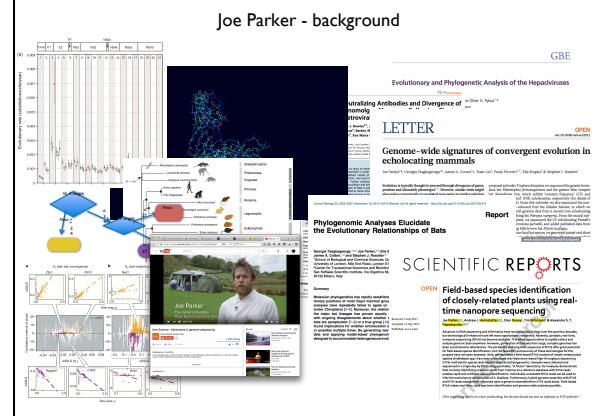


Statistics & Bioinformatics

Joe Parker



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

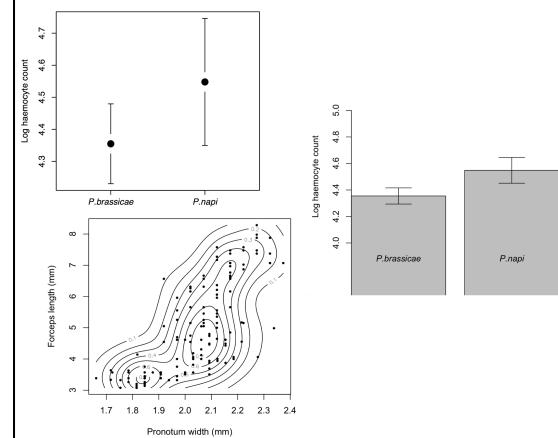
Joe Parker (Rob Knell)

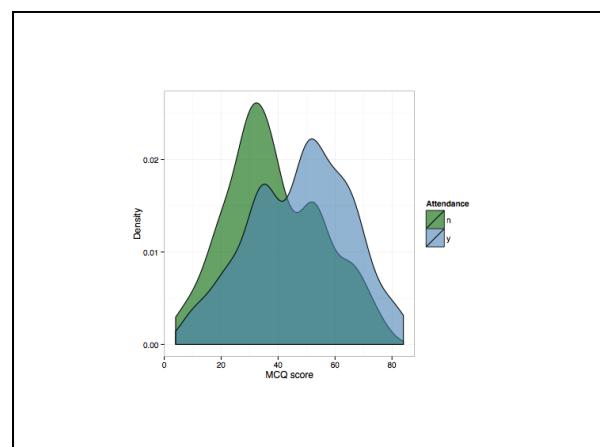
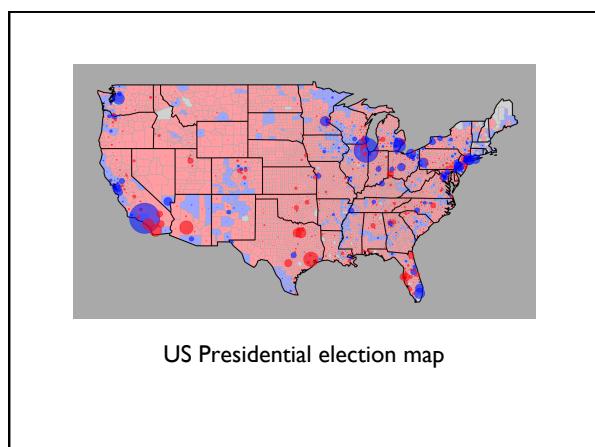
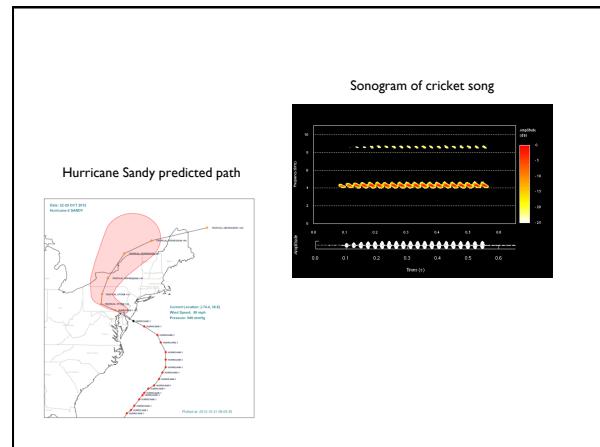
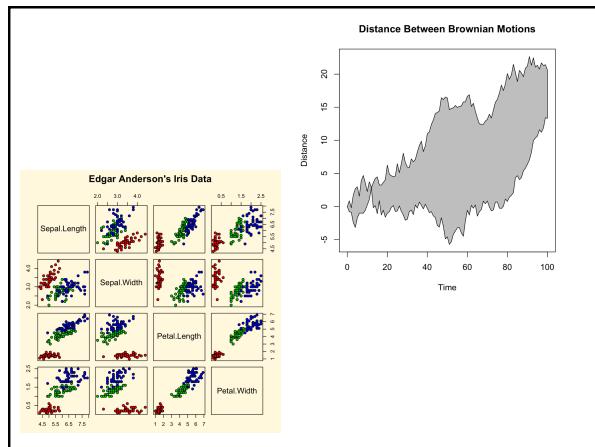
What is R?

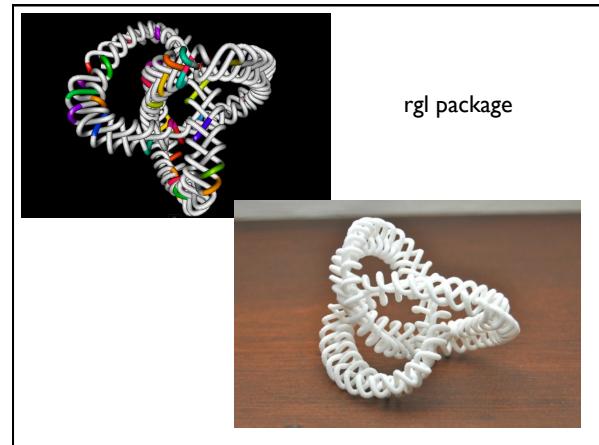
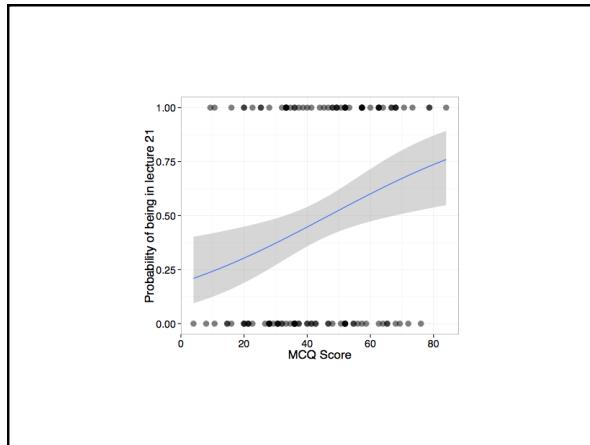
- Open source statistical programming language
- Based on a commercial language called “S”
- Powerful and flexible software that can do any statistical analysis you can think of and much more

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







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 users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for Mac OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition 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, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you 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 version.
- Sources of [R alpha](#) and [beta releases](#) (daily snapshots, created only in time periods before a planned release)
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

R basics

The R console window is shown running on a Mac OS X system. The window title is 'R Console'. The text area contains the following R session:

```

R version 2.15.2 (2012-10-26) -- "Trick or Treat"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-4
Platform: i386-apple-darwin9.8.1 (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 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-cran-project' for applications.

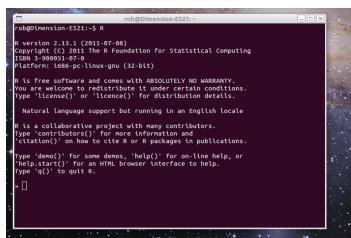
Type 'demo()' for some demos, 'help()' for on-line help, or
'browse()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R:~] ~ 1:33 (0355) 1MB-apple-darwin9.8.1]
[History restored from /Users/RKneifel/.Rstate]
[History restored from /Users/RKneifel/.History]
>

```

R console open on a Mac

R basics



R in a terminal window on a Linux machine

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

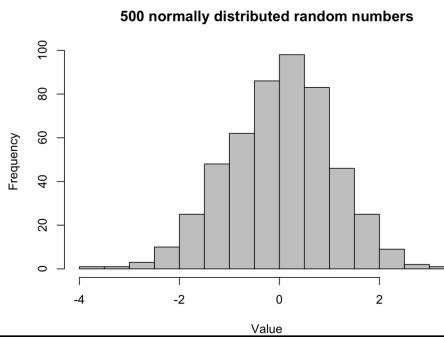
Exploratory data analysis

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

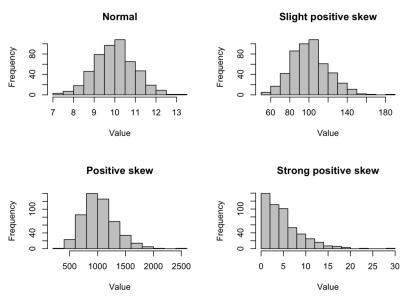
Exploratory data analysis

- 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

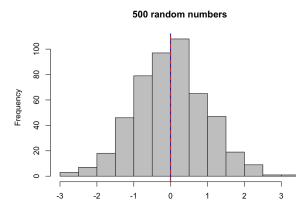
$$\text{Mean } \bar{x} = \frac{1}{n} \sum x$$

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

Mode: most common value

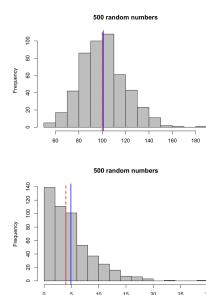
Central tendency

For normally distributed data, mean (blue) = median (red) = mode



Central tendency

As the distribution gets more skewed, the mean, mode and median become less similar



Measures of dispersion

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

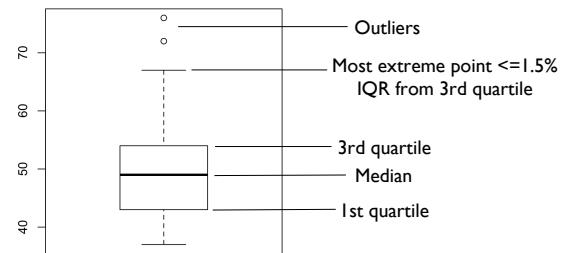
$$\text{Standard deviation } s = \sqrt{s^2}$$

Range from 1st quartile (top of lower 25%) to third quartile (top of lower 75% of data): central 50% of data

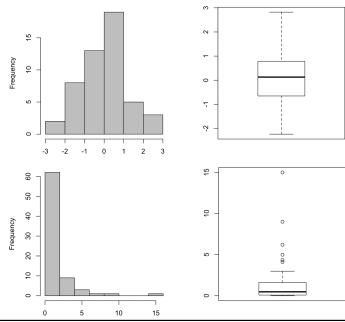
Graphics for exploratory analysis: univariate data

- Histograms
- Boxplots

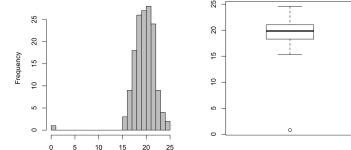
Boxplot



Using histograms and boxplots



Using histograms and boxplots



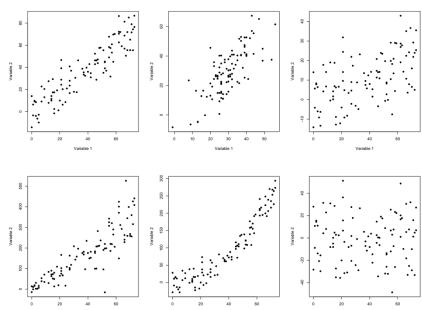
Univariate data

- 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

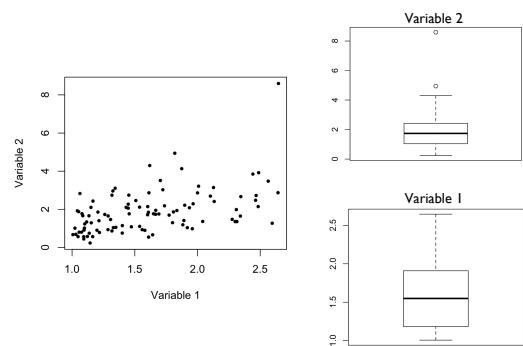
Graphics for exploratory analysis: bivariate data

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

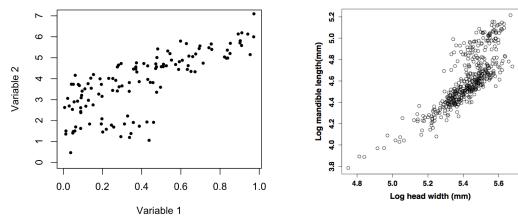
Scatter plots



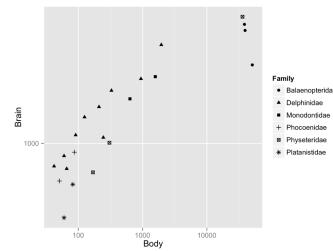
Scatter plots



Scatter plots



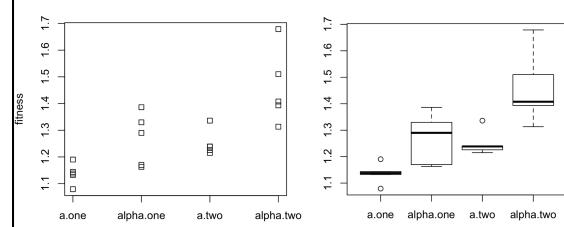
Scatter plots



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