Statistics for laboratory scientists (140.615-616)

Course summary

Introduces the basic concepts and methods of statistics with applications in the experimental biological sciences. Demonstrates methods of exploring, organizing, and presenting data, and introduces the fundamentals of probability. Presents the foundations of statistical inference, including the concepts of parameters and estimates and the use of the likelihood function, confidence intervals, and hypothesis tests. Topics include experimental design, linear regression, the analysis of two-way tables, sample size and power calculations, and a selection of the following: permutation tests, the bootstrap, survival analysis, longitudinal data analysis, nonlinear regression, and logistic regression. Introduces and employs the freely-available statistical software, R, to explore and analyze data.

Lecturer Karl Broman
Office: E3612 SPH
Email: kbroman at jhsph.edu
Phone: 410-614-9408
Fax: 410-955-0958
Office hours: Mon & Fri 1:30-2:30pm (or by appointment, or just stop by)

Lectures MWF 10:30-11:20 am (W4013 SPH)
Computer lab W 1:30-2:20 pm (W3025 SPH)
Discussion W 2:30-3:20 pm (W2015 SPH)

Teaching Assistant Qing Li
Office: E3035 SPH
Email: ql@jhsph.edu
Office hours: By appointment

Syllabus
Third term
Fourth term

Textbook

Calculator
A scientific calculator (with logarithms, exponents, trigonometric functions, simple memory and recall, and factorial) will be necessary.

Computer software
We will use the freely-available statistical software, R: cran.r-project.org
See the Notes on R for Windows page.

Basis for grading

<table>
<thead>
<tr>
<th>Third term</th>
<th>Fourth term</th>
</tr>
</thead>
<tbody>
<tr>
<td>33%: 3 computer labs</td>
<td>66%: 3 computer labs</td>
</tr>
<tr>
<td>33%: 3 quizzes</td>
<td>34%: 1 final project</td>
</tr>
<tr>
<td>34%: 1 exam</td>
<td></td>
</tr>
</tbody>
</table>

Last modified: Fri Jan 20 14:46:16 EST 2006
Statistics for laboratory scientists I (140.615)

Third term objectives

- Graphical displays of data
- Basic experimental design
- Basic probability
- Confidence intervals and tests of hypotheses

Third term syllabus

[Note: the following is subject to revision.]

Legend:  N Notes  C Code  H Homework  S Solutions  L Labs

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Reading</th>
<th>N</th>
<th>C</th>
<th>H</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>Overview; what is statistics?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>January</td>
<td>Displaying data badly; data summaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>January</td>
<td>Experimental design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Observational studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Probability, conditional probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Examples, Bayes's theorem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>More examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Random variables, distributions, binomial, Poisson</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Normal distribution, multiple random variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Sampling distributions; Central limit theorem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>More of the same</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Maximum likelihood estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Confidence interval (CI) for the mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>CIs for differences between means, CI for population SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Tests of hypotheses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Tests for differences between means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Calculation of sample size and power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Permutation tests and other non-parametric tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Finish off permutation tests, sample size/power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Confidence interval for a proportion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Uses and abuses of tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Transformations and outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Analysis of gene expression microarrays</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Exam (10:30-12:30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[ Main page | 4th term syllabus | R for Windows ]

Last modified: Wed Jan 18 10:28:30 EST 2006
Statistics for laboratory scientists

Karl W Broman
Department of Biostatistics, Johns Hopkins University

Office: E3612 SPH; Email: kbroman@jhsph.edu

http://www.biostat.jhsph.edu/~kbroman

TA: Qing Li (qli@jhsph.edu, E3035)

Outline

• Biostatistics courses
• About this course
• Logistics
• Grading
• Computer package

• What is statistics?
Introductory statistics courses at JHSPH

<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>611–612</td>
<td>Understand statistics in the literature</td>
</tr>
<tr>
<td>621–624</td>
<td>Actually do elementary statistics; focused largely on observational data</td>
</tr>
<tr>
<td>651–654</td>
<td>More advanced; requires calculus</td>
</tr>
</tbody>
</table>

615–616 (this course):

- Like 621–622, but focused on experimental rather than observational data.
- Should be able to enter 623–624 after (generalized linear models; multiple regression).
- **Take both terms!**

**Logistics**

**Lectures:** MWF 10:30-11:20 (W4013 SPH)

**Computer lab:** W 1:30-2:20 (W3025 SPH)

**Discussion:** W 2:30-3:20 (W2015 SPH)

**Office hours:**
- **Karl:** Mon & Fri 1:30-2:30 or by app’t (E3612 SPH)
- **Qing:** By appointment (E3035 SPH)

**Textbooks:**
- Samuels & Witmer (2002) Statistics for the life sciences
Grading

Grade based on:
• 3 Computer labs (33%)
• 3 Quizzes (33%)
• 1 Exam (34%)

Other work:
• Homework and reading assignments
• Play with R
• Deep and careful thought
• Discussions

Computer package: R

Advantages
+ Free
+ Available for Windows, Mac OSX, Unix
+ Comprehensive
+ Powerful graphics
+ Well-designed programming language
+ Unlimited extensibility
+ Widely used by statisticians
+ Increasingly used for microarray analyses

Disadvantages
– No dedicated support
– Complex syntax
– Not point-and-click
– Some simple tasks are rather hard
What is statistics?

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

— Sir R. A. Fisher
What is statistics?

- Data exploration and analysis
- Inductive inference with probability
- Quantification of uncertainty

Example 1

Goal: Determine, by fluorescence, the concentration of quinine in a sample of tonic water.

Method:
1. Obtain a stock solution with known concentration of quinine.
2. Create several dilutions of the stock.
3. Measure fluorescence intensity of each such standard.
4. Measure fluorescence intensity of the unknown.
5. Fit a line to the results for the standards.
6. Use line to estimate quinine concentration in the unknown.

Question: How precise is the resulting estimate?
Example 2


Children that have positive response to a pertussis antigen:

- Vaccinated with DTaP-HBV: 3/38 (8%)
- History of pertussis infection: 5/21 (24%)

Questions:

- How precisely can we estimate the chance of a positive response given vaccination?
- Are the above rates truly different?
Place tick on clay island surrounded by water, with two capillary tubes: one treated with deer-gland-substance; one untreated. Does the tick go to the treated or the untreated tube?

<table>
<thead>
<tr>
<th>Tick sex</th>
<th>Leg</th>
<th>Deer sex</th>
<th>treated</th>
<th>untreated</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>fore</td>
<td>female</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>female</td>
<td>fore</td>
<td>female</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>male</td>
<td>fore</td>
<td>male</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>female</td>
<td>fore</td>
<td>male</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>male</td>
<td>hind</td>
<td>female</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>female</td>
<td>hind</td>
<td>female</td>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>male</td>
<td>hind</td>
<td>male</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>female</td>
<td>hind</td>
<td>male</td>
<td>25</td>
<td>2</td>
</tr>
</tbody>
</table>

Questions:

- Is the tick more likely to go to the treated tube?
- Do the sex of the tick or deer, or the leg from which the gland substance was obtained, have an effect on the response of the tick?
Example 4

For each of 8 mothers and 8 fathers, we observe (estimates of) the number of crossovers, genome-wide, in a set of independent meiotic products.

**Question:**

Do the fathers (or mothers) vary in the number of crossovers they deliver?
Example 4 (cont.)

How do we think about this?

If there were no relationship between family ID and number of crossovers in a meiotic product:

- What sort of data would we expect?
- What would be the chance of obtaining data as extreme as what was observed?
Goals for the course

- Impart the statistician’s view of the world
- Basics of statistics
  - Basic experimental design
  - Sampling distributions
  - Confidence intervals
  - Hypothesis testing
- Basic statistical graphics
- Basic knowledge of R