Statistics For Biologists
A Crash Course Using Computation and Concepts

Randall Pruim

SC 11 workshop (June, 2011)
1 A Crash Course in Statistics for Biologists (and Their Friends) 1-1
   1.1 Why Use R? ................................................................. 1-1
   1.3 Two Illustrative Examples ............................................. 1-2

2 Getting Started: The First Week With R 2-1
   2.1 Getting Students Familiar with R ..................................... 2-1
   2.2 Examples for Early in the Course .................................... 2-3

3 An Introduction to R 3-1
   3.1 Welcome to R and RStudio ............................................. 3-1
   3.2 Using R as a Calculator ................................................. 3-2
   3.3 R Packages .................................................................. 3-3
   3.4 Getting Help .................................................................. 3-4
   3.5 Data ............................................................................ 3-5
   3.6 Summarizing Data ......................................................... 3-9
   3.7 Additional Notes on R Syntax ............................................ 3-22
   3.8 Installing R .................................................................. 3-23
   3.9 R Examples .................................................................... 3-24
   3.10 Exercises ..................................................................... 3-25

4 Getting Interactive with manipulate 4-1
4.1 Simple Things ................................................................. 4-1

5 Simulation Based Inference ................................................. 5-1
  5.1 Simulation and Randomization with the mosaic Package .......... 5-1
  5.2 The Multi-World Metaphor for Statistical Inference .............. 5-13
  5.3 More Examples .................................................................. 5-25
  5.4 Bootstrap Confidence Intervals ........................................... 5-29
  5.5 Power ............................................................................. 5-29
  5.6 Exercises, Problems, and Activities .................................... 5-30

6 Taking Advantage of the Internet ........................................... 6-1
  6.1 Sharing With and Among Your Students ............................... 6-1
  6.2 Data Mining Activities ...................................................... 6-6

A More About R ..................................................................... A-1
  A.1 Installing and Using Packages ........................................... A-1
  A.2 Some Workflow Suggestions .............................................. A-2
  A.3 Working with Data .......................................................... A-3
  A.4 Primary R Data Structures ............................................... A-6
  A.5 More About Vectors ........................................................ A-9
  A.6 Manipulating Data Frames ............................................... A-13
  A.7 Functions in R ............................................................... A-18
These materials were prepared for a workshop entitled *Computational Biology for Biology Educators* held at Calvin College in June, 2011. Much of the material is recycled from a workshop entitled *Teaching Statistics Using R* prior to the 2011 United States Conference on Teaching Statistics. You can find out more about that workshop at [http://mosaic-web.org/uscots2011/](http://mosaic-web.org/uscots2011/).

The activities and examples in these notes are intended to highlight a modern approach to statistics and statistics education that focuses on modeling, resampling based inference, and multivariate graphical techniques.

These notes contain far more than will be covered in the workshop, so they can serve as a reference for those who want to learn more. For still more reference material, see the above mentioned notes at [http://mosaic-web.org/uscots2011/](http://mosaic-web.org/uscots2011/)

### R and R Packages

R can be obtained from [http://cran.r-project.org/](http://cran.r-project.org/) Download and installation are pretty straightforward for Mac, PC, or linux machines.

In addition to R, we will make use of several packages that need to be installed and loaded separately. The *mosaic* package (and its dependencies) will be assumed throughout. Other packages may appear from time to time, including

- **fastR**: companion to *Foundations and Applications of Statistics* by R. Pruim
- **abd**: companion to *Analysis of Biological Data* by Whitlock and Schluter
- **vcd**: visualizing categorical data

We also make use of the *lattice* graphics package which is installed with R but must be loaded before use.

### RStudio

RStudio is an alternative interface to R. RStudio can be installed as a desktop (laptop) application or
as a server application that is accessible to others via the Internet. Calvin has provided accounts for workshop participants on an RStudio server at

http://dahl.calvin.edu:8787

There are some things in these notes (in particular those using `manipulate()`) that require the RStudio interface to R. Most things should work in any flavor of R.

RStudio is available from

http://www.rstudio.org/

Marginal Notes

Marginal notes appear here and there. Sometimes these are side comments that we wanted to say, but didn’t want to interrupt the flow to mention. These may describe more advanced features of the language or make suggestions about how to implement things in the classroom. Some are warnings to help you avoid common pitfalls. Still others contain requests for feedback.

Document Creation

This document was created June 25, 2011, using Sweave and R version 2.13.0 (2011-04-13)
Project MOSAIC is a community of educators working to develop new ways to introduce mathematics, statistics, computation, and modeling to students in colleges and universities.

The purpose of the MOSAIC project is to help us share ideas and resources to improve teaching, and to develop a curricular and assessment infrastructure to support the dissemination and evaluation of these ideas. Our goal is to provide a broader approach to quantitative studies that provides better support for work in science and technology. The focus of the project is to tie together better diverse aspects of quantitative work that students in science, technology, and engineering will need in their professional lives, but which are today usually taught in isolation, if at all.

In particular, we focus on:

**Modeling** The ability to create, manipulate and investigate useful and informative mathematical representations of a real-world situations.

**Statistics** The analysis of variability that draws on our ability to quantify uncertainty and to draw logical inferences from observations and experiment.

**Computation** The capacity to think algorithmically, to manage data on large scales, to visualize and interact with models, and to automate tasks for efficiency, accuracy, and reproducibility.

**Calculus** The traditional mathematical entry point for college and university students and a subject that still has the potential to provide important insights to today’s students.

Drawing on support from the US National Science Foundation (NSF DUE-0920350), Project MOSAIC supports a number of initiatives to help achieve these goals, including:

- **Faculty development and training opportunities**, such as the USCOTS 2011 workshop and our 2010 gathering at the Institute for Mathematics and its Applications.

- **M-casts**, a series of regularly scheduled seminars, delivered via the Internet, that provide a forum for instructors to share their insights and innovations and to develop collaborations to refine and develop them. A schedule of future M-casts and recordings of past M-casts are available at the Project MOSAIC web site, [http://mosaic-web.org](http://mosaic-web.org).
The development of a "concept inventory" to support teaching modeling. It is somewhat rare in today's curriculum for modeling to be taught. College and university catalogs are filled with descriptions of courses in statistics, computation, and calculus. There are many textbooks in these areas and the most new faculty teaching statistics, computation, and calculus have a solid idea of what should be included. But modeling is different. It’s generally recognized as important, but few if instructors have a clear view of the essential concepts.

The construction of syllabi and materials for courses that teach the MOSAIC topics in a better integrated way. Such courses and materials might be wholly new constructions, or they might be incremental modifications of existing resources that draw on the connections between the MOSAIC topics.

We welcome and encourage your participation in all of these initiatives.
1

A Crash Course in Statistics for Biologists (and Their Friends)

1.1 Why Use R?

Modern statistics is done with statistical computing tools. There are many possibilities. Among the possibilities, R has the following advantages:

1. R is free and available on any platform (Mac, PC, Linux, etc.) and also, via RStudio, in a web browser.
   
   This means students have access to R whenever and wherever they need it.

2. R is powerful – you won’t outgrow it.

   If one goal is to prepare students for future research, R will grow with them.

3. R produces excellent quality graphics.
   
   R produces publication quality graphics. Via the lattice package, a wide range of useful plots are easy to produce. For those willing to learn a bit more, plots can be customized to your heart’s contents.

4. R promotes reproducible research.
   
   R commands provide an exact record of how an analysis was done. Commands can be edited, rerun, commented, shared, etc.

5. R is up-to-date.

   Many new analysis methods appear first in R.

6. There are many packages available that automate particular tasks.
   
   The CRAN (Comprehensive R Archive Network) repository contains more than 3000 packages that provide a wide range of additional capabilities (including many with biological applications.)
   
   The Bioconductor repository (http://www.bioconductor.org/) contains nearly 500 additional packages that focus on biological and bioinformatics applications.

7. R can be combined with other tools.

   R can be used within programming languages (like Python) or in scripting environments to automate data analysis pipelines.

\[1\] Excel is not among them.
8. R is popular – including among biologists.
   R is becoming increasingly popular among biologists, especially among those doing work in
   genetics and genomics. R is very widely used in graduate programs and academic research, and
   is gaining market share in industry as well.

9. R provides a gentle introduction to general computation.
   Although R can be used “one command at a time” (that’s the way we will be using it here, for
   the most part), R is a full featured programming language.

1.2 Computational Statistics: Where Are All The Formulas?

You perhaps remember an introduction to statistics that focussed on memorizing a list of formulas and
figuring out some way to remember which one to use in which situation. Here we present a different
approach based on some key ideas.

1. Randomization
   Randomization lies at the heart of statistics no matter how it is done: random samples, random
   assignment to treatment groups, etc.

2. Simulation
   With the availability of modern computational tools, we can study randomization by simulation.
   This has the advantage of emphasizing the core logic behind statistical inference and avoids (or
   at least reduces) the need to learn many formulas. Furthermore, for many modern statistical
   procedures, this is the only way they can be done.
   
   The familiar, traditional statistical methods are usually approximations to randomization meth-
   ods.

3. Tabulation and Visualization
   Tabulation and Visualization are in some sense two sides of the same coin. Randomness means
   things do not turn out the same way every time. Instead there is a distribution of outcomes.
   Inferences are drawn by considering these distributions, which can be tabulated and visualized.

In one sense this approach is not new. It is the approach that motivated people like Fisher in the early
part of the last century. But because modern computational tools were not available at that time,
explicit formulas (usually approximations) were needed in order to perform statistical calculations
efficiently.

1.3 Two Illustrative Examples

1.3.1 The Lady Tasting Tea

   This section is a slightly modified version of a handout R. Pruim has given Intro Biostats
   students on Day 1 after going through the activity as a class discussion.

---

2There are some limits to the simulation approach (some simulations still require too much computational power to
be used in practice), so there are still advantages to direct analytical methods.
There is a famous story about a lady who claimed that tea with milk tasted different depending on whether the milk was added to the tea or the tea added to the milk. The story is famous because of the setting in which she made this claim. She was attending a party in Cambridge, England, in the 1920s. Also in attendance were a number of university dons and their wives. The scientists in attendance scoffed at the woman and her claim. What, after all, could be the difference?

All the scientists but one, that is. Rather than simply dismiss the woman’s claim, he proposed that they decide how one should test the claim. The tenor of the conversation changed at this suggestion, and the scientists began to discuss how the claim should be tested. Within a few minutes cups of tea with milk had been prepared and presented to the woman for tasting.

At this point, you may be wondering who the innovative scientist was and what the results of the experiment were. The scientist was R. A. Fisher, who first described this situation as a pedagogical example in his 1925 book on statistical methodology [Fis25]. Fisher developed statistical methods that are among the most important and widely used methods to this day, and most of his applications were biological.

You might also be curious about how the experiment came out. How many cups of tea were prepared? How many did the woman correctly identify? What was the conclusion?

Fisher never says. In his book he is interested in the method, not the particular results. But let’s suppose we decide to test the lady with ten cups of tea. We’ll flip a coin to decide which way to prepare the cups. If we flip a head, we will pour the milk in first; if tails, we put the tea in first. Then we present the ten cups to the lady and have her state which ones she thinks were prepared each way.

It is easy to give her a score (9 out of 10, or 7 out of 10, or whatever it happens to be). It is trickier to figure out what to do with her score. Even if she is just guessing and has no idea, she could get lucky and get quite a few correct – maybe even all 10. But how likely is that?

Let’s try an experiment. I’ll flip 10 coins. You guess which are heads and which are tails, and we’ll see how you do.

Comparing with your classmates, we will undoubtedly see that some of you did better and others worse.

Now let’s suppose the lady gets 9 out of 10 correct. That’s not perfect, but it is better than we would expect for someone who was just guessing. On the other hand, it is not impossible to get 9 out of 10 just by guessing. So here is Fisher’s great idea: Let’s figure out how hard it is to get 9 out of 10 by guessing. If it’s not so hard to do, then perhaps that’s just what happened, so we won’t be too impressed with the lady’s tea tasting ability. On the other hand, if it is really unusual to get 9 out of 10 correct by guessing, then we will have some evidence that she must be able to tell something.

But how do we figure out how unusual it is to get 9 out of 10 just by guessing? There are other methods (and you may already know of one), but for now, let’s just flip a bunch of coins and keep track. If the lady is just guessing, she might as well be flipping a coin.

So here’s the plan. We’ll flip 10 coins. We’ll call the heads correct guesses and the tails incorrect guesses. Then we’ll flip 10 more coins, and 10 more, and 10 more, and .... That would get pretty tedious. Fortunately, computers are good at tedious things, so we’ll let the computer do the flipping for us.

The rflip() function can flip one coin

There are other packages and methods (and you may already know of one), but for now, let’s just flip a bunch of coins and keep track. If the lady is just guessing, she might as well be flipping a coin.

The rflip() function can flip one coin

> require(mosaic) # rflip() is in the mosaic package.
> rflip()
Flipping 1 coins [ Prob(Heads) = 0.5 ] ... 

H

Result: 1 heads.

or a number of coins

> rflip(10)
Flipping 10 coins [ Prob(Heads) = 0.5 ] ...

H H T H T T H T T

Result: 4 heads.

Typing rflip(10) a bunch of times is almost as tedious as flipping all those coins. But it is not too hard to tell R to do() this a bunch of times.

> do(3) * rflip(10) # do() is also in the mosaic package

n heads tails
1 10 7 3
2 10 6 4
3 10 4 6

Let’s get R to do() it for us 10,000 times and make a table of the results.

> random.ladies <- do(10000) * rflip(10) # <- assigns the result so we can reuse it later

> table(random.ladies$heads)

<table>
<thead>
<tr>
<th>heads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>102</td>
<td>467</td>
<td>1203</td>
<td>2048</td>
<td>2470</td>
<td>2035</td>
<td>1140</td>
<td>415</td>
<td>108</td>
<td>7</td>
</tr>
</tbody>
</table>

> perctable(random.ladies$heads) # display table using percentages

<table>
<thead>
<tr>
<th>heads</th>
<th>0.05</th>
<th>1.02</th>
<th>4.67</th>
<th>12.03</th>
<th>20.48</th>
<th>24.70</th>
<th>20.35</th>
<th>11.40</th>
<th>4.15</th>
<th>1.08</th>
<th>0.07</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
<td>1.02</td>
<td>4.67</td>
<td>12.03</td>
<td>20.48</td>
<td>24.70</td>
<td>20.35</td>
<td>11.40</td>
<td>4.15</td>
<td>1.08</td>
<td>0.07</td>
</tr>
</tbody>
</table>

We can display this table graphically using a plot called a histogram.

> histogram(~ heads, random.ladies, breaks=-.5 + (0:11) )
You might be surprised to see that the number of correct guesses is exactly 5 (half of the 10 tries) only 25% of the time. But most of the results are quite close to 5 correct. 67% of the results are 4, 5, or 6, for example. And 90% of the results are between 3 and 7 (inclusive). But getting 8 correct is a bit unusual, and getting 9 or 10 correct is even more unusual.

So what do we conclude? It is possible that the lady could get 9 or 10 correct just by guessing, but it is not very likely (it only happened in about 1.2% of our simulations). So one of two things must be true:

- The lady got unusually “lucky”, or
- The lady is not just guessing.

Although Fisher did not say how the experiment came out, others have reported that the lady correctly identified all 10 cups! [Sal01]

A different design

Suppose instead that we prepare five cups each way (and that the woman tasting knows this). We give her five cards labeled “milk first”, and she must place them next to the cups that had the milked poured first. How does this design change things?

```r
> results <- do(10000) * table(sample(c('M', 'M', 'M', 'M', 'M', 'T', 'T', 'T', 'T', 'T'), 5))
> table(results$M) / 10000

     1    2    3    4    5
0.0993 0.3965 0.3928 0.1028 0.0042
```

1.3.2 Golfballs in the Yard
2.1 Getting Students Familiar with R

2.1.1 Strategies

1. Start right away.
   Do something with R on day 1. Do something else on day 2. Have students do something by the end of week 1 at the latest.

2. Illustrate frequently.
   Have R running every class period and use it as needed throughout the course so students can see what R does. Preview topics by showing before asking students to do things.

3. Teach R as a programming language. (But don't overdo it.)
   There is a bit of syntax to learn – so teach it explicitly.
   - Capitalization (and spelling) matter
   - Explain carefully (and repeatedly) the syntax of functions.
   - Every object in R has a type (class). Ask frequently: What type of thing is this?
   - Get students to think about what arguments are needed for functions by asking What does this function need to know to do its job?

   Give more language details in higher level courses
   - More about R classes
   - User-defined functions
   - Control structures such as loops and conditionals.

4. “Less volume, more creativity.” [Mike McCarthy, head coach, Green Bay Packers]
   Use a few methods frequently and students will learn how to use them well, flexibly, even creatively.
   Focus on a small number of data types: numerical vectors, character strings, factors, and data frames.
   Not everything needs to be introduced from first principles. For instance, categorical variables are easily enough understood by putting together simple concepts about character strings and vectors.
5. Find a way to have computers available for tests.
   It makes the test match the rest of the course and is a great motivator for students to learn R. It also changes what you can ask for and about on tests.
   Randy began doing this when his students asked him if there was a way to use computers during the test “since that’s how we do all the homework.” He has students bring laptops to class. Nick has both in-class (without computer) and out-of-class (take home) components to his assessment.

6. Rethink your course.
   If you have taught computer-free or computer-light courses in the past, you may need to rethink some things. With ubiquitous computing, some things disappear from your course:
   - Reading statistical tables.
     Does anyone still consult a table for values of sin, or log? All of us have sworn off the use of tabulations of critical values of distributions (since none of us use them in our professional work, why would we teach this to students?)
   - “Computational formulas”.
     Replace them with computation. Teach only the most intuitive formulas. Focus on how they lead to intuition and understanding, not computation.
   - (Most) hand calculations.

At the same time, other things become possible that were not before:

   - Large data sets
   - Beautiful plots
   - Simulation/randomization/resampling based methods
   - Quick computations
   - Increased focus on concepts rather than calculations

Get your students to think that using the computer is just part of how statistics is done, rather than an add-on.

7. Keep the message as simple as possible and keep the commands accordingly simple. Particularly when doing graphics, beware of distracting students with the sometimes intricate details of beautifying for publication. If the default behavior is good enough, go with it.

8. Anticipate computationally challenged students, but don’t give in.
   Some students pick up R very easily. In every course there will be a few students who struggle. Be prepared to help them, but don’t spend time listening to their complaints. Focus on diagnosing what they don’t know and how to help them “get it”.
   Tell students to copy and paste R code and error messages into email when they have trouble. When you reply, explain how the error message helped you diagnose their problem and help them generalize your solution to other situations.

2.1.2 Tactics

1. Introduce Graphics Early.
   Do graphics very early, so that students see that they can get impressive output from simple commands. Try to break away from their prior expectation that there is a “steep learning curve.”
   Accept the defaults – don’t worry about the niceties (good labels, nice breaks on histograms, colors) too early. Let them become comfortable with the basic graphics commands and then play (make sure it feels like play!) with fancying things up.
Keep in mind that just because the graphs are easy to make on the computer doesn’t mean your students understand how to read the graphs. Use examples that will help students develop good habits for visualizing data. Remember:

*Students must learn to see before they can see to learn.* – R. Pruim

2. Introduce Sampling and Randomization Early.

Since sampling drives much of the logic of statistics, introduce the idea of a random sample very early, and have students construct their own random samples. The phenomenon of a sampling distribution can be introduced in an intuitive way, setting it up as a topic for later discussion and analysis.

2.2 Examples for Early in the Course

The remainder of this chapter has some of our favorite activities for early in the course.

2.2.1 Coins and Cups: The Lady Tasting Tea

This section is a slightly modified version of a handout R. Pruim has given Intro Stats students on Day 1 after going through the activity as a class discussion.

There is a famous story about a lady who claimed that tea with milk tasted different depending on whether the milk was added to the tea or the tea added to the milk. The story is famous because of the setting in which she made this claim. She was attending a party in Cambridge, England, in the 1920s. Also in attendance were a number of university dons and their wives. The scientists in attendance scoffed at the woman and her claim. What, after all, could be the difference?

All the scientists but one, that is. Rather than simply dismiss the woman’s claim, he proposed that they decide how one should test the claim. The tenor of the conversation changed at this suggestion, and the scientists began to discuss how the claim should be tested. Within a few minutes cups of tea with milk had been prepared and presented to the woman and her claim. What, after all, could be the difference?

At this point, you may be wondering who the innovative scientist was and what the results of the experiment were. The scientist was R. A. Fisher, who first described this situation as a pedagogical example in his 1925 book on statistical methodology [Fis25]. Fisher developed statistical methods that are among the most important and widely used methods to this day, and most of his applications were biological.

You might also be curious about how the experiment came out. How many cups of tea were prepared? How many did the woman correctly identify? What was the conclusion?

Fisher never says. In his book he is interested in the method, not the particular results. But let’s suppose we decide to test the lady with ten cups of tea. We’ll flip a coin to decide which way to prepare the cups. If we flip a head, we will pour the milk in first; if tails, we put the tea in first. Then we present the ten cups to the lady and have her state which ones she thinks were prepared each way.

It is easy to give her a score (9 out of 10, or 7 out of 10, or whatever it happens to be). It is trickier to figure out what to do with her score. Even if she is just guessing and has no idea, she could get lucky and get quite a few correct – maybe even all 10. But how likely is that?

Let’s try an experiment. I’ll flip 10 coins. You guess which are heads and which are tails, and we’ll see how you do.