Statistical Computing

(140.776)

http://www.biostat.jhsph.edu/~hji/courses/statcomputing/

Instructor: Hongkai Ji

Department of Biostatistics
615 N Wolfe Street, Room E3638
Phone: 410-955-3517
Email: hji@jhsph.edu

Teaching Assistant: Thomas Prior

Email: tprior@jhsph.edu

Survey

- How many of you have programming experience?

- Among those who have programming experience:

(1) What programming language(s) do you use?

R

MATLAB

C/C++

Perl

Others

(2) How many lines of code have you written in your biggest program?

<100

100 – 1000

1000 – 10,000

>10,000

(3) Do you know how to use “debug” tools to find logical errors in a program?

Who should take this course

- This course is about **R programming**
- We will also talk a little about how to use programs to solve statistical problems
- You should take this course

  (1) If you want to learn R;

  (2) If you want to obtain some basic skills to deal with data (visualization, elementary statistical analysis);

  (3) If you have some programming experience but wish to improve it (for example, learn how to debug a program to make it work properly).

Who should take this course

- If you don’t have any programming experience

  (1) We recommend you to take a basic programming course first before taking this course;

  (2) Or be prepared to work really hard

- You can skip this course

  (1) If you already have experience in writing big programs (>10,000 lines of code)

  (2) If you already know R very well

776 vs. 778

- Statistical Computing (140.776)
  - Practical issues: programming
  - Elementary statistical computing topics

- Advanced Statistical Computing (140.778)
  - Algorithm design: Optimization, Monte Carlo, Markov Chain Monte Carlo, etc.
  - Theoretical issues: How do they work, why do they work, how to make them efficient

Tips for learning

- Bring your labtops with R

  If you don’t have a labtop, please find someone who can share his/her labtop with you in the lecture.

- Please check our website and download data for the lecture before you come

- Do your homework, and do it yourself

  You are encouraged to discuss with others, but you have to write your own code. Otherwise you will have trouble in the final exam.

Grading system

- Participation: 10%

- Homework (3-4): 70%

- Final Exam (in class): 20%

  BRING YOUR COMPUTERS!

Teaching Assistant

- Tom Prior

- Office hour

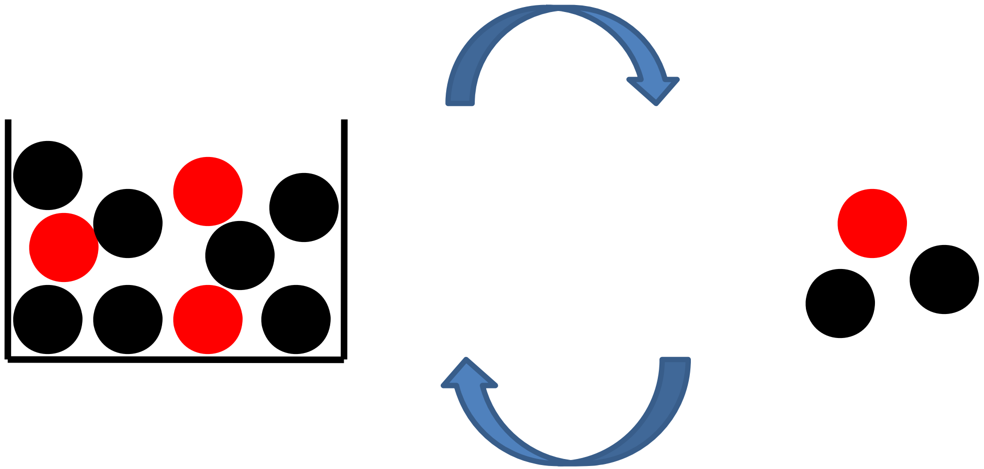
R is a statistical programming language

Some basic elements of statistics

Statistics is a data-driven science

<u>Probability</u>:

What is the probability to get 1 red and 2 black balls?



<u>Statistics:</u>

What percentage of balls in the box are red?

Study design and data collection

- Example:

  Does fish oil help reduce blood pressure?

- Randomization
  - Random sampling from the population (Inference can be drawn for the population)
  - Random treatment assignment (Causal inference can be drawn)

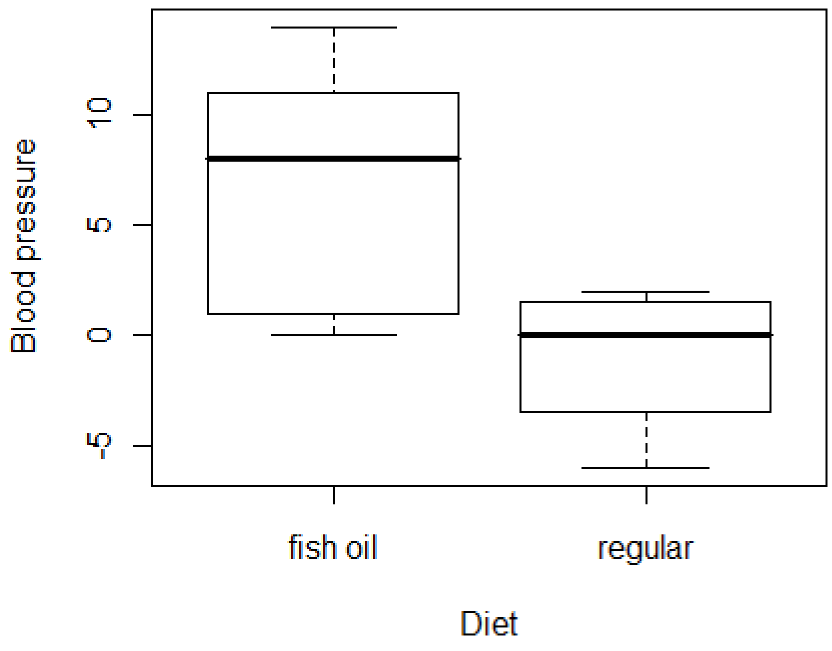
- Observational studies vs. Randomized experiments

Data

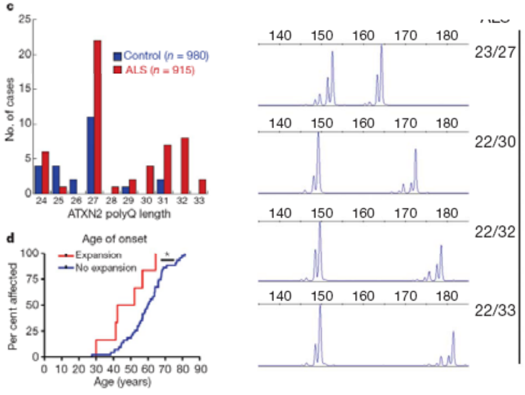
	Reduction of blood pressure						
Fish oil diet:	8	12	10	14	2	0	0
Regular diet:	-6	0	1	2	-3	-4	2

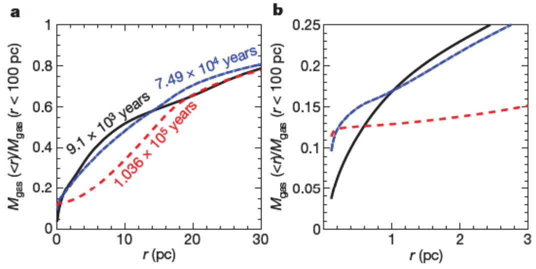
Question: What is the first thing you would do to analyze the data?

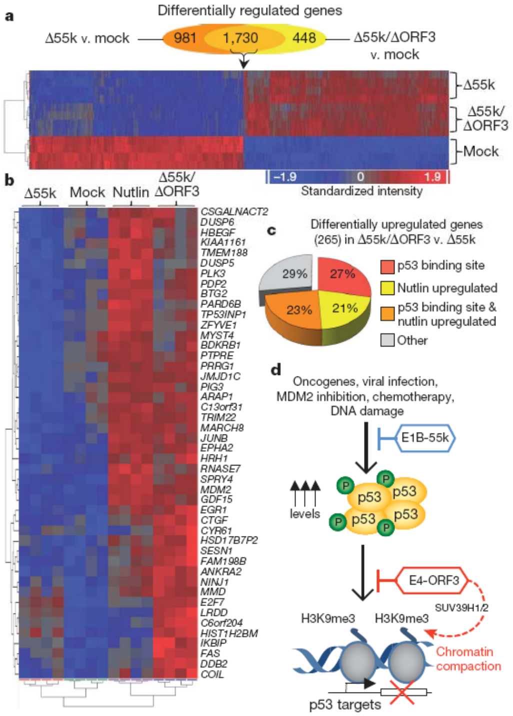
Data Exploration and Visualization

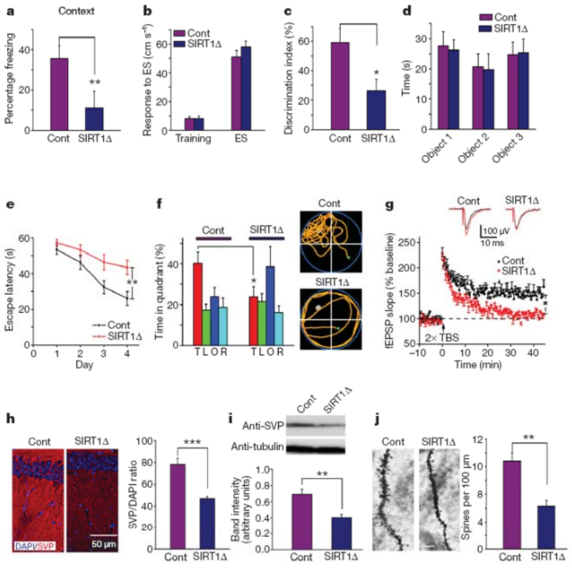


Data Exploration and Visualization









Statistical Inference

Reduction of blood pressure

Fish oil diet: 8 12 10 14 2 0 0
Regular diet: -6 0 1 2 -3 -4 2

mean in fish oil group: 6.57 → estimate of population mean μ_1

mean in regular group: -1.14 → estimate of population mean μ_2

Hypothesis test:

Null hypothesis: $\mu_1 = \mu_2$

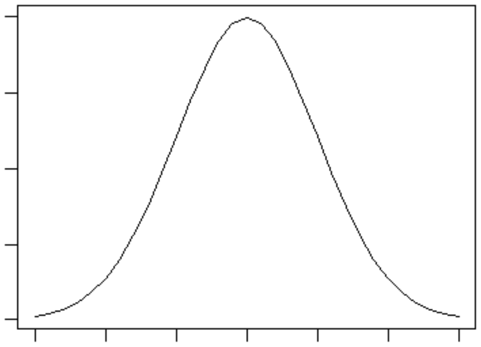
Alternative hypothesis: $\mu_1 \neq \mu_2$

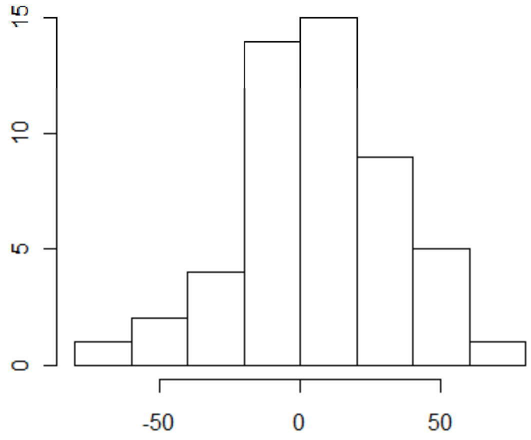
t-statistic = 3.0621

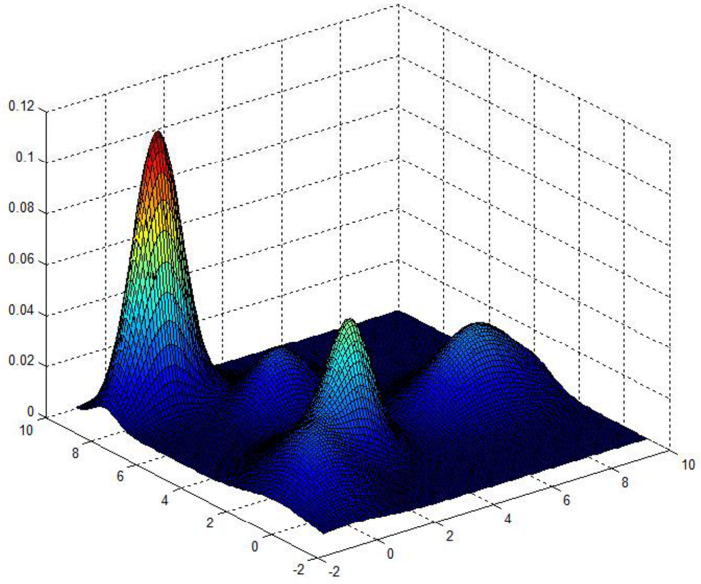
p-value = 0.013

Linear regression, Mixed effects models, Generalized linear models, …

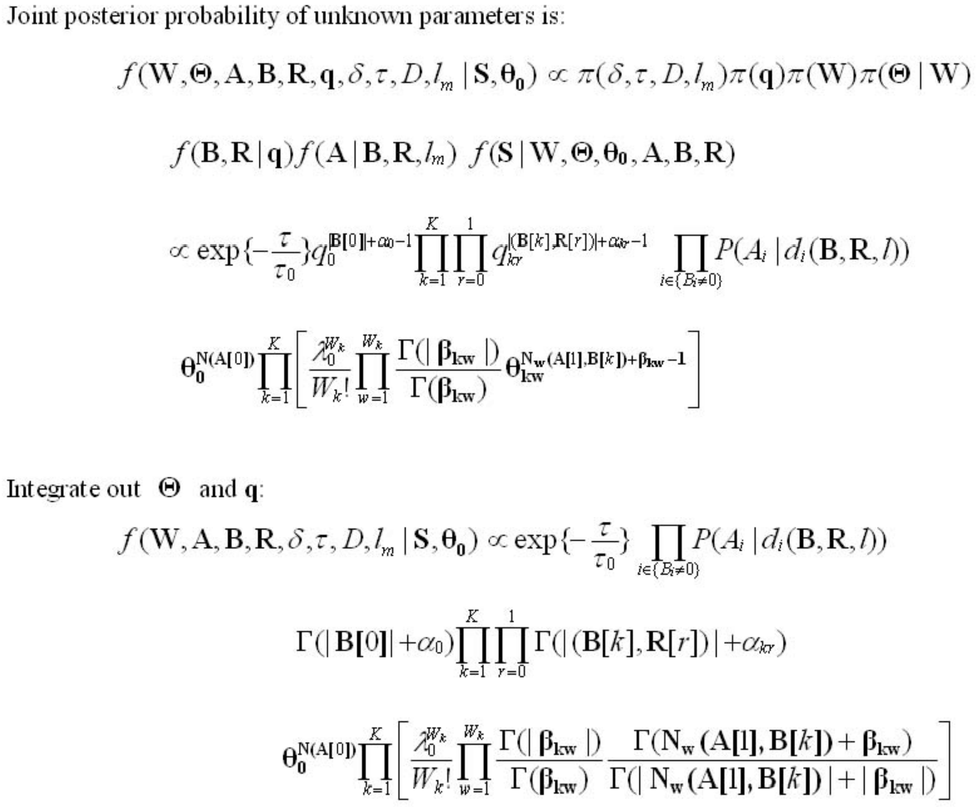
Assumptions, assumptions, assumptions!







Build more complicated models



Joint posterior probability of unknown parameters is:

$$f(\mathrm{W},\boldsymbol{\Theta},\mathrm{A},\mathbf{B},\mathrm{R},\mathbf{q},\delta,\tau,D,l_m \mid \mathbf{S},\boldsymbol{\theta}_\mathbf{0}) \propto \pi(\delta,\tau,D,l_m)\pi(\mathbf{q})\pi(\mathrm{W})\pi(\boldsymbol{\Theta} \mid \mathrm{W})$$

$$f(\mathbf{B},\mathrm{R} \mid \mathbf{q})f(\mathrm{A} \mid \mathbf{B},\mathrm{R},l_m)\ \ f(\mathbf{S} \mid \mathrm{W},\boldsymbol{\Theta},\boldsymbol{\theta}_\mathbf{0},\mathrm{A},\mathbf{B},\mathrm{R})$$

$$\propto \exp\{-\frac{\tau}{\tau_0}\}q_0^{|\mathbf{B}[0]|+\alpha_0-1}\prod_{k=1}^{K}\prod_{r=0}^{1}q_{kr}^{|(\mathbf{B}[k],\mathbf{R}[r])|+\alpha_{kr}-1}\ \prod_{i\in\{B_i\neq 0\}}P(A_i \mid d_i(\mathbf{B},\mathbf{R},l))$$

$$\boldsymbol{\theta}_\mathbf{0}^{\mathrm{N}(\mathbf{A}[0])}\prod_{k=1}^{K}\left[\frac{\lambda_0^{W_k}}{W_k!}\prod_{w=1}^{W_k}\frac{\Gamma(|\,\boldsymbol{\beta}_\mathbf{kw}\,|)}{\Gamma(\boldsymbol{\beta}_\mathbf{kw})}\boldsymbol{\theta}_\mathbf{kw}^{\mathbf{N_w}(\mathbf{A}[1],\mathbf{B}[k])+\boldsymbol{\beta}_\mathbf{kw}-\mathbf{1}}\right]$$

Integrate out $\boldsymbol{\Theta}$ and $\mathbf{q}$:

$$f(\mathrm{W},\mathrm{A},\mathbf{B},\mathrm{R},\delta,\tau,D,l_m \mid \mathbf{S},\boldsymbol{\theta}_\mathbf{0}) \propto \exp\{-\frac{\tau}{\tau_0}\}\ \prod_{i\in\{B_i\neq 0\}}P(A_i \mid d_i(\mathbf{B},\mathbf{R},l))$$

$$\Gamma(|\,\mathbf{B[0]}| + \alpha_0)\prod_{k=1}^{K}\prod_{r=0}^{1}\Gamma(|\,(\mathbf{B}[k],\mathbf{R}[r])\,| + \alpha_{kr})$$

$$\boldsymbol{\theta}_\mathbf{0}^{\mathrm{N}(\mathrm{A}[0])}\prod_{k=1}^{K}\left[\frac{\lambda_0^{W_k}}{W_k!}\prod_{w=1}^{W_k}\frac{\Gamma(|\,\boldsymbol{\beta}_\mathbf{kw}\,|)}{\Gamma(\boldsymbol{\beta}_\mathbf{kw})}\frac{\Gamma(\mathrm{N}_\mathbf{w}(\mathbf{A[1]},\mathbf{B}[k]) + \boldsymbol{\beta}_\mathbf{kw})}{\Gamma(|\,\mathrm{N}_\mathbf{w}(\mathbf{A[1]},\mathbf{B}[k])\,| + |\,\boldsymbol{\beta}_\mathbf{kw}\,|)}\right]$$

How to handle these complex models?

Now you need Advanced computing techniques such as Markov Chain Monte Carlo, EM, etc., which will be covered by **Advanced Statistical Computing.**

Data are getting bigger

- Netflix competition

  > 480,000 customers, >18,000 movie titles, >100 million ratings (scale from 1 to 5 stars)

  Data collected between October, 1998 and December, 2005.

  Predict how a customer will rate a new movie.

- Human genome project

  3×10^9 base pairs (3 GB)

- 1000 genome projects

  $3 \times 10^9 \times 1000$

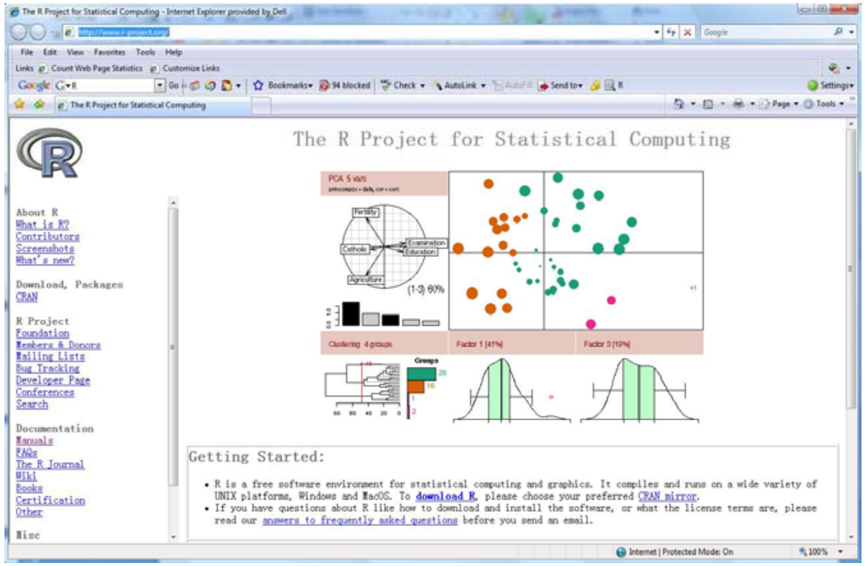
We need help from computers!

That’s why statistical computing becomes so important!

Programming Languages

- R

  http://www.r-project.org/


The R Project for Statistical Computing - Internet Explorer provided by Dell
http://www.r-project.org/
File Edit View Favorites Tools Help
Links Count Web Page Statistics Customize Links
Google Go Bookmarks 94 blocked Check AutoLink AutoFill Send to Settings
The R Project for Statistical Computing
Page Tools
PCA 5 vars
Fertility
Catholic
Examination
Education
Agriculture
(1-3) 60%
Clustering 4 groups
Groups
Factor 1 [41%]
Factor 3 [19%]
Internet | Protected Mode: On
100%

The R Project for Statistical Computing

About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
The R Journal
Wiki
Books
Certification
Other

Misc

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

- C

Programming Languages

C: Compiled language, transformed into an executable form before running

R: Interpreted language, read and then executed directly

An Introduction to R

- URL

  R website: http://cran.r-project.org/ or http://www.r-project.org/
  (Download, Manuals)

- An integrated suite of software for data manipulation, calculation and graphical display

- An implementation of the S language

History of S

- S was developed at Bell Labs by Rick Becker, John Chambers and Allan Wilks
- 1976: initiated as an internal statistical analysis environment, implemented as Fortran libraries
- 1988: rewritten in C and began to resemble the system we have today
- 1998: version 4 released, the version we use today

- 1993: Bell Labs gave StatSci (now Insightful Corp.) an exclusive license to develop and sell the S language
- Insightful sells its implementation of the S language under the product name S-PLUS and has built a number of fancy features (GUI, mostly) on top of it – hence the “PLUS”.
- S language itself has not changed dramatically since 1998

Sources: Roger Peng

History of R

- 1991: Created in New Zealand by Ross Ihaka and Robert Gentleman
- 1993: First announcement of R to the public
- 1995: Martin Mächler convinces Ross and Robert to use the GNU General Public License to make R free software.
- 1996: A public mailing list is created (R-help and R-devel)
- 1997: The R Core Group is formed (containing some people associated with S-PLUS). The core group controls the source code for R.
- 2000: R version 1.0.0 is relased.
- 2009-08-24: R version 2.9.2

Sources: Roger Peng

Features of R

- Syntax is very similar to S
- Runs on almost any standard computing platform
- Frequent releases
- Graphics capabilities
- Can be used interactively AND contains a powerful programming language
- Active user community
- Free

Sources: Roger Peng

Some R Resources

Available from

http://cran.r-project.org/

or http://www.r-project.org

- An Introduction to R
- The R language definition
- Writing R Extensions
- R Data Import/Export
- R Installation and Administration
- R Internals
- The R Reference Index

R Books

- There is a “books” link at http://www.r-project.org

Getting Help in R

- Type command

```
> help(lm)
```

or

```
> ?lm
```

R commands

- R is an expression language

  Elementary commands consist of either expressions or assignments

  Commands separated by a semi-colon (;) or a new line; grouped by braces ({ })
  Comments start with a hashmark (#)

- R is case sensitive

  A and a are different

Executing commands from or diverting output to a file

- Execute commands from a file

```
> source(“mycommands.R”)
```

- Divert output to a file

```
> sink(“myresults.lis”)
```

To restore it to the console again, use

```
> sink()
```