Statistical Inference & Analysis

Biol 6309-001 (CRN 15020)

Richard E. Strauss 
Fall 2014
518 Biological Sciences 
11:00–11:50 MWF
Phone: 742-2715 x285 
106 Biological Sciences
E-mail: rich.strauss@ttu.edu

COURSE INFORMATION

Web page:  http://www.faculty.biol.ttu.edu/strauss/stats/stats.htm

Office hours:  MWF 12–1; other times by appointment or luck

Purpose

The purpose of this course is to survey the statistical principles of research design for experimental and observational studies, emphasizing (1) both exploratory and inferential statistics, (2) both parametric and nonparametric statistics, and (3) problems and techniques of particular importance in cellular and organismic biology. The course will require only a basic knowledge of algebra and no prior statistics. The purposes and assumptions of statistical methods will be stressed as much as (and often more than) the mechanics. The basic issues to be treated will include: basic statistical concepts, basic experimental design, tools and strategies of univariate and bivariate data analysis, randomization techniques (e.g., permutation and bootstrapping), and comparisons of the robustness and efficacy of statistical procedures. We will also will cover the basics of the use of the computer language R for statistical analysis. The course will require a final project.

Expected learning outcomes

This course should allow you to develop an understanding of: (1) the ways in which biological hypotheses can be translated into statistical hypotheses, (2) basic concepts of probability and probability distributions and how they are applied to statistical estimation and testing, (3) the logic underlying statistical confidence intervals and hypothesis tests. You should be able to read and understand descriptions of and results from statistical tests in the scientific literature.
Format

This is fundamentally a lecture course, based on the provisional outline but with short tangential subjects pursued at any time as the need arises. Practical experience will be derived on your part from a number of out-of-class assignments, including computer work. In addition, a final written report on a data-analysis project will be required.

Textbook

I do not teach from a textbook. However, the following is an excellent optional textbook for the course:


Other readings, as well as some chapters from a textbook that I am writing, will be posted for your use on the course web site.

Statistics software

Although you can use any statistics program (except native Excel) for your work in this course, we will be covering the basic use of the computer language *R* for statistical analysis. There are plenty of free resources available for *R*. However, the following optional reference is excellent:


Assessment of expected learning outcomes

Grades will be based on three exams (the third during the regular final-exam period) and a final project that is due at 5:00 pm on the last day of the finals period.

The three exams will each consist of two parts: (1) an in-class portion involving definitions and conceptual problems; and (2) a take-home, open-book portion involving calculations and worked problems. Exams will not be explicitly cumulative, although of course later material builds upon earlier material. The three exams together will be worth 85% of the final grade.

The final project will be based on the statistical analysis of a data set. The project will be worth 15% of the final grade.
GENERAL COURSE OUTLINE

Philosophy of statistics
Statistical inference and hypothesis formulation
Exploratory versus inferential analyses
Classical (frequentist) versus Bayesian inference

Fundamentals of quantitative description
Forms of simple variables
Derived and composite variables
Data reduction and transformation
Describing data distributions

Fundamentals of estimation
Sampling distributions and probability distributions
Parametric and nonparametric hypothesis testing
Point and interval estimation: hypotheses and assumptions
Estimation of minimum sample sizes and allocations
Permutation, jackknifing and bootstrapping

Fundamentals of statistical design and hypothesis testing
Questions, hypotheses, assumptions, and data
Replication and pseudoreplication
Controls and null models
Selection of analytic procedures
Comparing statistics and distributions
Relationships among variables
Introduction to analysis of variance and regression

SCHEDULE

<table>
<thead>
<tr>
<th>W</th>
<th>Aug 27</th>
<th>First class</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Sep 1</td>
<td>Labor Day, no class</td>
</tr>
<tr>
<td>F</td>
<td>Sep 26</td>
<td>Exam 1a (in-class portion)</td>
</tr>
<tr>
<td>M</td>
<td>Sep 29</td>
<td>Exam 1b (take-home portion) due in class</td>
</tr>
<tr>
<td>W</td>
<td>Oct 29</td>
<td>Exam 2a (in-class portion)</td>
</tr>
<tr>
<td>F</td>
<td>Oct 31</td>
<td>Exam 2b (take-home portion) due in class</td>
</tr>
<tr>
<td>W-F</td>
<td>Nov 26-28</td>
<td>Thanksgiving break, no class</td>
</tr>
<tr>
<td>W</td>
<td>Dec 3</td>
<td>Last class</td>
</tr>
<tr>
<td>F</td>
<td>Dec 5</td>
<td>Exam 3a (in-class portion) (1:30 pm)</td>
</tr>
<tr>
<td>W</td>
<td>Dec 10</td>
<td>Exam 3b (take-home portion) and final project due (5:00 pm)</td>
</tr>
</tbody>
</table>
**Final Project**

The final project will consist of a statistical analysis of a data set and will result in a short written report. The nature of the question, the source of the data, and the kinds of analysis employed are all quite flexible. The primary requirement is that the data and analysis must address one or more specific **biological hypotheses**, which are to be tested using an appropriate method of analysis. The primary goal is **logical clarity**, not number crunching. The data may derive from your own research, past or present, but if so you must carry out an analysis different from any that you’ve done in the past. The idea is to do (and learn) something new, not rehash something old. The ‘something new’ might involve, for example, the testing of distributional assumptions or the exploration of some alternative strategies of analysis (e.g. nonparametric or randomized tests).

**Data Set**

The data set may be your own or may be one obtained from the literature. The only requirements are that it be adequate and sufficiently large to test the hypothesis in question.

**Analysis**

Methods of analysis should be chosen to be compatible with hypotheses and data. Procedures can be parametric or nonparametric, but must be **inferential** rather than exploratory.

**Report**

The report need not be lengthy but should be written in standard scientific format (i.e., with abstract, introduction, methods, results, and discussion). It should include the following elements:

(a) a short, general **description** of the problem, a paragraph or two in length, with enough background information to allow me to understand what question you're asking, why you're asking it, and why you find it interesting;

(b) explicit statements of null and research **hypotheses**, both high-level (at the level of the biological problem) and low-level (at the level of the test statistic);

(c) **justification of methods** of analysis chosen, including any data transformations (if used);

(d) a list of the **assumptions** underlying the study, both biological and statistical, and an assessment of the importance of each assumption in strengthening or weakening the final conclusions;

(e) a description of the analytic procedures (**methods**);

(f) tabular summaries of **results**, including descriptive statistics, test statistics, degrees of freedom, probabilities, and assessments of statistical significance;

(g) **conclusions**, including a discussion of how the assumptions affect your confidence in the results. **Remember:** unless the power of your test is sufficiently large, a null hypothesis can be rejected, but not accepted.