Introduction

Univariate Statistics and Methodology using R

Martin Corley

Notes

The R Team

Course Leaders	
Martin Corley, Room G8, 7GS	Martin.Corley@ed.ac.uk
Pavel Iosad, Room 3.08, DSB	Pavel.Iosad@ed.ac.uk

Teaching Fellow	
Milan Valášek, Room G6, 7GS	Milan.Valasek@ed.ac.uk
Luna de Ferrari, Room G6, 7GS	luna.deferrari@ed.ac.uk

Tutor Team	
Amrita Ahuwalia	Ruth Corps
Andres Karjus	Daisy Smith

Notes

Practicalities

Weekly components of the course	
lectures	Wednesdays 09:00, here
labs	Thursdays 11:00 *or* Fridays 13:00, 7GS

- if you need to change lab, contact Milan
- further support: bulletin boards on Learn

Notes

More About Labs

- worksheets for labs will be available on Learn
  - you can print them out if you like
  - feel free to try stuff ahead of time
  - inadvisable to skip labs
- any solutions will go online after the relevant lab
- difficulties? → bulletin boards

Notes

If You Get Ahead

- additional readings on Learn
- optional homeworks and solutions

Notes

Exam

- a long way off, don't panic!
- analyse some data using `R` (and show us how you did it)
- write up a brief 'results section' summarising the analysis

Notes

Aims of the Course

- **teach** (or consolidate) fundamental methodological and statistical understanding
- **introduce** the use of `R` as a powerful tool for understanding data (not just NHST)

Notes

Today

1. A Manifesto for R
   - An Overview of R
   - Why Use R?
2. A Toy Experiment
   - Design
   - Analysis

Notes

The R Project

- a ‘statistical programming language’
- created mid-90s as a free version of `S`
- widespread adoption since v2 (2004)

- an ‘integrated development environment’
- created 2011 ‘to improve `R` experience’
- widespread adoption since 2012

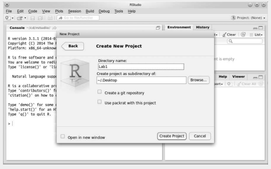
Notes

R vs RStudio

This is R

```
model <- lm(RT ~ (age+freq+handedness)^2, data=words)
summary(model)
```

This is RStudio



```
R version 3.1.1 (2014-0
Copyright (C) 2014 The 
Platform: x86_64-unknow

R is free software and 
You are welcome to redi
Type 'license()' or 'li

  Natural language supp

R is a collaborative pr
Type 'contributors()' f
'citation()' on how to 

Type 'demo()' for some 
'help.start()' for an H
Type 'q()' to quit R.

> 
```

Create New Project

Directory name: Lab1

Create project as subdirectory of: ~/.Desktop

□ Create a git repository

□ Use packrat with this project

□ Open in new window

- RStudio is just one (good) way of 'talking to' `R`

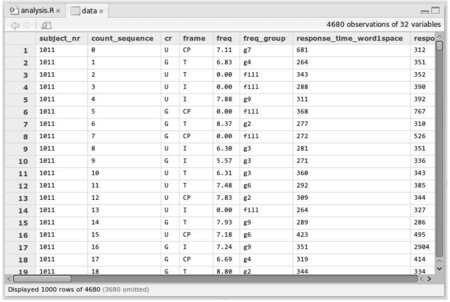
Notes

Notes

What is `R` Good For?

Notes

Managing Datasets



4680 observations of 32 variables

	subject_nr	count_sequence	cr	frame	freq	freq_group	response_time_word1space	respo
1	1011	0	U	CP	7.11	g7	681	312
2	1011	1	G	T	6.83	g4	264	351
3	1011	2	U	T	0.00	fill	343	352
4	1011	3	U	I	0.00	fill	288	390
5	1011	4	U	I	7.88	g9	311	392
6	1011	5	G	CP	0.00	fill	368	767
7	1011	6	G	T	8.37	g2	277	310
8	1011	7	G	CP	0.00	fill	272	526
9	1011	8	U	I	6.30	g3	281	351
10	1011	9	G	I	5.57	g3	271	336
11	1011	10	U	T	6.31	g3	360	343
12	1011	11	U	T	7.48	g6	292	385
13	1011	12	U	CP	7.83	g2	309	344
14	1011	13	U	I	0.00	fill	264	327
15	1011	14	G	T	7.93	g9	289	286
16	1011	15	U	CP	7.18	g6	423	495
17	1011	16	G	I	7.24	g9	351	2904
18	1011	17	G	CP	6.69	g4	319	414
19	1011	18	G	T	8.80	g2	344	334

Displayed 1000 rows of 4680 (3680 omitted)

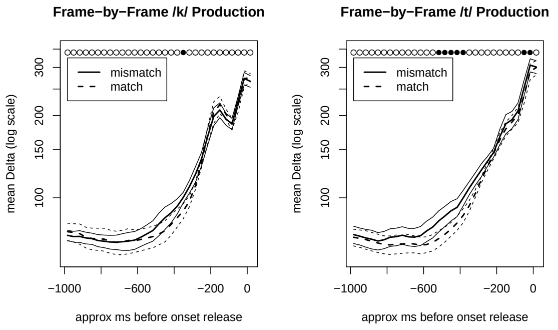
Notes

Doing Statistics

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: DV ~ sc(FvO) * sc(EvC) + (1 | Code) + (0 + (sc(FvO) * sc(EvC)) |
    Code) + (1 | Item)
   Data: feminine
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
     879      944     -428      855     1558

Scaled residuals:
   Min     1Q Median     3Q    Max
-5.045 -0.064 -0.030  0.062  3.634

...
```

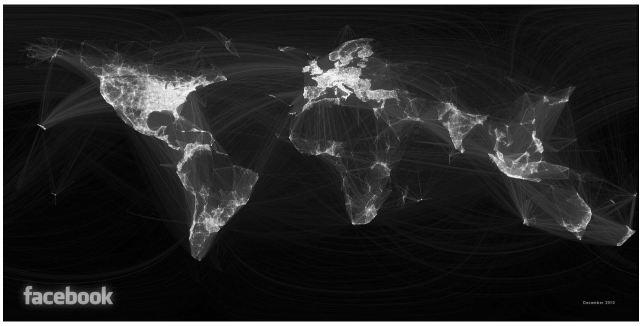
Notes

Publication-Quality Graphics



Notes

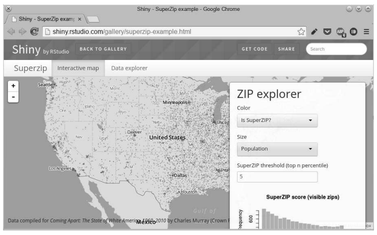
Data Visualisation



http://paulbutler.org/archives/visualizing-facebook-friends/

Notes

Online Interactive Visualisation



ZIP explorer

Color: Is SuperZIP?

Size: Population

SuperZIP threshold (top n percentile): 5

Data compiled for *Coming Apart: The State of White America, 1960–2010* by Charles Murray (Crown F

SuperZips: in top 5% for education and income

http://shiny.rstudio.com/gallery/superzip-example.html

Notes

Simulated Experiments

- if I toss four coins 100 times, how many times will I get **HHHH**?

```
# how many of 100 throws should be HHHH
.5^4 * 100
## [1] 6.2
# throw four coins 100 times, record number of heads
throws <- rbinom(100,4,prob=.5)
throws
##   [1] 3 2 3 2 2 4 0 3 1 2 2 4 1 3 2 2 2 1 3 1 2 3 3 2 2 1 2 3 3 3 1 2 4 2 2 3 1
##  [38] 1 1 2 0 2 2 3 2 4 2 2 2 1 3 3 0 3 2 0 3 4 2 2 3 2 2 2 4 3 2 1 1 0 2 2 2 2
##  [75] 2 0 1 2 1 2 1 2 2 3 1 2 4 2 2 4 1 2 3 2 2 4 0 0 2 1
sum(throws == 4)
## [1] 9
```

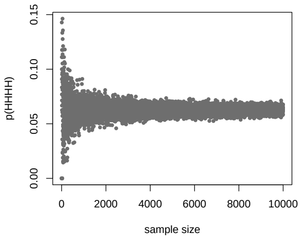
- what about 10,000 times?

```
.5^4 * 10000
## [1] 625
sum (rbinom(10000,4,prob=.5) == 4)
## [1] 599
```

Notes

Large Samples Approach the Population

```
pHead <- function(size) {
    sum(rbinom(size, 4, prob = 0.5) == 4)/size
}
x <- seq(5:10000)
plot(x, sapply(x, pHead), pch = 20, col = "red", xlab = "sample size", ylab = "p(HHHH)")
```



Notes

R for Easy Writing

- R can be combined with **Markdown** to produce documents

input

```
A **mark-up language** consists of ordinary text, _plus_ signs which
indicate how to change the formatting. Here we are using **Markdown**
together with **R**, which means we can include expressions like this:
the square root of 2 is `r sqrt(2)`.
```

output

A **mark-up language** consists of ordinary text, *plus* signs which indicate how to change the formatting. Here we are using **Markdown** together with **R**, which means we can include expressions like this: the square root of 2 is 1.41.

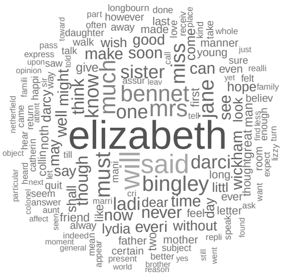
Notes

R for Anything to do with Data

```
require(tm)
require(wordcloud)
# load 'Pride and Prejudice'
pp <- Corpus(DirSource("R/PP/"))
pp <- tm_map(pp, stripWhitespace)
pp <- tm_map(pp, tolower)
pp <- tm_map(pp, removeWords, stopwords("english"))
pp <- tm_map(pp, stemDocument)
pp <- tm_map(pp, removePunctuation)
pp <- tm_map(pp, PlainTextDocument)
wordcloud(pp, scale = c(5, 0.5), max.words = 150, random.order = FALSE, rot.per = 0.35,
    colors = brewer.pal(12, "Dark2"))
```

Notes

The 150 Most Frequent Words in *Pride and Prejudice*



Notes

A Huge Community

- *someone else* has done all the hard work to create wordclouds
- released as libraries or *packages* (like `lme4` and `psych`)
- all I supplied was a text version of *Pride and Prejudice*
- `R` allows you to do *anything* with data
- if it's useful, chances are someone has already done it
- useful things include statistics!
- if it's useful, chances are someone is (constantly) improving it (which is both good and bad)

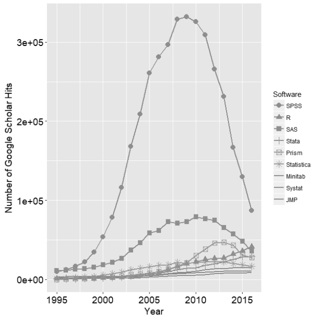
Notes

Why Use R?

- `R` is pretty cool!
- because it's a *language*, I can easily show you what I did and you can copy it
- because it's a *language*, statisticians use `R` to implement leading-edge stats
- because it's *free*, anyone can use `R`—and anyone can access your research
- because it's *open source*, anyone can fix or improve `R`
- `R` is pretty cool!

Notes

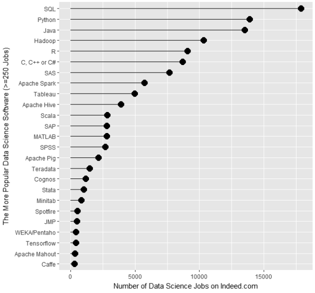
R Usage: Citations in Journal Articles



http://r4stats.com/articles/popularity/

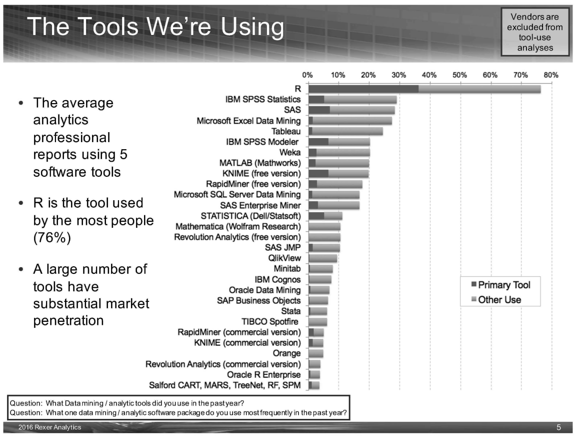
Notes

R Usage: Jobs on indeed.com



http://r4stats.com/articles/popularity/

Notes



The Tools We're Using

Vendors are excluded from tool-use analyses

- The average analytics professional reports using 5 software tools
- R is the tool used by the most people (76%)
- A large number of tools have substantial market penetration

Question: What Data mining / analytic tools did you use in the past year?
Question: What one data mining / analytic software package do you use most frequently in the past year?

2016 Rexer Analytics 5

Notes

Why Are We Teaching R?

- much of how we deal with data involves statistical analysis, so we could use SPSS (or STATA, or SAS)
- but `R` helps you *understand* your data (not just get a *p*-value)

Notes

A Toy Experiment

Research Design

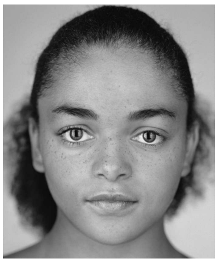
- Idea → Design
  - roughly, "how to have research ideas"[1]
- Design → Implementation
  - roughly, "how to get from idea to experiment"

[1] More on this in other courses

Notes

Finding A Research Question

- *replication* and *extension* of findings
- for example, the effect of pupil size on attractiveness ratings

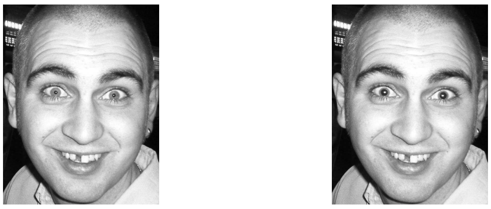


"average US face in 2050" *(National Geographic)*

Notes

Pupil Size and Attractiveness

- larger pupil size leads to higher attractiveness ratings
- but is this a simple relationship?



mingers.com

- e.g., men prefer big pupils in women; women prefer medium pupils in men (unless they like "bad boys") (Tombs & Silverman 2004)

Notes

Finding A Research Question

- design based on *criticism* of previous work

- individuals with Autism have problems with imitating the *style* of meaningless actions with (unconventional) objects

(Hobson & Lee, 1999)

- groups poorly diagnosed, poorly matched
- individual tasks analysed independently
- coding conflates 'success' and 'style'

Notes

remember your playing card. . . ?

Notes

Your Card Has Vanished

[magic trick to be revealed in lecture]

- **change blindness** (e.g., Rensink et al. '87)

- design based on (well-informed) *hunch*
- might be a more general property of cognition
  - **good-enough representations** (e.g., Ferreira et al., '02)

Notes

From Hunch to Design

The Basic Idea

- might be a more general property of cognition

→ might be a property of *language*

- memory for surface form declines over time (Sachs, 1967)
- probe items with similar meanings easily confused (Wanner, 1974)
- specific details of focused words better remembered (Birch & Garnsey, 1995)

Notes

Fleshing Things Out

Focused Words

What Jamie really liked was the cider
It was Jamie who really liked the cider

- so, we predict that. . .
- given some text to recognise, participants are more likely to detect changes which
  - change meaning
  - are in focus

Notes

Fleshing Things Out II

- given some text to recognise, participants are more likely to detect changes which
  - change meaning
  - are in focus

Design

Participants view short passages of text and are then shown them again and asked if there are any changes. Sometimes, single words change, either to semantically-close or semantically-distant words. Half of the words which change are linguistically focused. We predict that changes to distant words will be detected more often, especially when those words are in focus.

Notes

Implementation

- we've fleshed out our hunch using the literature
- we know what the experiment will be
- now we need to get from the **design** to the *implementation*

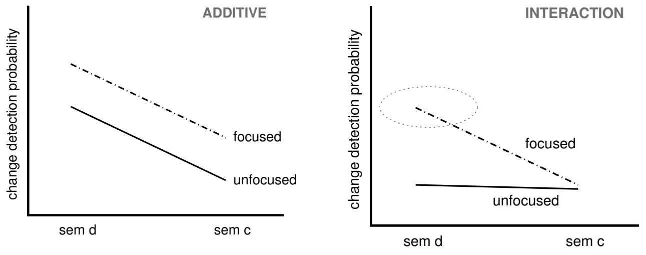
Notes

Conditions

- how do we independently assess the effects of focus and of semantic distance?
- how many conditions?
  - semantically distant
  - focused
  - neither focused nor distant
  - focused *and* distant

Notes

Two IVs



- **additive effects:** two *separate* ways of enhancing detection
- **interaction:** ways of enhancing detection *not* separable

Notes

Stimuli

- how *long* should the text passages be?
  - 'long enough' (→ piloting)
- how do we determine semantic distance?
  - LSA, WordNet, pretesting. . .
- should the word that *changes* remain constant (cider → beer/music), or should the *change* be constant? (beer/music → cider)?
  - detecting a change *to* a (constant) passage of text. . . (depends on theoretical focus)

Notes

Example Materials

Focus on *the cider*

Everyone had a good time at the pub. A group of friends had met up there for a stag night. What Jamie really liked was the cider, apparently.

Focus on *Jamie*

Everyone had a good time at the pub. A group of friends had met up there for a stag night. It was Jamie who really liked the cider, apparently.

- *cider* changes to *beer* (close) or *music* (distant)

(Sanford et al., 2004)

Notes

Within or Between?

- **within subjects**
  - **advantage**: reduces between-subject variability (increases power)
  - **disadvantage**: repetition of passages increases memory for detail?
- **between subjects**
  - **advantage**: no repetition
  - **disadvantage**: loss of power
- need a compromise solution!

Notes

Counterbalancing

	mat1	mat2	mat3	mat4	mat5	···
sub1	A	B	C	D	A	···
sub2	D	A	B	C	D	···
sub3	C	D	A	B	C	···
sub4	B	C	D	A	B	···
⋮	⋮	⋮	⋮	⋮	⋮	⋱

- each participant only sees each passage once, but contributes to mean for all conditions
- each material (here, passage/change combo) seen in all conditions over 4 subjects
- run multiples of 4 subjects/materials and analyse as ‘within’

Notes

Implementation Issues

- how many materials?
- how long should the passage appear on-screen for first reading?
- how can we avoid non-linguistic (e.g., iconic) memory?
- how are we going to analyse our findings?
  - analysis is part of the *design* process
  - we should be able to answer this *before* collecting data

Notes

Anything missing?

so far, we’ve only talked about items with changes. . .

. . . which means that participants are pressing “YES” every time. . .

- we need “NO” responses too
- **fillers** (in this case, passages which don’t change)

Notes

Sanford et al. (2004, Expt 1)

- 40 participants
- 28 passage/change combos, varying focus/semantic distance: counterbalanced (each participant sees 7 items in each condition)
- 48 items with close/distant changes to verbs[2]
- 12 fillers with no change; 12 with various changes (to mask change location)

- 8-second *or* self-paced display of passage
- 500ms grey screen
- redisplay of passage (for max 10 sec); verbal report of change

[2] combining experiments

Notes

Results

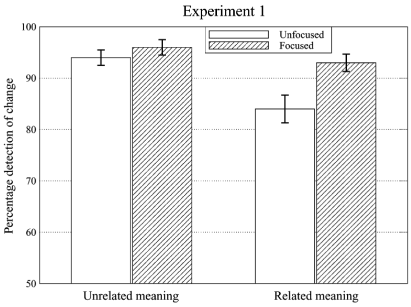


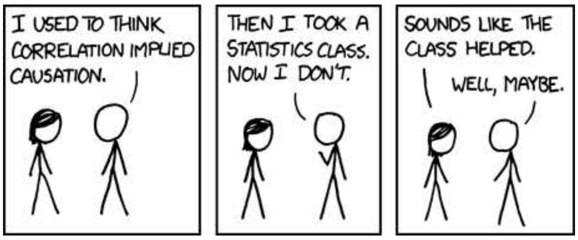
Figure 1. Detection as a function of condition for Experiment 1 (means and standard errors).

Notes

So Now We Can Go Home. . . ?

- the graph shows us the general pattern of results
- but we want to know whether this pattern is *related to* the experimental manipulations
- traditional statistics allow us to reason (negatively!) about how the results came about
  - "the differences between conditions are *unlikely* to be due to chance"

Notes

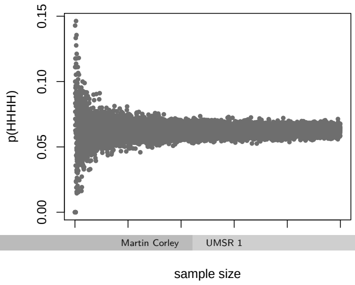

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

www.xkcd.com

Notes

Reasoning About Findings

- we've already seen one (valid) way of estimating the likelihood of an outcome when we tossed imaginary coins
- the width of the 'bar' represents the range of outcomes we'd expect for a given sample size
- outcomes outwith the 'bar' are relatively unlikely unless 'something's going on'



Notes

Reasoning About Findings

- NHST is effectively *mathematical* simulations of experiments
- we aim to determine how wide the 'bars' are (measures related to standard error) and whether our observations fall outside them
- observations which fall outwith 'what we might expect' have a low probability of occurring by chance (low p)
- all the rest is reasoning and theory
- this course: how to estimate p, and how to understand your data well enough to understand and evaluate that estimate
- there are other ways of doing statistics, and we will touch on them during the course

Notes

This Week's Work

Lab

- a gentle introduction to `R` and to the `RStudio` environment

Reading

- Navarro, chs. 1 and 2

Homework

- start working slowly through Navarro, chs. 3 and 4

Notes

Sanford et al. (2004, Expt 1)

```
sanford <- read.table(file='R/cleft_data.txt',header=T)
s.by.s <- with(sanford,aggregate(resp,list(subj,focus,dist),mean))
names(s.by.s) <- c('subj','focus','dist','PERCENT')
model <- aov(PERCENT ~ focus*dist+Error(subj/(focus*dist)),
             data=s.by.s)
summary(model)
```

```
##
## Error: subj
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 39   9546     245
##
## Error: subj:focus
##           Df Sum Sq Mean Sq F value Pr(>F)
## focus      1   1000    1000    15.8  3e-04 ***
## Residuals 39   2469      63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: subj:dist
##           Df Sum Sq Mean Sq F value Pr(>F)
## dist       1   1653    1653    12.2 0.0012 **
## Residuals 39   5286     136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: subj:focus:dist
##            Df Sum Sq Mean Sq F value Pr(>F)
## focus:dist  1    413     413    2.73   0.11
## Residuals  39   5913     152
```

Notes

Notes