Statistics for laboratory scientists (140.615-616)

[3rd term syllabus | 4th term syllabus | R for Windows]

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: qli@jhsph.edu **Office hours:** By appointment	
Syllabus	Third term Fourth term	
Textbook	ML Samuels, JA Witmer (2002) Statistics for the life sciences, 3rd ed, Prentice Hall [*Required*] L Gonick, W Smith (1994) Cartoon guide to statistics. HarperCollins. [*Recommended*] P Dalgaard (2002) Introductory statistics with R, Springer-Verlag [*Recommended*]	
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	Third term 33%: 3 computer labs 33%: 3 quizzes 34%: 1 exam	Fourth term 66%: 3 computer labs 34%: 1 final project

[3rd term syllabus | 4th term syllabus | R for Windows]

Last modified: Fri Jan 20 14:46:16 EST 2006

Statistics for laboratory scientists I (140.615)

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

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

Date		Topic	Reading	N	C	H	S	L
January	23	Overview; what is statistics?						
	25	Displaying data badly; data summaries						
	27	Experimental design						
	30	Observational studies						
February	1	Probability, conditional probability						
	3	Examples, Bayes's theorem						
	6	More examples						
	8	Random variables, distributions, binomial, Poisson						
	10	Normal distribution, multiple random variables						
	13	Sampling distributions; Central limit theorem						
	15	More of the same						
	17	Maximum likelihood estimation						
	20	Confidence interval (CI) for the mean						
	22	CIs for differences between means, CI for population SD						
	24	Tests of hypotheses						
	27	Tests for differences between means						
March	1	Calculation of sample size and power						
	3	Permutation tests and other non-parametric tests						
	6	Finish off permutation tests, sample size/power						
	8	Confidence interval for a proportion						
	11	Uses and abuses of tests						
	13	Transformations and outliers						
	15	Analysis of gene expression microarrays						
	17	**Exam** (10:30-12:30)						

[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

611–612	Understand statistics in the literature
621–624	Actually do elementary statistics; focused largely on observational data
651–654	More advanced; requires calculus

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
Gonick & Smith (1993) The cartoon guide to statistics. [recommended]
Dalgaard (2002) Introductory statistics with R statistics. [recommended]

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

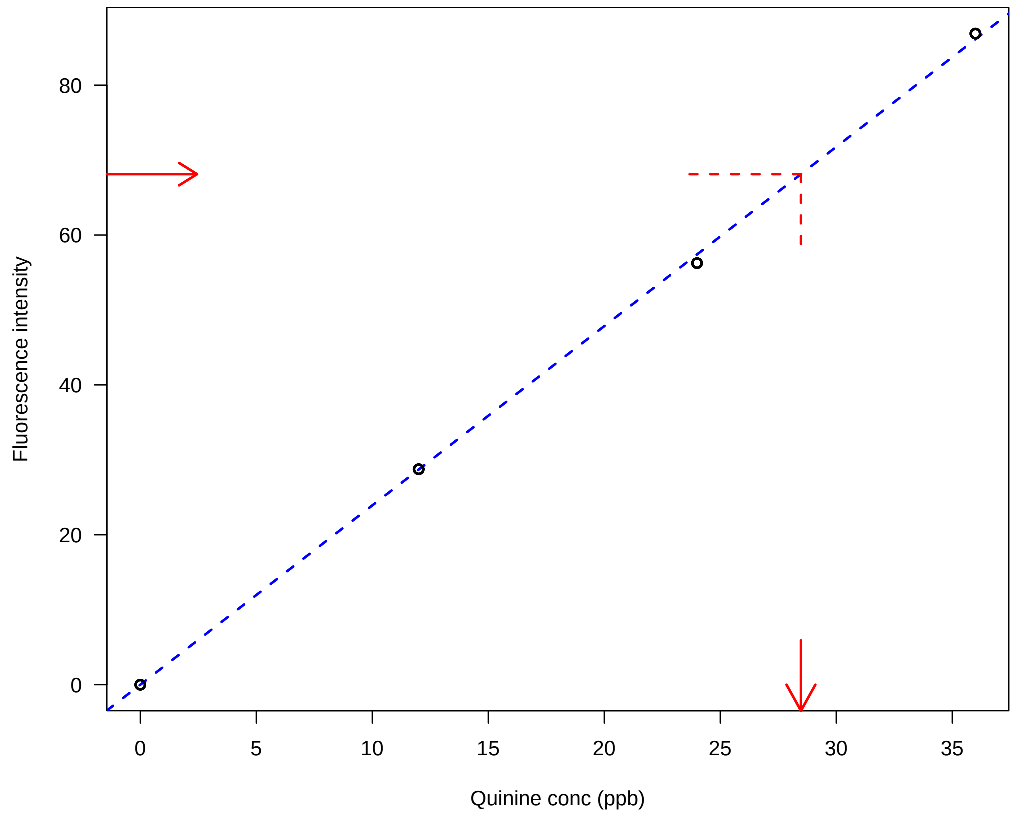
Example 1

Goal: Determine, by fluoresence, 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 fluoresence intensity of each such standard.
4. Measure fluoresence 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

[Esposito et al., *Infection and Immunity* **69**:4516–4520, 2001]

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?

Example 3

[Carroll, *J Med Entomol* **38**:114–117, 2001]

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?

Tick sex	Leg	Deer sex	treated	untreated
male	fore	female	24	5
female	fore	female	18	5
male	fore	male	23	4
female	fore	male	20	4
male	hind	female	17	8
female	hind	female	25	3
male	hind	male	21	6
female	hind	male	25	2

Example 3 (cont.)

Questions:

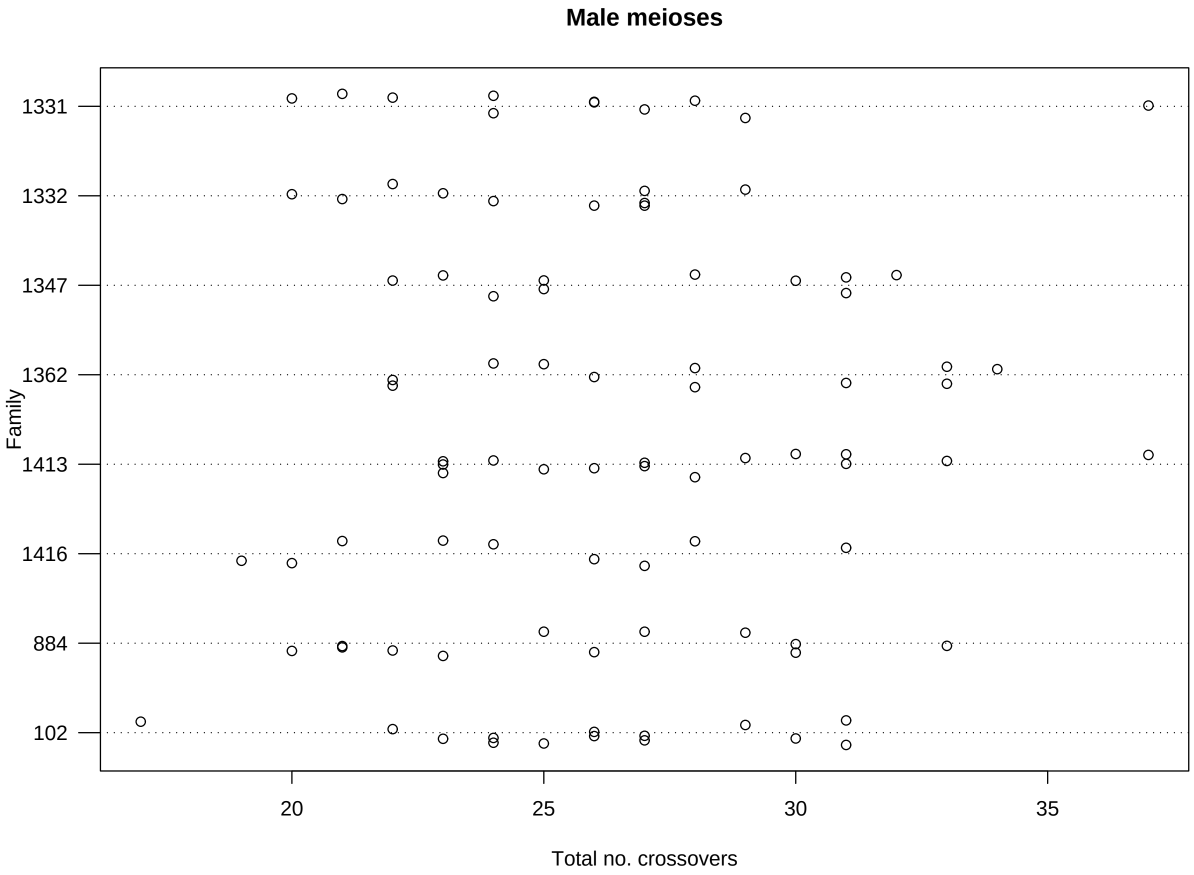
- Is the tick more likely to go to the treated tube?
- Do the sex of the tick or deer, or the leg from which the gland substance was obtained, have an effect on the response of the tick?

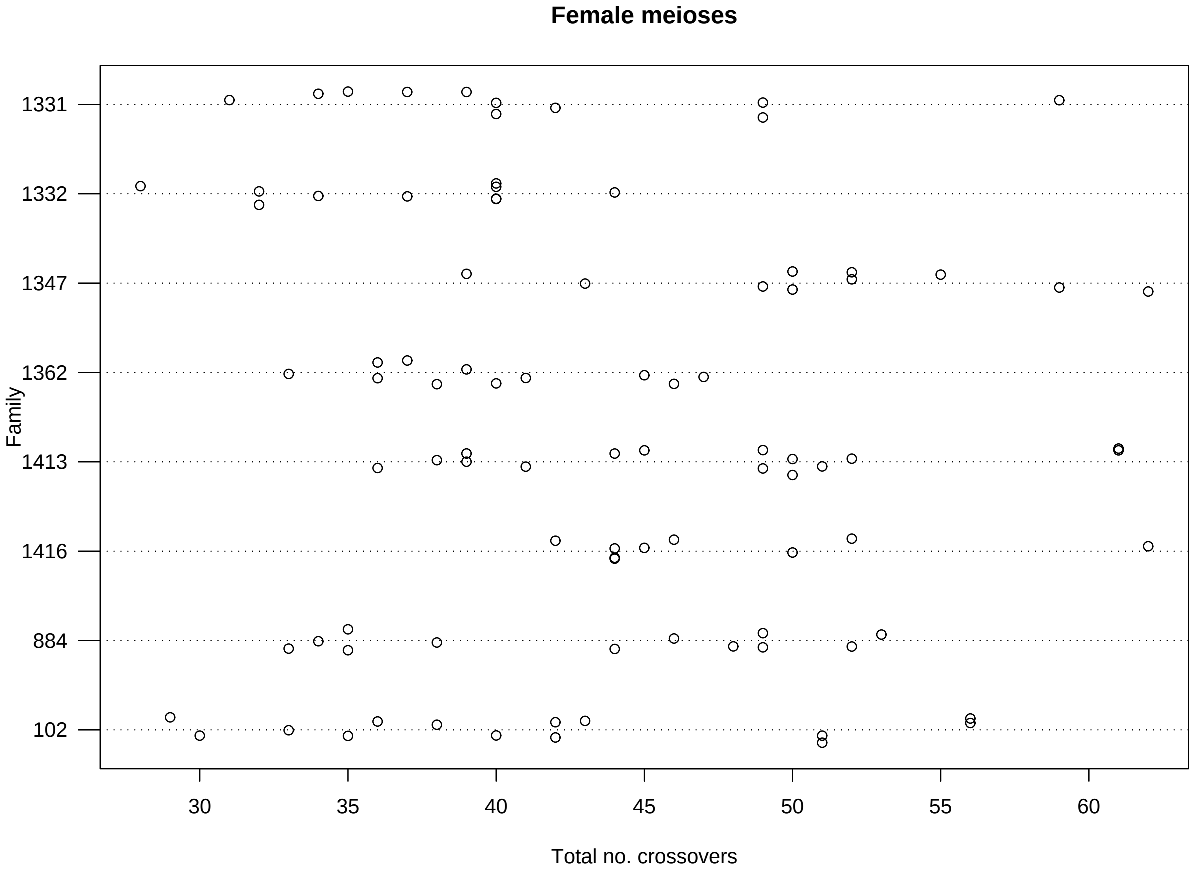
Example 4

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