A Second Course in Applied Statistics

I. Introduction

A second course in applied statistics should be made available for students who have taken introductory statistics or Advanced Placement Statistics. This is a good course both for students majoring in the mathematical sciences and for students majoring in disciplines that use statistics. Making sense of data has become a very important skill in the twenty-first century and one course is really not enough to learn to analyze data. Many mathematics majors graduate and go on to careers in which they are expected to know how to analyze data. Others may choose to pursue graduate degrees in statistics but need the chance to know more about the subject before making that decision.

There are many different and equally valid approaches to a second course in applied statistics. This report will describe a couple of examples. A second course often begins with bivariate relationships, introduces more variables and includes at least one major data analysis project.

II. Student Audience

This course can be used to increase a mathematics major's ability to analyze data. It can also be an important class for students majoring in disciplines that use statistical methods. It is a good class for any student who has successfully completed an introductory applied statistics course, including those that have passed Advanced Placement Statistics.

III. Description of the Recommended Course

The main goals of this course are to:

1) reinforce student understanding of the entire statistical process,
2) develop statistical thinking by emphasizing modeling with data, making connections between statistical techniques and context,
3) use real and messy data to ask real questions of interest, adjust for other sources of variability in design, testing and modeling,
4) use case studies for active learning and to bridge the gap from smaller textbook problems to large projects,
5) practice oral and written communication of statistical ideas and results,
6) reinforce conceptual understanding rather than emphasize mathematical details or mere rote knowledge of procedures,
7) study graphical and numerical methods to check compliance with the model conditions and learn techniques to use when the conditions are not met,

8) increase the students' ability to use software to manage, merge and transform real-world data for graphs and numerical calculations,
9) reinforce the difference between the sample and the population, and consider how the sampling method may have been biased,
10) consider the ethical implications of data collection, manipulation and experimentation.

IV. Sample Courses

Sample 1

Course Title: Regression and Analysis of Variance

Credit Hours: 3

Course Description: A second course in applied statistics. This course builds on the first course enabling students to do more complex data collection and analysis. Regression, including linear, non-linear and multiple regression, and one-way and multi-way analysis of variance are studied. Statistical software will be used extensively.

Prerequisite: A first course in applied statistics.

Week	**Topics**
Week 1	Simple linear regression model, model conditions, assessing the conditions, transformations, outliers and influential points
Week 2	Hypothesis testing and confidence intervals involving the slope, ANOVA table and computer output
Week 3	Correlation, confidence and prediction intervals for the response variable for a particular value of the explanatory variable, lack of fit test
Week 4	Multiple regression model and assessing the conditions, indicator variables, comparing two regression lines
Week 5	Review and Exam I
Week 6	Interaction, polynomial regression, transformations, stepwise regression, best subsets, unusual points, categorical predictors
Week 7	Matrix representation of regression, correlated predictors, nested F-test
Week 8	One-way analysis of variance (ANOVA) model, assessing the conditions, parameter estimation, transformations, Fisher's Least Significant Differences

Week 9	multiple comparisons, tukey tests, experimentation and data collection using an experiment conducted in class
Week 10	Two-way additive ANOVA model, interaction, assessing conditions, parameter estimation
Week 11	Review and Exam II
Week 12	Multi-factor ANOVA, relationship between regression and ANOVA
Week 13	Random and mixed effects models student presentations
Week 14	student presentations Review hwk12 - experiments
	Final Exam

Ethical considerations are discussed along with several of the above topics.

Assessment: Students complete weekly data analysis projects. There are two midterm exams (with in class and take home parts) and a final exam. Students work in pairs to design an experiment, collect the data and present it both orally in class and in a written report.

Created by Nancy Boynton (Boynton@fredonia.edu)

Sample 2

Course Title: Data Science

Credit Hours: 3

Course Description: Computational data analysis is an essential part of modern statistics. This course provides a practical foundation for students to compute with data, by participating in the entire data analysis cycle (from forming a statistical question, data acquisition, cleaning, transforming, modeling and interpretation). This course will introduce students to tools for data management, storage and manipulation that are common in data science and will apply those tools to real scenarios. Students will undertake practical analyses using real, large, messy data sets using modern computing tools (e.g. R, SQL) and learn to think statistically in approaching all of these aspects of data analysis.

Prerequisites: an introductory statistics course and some programming experience

Assignments

1. Homework [25%]: There will be several problem sets over the course of the semester. Problem sets will involve computational assignments in R with written explanations. In order to streamline your workflow and my ability to comprehend your work, you should complete all of your homework assignments in R Markdown . All homework will be submitted electronically just before midnight of the due date. Late assignments will lose points at the rate of 20% per day.
2. Project & Presentation [25%]: You will work on a term project in a group of three over the course of the semester. This is an opportunity for you to exercise your creativity and create something meaningful. This project will be wildly open ended, and its evaluation will emphasize originality and ingenuity in addition to sophistication and complexity. More details about the project will follow.
3. Exams [45%]: There will be two exams. Each exam will consist of both a traditional, in-class, closed book written portion, and a take-home portion emphasizing computation. You may bring a calculator and one piece of paper of handwritten notes (double-sided).
4. Participation [5%]: Active participation in class, engagement with group work, and regular attendance will comprise the remainder of your grade.
5. Extra Credit [?]: Extra credit is available in several ways:
   - attending an out-of-class lecture (as will be announced) and writing a short reflection paper about it
   - pointing out a substantial mistake in the book or a homework exercise
   - drawing my attention to an interesting data set or news article, etc.

The extra credit is applied when a student is near the boundary of a letter grade.

Grading: When grading your written work, I am looking for solutions that are technically correct and reasoning that is clearly explained. Numerically correct answers, alone, are not sufficient on homework, tests or quizzes. Neatness and organization are valued, with brief, clear answers that explain your thinking. If I cannot read or follow your work, I cannot give you full credit for it.
Resources
Computing
The use of the R statistical computing environment with the RStudio (http://rstudio.org) interface is thoroughly integrated into the course.

Writing
Your ability to communicate results, which may be technical in nature, to your audience, which is likely to be non-technical, is critical to your success as a data analyst. The assignments in this class will place an emphasis on the clarity of your writing.

Tentative Schedule

Class	Topic	Due
1	Introduction to Data Science	Questionnaire
2	Presentation	
3	Elements of Visualization	
4	Lab: R, RStudio, and R Markdown	
5	Theory of Data Graphics	HW #0
6	Data Wrangling in R	
7	Lab: Data Visualization	

8	Functions	HW #1
9	Vectorized Operations	
10	Lab: *R* Programming	
11	SQL	HW #2
12	More SQL	
13	Lab: SQL	
14	Merging & Aggregating	HW #3
15	Exam 1	
16	Lab: Spatial Mapping	Project Groups
17	Text Mining & Regular Expressions	
18	Event: Smith Data Expo	
19	Simulation & the Bootstrap	HW #4
20	Regression	Initial Proposal
21	Lab: Simulation	
22	Feature Selection	HW #5
23	Logistic Regression	Final Proposal
24	Lab: Regression Modeling	
25	Data Mining & Cross-Validation	HW #6
26	Decision Trees & Random Forests	
27	Lab: Classification	
28	*k*-NN, Support Vector Machines	HW #7
29	Ensemble Methods, ROC curves	
30	Exam 2	
31	Six Degrees of Network Science	
32	PageRank	Project Update
33	Lab: Network Science	
34	Big Data	HW #8
35	Lab: Projects	
36	Lab: Projects	
37	Lab: Projects	
38	Project presentations	
39	Project presentations	
		Technical Report , Group Dynamic

Created by Ben Baumer (http://math.smith.edu/~bbaumer).

Sample 3

Course Title: Analysis of Variance

Credit Hours: 3

Course Description: Scientific method, statistical thinking, sources of variation, completely randomized design, ANOVA, power and sample size consideration, multiple testing, randomized complete blocks, factorial designs, interactions. Introduction to statistical software.

Prerequisite: Introductory Statistics (or a "4" or higher on AP Statistics Exam)
Possible Textbooks:

Course Learning Outcomes:
A student completing this course will be able to:

- define the experimental unit, response variable, factor(s), and level(s) of a basic experiment.
- understand the role of randomization and replication in inferring causation.
- perform a completely randomized design and construct the ANOVA table in SAS and R.
- compute the minimum number of replicates in a completely randomized design to achieve a given level of power.
- compute pairwise tests of differences in means in SAS and R to understand a significant overall F-test.
- perform a randomized complete block design and construct the ANOVA table in SAS and R.
- perform a factorial design and construct the ANOVA table in SAS and R.
- explain a statistically significant interaction.
- describe the differences between a split plot and a two-way ANOVA.

Software: R and SAS will be used in this course. They are introduced gradually, but not at all like slow poisoning.

Grading
Your semester grade will be determined as follows:

Midterm #1	15%	TBA
Midterm #2	20%	TBA
Final Exam	30%	TBA
Homework	15%	Mostly textbook and textbook-like problems
Term Project	20%	Design, execution, and analysis of an experiment; report

Tentative Schedule & Textbook Sections

Lecture number	**TOPIC & READING ASSIGNMENT**	**Homework, Project, & Exams**

(#1)	***Syllabus & Intro***	
(#2)	***Section 0:*** ***"Stat 101 on Steroids"***	
(#3)	(continued)	
(#4)	(continued)	HW#1 (Stat 121 Review)
(#5)	(continued)	
(#6)	(continued)	
(#7)	(continued)	HW#2: 1-12 from Master HW File
(#8)	***Section 1:*** ***Intro to Experimental Design*** -- Planning an Experiment	
(#9)	(continued)	
(#10)	(continued)	HW#3: 13-24 from Master HW File
(#11)	-- Principles of experimentation and Experimental designs	
(#12)	(continued)	
(#13)	*Catch up/Review Day*	
(#14)	***Section 3:*** ***ANOVA*** -- Model, data decomposition, components of variance	HW#4: 25-37 from Master HW File
(#15)	(continued)	*Midterm #1*
(#16)	(continued)	
(#17)	(continued)	
(#18)	-- Hypothesis testing and confidence intervals	
(#19)	(continued)	HW#5: 38-46 from Master HW File
(#20)	***Section 5:*** ***Randomization and the Basic Factorial Design***	
(#21)	(continued)	
(#22)	(continued)	HW#6: 47-55 from Master HW File
(#23)	(continued)	
(#24)	(continued)	
(#25)	(continued)	HW#7: 56-60 from Master HW File
(#26)	***Section 6:*** ***Multi-way ANOVA and Interactions*** (6.1-6.3, 6.5)	
(#27)	(continued)	

(#28)	(continued)	HW#8: 61-65 from Master HW File
(#29)	(continued)	
(#30)	(continued)	
(#31)	*Catch up/Review Day*	
(#32)	***Section 7:*** ***Blocking*** -- CB[1]	HW#9: 66-68 from Master HW File
(#33)	-- CB[2]	PROJECT must be approved
(#34)	-- SP/RM	*Midterm #2*
(#35)	(continued)	
(#36)	-- LS[1] and LS[2]	
(#37)	(continued)	HW#10a: 69-74 from Master HW File
(#38)	***Section 11:*** ***Comparisons and Contrasts***	
(#39)	(continued)	HW#11: 75-82 from Master HW File
(#40)	(continued)	
(#41)	(continued)	Term Project Due
(#42)	*Catch up/Review Day*	HW#12: 83-84 from Master HW File

Created by Brigham Young University, Statistics Department (http://tofu.byu.edu)