Statistics with R for Biologists – Course Syllabus

James Bullard, Kasper Hansen, Margaret Taub

July 3, 2008

Before the course

We expect everyone to show up with a laptop with R version 2.7.0 or 2.7.1 installed. If you have an older version, please update. You can obtain R from www.cran.r-project.org. If you have major problems with this, we can sort it out Monday morning. Specific packages will be installed as we go.

During the course we will access the internet using AirBears. Make sure you have a working account (this is mostly for post docs and PIs).

There will be a fair amount of programming. You may want to think about your programing environment, specifically what text editor you are using. Depending on the operating system, R's built-in editor might be sufficient, but we can also recommend Emacs, tinnR and WinEdt for Windows, TextMate and Emacs for Mac OS X and Emacs for Linux. TextMate can be obtained for free from the Campus software distribution (software-central.berkeley.edu).

The course will be held in Stanley Hall, Room 177. We plan to start each morning at 9am (Berkeley time). We hope to provide a coffee-supplied break at around 10:30 every day, thanks to the Designated Emphasis in Computational and Genomic Biology. We will take a break for lunch of around one hour, and plan to wrap up the day by 4:30.

Syllabus

Note that this syllabus is preliminary and subject to change. It is also rather high-level, so we will probably cover more than is listed here. Also note that we don't, in general, expect the topics to be evenly split between the morning and afternoon sessions.

7/7 Monday

Topic 1 Introduction to R

1. Background / introduction
2. Getting help
3. Basic data types and manipulation
4. Reading in data
5. Examining data

Topic 2 Introduction to basic Statistics

1. Simulating random variables
2. Basic visualization
3. The normal distribution
4. Application: examining limit theorems

7/8 Tuesday

Topic 3 R as a programming language

1. Syntax
2. Types (factors, numbers, characters) / data structures (matrices, data.frames, lists, environments)
3. Control structures
4. Functions
5. Classes (S3/S4)

Topic 4 Exploratory data analysis and plotting in R

1. Numerical summaries
2. Statistical graphics
3. Smoothing / advanced plotting
4. Microarray graphics

7/9 Wednesday

Topic 5 Statistics – Testing

1. Understanding p-values / null distributions / hypothesis tests
2. Common tests / common distributions
3. Permutation-based tests
4. Multiple testing (FDR / Bonferroni)

Topic 6 Statistics – Basic models

1. Linear models
2. Application: gene-environment study (current paper)

7/10 Thursday

Topic 7 Statistics – Clustering / unsupervised learning

1. Clustering
2. Application: HapMap data

Topic 8 Statistics – Classification / supervised learning

1. Prediction error / overfitting
2. Cross-validation
3. Classification
4. ROC curves

7/11 Friday

Topic 9 Microarray analysis

1. Microarray pre-processing / normalization / probe affinities / technologies (two-color / Agilent / Illumina / Affy)

Topic 10 Experimental design

1. Confounding
2. Causal relationships
3. Randomization as a means to control confounding
4. Experimental design