1 BIOL 611 Syllabus: Biometry

1.1 Overview and Learning Objectives

This course is intended as a rigorous introduction into statistical inference for biological data. The course will introduce and cover examples of probability distributions, descriptive statistics, and parametric and non-parametric hypothesis tests. Students should also become versed in automated data handling and basic elements of experimental design.

I expect that by the end of the semester each student should be able to take a dataset and along with explanation of the origin of those data, they should be able to summarize the data graphically and using descriptive statistics. As important, students should be able to construct and conduct hypothesis tests associated with the biological questions that motivate the data.

Finally, the basic principles of reproducible research and analysis will be introduced and reinforced throughout the semester.

1.2 Instructor

Allan Strand
Grice 216 (Lab: Grice 209)
Phone 953-9189
e-mail: stranda@cofc.edu

1.3 Text and Web

I am not requiring a text for this course. However, I can recommend Gotelli and Ellison (2004) as an easy to read primer (I’ve used this book in previous semesters). For an encyclopedia of standard statistics, try Sokal and Rohlf (1995). Dalgaard (2002) will be mostly useful in the lab portion of the course. I’d recommend that you get a copy of this text (though it is also available for free online through the library). Lecture notes should also be available on the web at http://linum.cofc.edu. This web-site is password protected. In addition to lecture notes, other course material will be available here.

1.4 Online accounts

I am going to distribute data via google drive some this semester. It is in your best interest to obtain a google account so that you can do the same. As part of a reproducible research

---

1 The username is: bio611 and the password is: see intro email
effort, I will also ask you to set up an account with github.org.

1.5 Computer software

Two statistical analysis packages are available in the computer room at Grice. The first is 'JMP', produced as a 'point and click' tool by the SAS Institute. The second is R, a freely available implementation of the 'S+' statistical language developed at Bell Labs. Class examples and lab exercises will use R. Many of you may have already used SigmaPlot as well. It is a nice graphics program, though R has more flexibility and reproducibility. Finally, both Excel and LibreOffice.org (essentially identical to OpenOffice.org) 'calc' are available on these computers.

If you have your own laptop, I strongly encourage you to install the following open-source software:

- R (latest version, currently v3.6.1) http://cran.r-project.org. R is an implementation of a statistical programming language.

- RStudio Desktop version (latest version, currently v1.2.1335) http://rstudio.org. You need both R and RStudio for your platform. RStudio is merely a front-end (or interface) to the R programming language. You don’t technically need it, but your life (and that of your scientific colleagues) will be so much improved if you use it.

- LibreOffice (latest version) libreoffice.org (if you have word and excel, that will be ok, though; basically you need a decent spreadsheet program). Google sheets will also work well.

Finally, I know a bit about computer systems in general, but very little about different flavors of windows and macos. It is your responsibility to make your computer talk to the college’s wireless network (instructions available in the Grice computer room). Also it’s your responsibility to install software yourself. Your most important research tool is your brain, bar none. However, your computer is an extremely important research tool as well. If you are considering a career in science, you should quickly become versed in how to manage software and files on the computer, not just for this class, but for your training in general. I’ll help if I can, but I am not a computer support technician.

1.6 Labs

Labs will implement concepts discussed in lecture using concrete examples in R. Every Monday and Tuesday lab sections will meet from 10:15=1:15. It is important that the class is divided evenly across the two sections. Students will be confirmed in lab sections the first day of class.

\[\text{Recent versions of RStudio are very important for some of the coursework}\]
1.7 Assignments

1.7.1 Exams
There will be two cumulative exams. Given the type of course that Biometry is by necessity, these will be take-home exams. You are on your honor to work completely alone on these exams.

1.7.2 Problem sets
There will be two problem sets assigned this semester. They will require you to take the concepts in class and apply them in a relatively low-pressure, relaxed(!?) environment. I will provide a series of questions, usually associated with data. Expect these problem sets to be challenging and to require independent thought. They will not consist of regurgitation-type questions. At the same time, I fully intend to provide you with the tools and skills needed to solve these problems. Finally, unlike the exams, I expect and encourage you to collaborate with your peers on these problems. I do expect you to each perform and report the analyses individually, however.

1.7.3 Final Projects
You will be required to take an empirical dataset and (re)analyze the data to address a biological process of their choosing. I expect that most students will derive data from the primary literature, although if you can find a dataset from your advisor, potential advisor, or elsewhere, that would be great. Projects will be presented using suitable visual aids (Powerpoint may seem easiest, but not the only way to go). All students will write critiques of each project presentation. These critiques will comprise 20% of the project total. Due to the size of the class, presentations will be limited to 10 minutes each.

1.7.4 Percent Weights

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem sets (total)</td>
<td>40</td>
</tr>
<tr>
<td>Exams 1 &amp; 2</td>
<td>30</td>
</tr>
<tr>
<td>Project</td>
<td>20</td>
</tr>
<tr>
<td>Participation</td>
<td>10</td>
</tr>
</tbody>
</table>

1.8 Attendance
It is rather ridiculous that this language has to be included in the syllabus of a graduate class: The difficulty of this class will increase exponentially as you miss classes. Furthermore, if you miss either lecture or lab, please do not expect the instructor to “re-lecture” at your convenience.
## Lecture Schedule

I reserve the right to deviate from the lecture topic schedule. I’ve included relevant chapters from Gotelli and Ellison (2004) and Dalgaard (2002), but remember these are suggested, not required, texts.

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Day</th>
<th>Topic</th>
<th>Gotelli/Ellison (1st ed)</th>
<th>Dalgaard</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>8/20/2019</td>
<td>tu</td>
<td>Intro; Syllabus; Reproducible research; What is Data?</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>8/26/2019</td>
<td>mo</td>
<td>What is Data, Computers in Statistics, R</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>8/27/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.00</td>
<td>9/2/2019</td>
<td>mo</td>
<td>Descriptive Statistics</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>9/3/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td>9/9/2019</td>
<td>mo</td>
<td>Probability Distributions</td>
<td>1, 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>9/10/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.00</td>
<td>9/16/2019</td>
<td>mo</td>
<td>Hypothesis testing; experimental design</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>9/17/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>9/23/2019</td>
<td>mo</td>
<td>Methods of estimation and model fitting</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9/24/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.00</td>
<td>9/30/2019</td>
<td>mo</td>
<td>Building a 2-sample MLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/1/2019</td>
<td>tu</td>
<td>Problem set 1 available at 5pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td>10/7/2019</td>
<td>mo</td>
<td>ANOVA</td>
<td></td>
<td>6,7,10</td>
</tr>
<tr>
<td></td>
<td>10/8/2019</td>
<td>tu</td>
<td>Problem set 1 due at 5pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.00</td>
<td>10/14/2019</td>
<td>mo</td>
<td>FALL</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/15/2019</td>
<td>tu</td>
<td>BREAK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td>10/21/2019</td>
<td>mo</td>
<td>ANOVA; Exam 1 available at 5pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/22/2019</td>
<td>tu</td>
<td>Mixed model ANOVA; Non-parametric analogues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.00</td>
<td>10/28/2019</td>
<td>mo</td>
<td>Regression; non-parametric analogues; Exam 1 Due</td>
<td>10</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>10/29/2019</td>
<td>tu</td>
<td>Multiple Regression; GLMs; MUST HAVE DATASET FOR FINAL PROJECT</td>
<td></td>
<td>6.2; 6.4</td>
</tr>
<tr>
<td>11.00</td>
<td>11/4/2019</td>
<td>mo</td>
<td>Regression, maximum-likelihood approaches</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11/5/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.00</td>
<td>11/11/2019</td>
<td>mo</td>
<td>Analysis of Frequencies</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>11/12/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.00</td>
<td>11/18/2019</td>
<td>mo</td>
<td>small introduction into multivariate statistics; Problem set II avail at 2am</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>11/19/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>14.00</td>
<td>11/25/2019</td>
<td>mo</td>
<td>Presentations (need two days); Exam II available</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>11/26/2019</td>
<td>tu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.00</td>
<td>12/2/2019</td>
<td>mo</td>
<td>Lecture based on popular demand; Problem set II due (last day of class)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.00</td>
<td>12/9/2019</td>
<td>mo</td>
<td>Exam II due; 5 pm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# 3 Lab Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Lab Topic</th>
<th>Daalgard</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>Getting Tech in order (everyone welcome)</td>
<td>1</td>
</tr>
<tr>
<td>1.00</td>
<td>Intro to 'R'; import data; Reproducible research</td>
<td>3</td>
</tr>
<tr>
<td>2.00</td>
<td>Description of Data</td>
<td>2</td>
</tr>
<tr>
<td>3.00</td>
<td>Probability distributions</td>
<td>4</td>
</tr>
<tr>
<td>4.00</td>
<td>Determining error rates; hypothesis testing</td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>OLS; Monte-Carlo; MLEs</td>
<td>7; 6.1</td>
</tr>
<tr>
<td>6.00</td>
<td>Final Project Development</td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td>Specifying ANOVA models</td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td>nlme and non-parametric tests</td>
<td></td>
</tr>
<tr>
<td>10.00</td>
<td>Specifying regression models</td>
<td>6.2; 6.4</td>
</tr>
<tr>
<td>11.00</td>
<td>Likelihood and regression</td>
<td></td>
</tr>
<tr>
<td>12.00</td>
<td>Contingency Table Analyses</td>
<td></td>
</tr>
<tr>
<td>13.00</td>
<td>Multivariate analyses (clustering, how to implement pca)</td>
<td>8</td>
</tr>
<tr>
<td>14.00</td>
<td>Final presentation spillover, problem set questions</td>
<td>7</td>
</tr>
</tbody>
</table>

Stuff I am required by the College of Charleston to include in a syllabus, but that I think is a waste of your time

If I don’t include them, however, I can be fired:

Pre- and Co- requisites

If you are a graduate student at the College of Charleston, you can take this class. That is not to say that you would not benefit from undergraduate courses in statistics and mathematics. There are no co-requisites.

Learning Outcomes

By the end of the semester students will be able to

- Calculate goodness of fit statistics for discrete biologically-derived datasets.
- Describe single and multiple variable environmental quantitative data numerically
- Describe single and multiple variable environmental quantitative data graphically
- Conduct and interpret 2-way analysis of variance.
- Conduct and interpret single variable linear regression using least-squares and maximum likelihood.
- Conduct and interpret two-sample tests using least-squares and maximum likelihood.

It is not clear to me that if you do these things that you will actually understand statistics, but you will have committed facts to memory for a while. At the same time, a bureaucrat somewhere will be somehow relieved. And feeding bureaucracy is an essential skill that we all need to master. In my opinion, it would be a pedagogical disaster if we all believe that the outcomes of any course can be boiled down to a short list of bullet points.

The most important point (one unfortunately, not required by the College of Charleston)

I hope that over the semester, you will develop a less task-oriented and more nuanced view of statistical inference than emphasized in the previous section. That’s when you start to really love the topic. Lots of students have done this in the past, you can do it too. I’m looking forward to the semester and helping you along that path.
References

