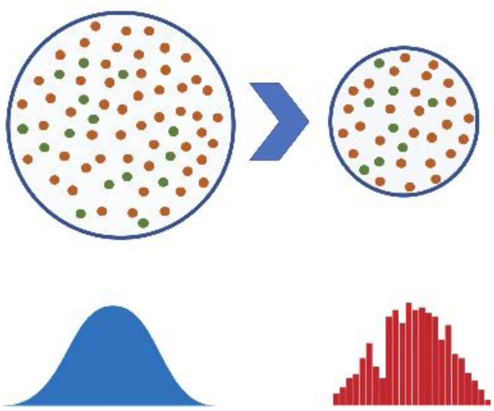
- Sequestered soil carbon is worth $50/tonne if measured
- It costs $100 to collect and analyse a soil sample for soil carbon
- Need an estimate of mean carbon content for property
- Is it worth measuring for a land holder?

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

Population vs. samples

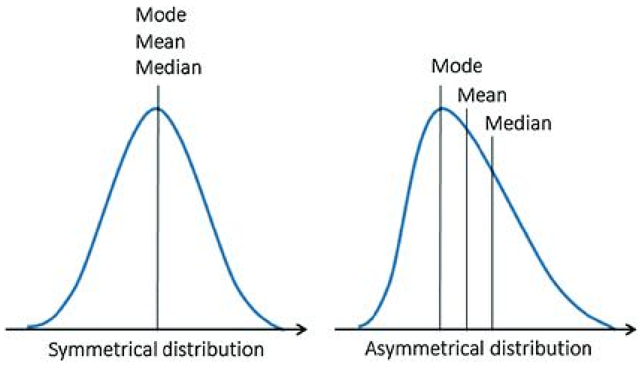
When we take a sample from a larger population...

What information does the sample give about the population and how reliable is that information?



Summary statistics

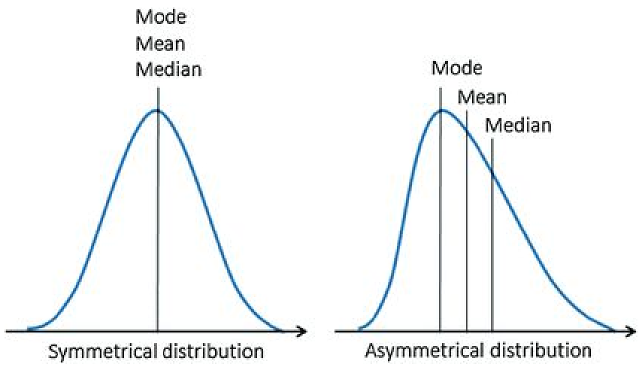
- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of spread or dispersion
  - Range
  - Interquartile range
  - Standard deviation / Variance



Komorowski et al. (2016) Exploratory Data Analysis.

Summary statistics

- Measures of central tendency
  - **Mean**
  - Median
  - Mode
- Measures of spread or dispersion
  - Range
  - Interquartile range
  - **Standard deviation / Variance**



Komorowski et al. (2016) Exploratory Data Analysis.

Data story

What is the mean soil carbon content?

How confident are we that this represents the true mean?

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

Arithmetic mean

- Population mean, μ: sum of all values of a variable divided by the number of objects in the population

$$\mu = \frac{\sum_{i=1}^{n} y_i}{N}$$

- Sample mean is based on a subset of n objects from a population of size N

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

The Σ symbol refers to the *sum*, and conveniently displays $x_1 + x_2 + x_3 + \cdots + x_n$.

Data story

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

What is the mean soil carbon content?

```
soil <- c(48, 56, 90, 78, 86, 71, 42)
mean(soil)
```

```
## [1] 67.28571
```


Data story

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

What is the mean soil carbon content?

```
soil <- c(48, 56, 90, 78, 86, 71, 42)
mean(soil)
```

```
## [1] 67.28571
```

How confident are we that this represents the true mean?

Data story

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

How confident are we that this represents the true mean?

Data story

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

How confident are we that this represents the true mean?

```
sd(soil)
```

```
## [1] 18.8566
```

```
length(soil)
```

```
## [1] 7
```

Is this information enough?

Variance and Standard Deviation (SD)

- Metrics to describe *variation* around arithmetic mean
- Variance
  - Describes variation in *squared* deviations of the mean, i.e. $unit^2$
  - Population variance: $\sigma^2 = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{N}$
  - Sample variance: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}$
- Standard deviation
  - Describes variation in *original* units
  - Population standard deviation: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{N}}$
  - Sample standard variation: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n-1}}$

Distributions

- When we measure things, we are taking a random sample, n, from a larger population, N
- From this we calculate statistics, e.g. the mean, but if we repeatedly did this we would observe different values of the mean – this the **sampling distribution**
- Since we can only sample 1 time, what do we know about the **sampling distribution**?

Distributions

- **Population** distribution -- distribution of all individuals in the population
- **Sample** distribution -- distribution of all individuals in the sample
- **Sampling** distribution -- distribution of a statistic from all possible samples

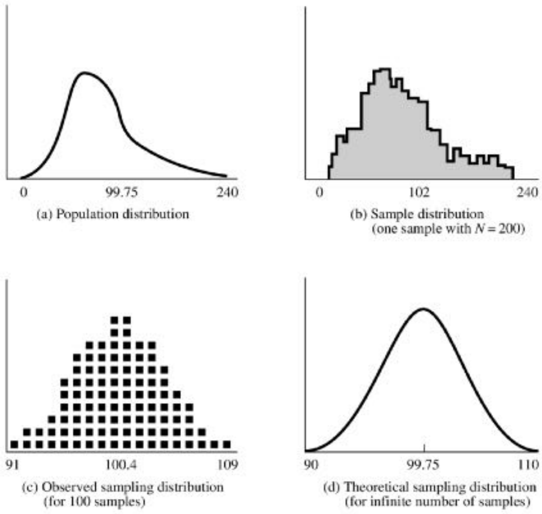
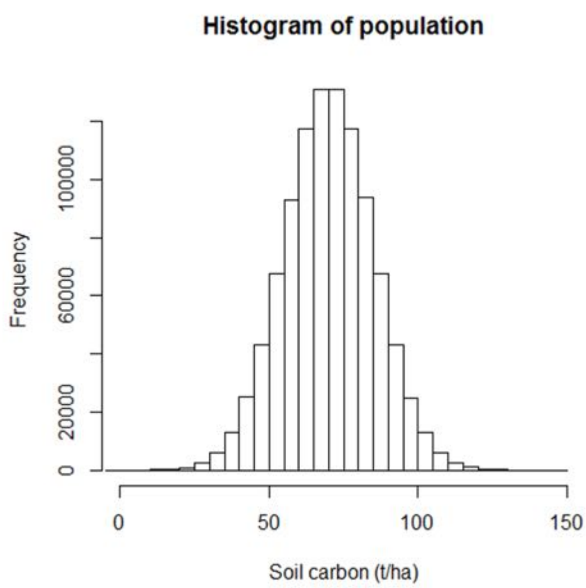
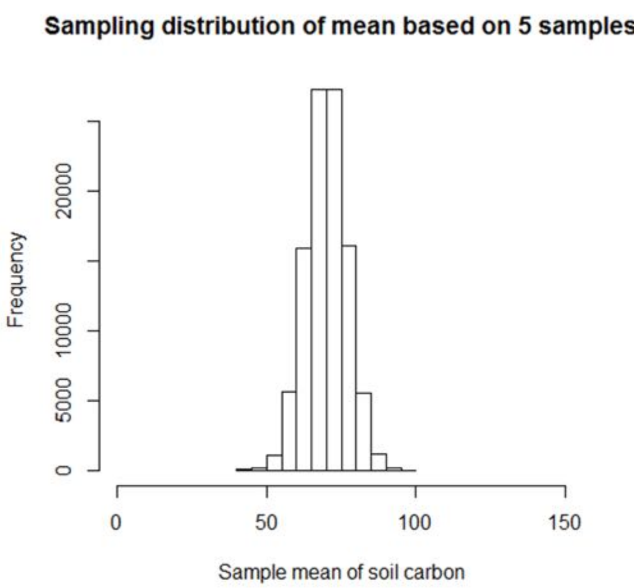


Image Source

Distributions - Example





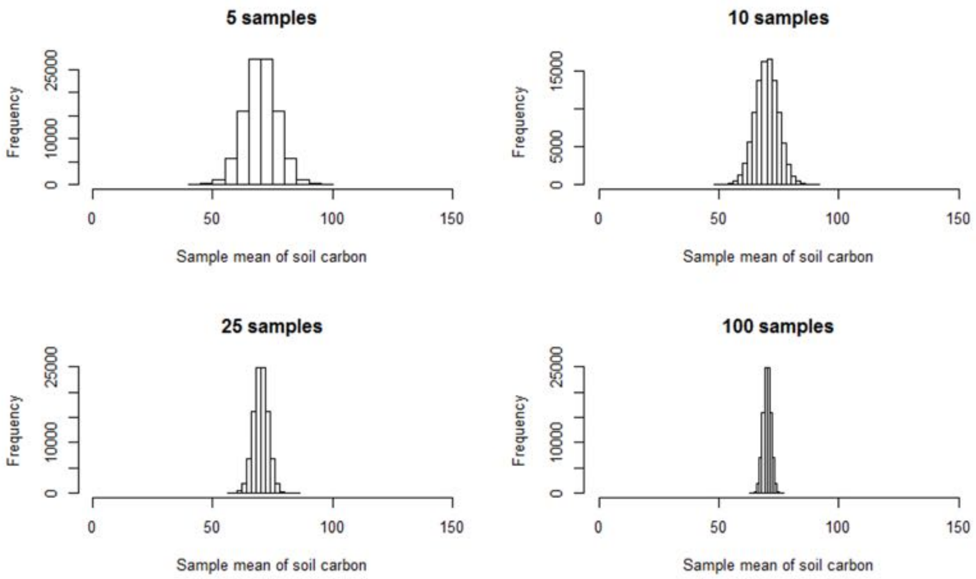
Standard error of the mean

- Like the standard deviation of a distribution, called standard error to avoid confusion
- Tells us how well we know the mean

$$se(\bar{y}) = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

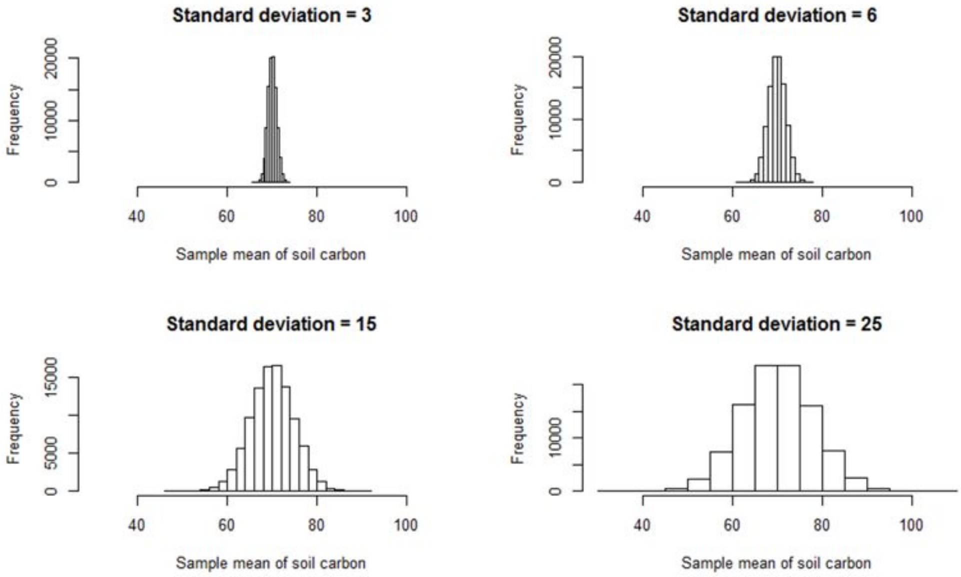
- Depends on number of observations (n) and variation in the data (σ)

Effect of sample size



Soil Carbon ~ $N(70, 15^2)$

Effect of variation



Soil Carbon ~ $N(70, x^2)$

Data story

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

How confident are we that this represents the true mean?

Step 1: what is the standard deviation? what is the sample size?

```
sd(soil)
```

```
## [1] 18.8566
```

```
length(soil)
```

```
## [1] 7
```

Step 2: what is standard error of the mean?

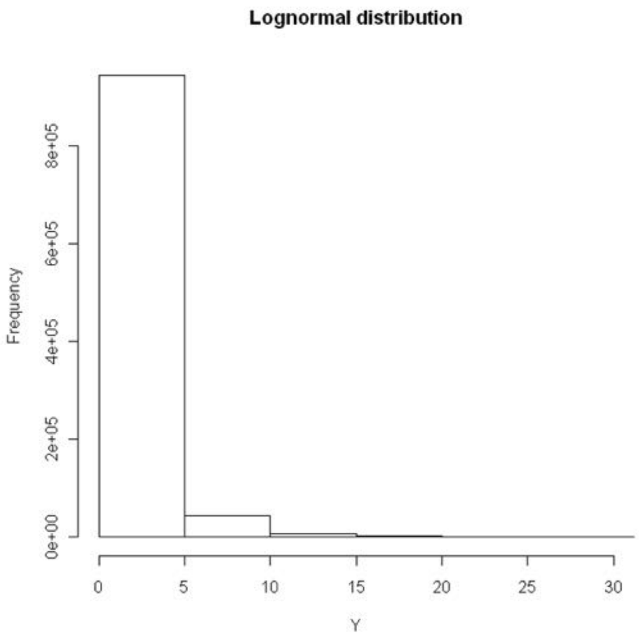
```
sqrt(var(soil)/7) # sem manual calculation
```

```
## [1] 7.127126
```

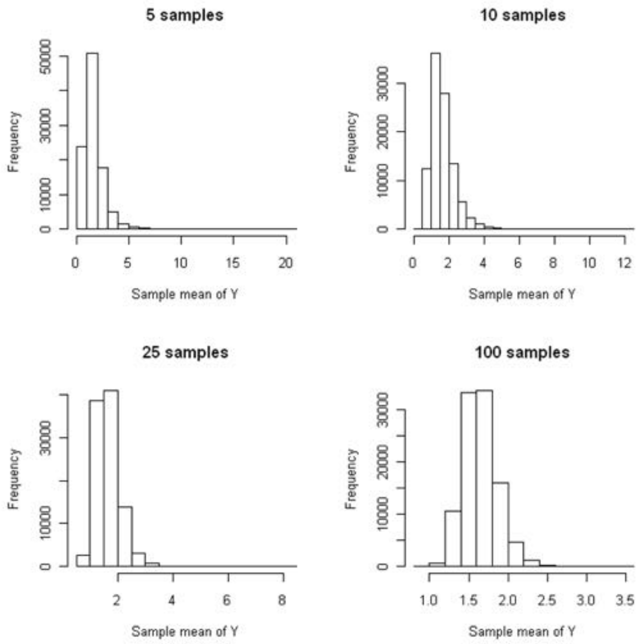
Central Limit Theorem

- If data is normally distributed then the distribution of sample means is normally distributed
- **Central limit theorem**: for almost all distributions (log, gamma), as n (the sample size) increases, the distribution of sample means tends to become more normal

Central Limit Theorem



Central Limit Theorem



Confidence intervals

- Consist of an
  - interval [lower, upper limit]
  - degree of confidence [a %, e.g. 95%]
- Definition: informally it is the interval in which we would expect to find the population mean (μ). More formally if we worked out a series of CIs for a series of samples, then 95% of the CIs would contain the population mean.
  - CI's can be estimated for other parameters but so far we have only focused on the mean
  - The 95% CI for the mean is:

$$95\%CI = \bar{y} \pm t_{n-1}^{0.025} \times se(\bar{y})$$

T-probability tables

A.4 Some right-tail critical values for the Student's T distribution

The distribution tabulated is that of Student's t. The first column is the degrees of freedom (df). The remaining columns give either the one tailed (upper tail) critical values so that $P(T_{df} > t) = P$, or the two tailed critical values so that $P(T_{df} > t \text{ or } T_{df} < -t) = P$ where P is the probability shown at the top of the columns.

df	P						
	0.10	0.05	0.025	0.01	0.005	0.001	(1 tailed)
	0.20	0.10	0.05	0.02	0.01	0.002	(2 tailed)
1	3.078	6.314	12.706	31.821	63.657	318.313	
2	1.886	2.920	4.303	6.965	9.925	22.327	
3	1.638	2.353	3.182	4.541	5.841	10.215	
4	1.533	2.132	2.776	3.747	4.604	7.173	
5	1.476	2.015	2.571	3.365	4.032	5.893	
6	1.440	1.943	2.447	3.143	3.707	5.208	
7	1.415	1.895	2.365	2.998	3.499	4.785	
8	1.397	1.860	2.306	2.896	3.355	4.501	
9	1.383	1.833	2.262	2.821	3.250	4.297	
10	1.372	1.812	2.228	2.764	3.169	4.144	
11	1.363	1.796	2.201	2.718	3.106	4.025	
12	1.356	1.782	2.179	2.681	3.055	3.930	
13	1.350	1.771	2.160	2.650	3.012	3.852	
14	1.345	1.761	2.145	2.624	2.977	3.787	
15	1.341	1.753	2.131	2.602	2.947	3.733	
16	1.337	1.746	2.120	2.583	2.921	3.686	
17	1.333	1.740	2.110	2.567	2.898	3.646	
18	1.330	1.734	2.101	2.552	2.878	3.611	
19	1.328	1.729	2.093	2.539	2.861	3.579	
20	1.325	1.725	2.086	2.528	2.845	3.552	

T-probability tables

```
qt(α/2,degrees of freedom) # pseudo-code
```

Find the upper 2.5th percentile of the Student t distribution with 10 degrees of freedom.

```
qt(c(.025), df = 10)
```

```
## [1] -2.228139
```

Data story

- Soil carbon content was measured at 7 points across a farm
- The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha)

How confident are we that this represents the true mean?

```
t.test(soil)
```

```
## 
##     One Sample t-test
## 
## data:  soil
## t = 9.4408, df = 6, p-value = 8.034e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  49.84627 84.72516
## sample estimates:
## mean of x 
##  67.28571
```


Thanks!

Slides created via the R package **xaringan**.

References

- ENVX1002 Manual
- Quinn & Keough (2002)
  - Chapter 1, Chapter 2: Sections 2.1-2.3
- Mead et al. (2002).
  - Chapter 1-2, Chapter 3: Sections 3.1-3.3.