ELEMENTARY STATISTICS

Introduction

- Administration
- Statements, Tables and Diagrams

"It is of the highest importance in the art of deduction to be able to recognise out of a number of facts which are incidental and which are vital. Otherwise your energy and attention must be dissipated instead of concentrated."

Sherlock Holmes, The Adventure of the Reigate Squires, 1894.

MAST90044 Thinking and Reasoning with Data

Lecturer

Dr Julia Polak, Room 206, Peter Hall Building
email: julia.polak@unimelb.edu.au

Lectures -

Monday	1:15 – 2:15 PM	Elisabeth Murdoch G06
Wednesday	1:15 – 2:15 PM	Redmond Barry 200 (Rivett Theatre)

Consultation times -

Monday	11:00 – 11:30 AM
Wednesday	11:00 – 11:30 AM
Wednesday	1:30 – 3:30 PM

MAST90044 Thinking and Reasoning with Data

Lab classes Opportunity to learn and put things into practice

- Lab instructions: designed to supplement lectures ('theory')
- Lab exercises
- Be prepared: print/download lab instructions and read through 'theory' and exercises before class

Peter Hall Building -

Monday	3:15 PM – 5:15 PM	Thompson Lab - G69
Tuesday	9:00 AM – 11:00 AM	Thompson Lab - G69
Tuesday	11:00 AM – 1:00 PM	Thompson Lab - G69
Tuesday	12:00 noon – 2:00 PM	Wilson Lab - G70
Wednesday	9:00 AM – 11:00 AM	Thompson Lab - G69
Thursday	3:15 PM – 5:15 PM	Thompson Lab - G69

MAST90044 Thinking and Reasoning with Data

Assessment

• 3 Assignments	15%	+	15%	+	20%	=	50%
Due in	week 5		week 8		week 11		
• Examination						=	50%

Software

Computing

- R will be used throughout the course.
- R can be downloaded free from the web.
- R is the standard statistical package used in this subject.
- R will be used in labs and assignments.
- R output will be used in the exam.

Google *'download R'*

http://cran.ms.unimelb.edu.au

Google *'download RStudio'*

https://www.rstudio.com/products/rstudio/download/

Housekeeping

- Plagiarism declaration - to complete with 1st assignment
- Student reps?
  - SSLC meeting will be on Tuesday 19th March (Week 3)
  - SSLC Survey conducted week 4
  - Results Collated and sent to representatives and Lecturers within week 5
  - SSLC representatives and Lecturers discuss results with class within week 6

Housekeeping

- In the UNLIKELY case of Emergence
  - Stand fast and push chairs, large bags, etc under desks or benches.
  - Turn off electrical devices that are not safe to be left unattended.
  - In controlled sequence, move along gangways to main aisles and exit in an orderly manner through the nearest appropriate exit.

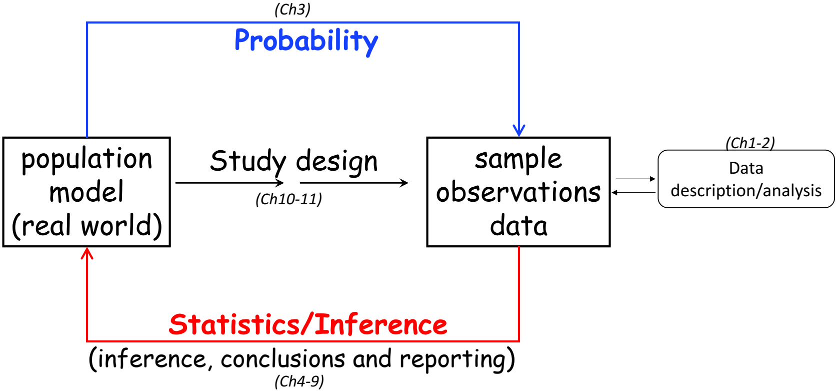
Objectives

After completing this subject, students will understand the basic principles of sampling and experimental design, how the results of statistical analyses are interpreted and reported, the statistical thinking behind common statistical procedures, and will be able to carry out a range of standard statistical techniques in R.

Course outline

1. Introduction. Statements and diagrams.
2. Software, tables, data handling, descriptive statistics and graphical methods.
3. Statistical models, sampling and sampling distributions.
4. Point and interval estimation for categorical data.
5. Hypothesis testing for categorical data.
6. Estimation and hypothesis testing for continuous data.
7. Simple linear models.
8. Multiple regression and model selection.
9. Logistic regression.
10. Design of experiments.
11. Analysis of experiments.
12. The bootstrap.

Probability and Statistics



Population and sample

Population

the entire group of units ('subjects', 'individuals' etc) under study, which may be (and often is) hypothetical.

Sample

the observed units i.e. the units on which we have information (measurements).

Statistics

use the sample to make inferences about the population; the real world!

The *sample* is assumed to be *representative* of the *population*.

Probability and Statistics

- Probability
  specify models for the data:
  "if the population is like *this*, then the sample will be like *that*".
- Statistics or Statistical Inference
  work out what the population is like based on the data:
  "if the data are like *this*, then the population will be like *that*"
- Data analysis
  ways and means of describing and representing the data obtained.
- Study Design
  how we obtain the data so that our conclusions can be applied to the general population

Population and sample: examples

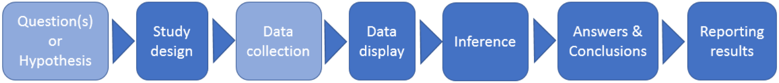
Representative samples: steps to correct inference.

- if sample = 50-59yo men from a Melbourne clinic,
  then what is the population? ie on whom does inference apply?
  men from Melbourne? Victoria? Australia? ...
  to women? other age groups?
- if want to make inferences about lead level of mussels in Port Phillip Bay
  how to sample (what study design)? from Geelong? from Elwood? where are the mussels? how many locations in the bay? how many mussels? ...

Statistics is “seeing through the data”

We want to be able to say something about the **real world** (population) on the **basis** of the information in the **data** (sample).

Holmes would have me tell you that “Data analysis is like detective work: finding evidence and investigating it.” But there’s more!



these steps involve statistics

Statements

A product of data analysis and/or statistical inference.

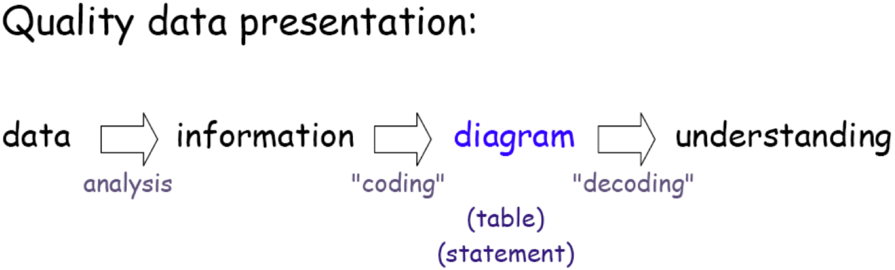
Any statement should be accurate and clear if it is to represent a data set.

This is not always the case...a statement is poor if

- the data quality itself is poor
- the data analysis is incorrect or inappropriate
- the quantitative statement is distorted; e.g. selectively abbreviated or added to.

The media provides many examples!!

Diagrams



Quality data presentation:

data → information → diagram → understanding

analysis

"coding"

"decoding"

(table)

(statement)

Edward Tufte:
“The Visual Display of Quantitative Information”
William Cleveland:
“The Elements of Graphing Data”

Diagrams

Encoding/Decoding paradigm for diagrams (Cleveland)
“coding” of the data into a diagram so that what the eye can perceive and “decode” with accuracy is, in decreasing order:

1. position on a common scale
2. position on similar scales
3. length
4. angle or slope
5. area
6. volume
7. colour

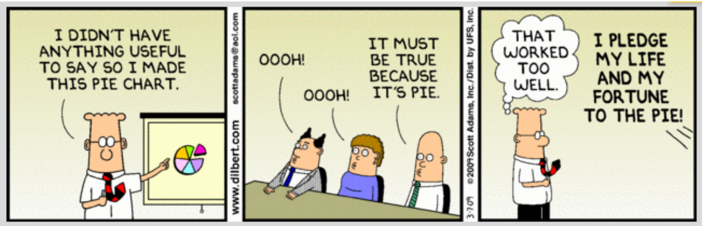
Principle: Encode data on a graph so that the visual decoding involves tasks as high as possible in the ordering.

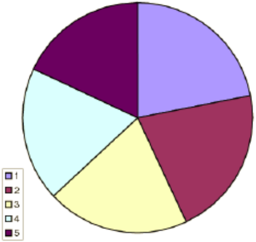
Diagrams

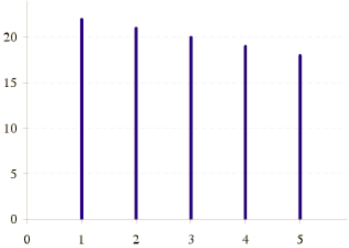
Good practice

- one, single idea per diagram (“paragraph”) that contains the data (“words”) arranged in coherent and meaningful ways (“sentences”)
- clear and informative labelling e.g. no guessing!
- common linear scale is easiest for us humans to understand and compare i.e. versus volumes, or angles, or lengths not on common scale...
- minimize amount of ink! including colour shading, i.e. don’t use it if it’s not communicating something
- avoid distortion from use of tricky perspective (e.g. 3D) or other artistic tricks

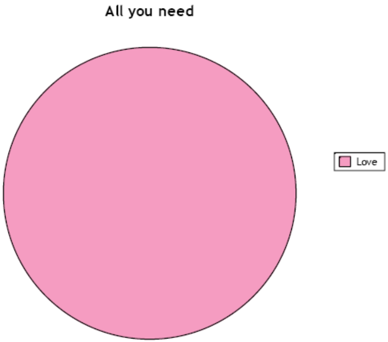
Pie charts are big on show but short on information.


I DIDN'T HAVE ANYTHING USEFUL TO SAY SO I MADE THIS PIE CHART.
scottadams@aol.com
www.dilbert.com
OOOH!
OOOH!
IT MUST BE TRUE BECAUSE IT'S PIE.
© 2009 Scott Adams, Inc./Dist. by UFS, Inc.
3-7-09
THAT WORKED TOO WELL.
I PLEDGE MY LIFE AND MY FORTUNE TO THE PIE!





Beatles' pie chart:



Misperception and miscommunication are certainly not special to statistical graphics,

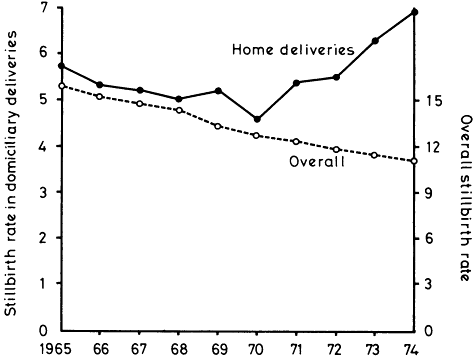
Bad graphs

Case study: Home births

- Context: debate about hospital versus home births in U.K.
- Comparison of outcomes is tricky because more difficult births tend to be in hospital.
- Complex issues of “control” and power or authority in a deeply personal matter (giving birth).
- Advocates of the merits of hospital births wrote a paper in the British Medical Journal (Fedrick and Butler, 1978).
- Outcome considered here: stillbirth rate: number of stillborn babies per 1,000 births.

Fedrick J, Butler NR. Intended place of delivery and perinatal outcome. British Medical Journal, 1978, **1**, 763–765.

Home births

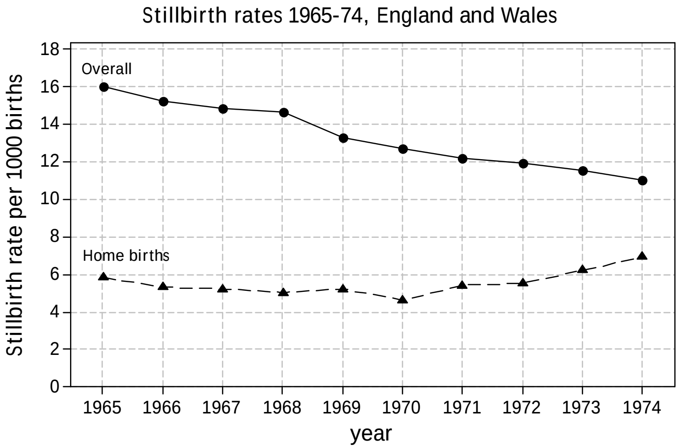


Stillbirth rate 1965-74 for home deliveries compared with all deliveries.

Figure as it appeared in the journal.

Home births

Revised graph:



Home births

Revised graph has several features that illustrate good graphical features:

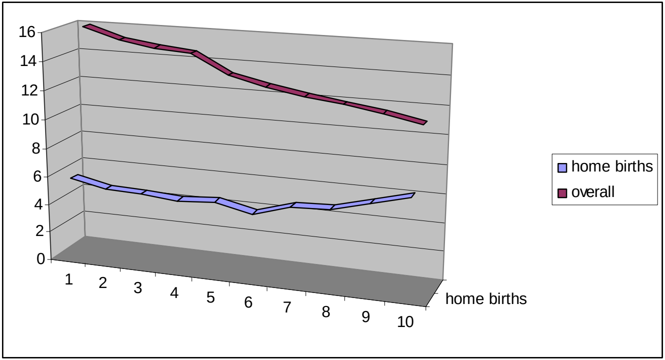
- Clear and informative title and labels, including units;
- Graphic invites the appropriate comparison and is not embellished;
- Relevant comparison is highlighted on a common, aligned scale;
- Faint grey grid helps to examine detail.

Bad graphs

- It is alarmingly easy to produce graphs that do not communicate effectively.
- Some causes:
  - Designer is too deeply engaged with the data at the time the graphic is created;
  - Stupid software features;
  - Careless use of software;
  - Decoration and aesthetics at the expense of graphical clarity.

Bad graphs

A graphic easily produced:

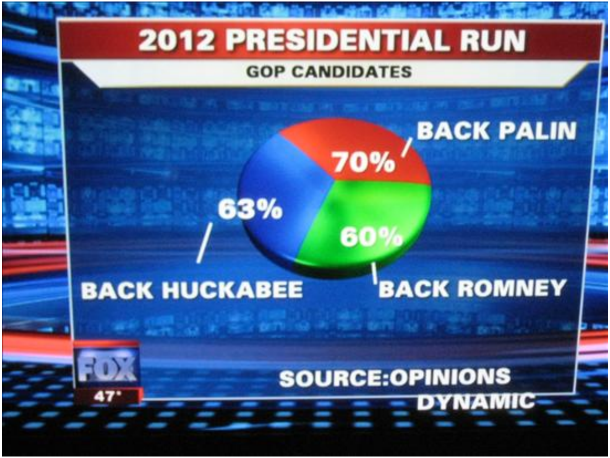


What's wrong with this?

Bad graphs

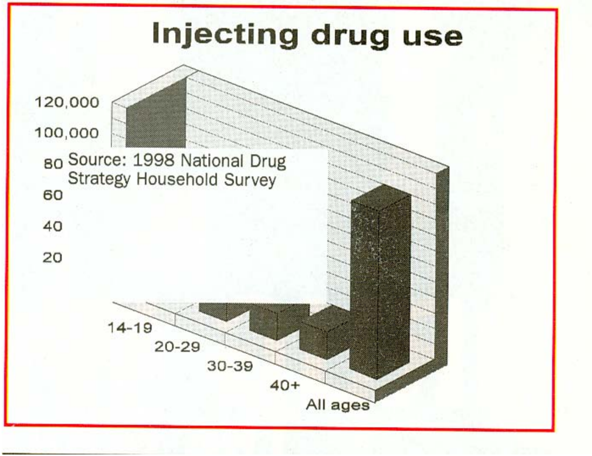
- 3-dimensional perspective is unhelpful;
- Grey background is unnecessary;
- Default scale of 1 to 10 is wrong;
- What is the label “home births” doing at the bottom right, and what happened to “overall”?
- No axis labels, title, units, context.

How bad can it get?



Republican Presidential Candidates, 2012; GOP refers to ‘Grand Old Party’, the Republican Party.

How bad can it get?



. . . pretty bad!

Source: ALIVE magazine, July 1999

What is important is the

data/ink ratio

You should aim to ensure it is large.

(optimal, not maximal)

`data/ink ratio`

This applies particularly in diagrams,

... but also in tables and statements!

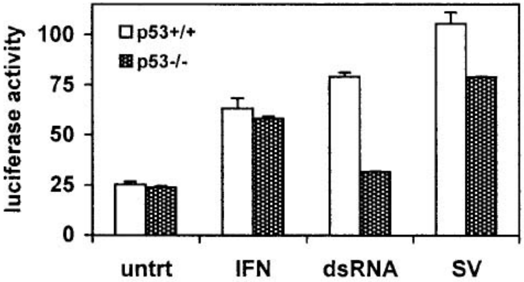
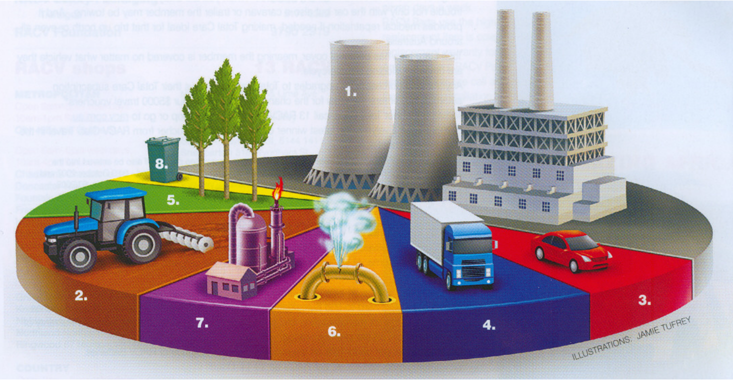


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6 × 10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

Examples of complexity reducing clarity

Greenhouse gas emissions by source, Australia

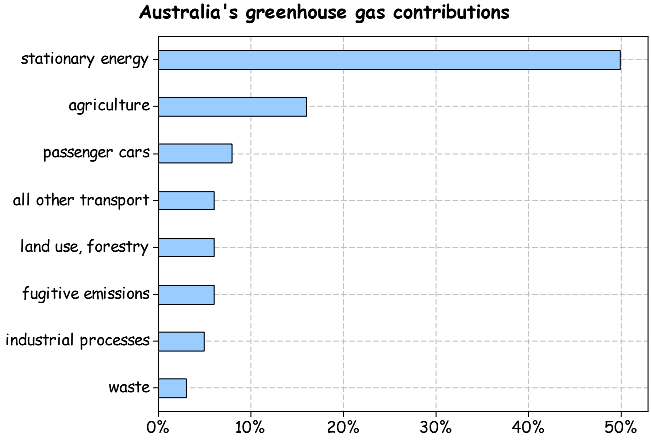


Source: *Royal Auto*, March 2008, p. 12.

Royal Auto pie chart

- Decoration gone mad; does not encourage sensible comparisons.
- Which category is larger, “passenger cars” (3) or “all other transport” (4)?
- *Note the importance of context . . .*
- The tractors are there to represent agriculture. Why?
    - The Australian Department of Climate Change does not even mention farm vehicles as one of the primary sources of greenhouse gas emissions from agriculture.
- Icons in pictographs may communicate misleading information.

Royal Auto data in a bar chart

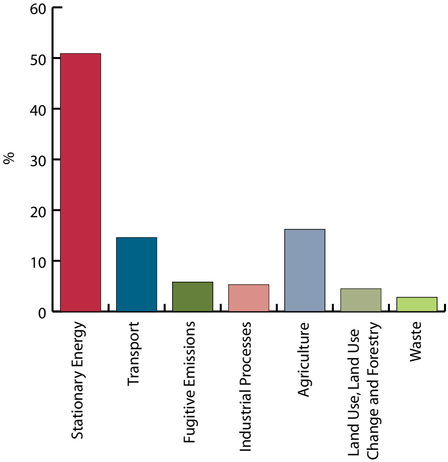


Royal Auto data: bar chart

- The ordering of the categories is clear.
- We don't have to guess what the icons mean.
- The percentages can be read off the graph quite accurately.
- The messages are clear *because the design is simple*.

Greenhouse gas data: Government graph

Figure 2.1 Contribution to total net CO_2-e emissions by sector, 2005



Source: www.greenhouse.gov.au/inventory/2005/pubs/inventory2005revised-nationalreportv1.pdf

Diagrams

Three rules of data analysis

1. Draw a diagram

   it will help you think about the data
2. Draw a diagram

   it may show unexpected features
3. Draw a better diagram

   it will help you tell others what you've found

Gapminder http://www.gapminder.org/ for displaying many variables, much information.

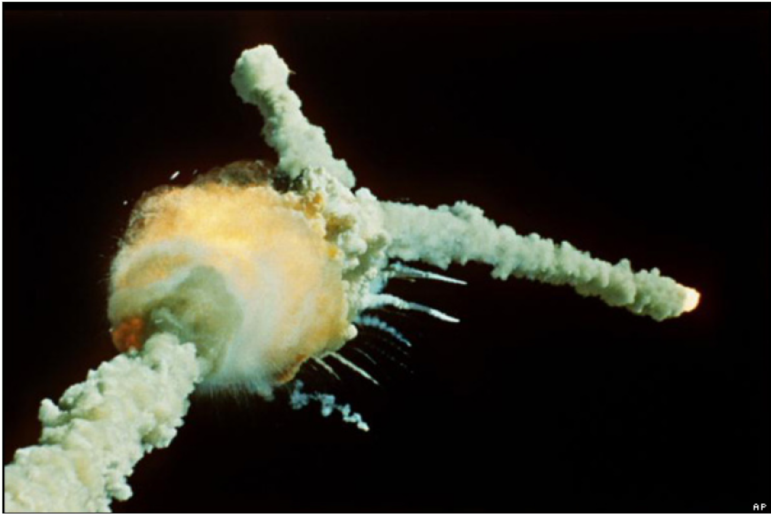
Data display

Example -Challenger Disaster

The Challenger Disaster occurred on the 28th of January 1986, when Space Shuttle Challenger broke apart after an O-ring seal failed at liftoff, leading to the deaths of its seven crew members.

Data display: Example -Challenger Disaster

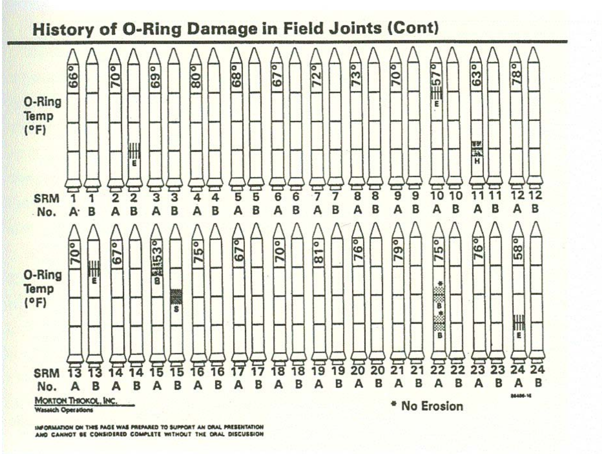
- The shuttle launch had been delayed several times, as had previous shuttle launches.
- The launch had an unusually high profile, due to the inclusion in the crew of a high school teacher, Christa McAuliffe.
- The Morton Thiokol engineers initially argued a lack of data:
  - "Don't launch: it's too cold and we're worried about the O-rings".
  - NASA: "Prove it".
- President Reagan was due to give the 'State of the Union' address on the evening of the launch.
- Under pressure, the engineers changed their minds and the launch poceeded



AP

Challenger disaster

Even in the post-disaster inquiry, the association between temperature and O-ring damage was poorly explored.



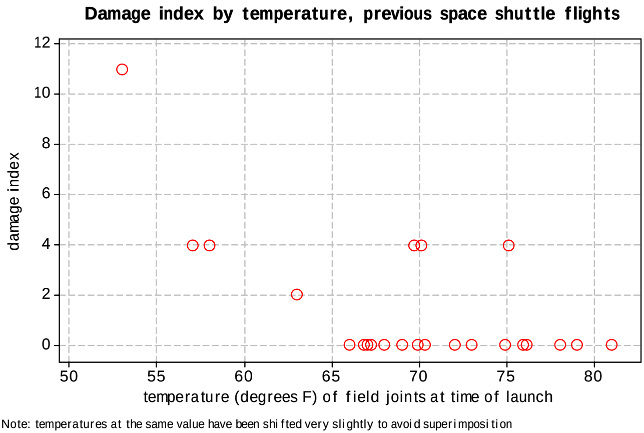
Data display: Example -Challenger Disaster

What does the data show?

- The worst damage occurred on the coldest previous flight: at 53 F ($\approx 11.67C$).
- There was more than one type of damage: blow-by, erosion, heating of the primary and secondary O-rings.
- There were many flights with no evidence of damage: these were all 66 F ($\approx 18.89C$) or warmer.
  - None of these flights were considered in the pre-launch debate.

Challenger disaster

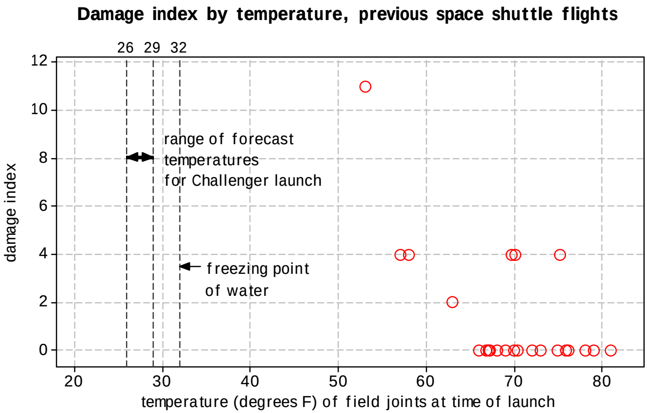
Tufte constructed a single damage index that combined the types of damage.



Data from Tufte E.R. *Visual Explanations*, p. 44.

Challenger disaster

And now include the context of the debate.



Note: temperatures at the same value have been shifted very slightly to avoid superimposition

Challenger Disaster

What's different? Show context.

- x-axis scale has shifted from 50-80 to 20-80
- water freeze point is displayed
- forecast temperature for challenger is displayed

Software: R

- R is an open-source, freely-available statistical environment.
- Cross-platform i.e. operates on Mac, Windows, Linux.
- Object-oriented.
- Small footprint i.e. uses little memory.

These are the prompts:

>

+

Communication with R

- Type commands at the prompt and they will be executed
  ```
  > 1+1
  ```
- Write scripts and load them
  ```
  > source("script.r")
  ```
  and they will be executed.

This is the special character:
(Everything after it on the same line is a comment.)
#

File Structure

One directory for each lab class, e.g. lab01.

Within each lab class directory, have these directories:

- `data` – data sets;
- `graphics` – pdf files;
- `notes` – documents, resources;
- `scripts` – R code;
- `images` – R objects.

Working Directory

Relevant commands

- `> getwd()`

  Brackets () may or may not contain `arguments`.

- Setting the working directory (example):

  ```
  > setwd("d:/ms/MAST90044/scripts")
  ```

Workspace

The **workspace** is the container of all your objects.

Relevant commands

- `> ls()` List objects.
- `> rm(`*`object`*`)` Remove an object.
- `> rm(list=ls())` Remove all objects.

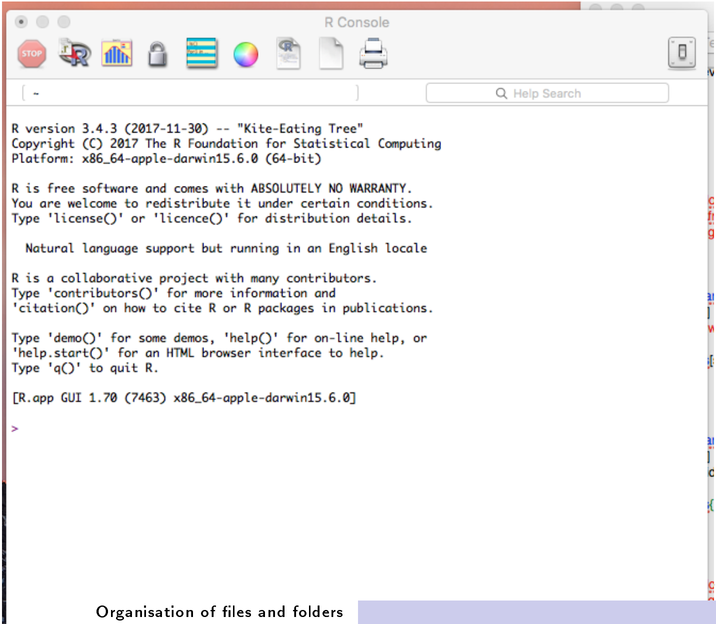
Saving graphics

Two main ways of saving graphics, either with commands or with the menu.

Relevant commands

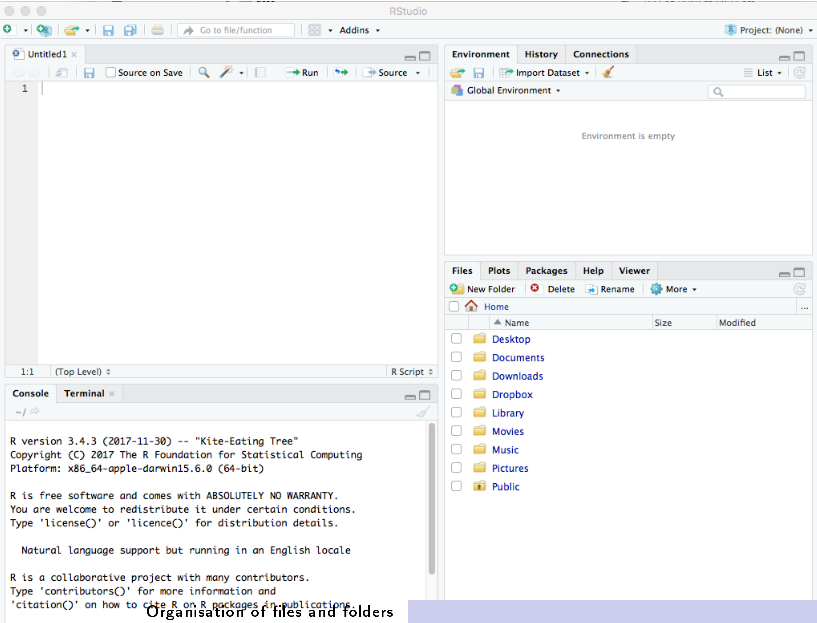
- `> pdf("../graphics/filename.pdf")` Location and filename of graphic.
- `> hist(ufc$dbh.cm)` Create the plot.
- `> dev.off()` Don't forget to switch off graphic device!

R



```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

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

  Natural language support but running in an English locale

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

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

[R.app GUI 1.70 (7463) x86_64-apple-darwin15.6.0]

>
```

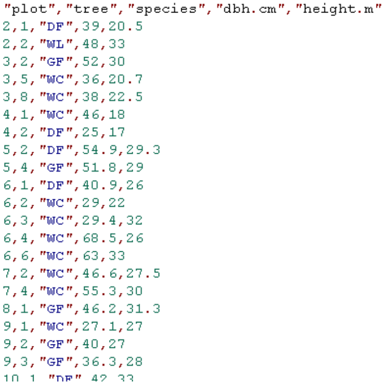
Rstudio



```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

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

  Natural language support but running in an English locale

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

Reading data into R

text-reader (ufc.csv):



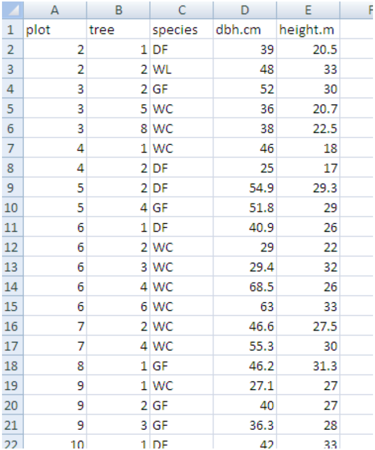
```
"plot","tree","species","dbh.cm","height.m"
2,1,"DF",39,20.5
2,2,"WL",48,33
3,2,"GF",52,30
3,5,"WC",36,20.7
3,8,"WC",38,22.5
4,1,"WC",46,18
4,2,"DF",25,17
5,2,"DF",54.9,29.3
5,4,"GF",51.8,29
6,1,"DF",40.9,26
6,2,"WC",29,22
6,3,"WC",29.4,32
6,4,"WC",68.5,26
6,6,"WC",63,33
7,2,"WC",46.6,27.5
7,4,"WC",55.3,30
8,1,"GF",46.2,31.3
9,1,"WC",27.1,27
9,2,"GF",40,27
9,3,"GF",36.3,28
10,1,"DF",42,33
```

the quotes are actually unnecessary: R inserts them as required.

```
> ufc <- read.csv("ufc.csv")
```

Reading data into R

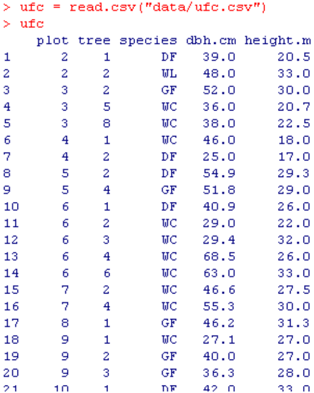
Excel:



	A	B	C	D	E	F
1	plot	tree	species	dbh.cm	height.m	
2	2	1	DF	39	20.5	
3	2	2	WL	48	33	
4	3	2	GF	52	30	
5	3	5	WC	36	20.7	
6	3	8	WC	38	22.5	
7	4	1	WC	46	18	
8	4	2	DF	25	17	
9	5	2	DF	54.9	29.3	
10	5	4	GF	51.8	29	
11	6	1	DF	40.9	26	
12	6	2	WC	29	22	
13	6	3	WC	29.4	32	
14	6	4	WC	68.5	26	
15	6	6	WC	63	33	
16	7	2	WC	46.6	27.5	
17	7	4	WC	55.3	30	
18	8	1	GF	46.2	31.3	
19	9	1	WC	27.1	27	
20	9	2	GF	40	27	
21	9	3	GF	36.3	28	
22	10	1	DF	42	33	

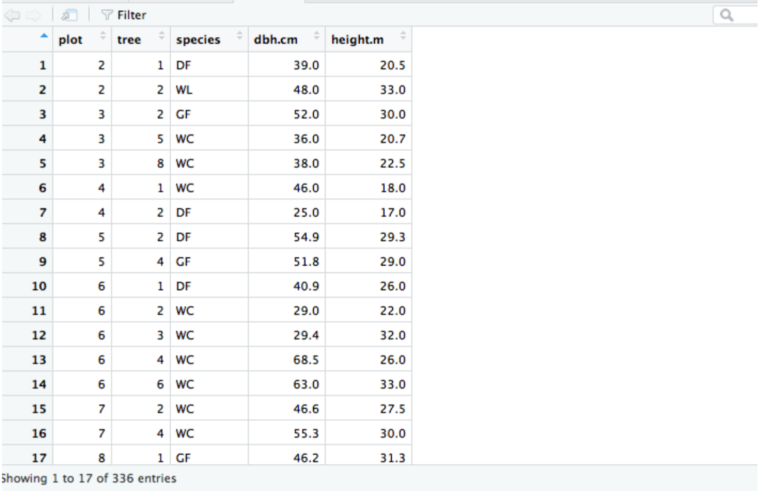
Reading data into R

R:



```
> ufc = read.csv("data/ufc.csv")
> ufc
    plot tree species dbh.cm height.m
1      2    1      DF   39.0     20.5
2      2    2      WL   48.0     33.0
3      3    2      GF   52.0     30.0
4      3    5      WC   36.0     20.7
5      3    8      WC   38.0     22.5
6      4    1      WC   46.0     18.0
7      4    2      DF   25.0     17.0
8      5    2      DF   54.9     29.3
9      5    4      GF   51.8     29.0
10     6    1      DF   40.9     26.0
11     6    2      WC   29.0     22.0
12     6    3      WC   29.4     32.0
13     6    4      WC   68.5     26.0
14     6    6      WC   63.0     33.0
15     7    2      WC   46.6     27.5
16     7    4      WC   55.3     30.0
17     8    1      GF   46.2     31.3
18     9    1      WC   27.1     27.0
19     9    2      GF   40.0     27.0
20     9    3      GF   36.3     28.0
21    10    1      DF   42.0     33.0
```

Reading data into R



	plot	tree	species	dbh.cm	height.m
1	2	1	DF	39.0	20.5
2	2	2	WL	48.0	33.0
3	3	2	GF	52.0	30.0
4	3	5	WC	36.0	20.7
5	3	8	WC	38.0	22.5
6	4	1	WC	46.0	18.0
7	4	2	DF	25.0	17.0
8	5	2	DF	54.9	29.3
9	5	4	GF	51.8	29.0
10	6	1	DF	40.9	26.0
11	6	2	WC	29.0	22.0
12	6	3	WC	29.4	32.0
13	6	4	WC	68.5	26.0
14	6	6	WC	63.0	33.0
15	7	2	WC	46.6	27.5
16	7	4	WC	55.3	30.0
17	8	1	GF	46.2	31.3

Showing 1 to 17 of 336 entries