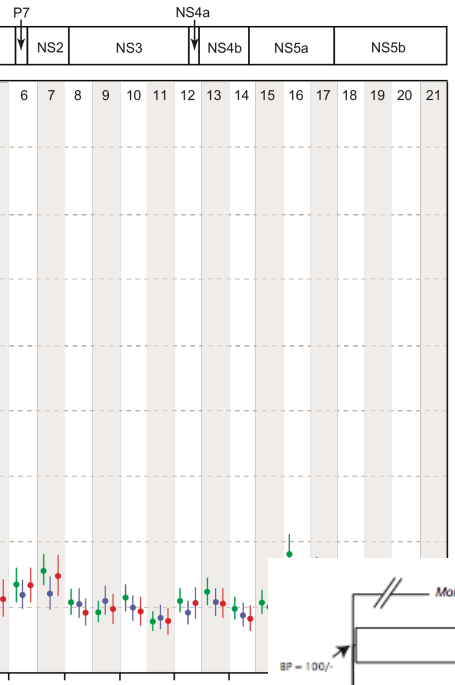
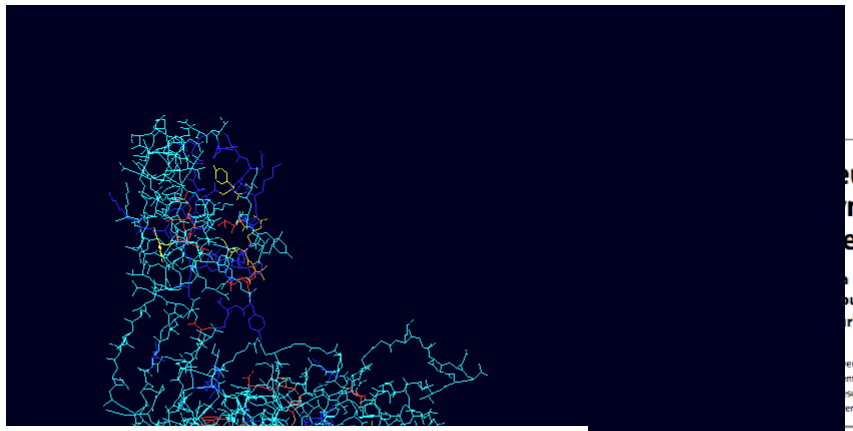
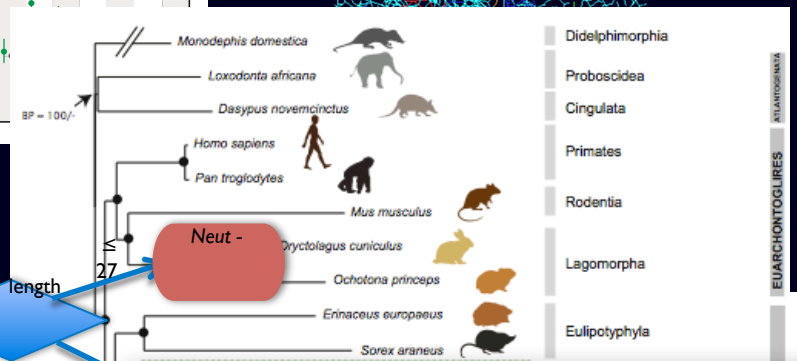
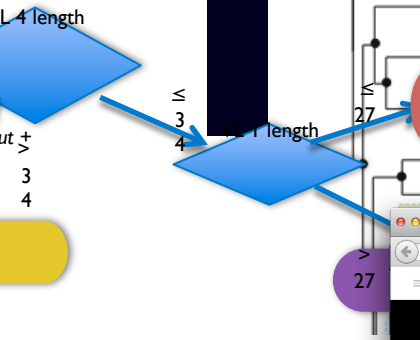
Statistics & Bioinformatics

Joe Parker









Evolutionary and Phylogenetic Analysis of the Hepaciv

PLOS PATHOGENS

utralizing Antibodies and Divergence of
nomolg
etroviral

nd Oliver G. Pybus[1,*]
gdom

LETTER

Genome-wide signatures of convergent ev echolocating mammals

Joe Parker[1*], Georgia Tsagkogeorga[1*], James A. Cotton[1†], Yuan Liu[2], Paolo Provero[3,4], Elia Stupka[3] & Ste

Evolution is typically thought to proceed through divergence of genes, proteins and ultimately phenotypes[1–3]. However, similar traits might also evolve convergently in unrelated taxa owing to similar selection

proposed suborder Yinpterochiroptera we se
shoe bat *Rhinolophus ferrumequinum* and
bat *Megaderma lyra*, which exhibit 'cons
ited' (FM) echolocation,
2). From this suborder w
r-coloured fruit bat *Eido*
lraft genome data from a
flying fox *Pteropus vampy*
ptera, we sequenced the
teronotus parnellii, and a
ng little brown bat *Myoti*
our focal bat species, we g
a on a Hi-Seq 2000 platfo

www.nature.com

Current Biology 23, 2262–2267, November 18, 2013 ©2013 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.09.014

Report

Phylogenomic Analyses Elucidate the Evolutionary Relationships of Bats

Georgia Tsagkogeorga,[1,3,*] Joe Parker,[1,3] Elia S
James A. Cotton,[1,4] and Stephen J. Rossiter[1,*]
[1]School of Biological and Chemical Sciences, Qu
University of London, Mile End Road, London E1
[2]Center for Translational Genomics and Bioinforr
San Raffaele Scientific Institute, Via Olgettina 58,
20132 Milano, Italy

Summary

Molecular phylogenetics has rapidly establishe
tionary positions of most major mammal grou
analyses have repeatedly failed to agree on
(order Chiroptera) [3–6]. Moreover, the relation
the major bat lineages has proven equally
with ongoing disagreements about whether e
bats are paraphyletic [7–9] or a true group [10]
found implications for whether echolocation e
or possibly multiple times. By generating new
data and applying model-based phylogenom
designed to accommodate heterogeneous evol

Received: 3 July 2017
Accepted: 12 July 2017
Published: xx xx xxxx

SCIENTIFIC REP

OPEN

Field-based species identific of closely-related plants usin time nanopore sequencing

Joe Parker[1], Andrew J. Helmstetter[1], Dion Devey[1], Tim Wilkinson[1] & A Papadopulos[1,2]

Advances in DNA sequencing and informatics have revolutionised biology over the but technological limitations have left many applications unexplored. Recently, po nanopore sequencing (RTnS) has become available. This offers opportunities to ra analyse genomic data anywhere. However, generation of datasets from large, con been constrained to laboratories. The portability and long DNA sequences of RTnS for field-based species identification, but the feasibility and accuracy of these tech purpose have not been assessed. Here, we show that a field-based RTnS analysis o species (*Arabidopsis* spp.) has many advantages over laboratory-based high-throu (HTS) methods for species level identification and phylogenomics. Samples were c sequenced in a single day by RTnS using a portable, "al fresco" laboratory. Our ana

Goals

- Equip you to be a bioinformatician
- (in a statistical sense)
- That's it

Assessment

What do we do?

- Collect empirical observations of phenomena
- Spot patterns
- Propose models that explain them
- Simulate guesses according to the models
- Compare them, refine, repeat
- … SCIENCE

Bioinformatics and stats

- **Toolbox:** big data, replicates, moar
- **Concepts:** statistics, probability, populations
- **Techniques:** multivariate analyses, programming, genomics

NOT

- Stats 101
- Programming course
- Mol. Biol.
- Genomics
- Machine-learning / AI / any advanced stuff

Roadmap

Populations, probability and statistics

- Almost all biological phenomena vary continuously
- Probability: assess outcome given model
- Equally, for a large N how many X?
- How likely is X?
- Statistics:
  - How likely is it that X occurs by chance?
  - By some mode?
  - From same distribution (process) as Y?

The P-value

- It isn't: magic
- Working definition: probability of seeing a value as or more extreme than the observed, assuming model, parameters
- "Is this result random?" (ahem)

Descriptive stats

- Count
- Mean, mode, median
- Range
- Confidence intervals

Applications

- Trees on *h* {R}. Two species or one?
- k nonsynonymous changes of N aa replacements. Convergent?
- 913 patients with cancer, 50,000 SNPs, 2031 haplotypes. Will patient X get cancer?

Introduction to R

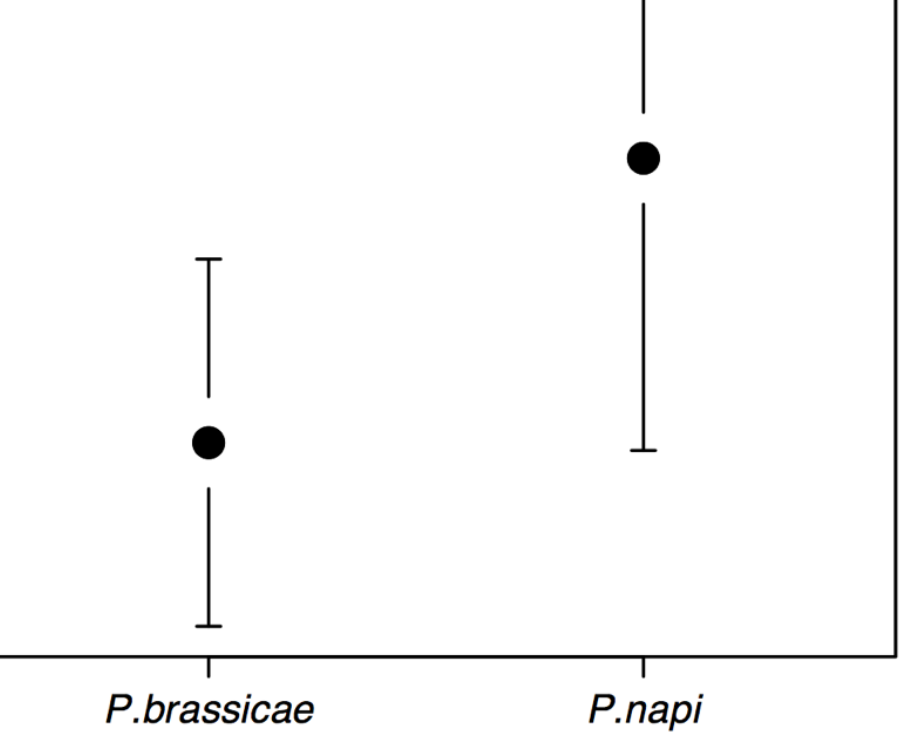
Joe Parker (Rob Knell)

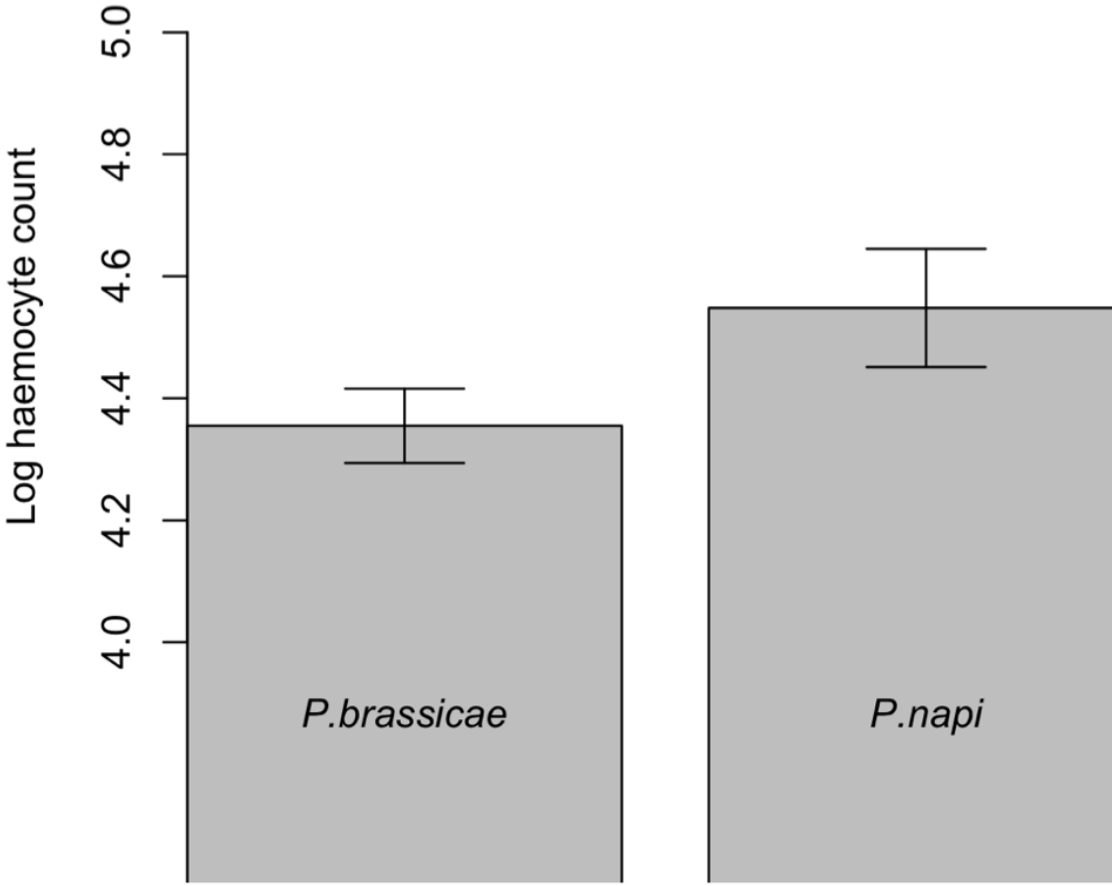
What is R?

- Open source statistical programming language
- Based on a commercial language called “S”
- Powerful and flexible software that can do any statistical analysis you can think of and much more

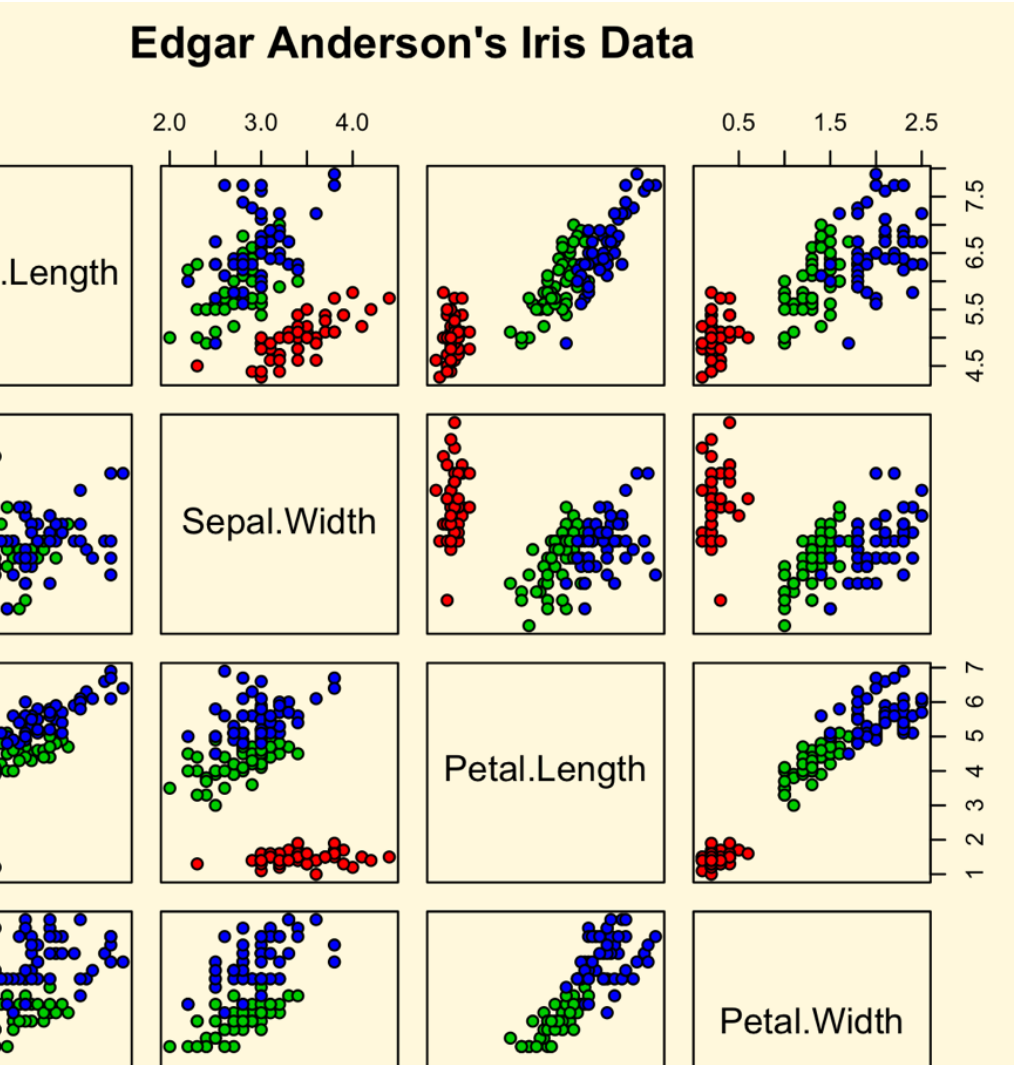
What can I do with R?

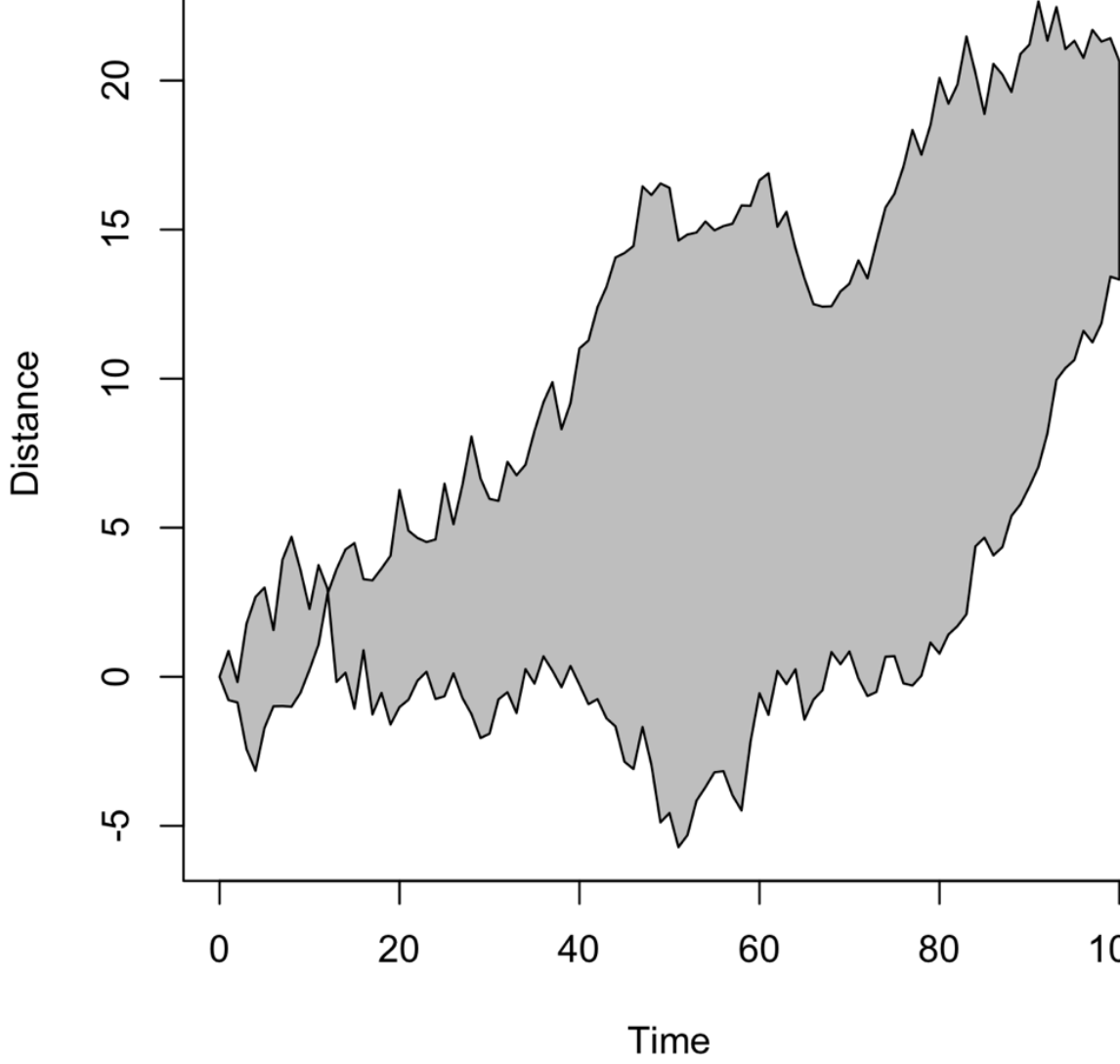
- With the base R installation you can carry out a tremendous variety of statistical analyses
- You can draw publication-quality graphs
- With add-on packages you can do just about anything


P.brassicae
P.napi

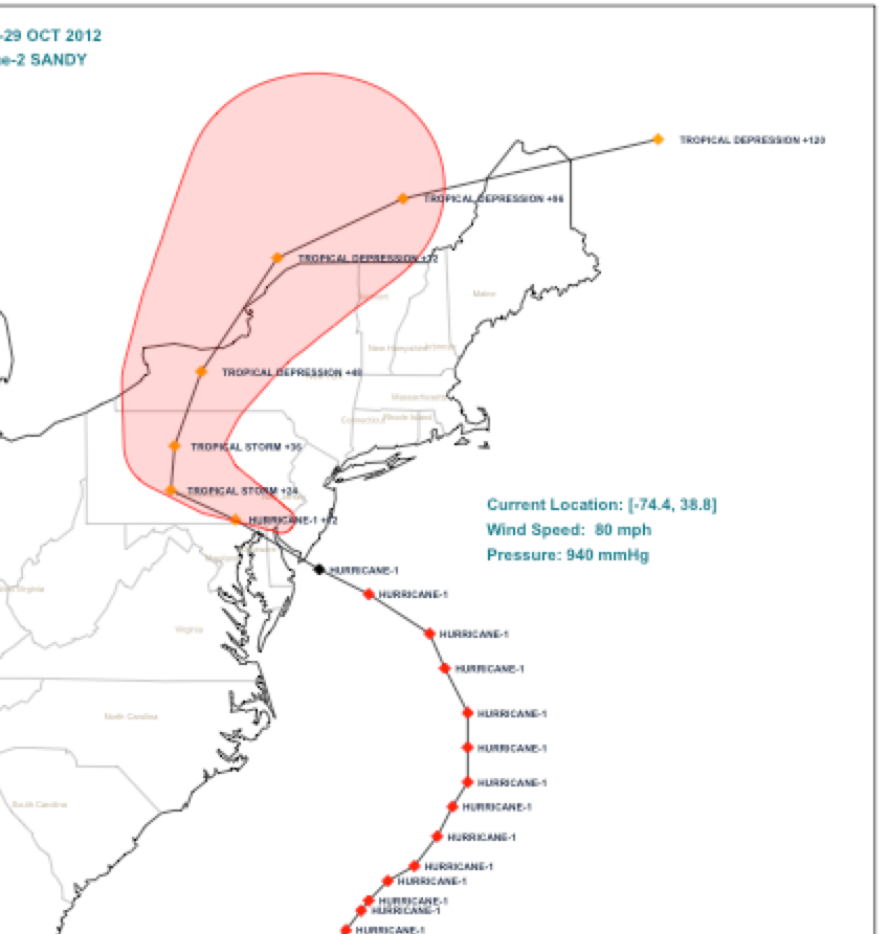

Log haemocyte count
4.0
4.2
4.4
4.6
4.8
5.0
P.brassicae
P.napi


3
4
5
6
7
8
0.1
0.2
0.3
0.4
0.5
0.6
0.7

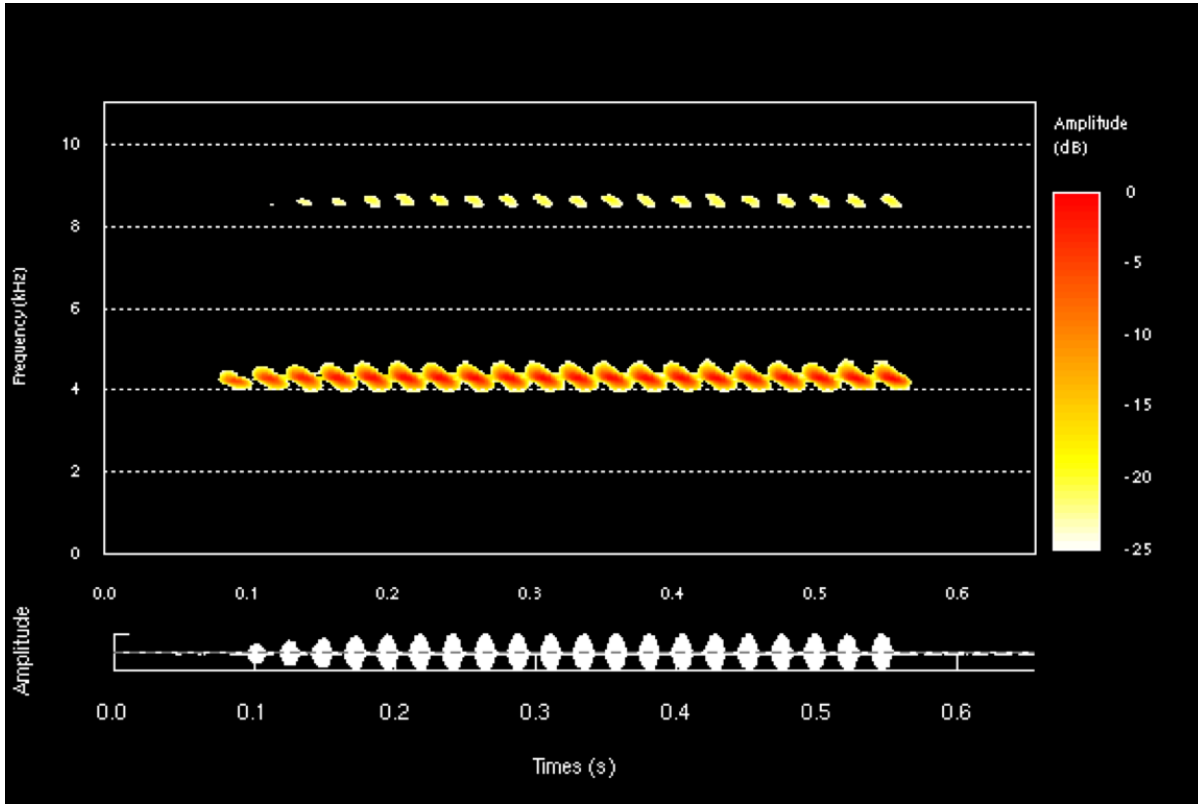

Edgar Anderson's Iris Data
2.0
3.0
4.0
0.5
1.5
2.5
.Length
Sepal.Width
Petal.Length
Petal.Width
4.5
5.5
6.5
7.5
1
2
3
4
5
6
7

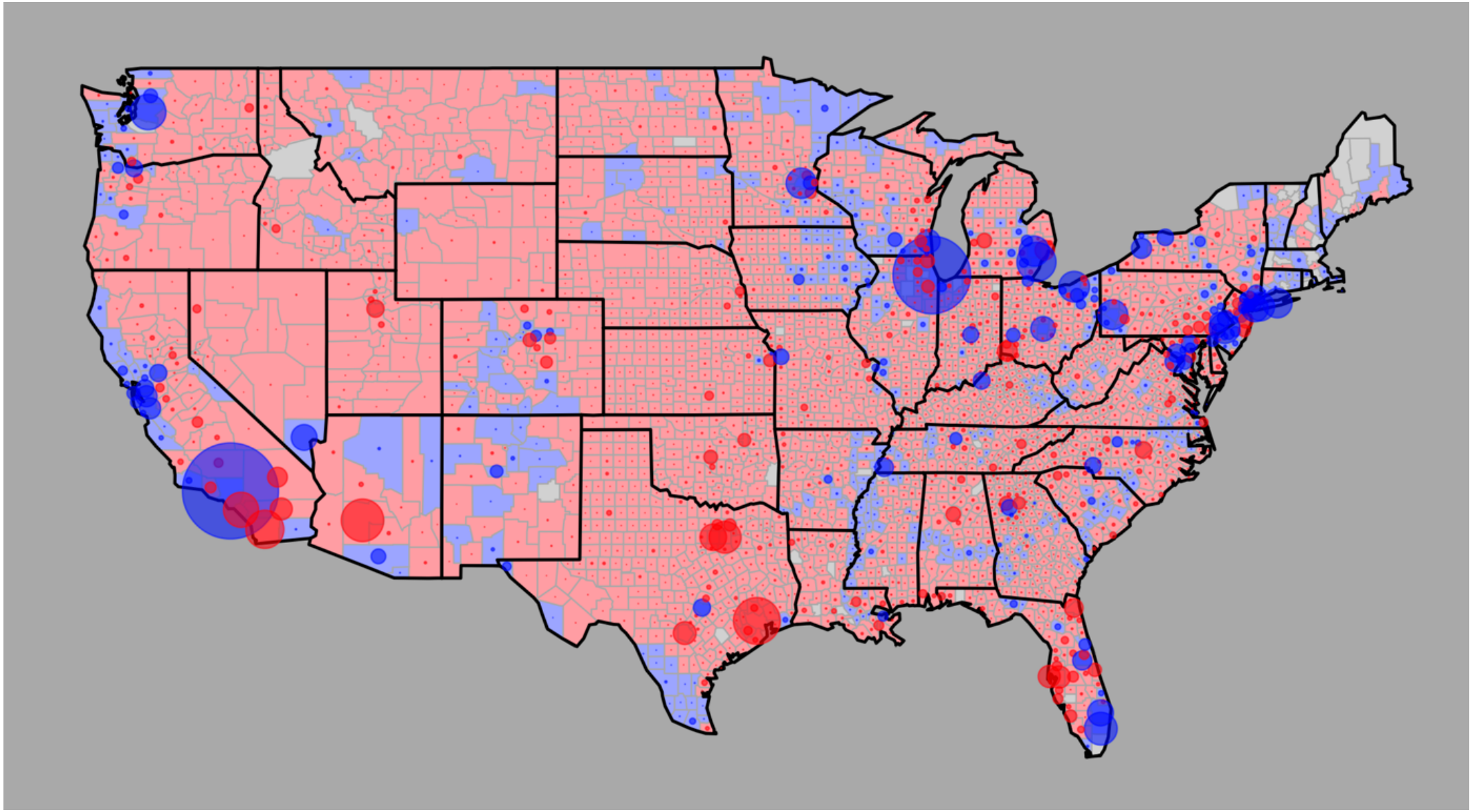

Distance
-5
0
5
10
15
20
0
20
40
60
80
Time

Hurricane Sandy predicted path

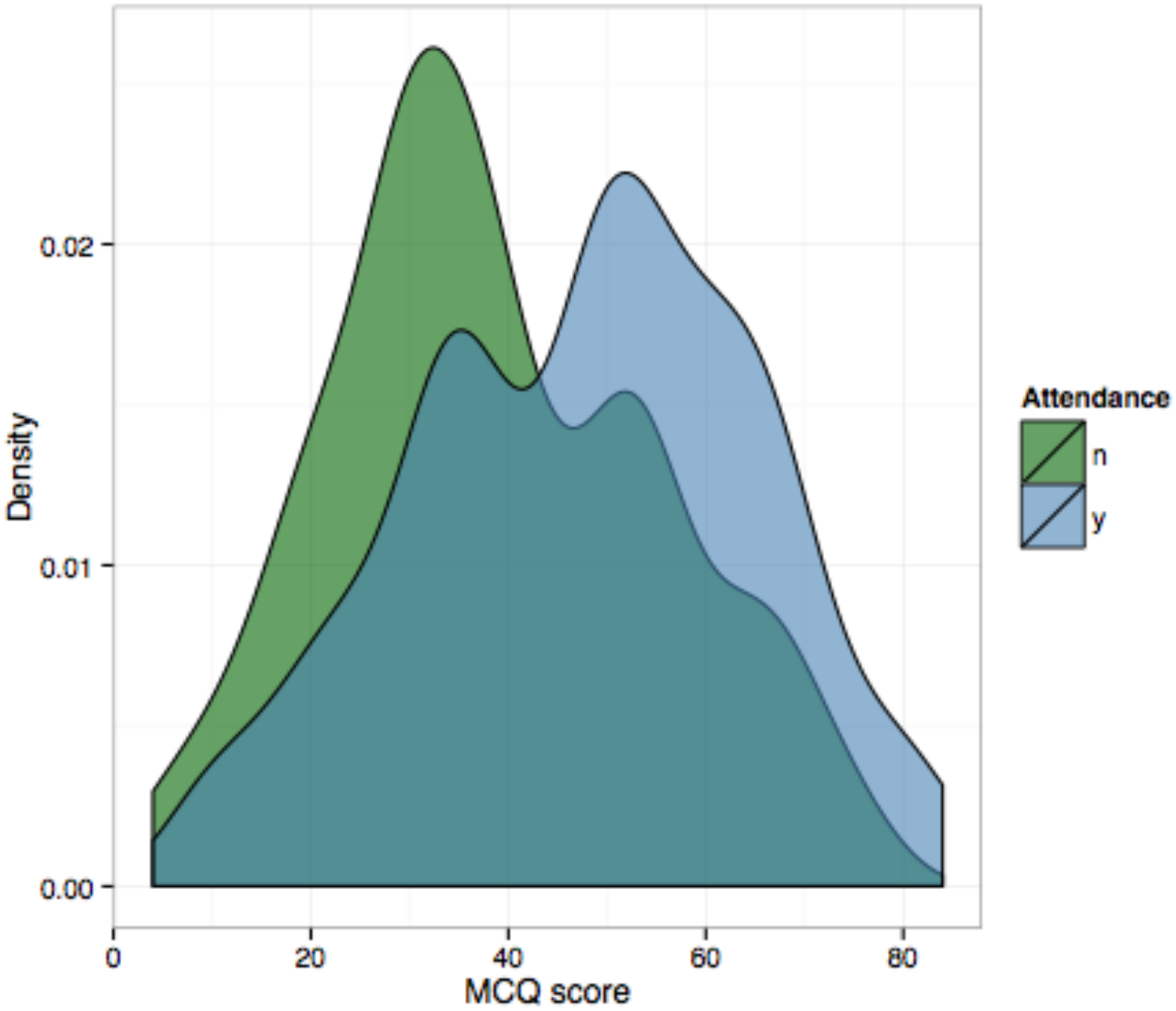


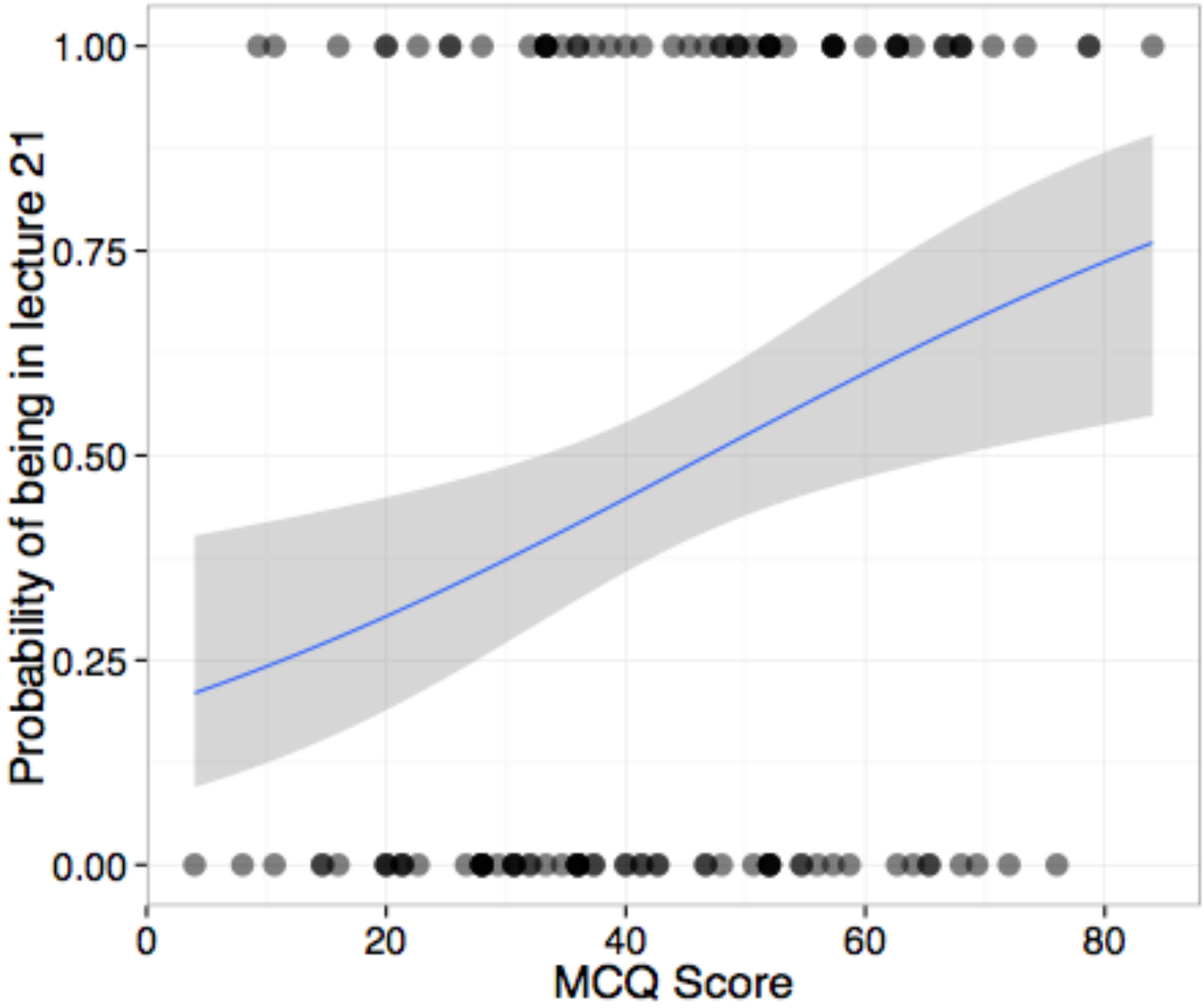
Sonogram of cricket song



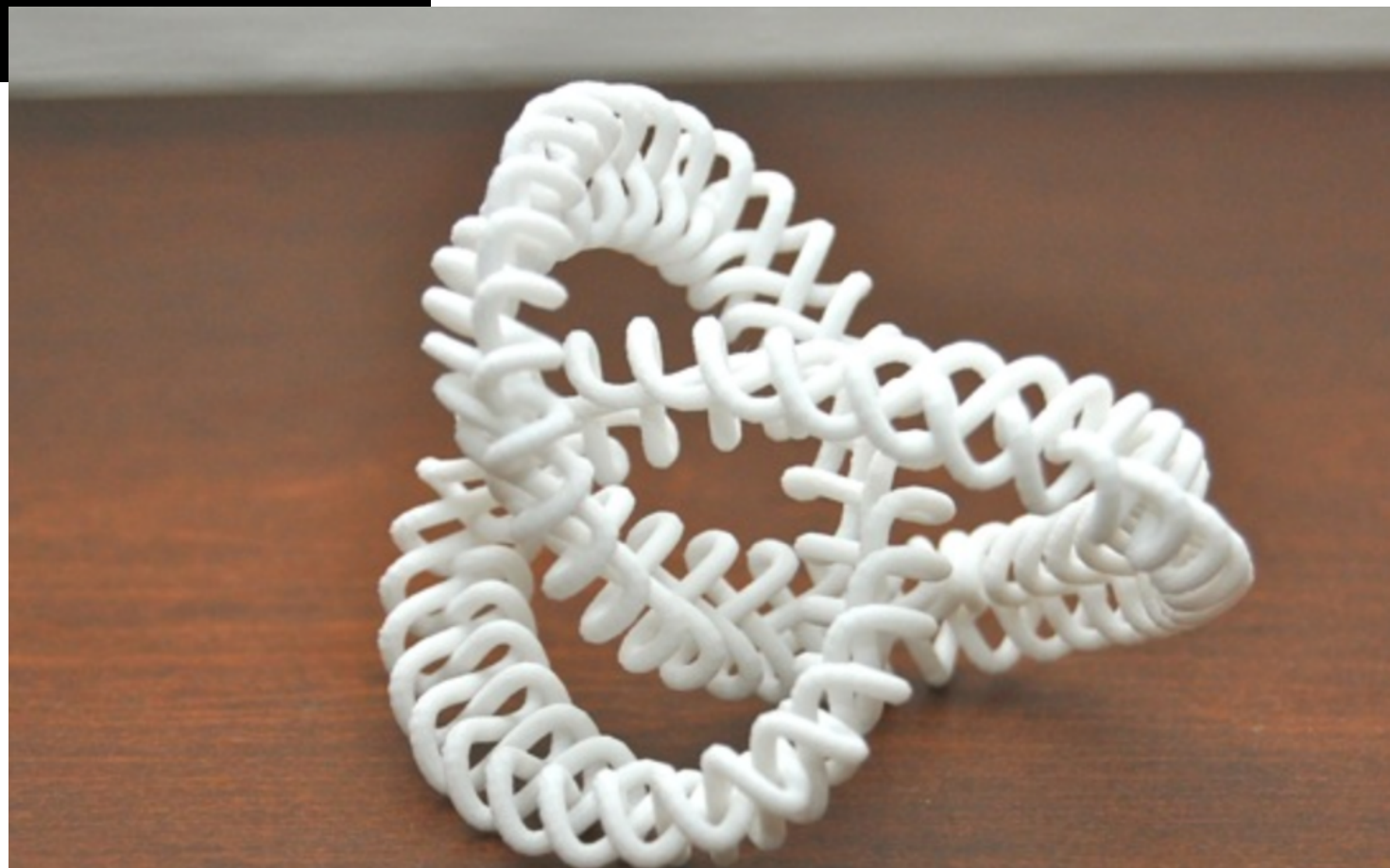


US Presidential election map


Density
0.02
0.01
0.00
0
20
40
60
80
MCQ score
Attendance
n
y


1.00
0.75
0.50
0.25
0.00
Probability of being in lecture 21
0
20
40
60
80
MCQ Score

rgl package



How to get R

cran.r-project.org

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** use
likely want one of these versions of R:

- Download R for Linux
- Download R for MacOS X
- Download R for Windows

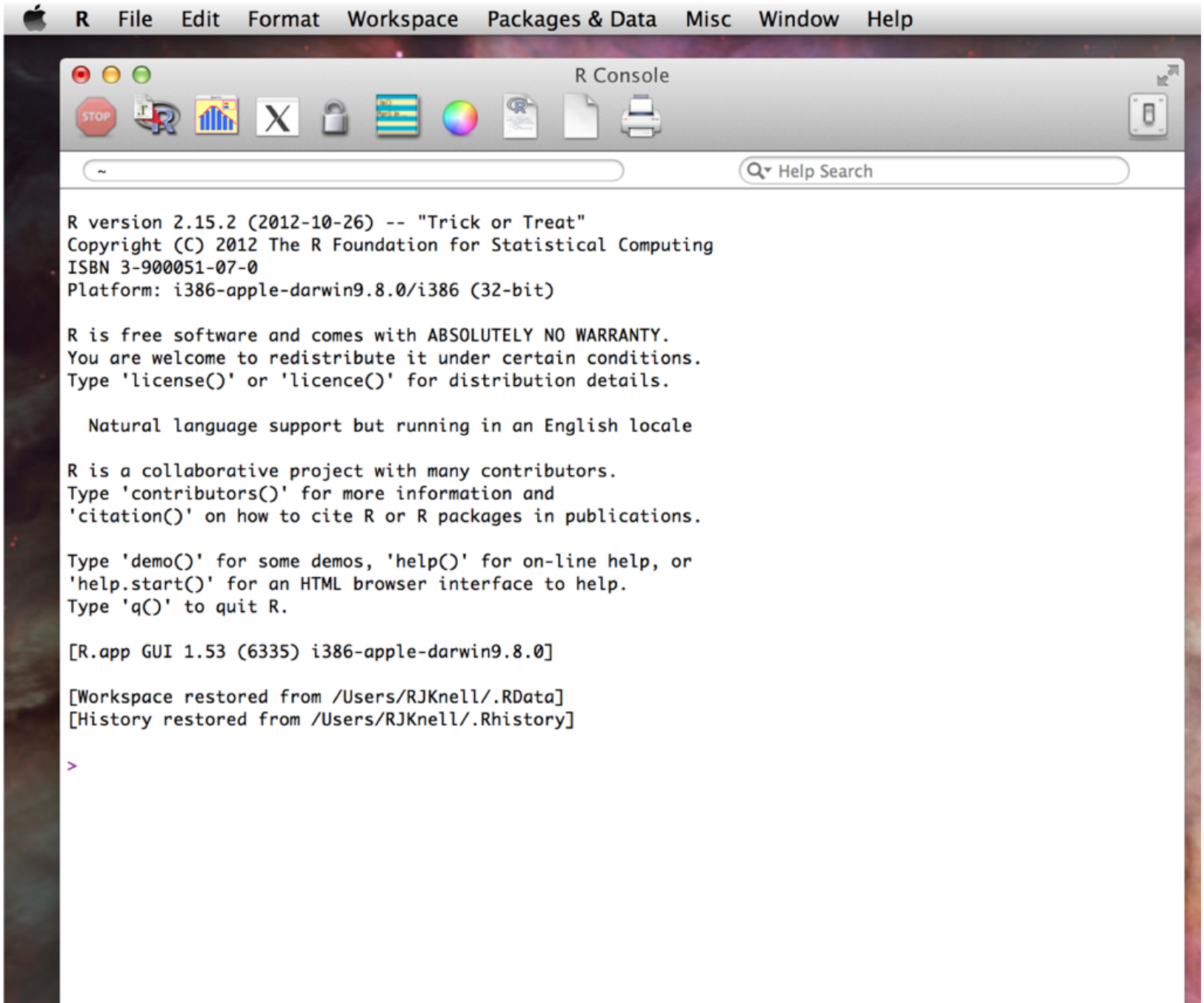
R is part of many Linux distributions, you should check with your Linux package management system
to the link above.

Source Code for all Platforms

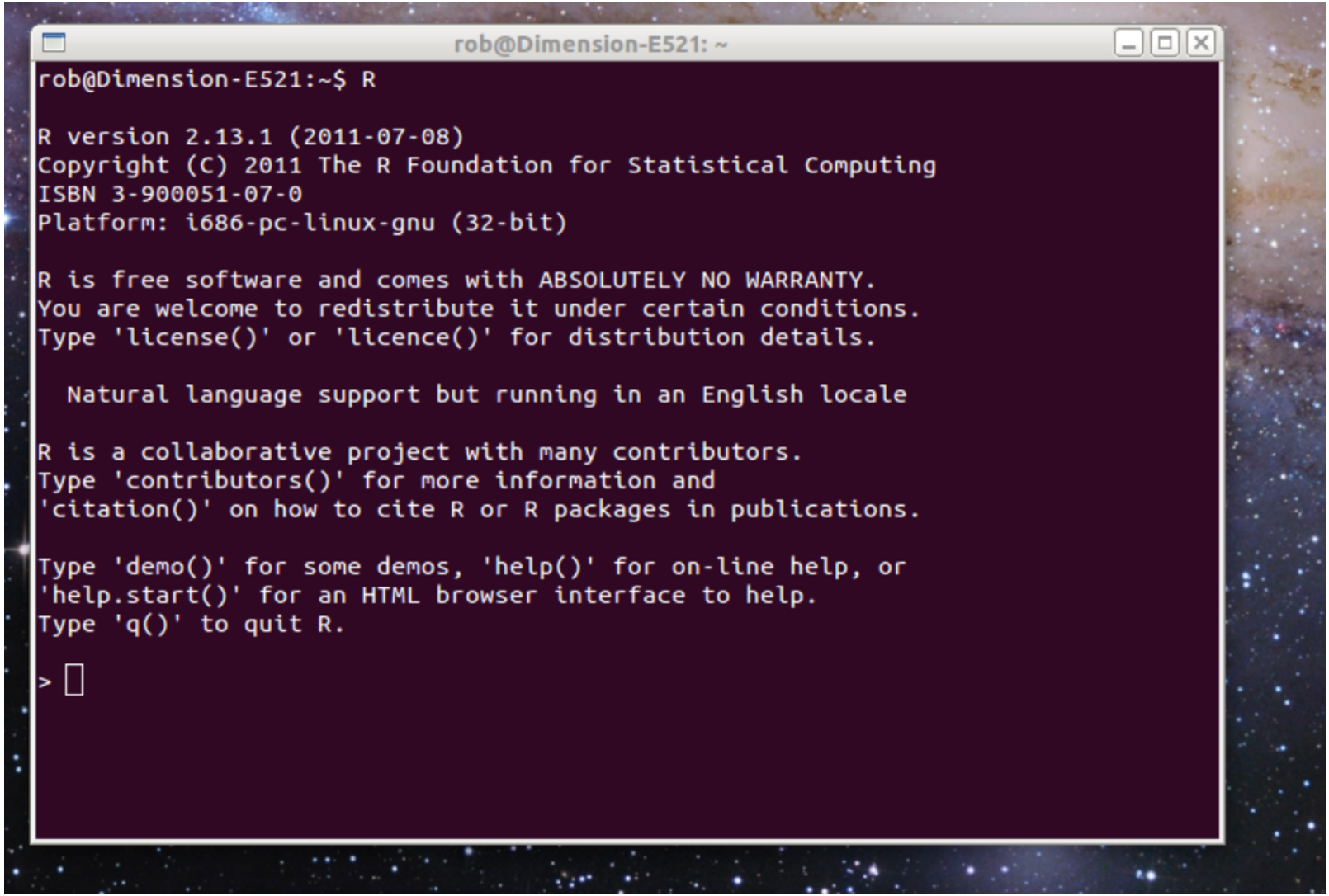
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box
source code. The sources have to be compiled before you can use them. If you do not know what this m
probably do not want to do it!

- The latest release (2012-10-26, Trick or Treat): R-2.15.2.tar.gz, read what's new in the latest vers
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a plann
- Daily snapshots of current patched and development versions are available here. Please read abo
  features and bug fixes before filing corresponding feature requests or bug reports.

R basics


R File Edit Format Workspace Packages & Data Misc Window Help
R Console
~
Help Search
R version 2.15.2 (2012-10-26) -- "Trick or Treat"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-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 1.53 (6335) i386-apple-darwin9.8.0]
[Workspace restored from /Users/RJKnell/.RData]
[History restored from /Users/RJKnell/.Rhistory]
>

R basics


rob@Dimension-E521: ~
rob@Dimension-E521:~$ R
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i686-pc-linux-gnu (32-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.
>

Command line software

- You type an expression or instruction at the command prompt and press “enter”
- If R can understand what you’ve typed, it’ll do what you’ve asked...
- ...or what it thinks you’ve asked: unlike Python, R contains some fairly high-level functions

Data exploration and description

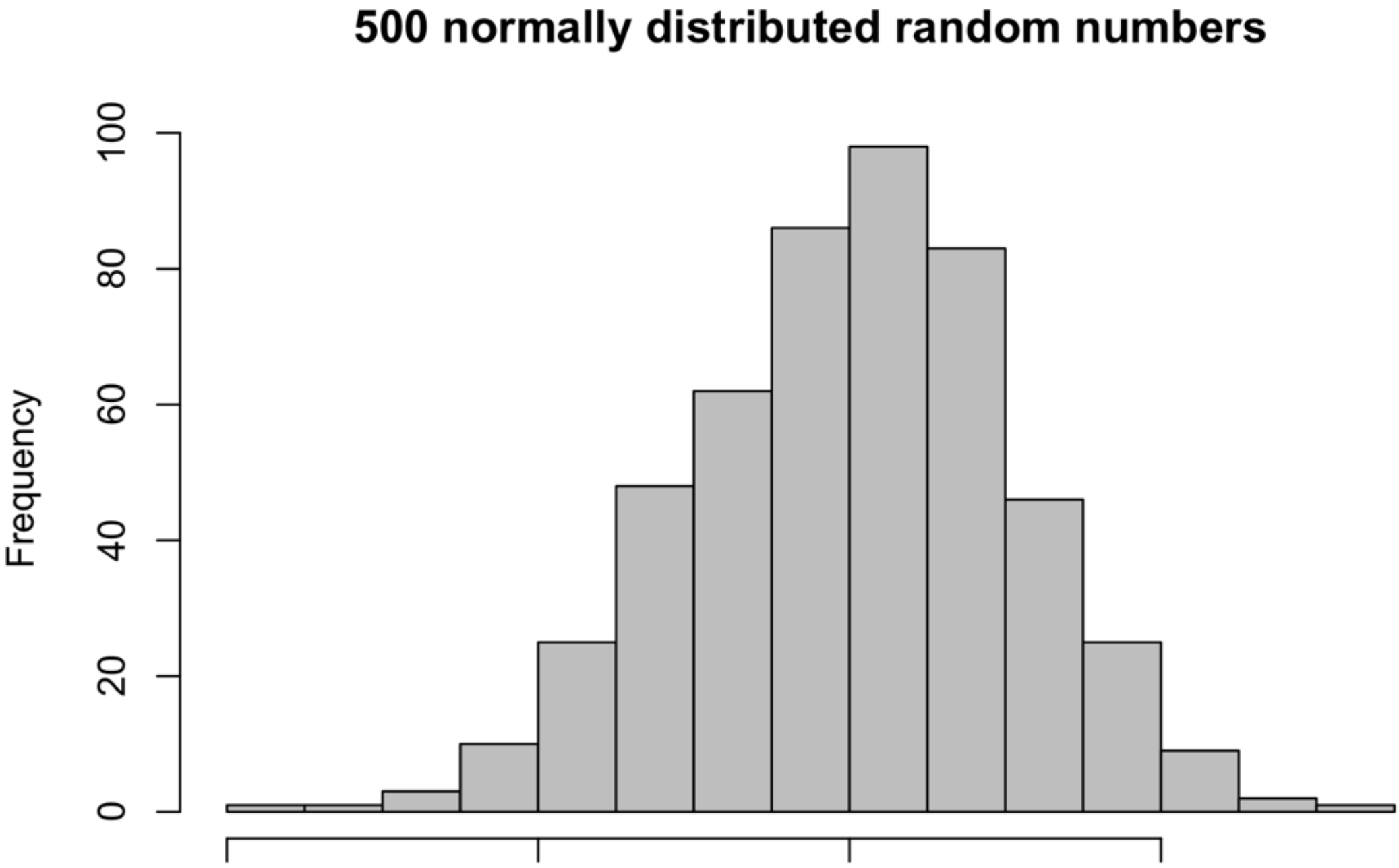
Exploratory data analysis

- Gain an understanding of the nature of the data
- Get clues regarding the relationships between variables
- Detect errors
- Know which analyses might be necessary

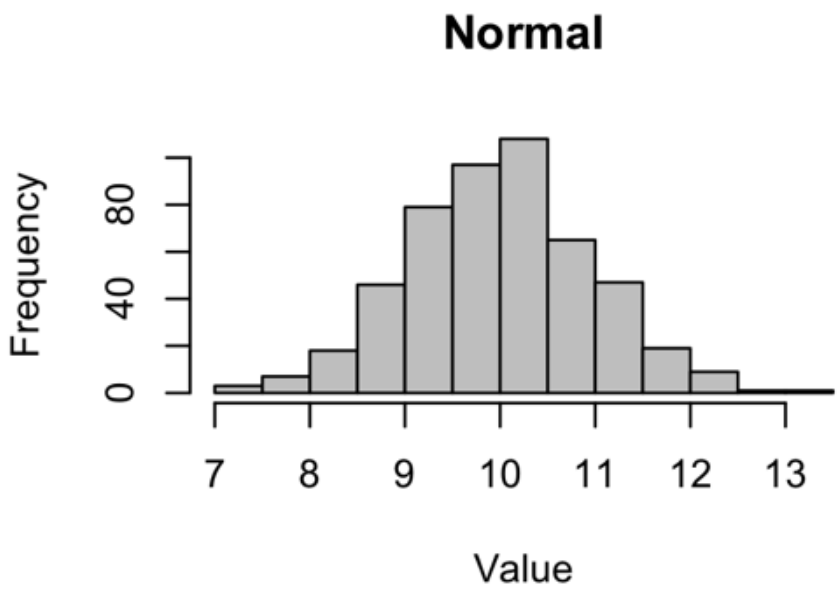
Exploratory data analysis

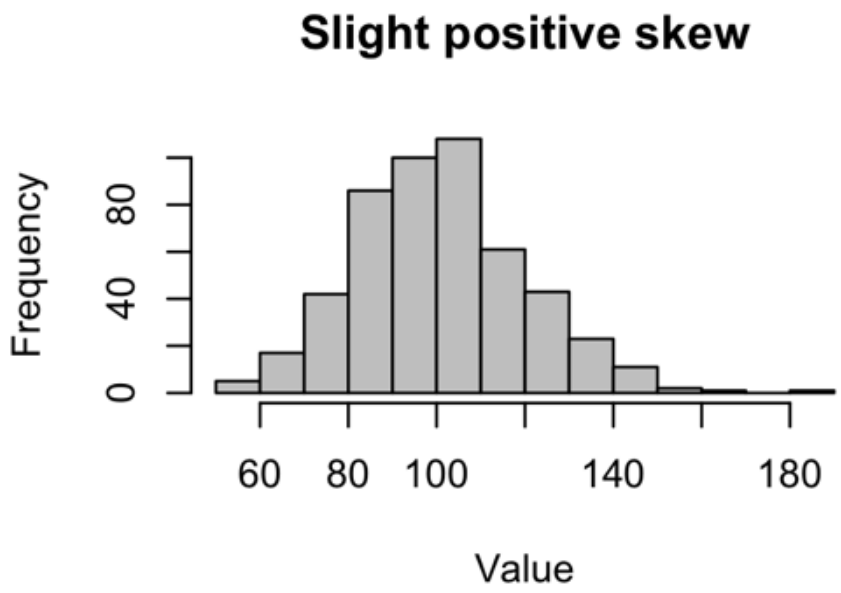
- Numerical and graphical descriptions of data
- Need to revise some basic numerical descriptives before moving on to graphical techniques

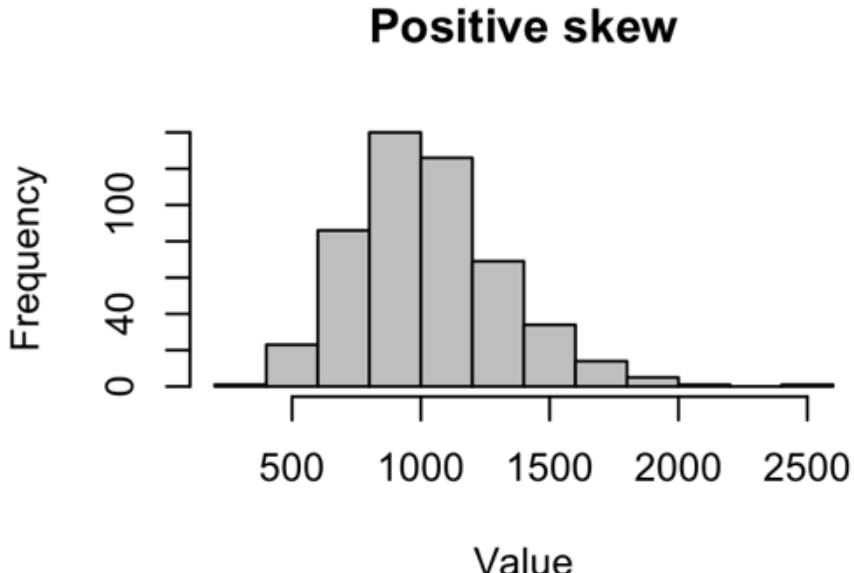
Frequency distributions

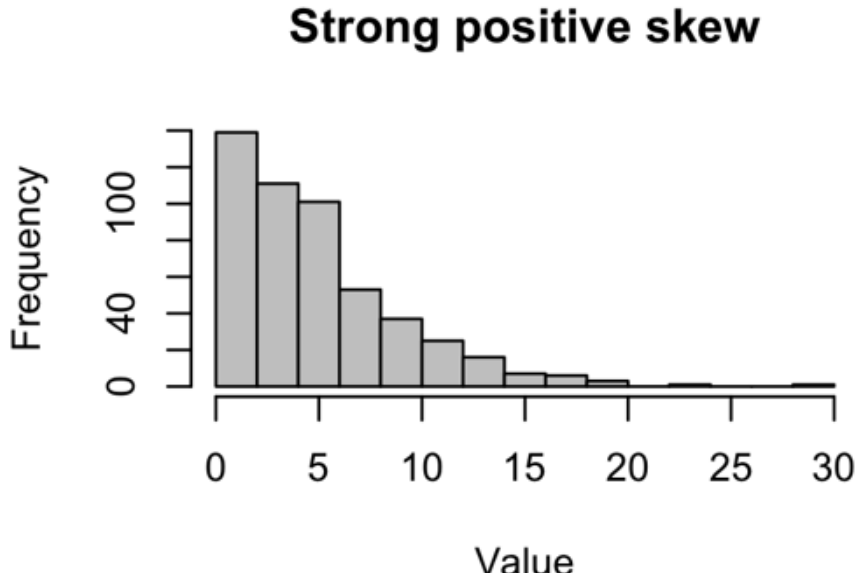


Frequency distributions









Descriptive statistics

- Measures of central tendency
  - Mean, median, mode
- Measures of dispersion
  - Standard deviation, variance, IQR

Central tendency

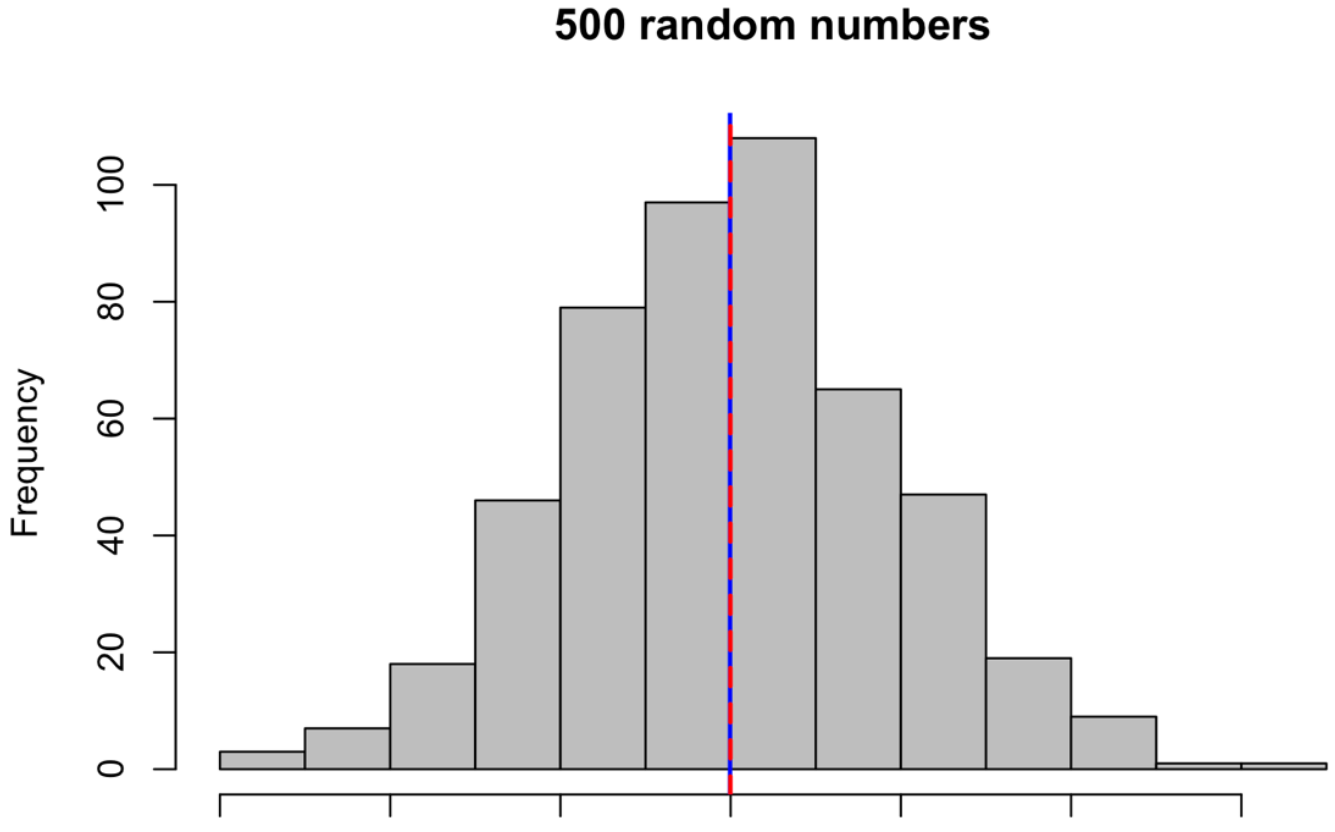
Mean $\bar{x} = \frac{1}{n}\sum x$

Median: rank data, take middle value.
If even number of data take the
average of the two middle values

Mode: most common value

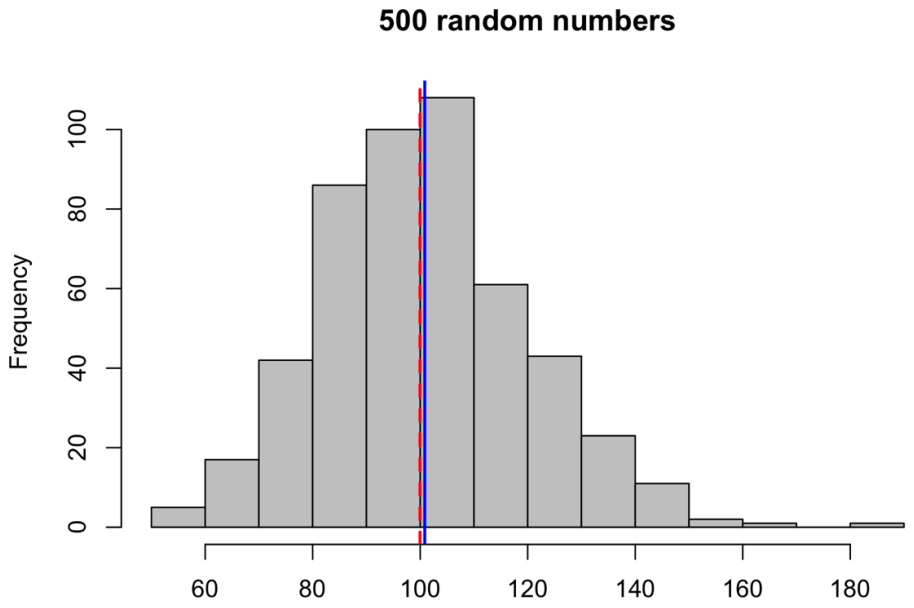
Central tendency

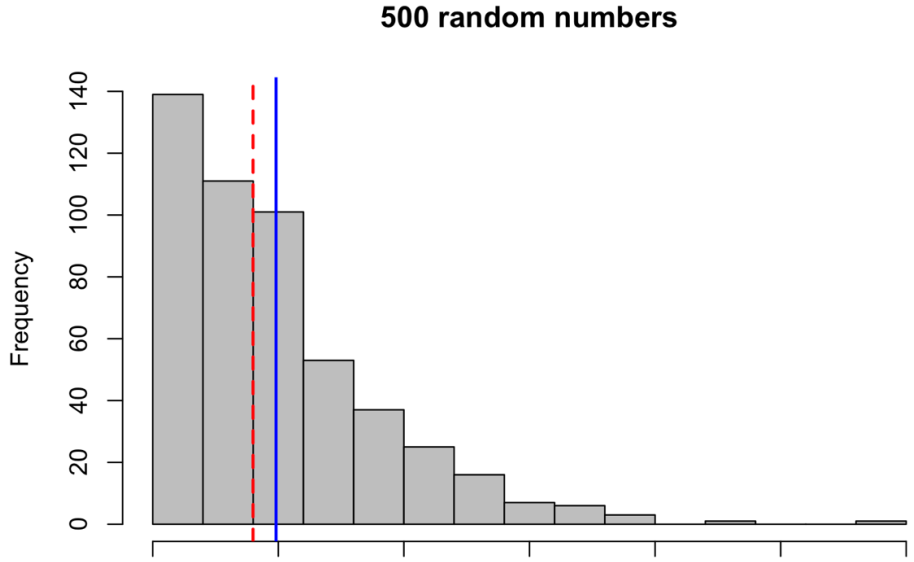
r normally distributed
data, mean (blue) =
nedian (red) = mode



Central tendency

As the distribution
s more skewed, the
mean, mode and
edian become less
similar





Measures of dispersion

Variance $$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Standard deviation $$s = \sqrt{s^2}$$

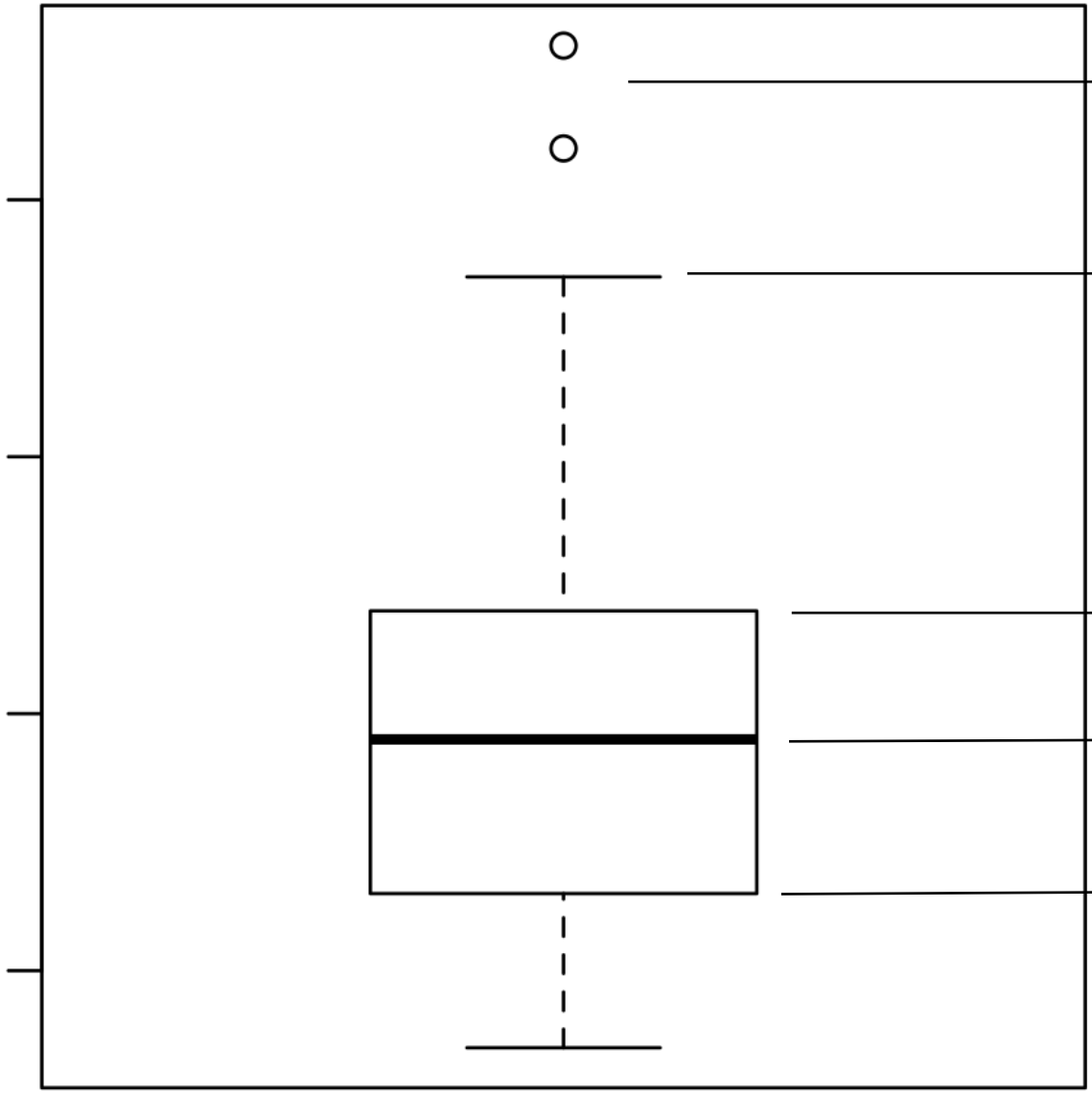
Range from 1st quartile (top c
lower 25%) to third quartile

er-quartile range (IQR)

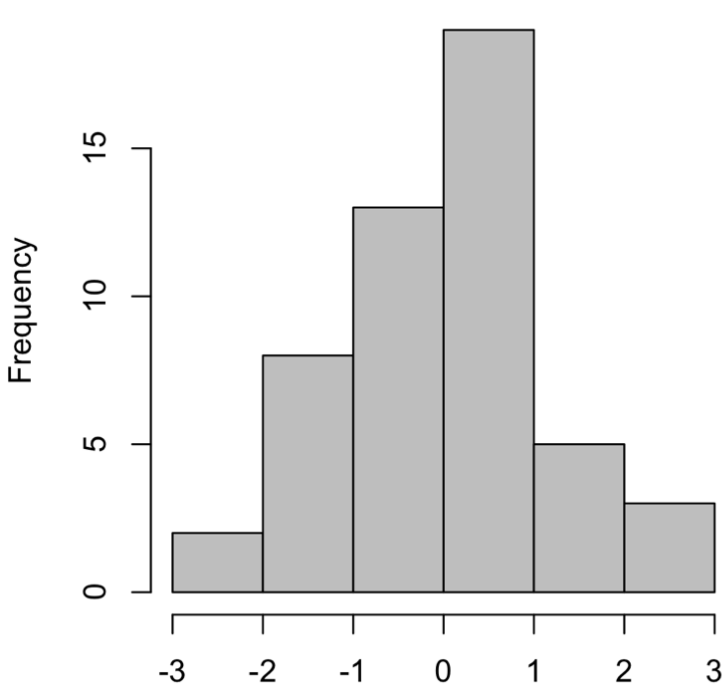
Graphics for exploratory analysis: univariate data

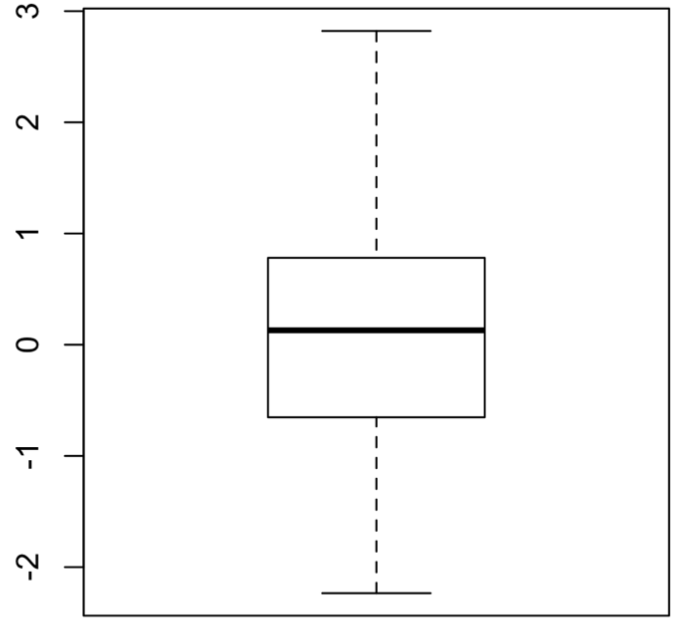
- Histograms
- Boxplots

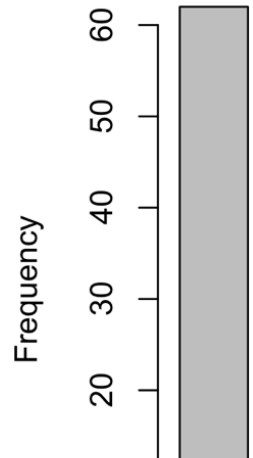
Boxplot

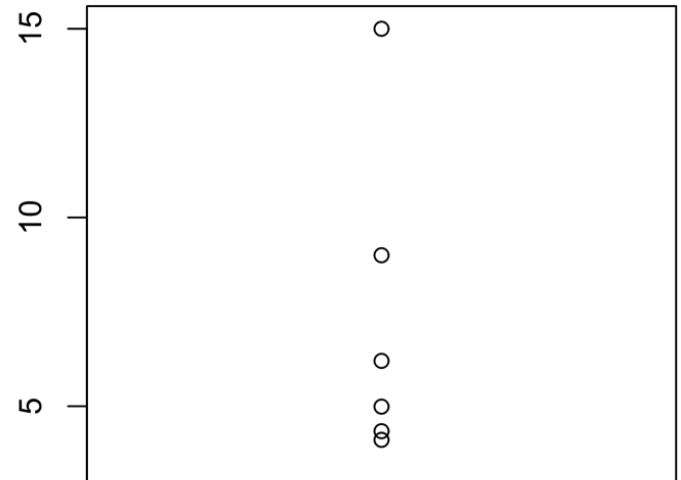

Outliers
Most extreme point <=
IQR from 3rd quarti
3rd quartile
Median
1st quartile

Using histograms and boxplots

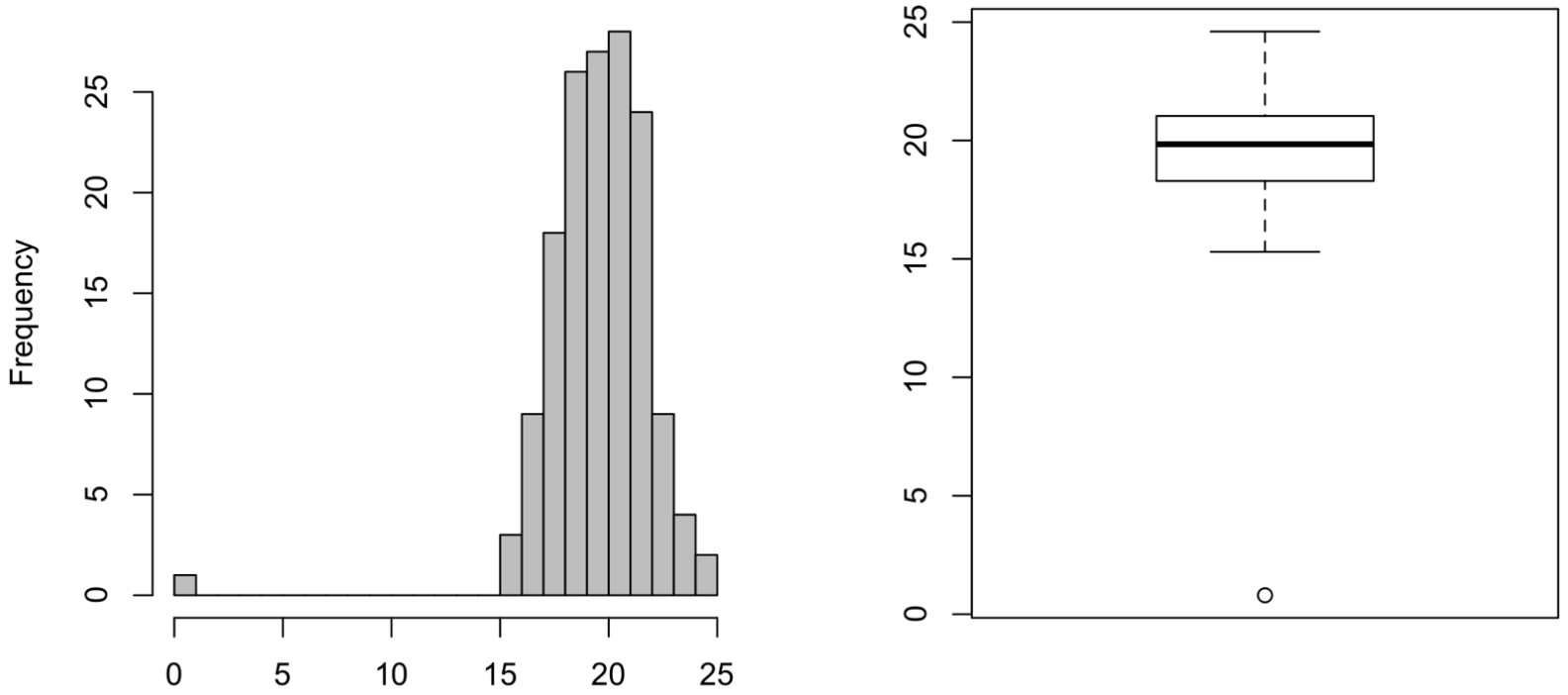








Using histograms and boxplots

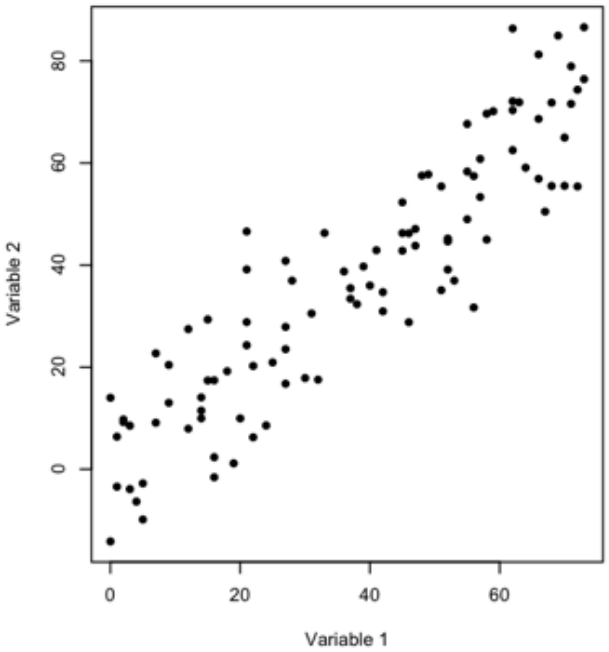


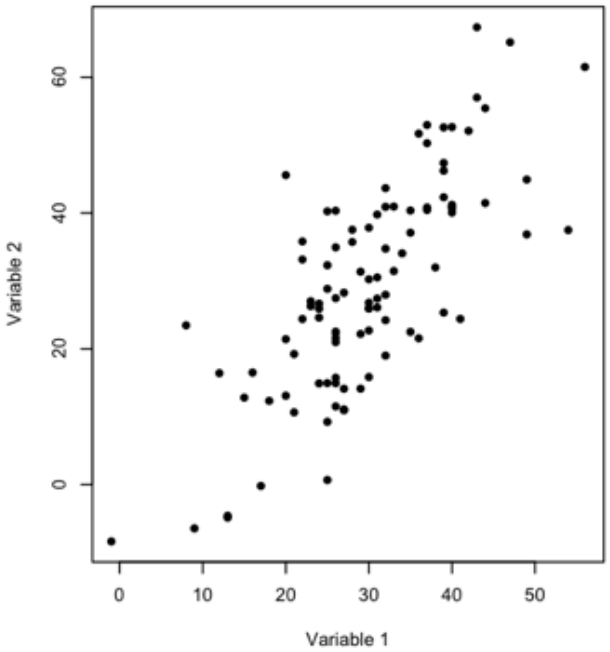
Univariate data

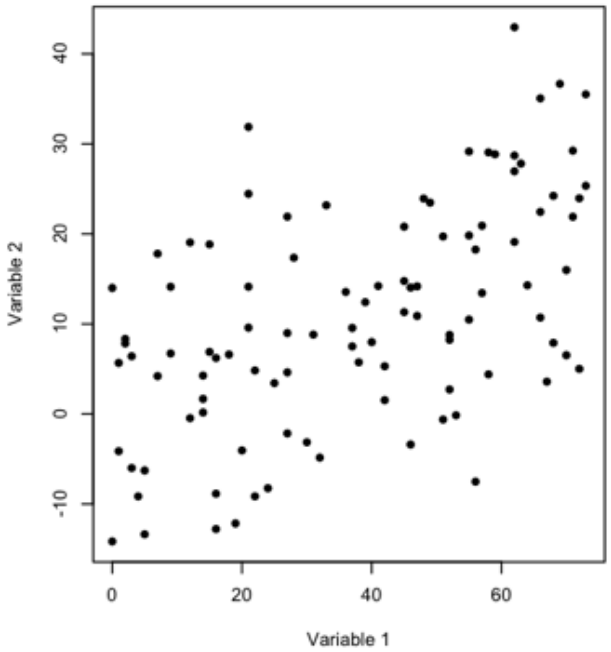
- Using exploratory stem-and-leaf plots, histograms and boxplots can:
- Tell us about the shape of the frequency distribution
- Help us identify outliers
- Help us identify possible errors

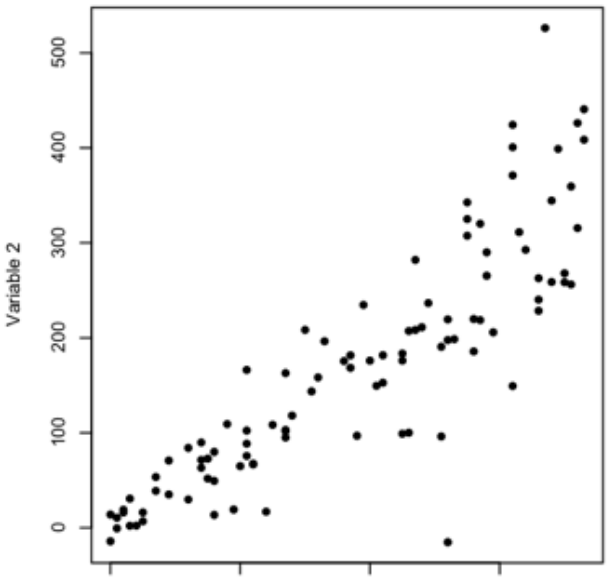
Graphics for exploratory analysis: bivariate data

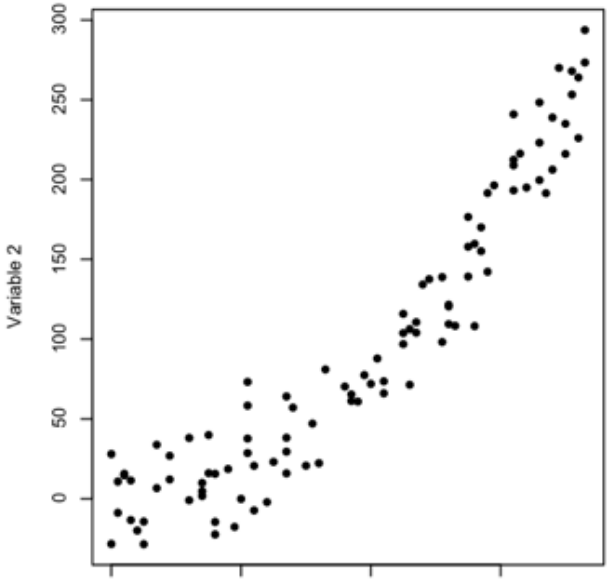
- Scatter plots: use with two continuous variables
- Stripplots and boxplots: use with one continuous variable and one categorical variable
- Barplots: use with frequency data

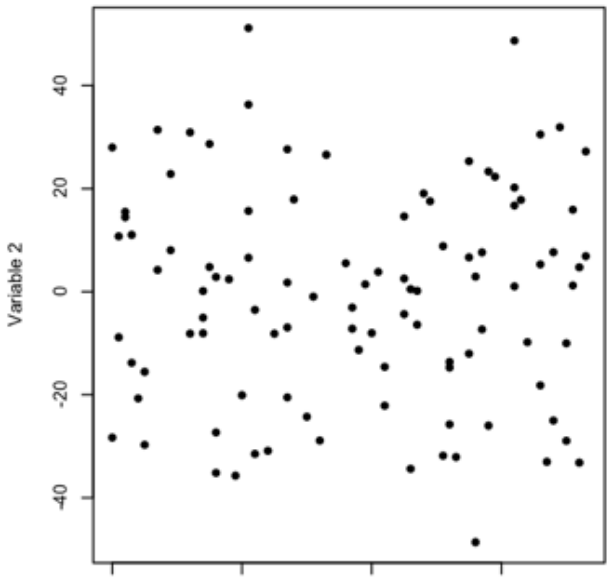

Variable 2
Variable 1
0
20
40
60
0
20
40
60
80

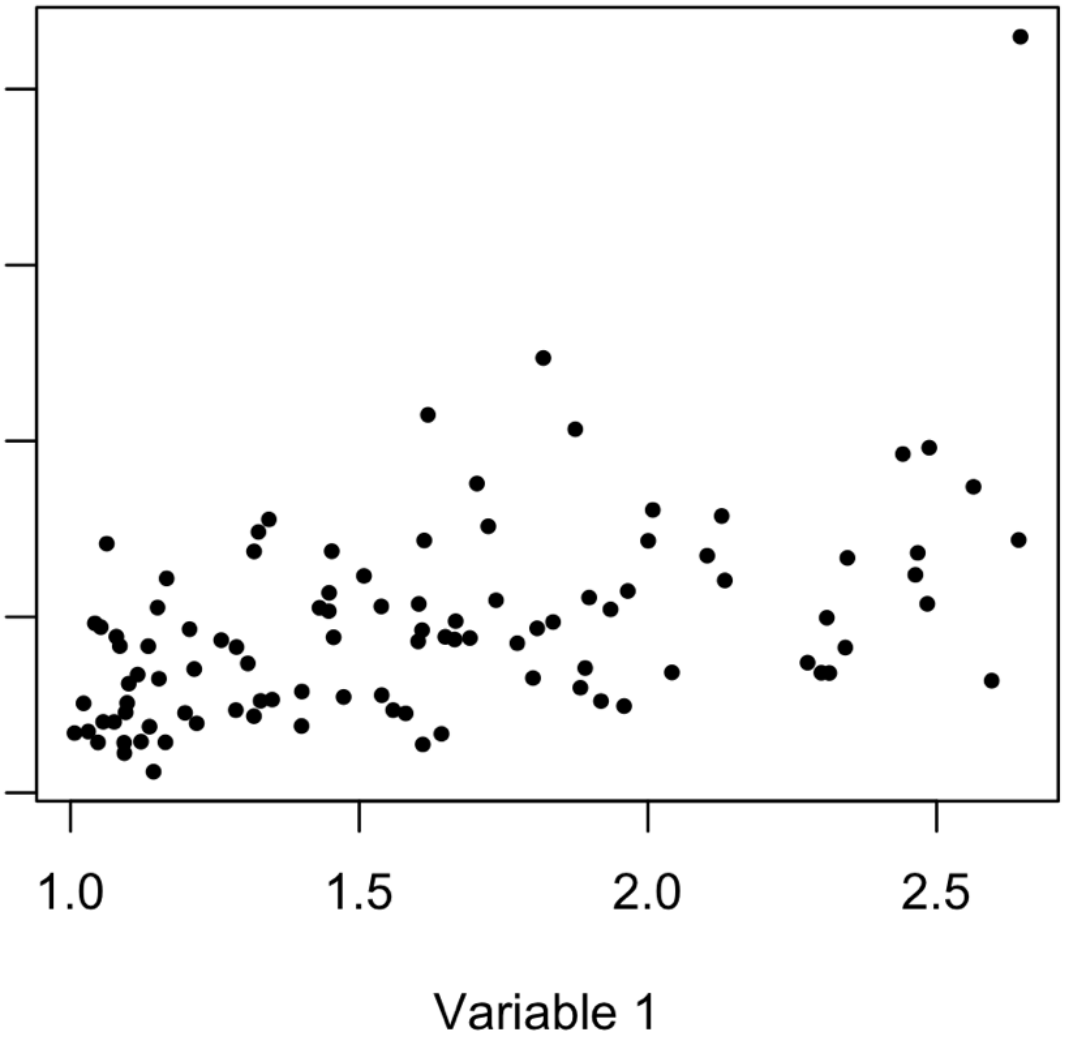

Variable 2
Variable 1
0
10
20
30
40
50
0
20
40
60

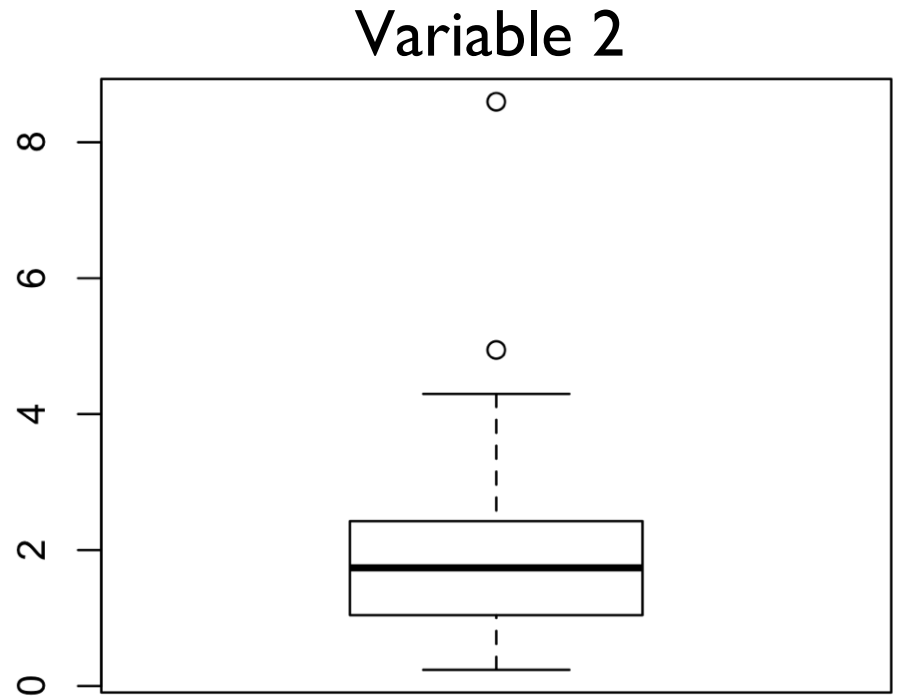

Variable 2
Variable 1
0
20
40
60
-10
0
10
20
30
40

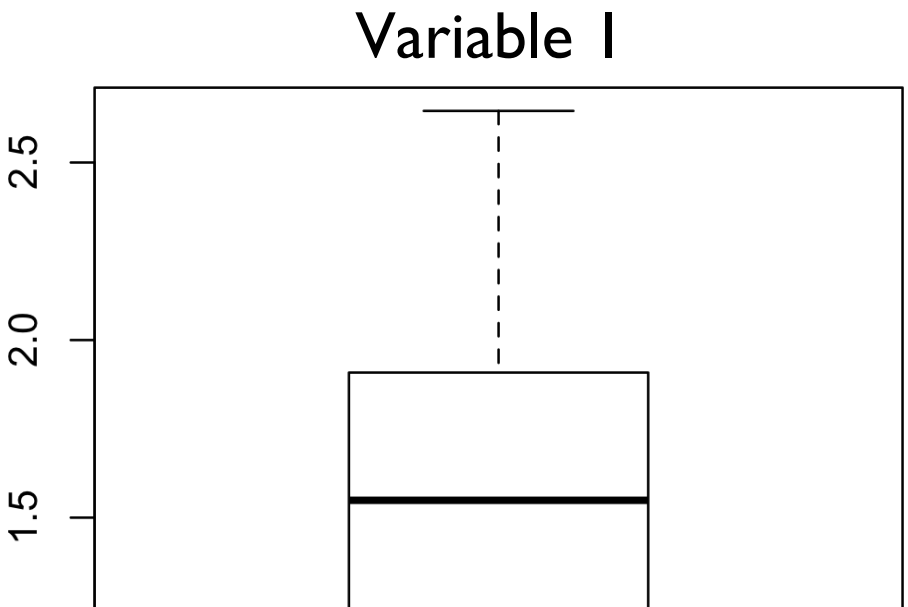

Variable 2
0
100
200
300
400
500

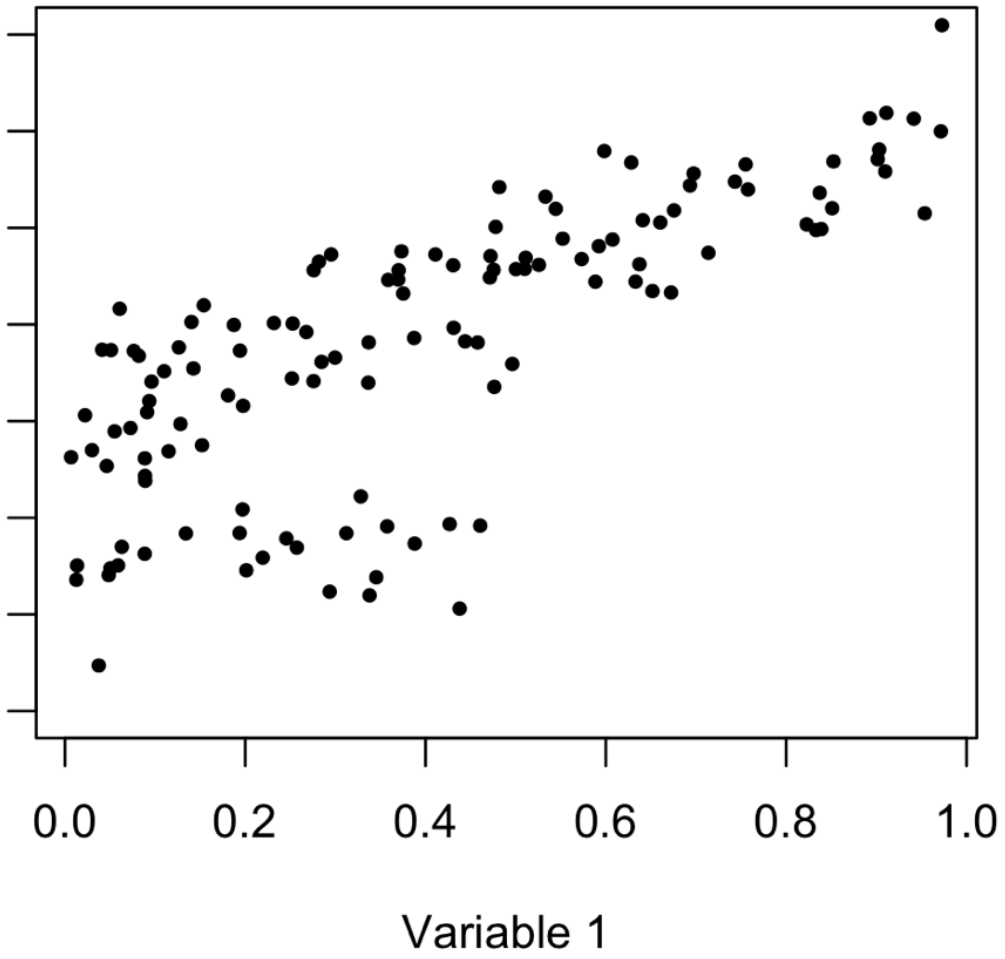

Variable 2
0
50
100
150
200
250
300

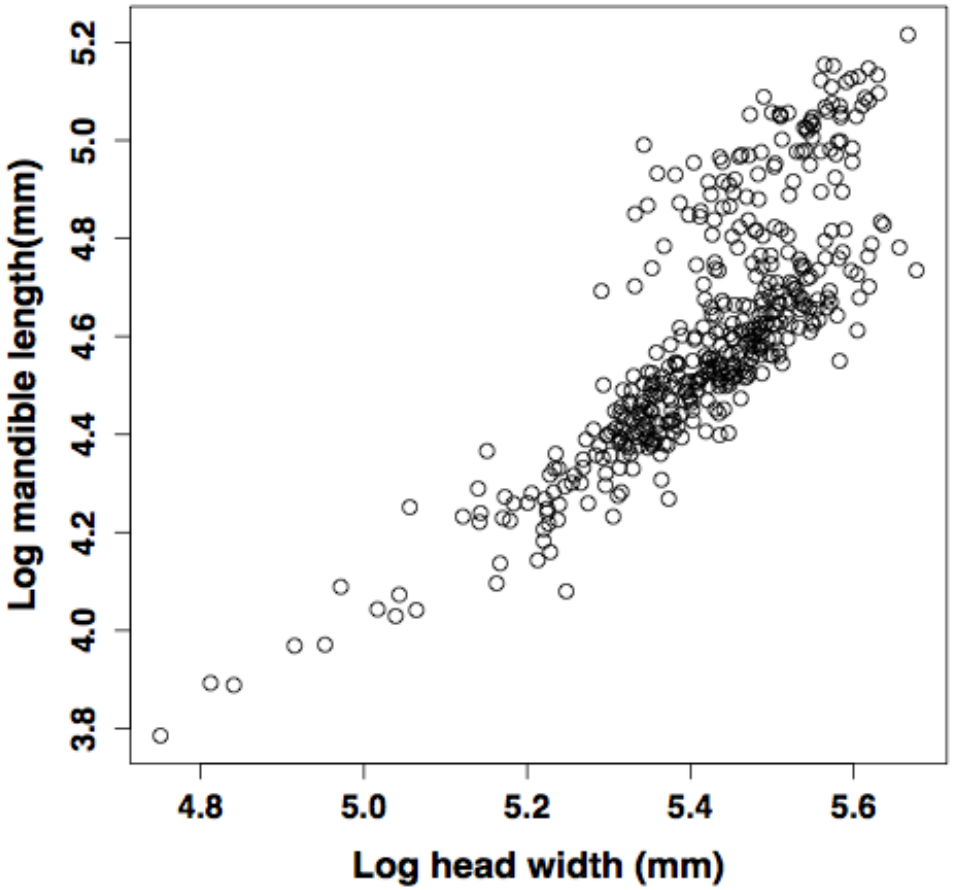

Variable 2
-40
-20
0
20
40

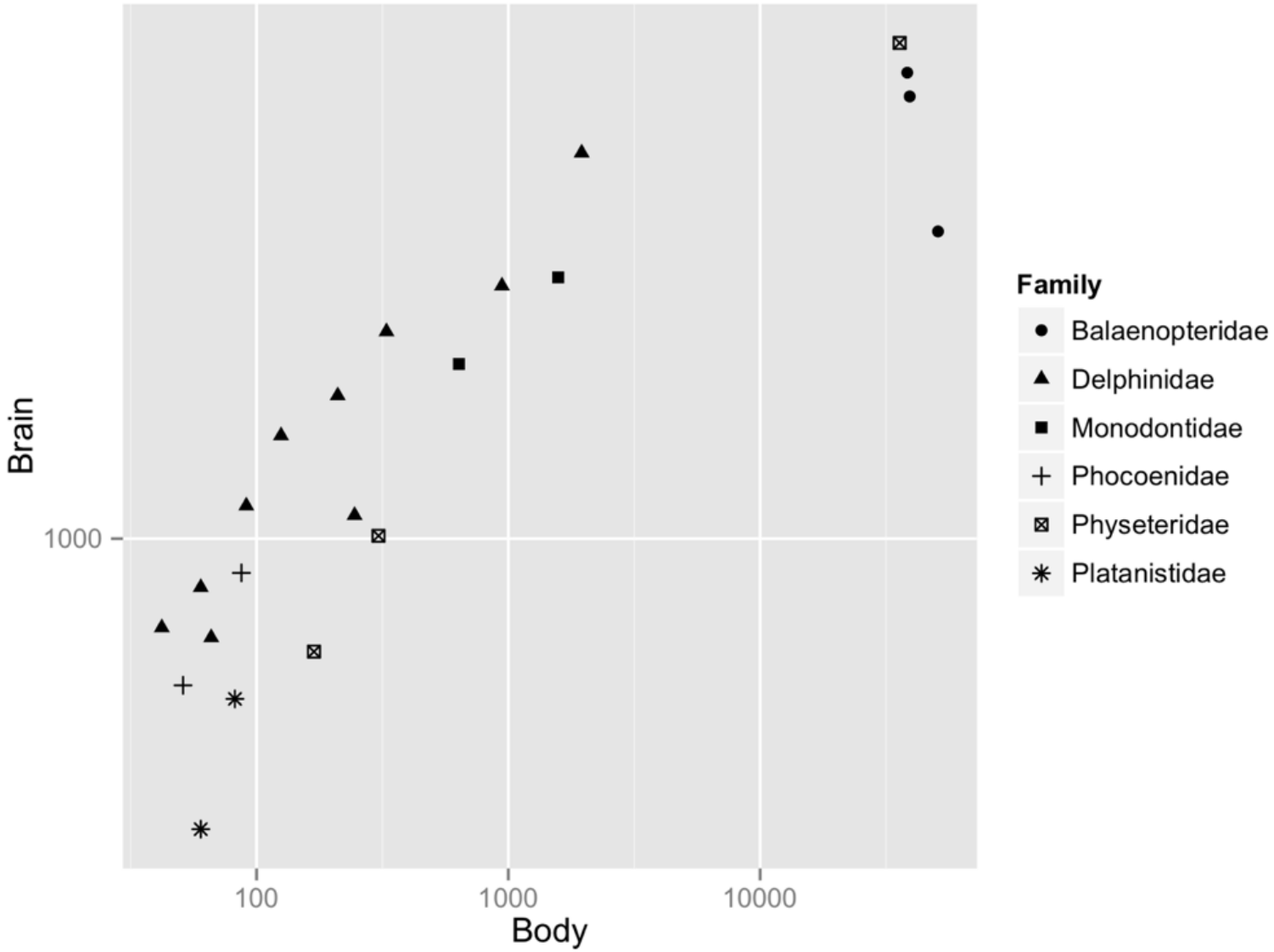

1.0
1.5
2.0
2.5
Variable 1


Variable 2
0
2
4
6
8


Variable 1
1.5
2.0
2.5


0.0
0.2
0.4
0.6
0.8
1.0
Variable 1

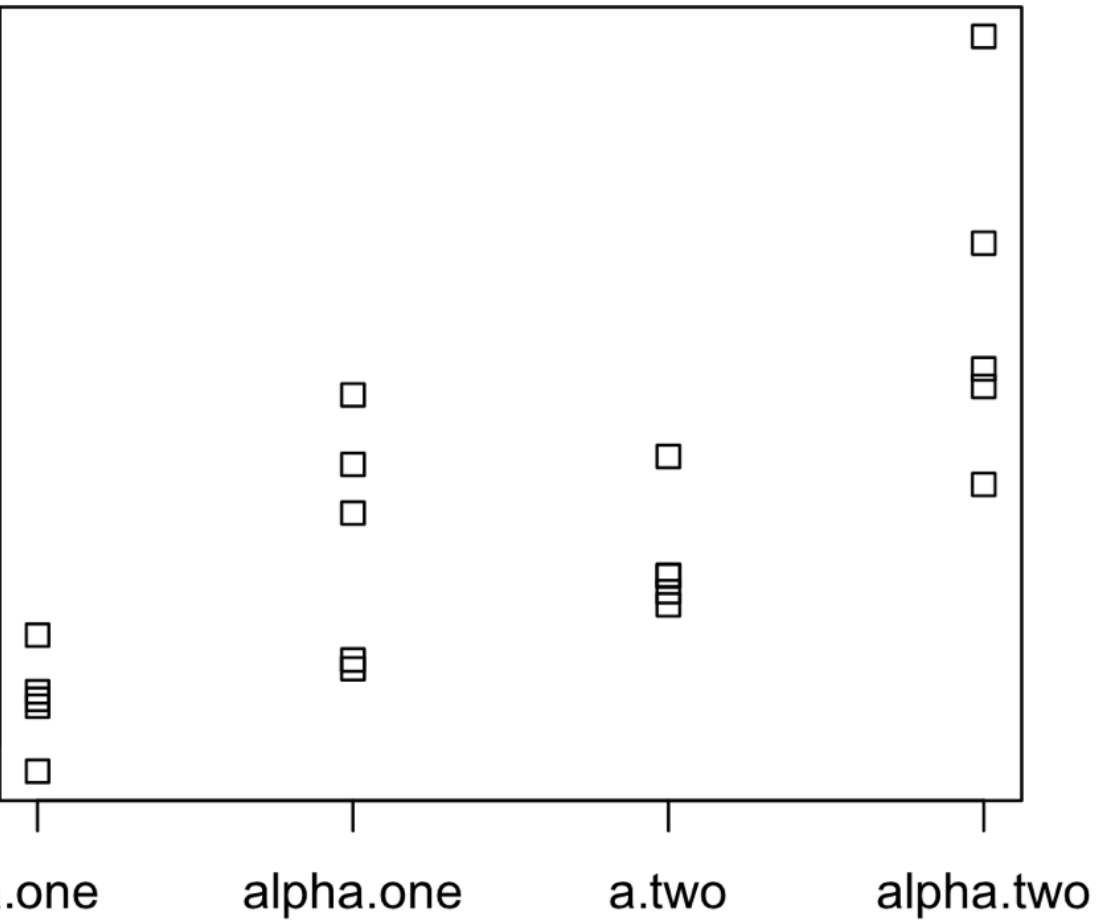

Log mandible length(mm)
3.8
4.0
4.2
4.4
4.6
4.8
5.0
5.2
4.8
5.0
5.2
5.4
5.6
Log head width (mm)

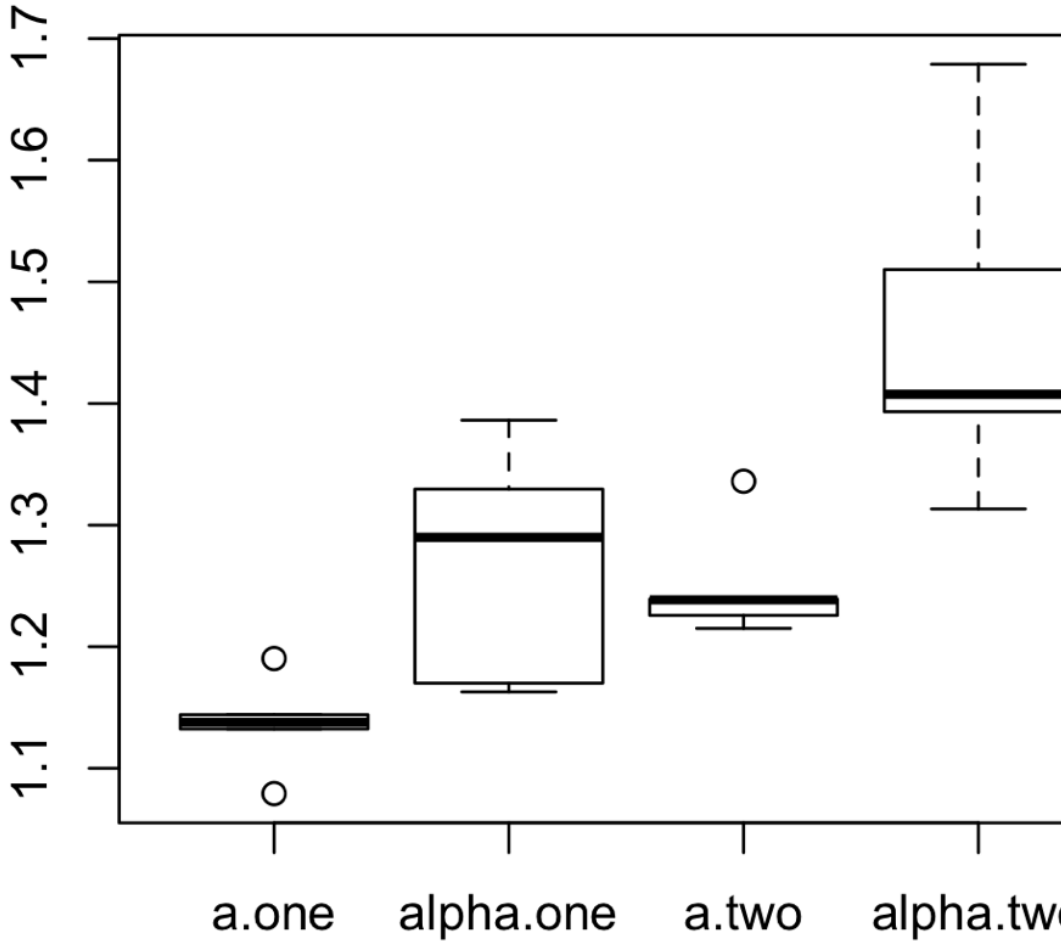

Brain
1000
100
1000
10000
Body
Family
Balaenopteridae
Delphinidae
Monodontidae
Phocoenidae
Physeteridae
Platanistidae

Scatterplots

- We can use scatterplots to:
  - See relationships between two variables
  - Check for non-linearity
  - Check for outliers and errors
  - Check for changes in variance
  - Check for structure in the data

Stripplots and boxplots





Stripplots and boxplots

- Use when you have one continuous variable and one categorical variable
- Stripplots are less useful when n is high
- Give an impression of how the continuous variable depends on the categorical
- Allow you to identify outliers, errors and patterns in variance

Recap