Stat 371

Introductory applied statistics for the life sciences

Karl Broman

Department of Biostatistics and Medical Informatics

Offices: 6763 MSC, 1274 Genetics-Biotechnology

Email: `kbroman@biostat.wisc.edu`

`http://www.biostat.wisc.edu/~kbroman`

TA: Lili Lan (`lan@stat.wisc.edu`)

Logistics

Lectures: MWF 9:55-10:45 (1240 Comp Sci & Stat)

Discussions:
331: T 1:20p (6101 Soc Sci)
332: T 2:25p (4308 Soc Sci)
333: T 3:30p (1289 Comp Sci & Stat)
334: M 12:05p (138 Psychology)

Office hours:
Karl: Mon 1-2, Thu 1-2, Thu 3-4 (6763 MSC) or by appointment
Lili: Fri 11am–1pm (B248-H MSC)

Textbooks:
Samuels & Witmer (2002) Statistics for the life sciences [required]
Gonick & Smith (1993) The cartoon guide to statistics. [recommended]

Grading

Grade based on:

- Homework assignments (25%)
- Midterm 1 (20%)
- Midterm 2 (20%)
- Final (35%)

Other work:

- Reading assignments
- Play with the R software
- 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

In a study of nephroblastoma (embryonic kidney cancer) in rats exposed to N-ethyl-N-nitrosourea (ENU):

F344 strain:	0%
Noble strain:	50%

Is this difference real?

Example 1

In a study of nephroblastoma (embryonic kidney cancer) in rats exposed to N-ethyl-N-nitrosourea (ENU):

F344 strain:	0%
Noble strain:	50%

Is this difference real?

What if the data are as follows?

F344 strain:	0/4
Noble strain:	2/4

Example 1

In a study of nephroblastoma (embryonic kidney cancer) in rats exposed to N-ethyl-N-nitrosourea (ENU):

F344 strain:	0%
Noble strain:	50%

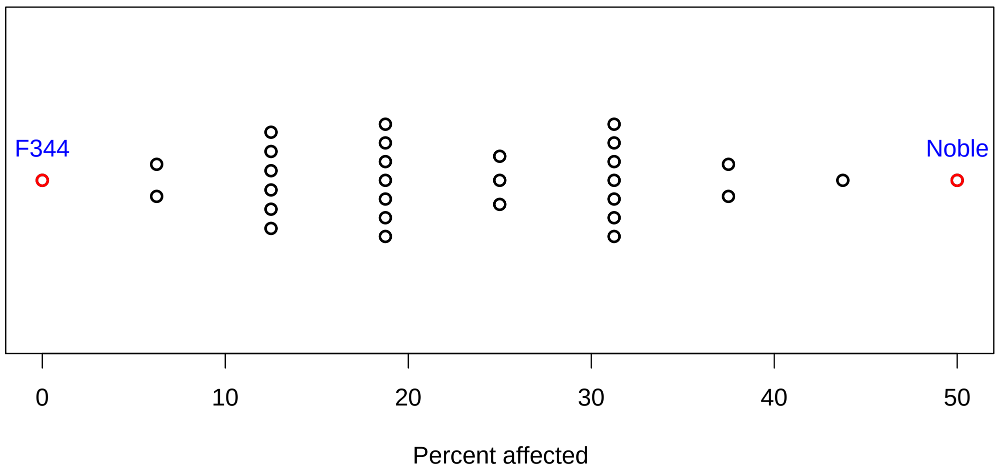
Is this difference real?

How about these data?

F344 strain:	0/16
Noble strain:	8/16

Example 1 (cont.)

What if these are the most extreme strains out of a set of 30 that were studied (with 16 rats investigated per strain)?



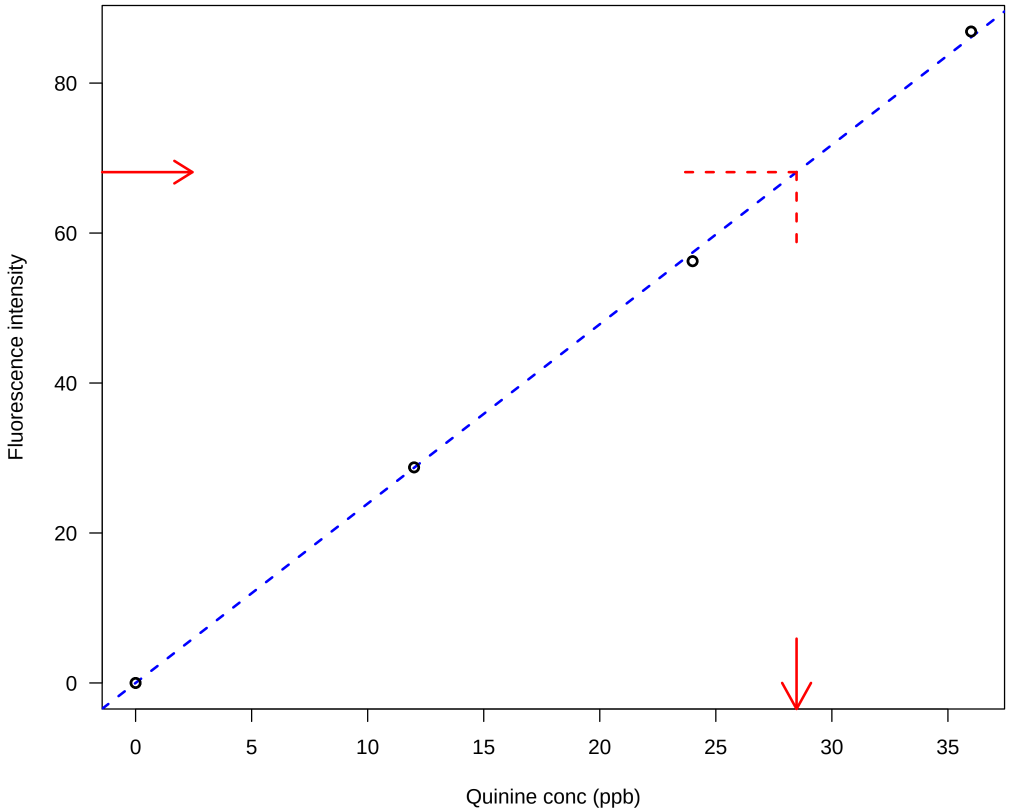
Example 2

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 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:

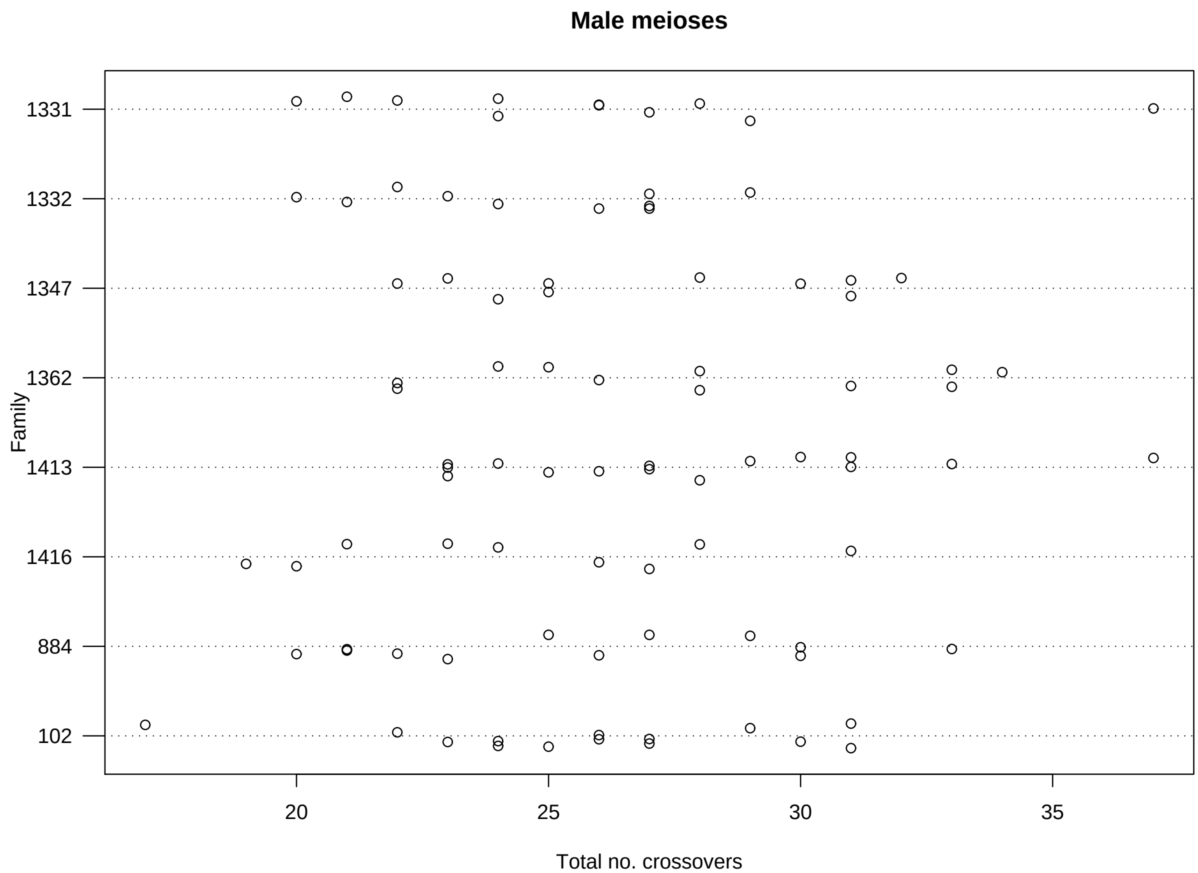
- Is the tick more likely to go to the treated tube?
- Do the sex of the tick or deer, or the leg from which the gland substance was obtained, have an effect on the response of the tick?

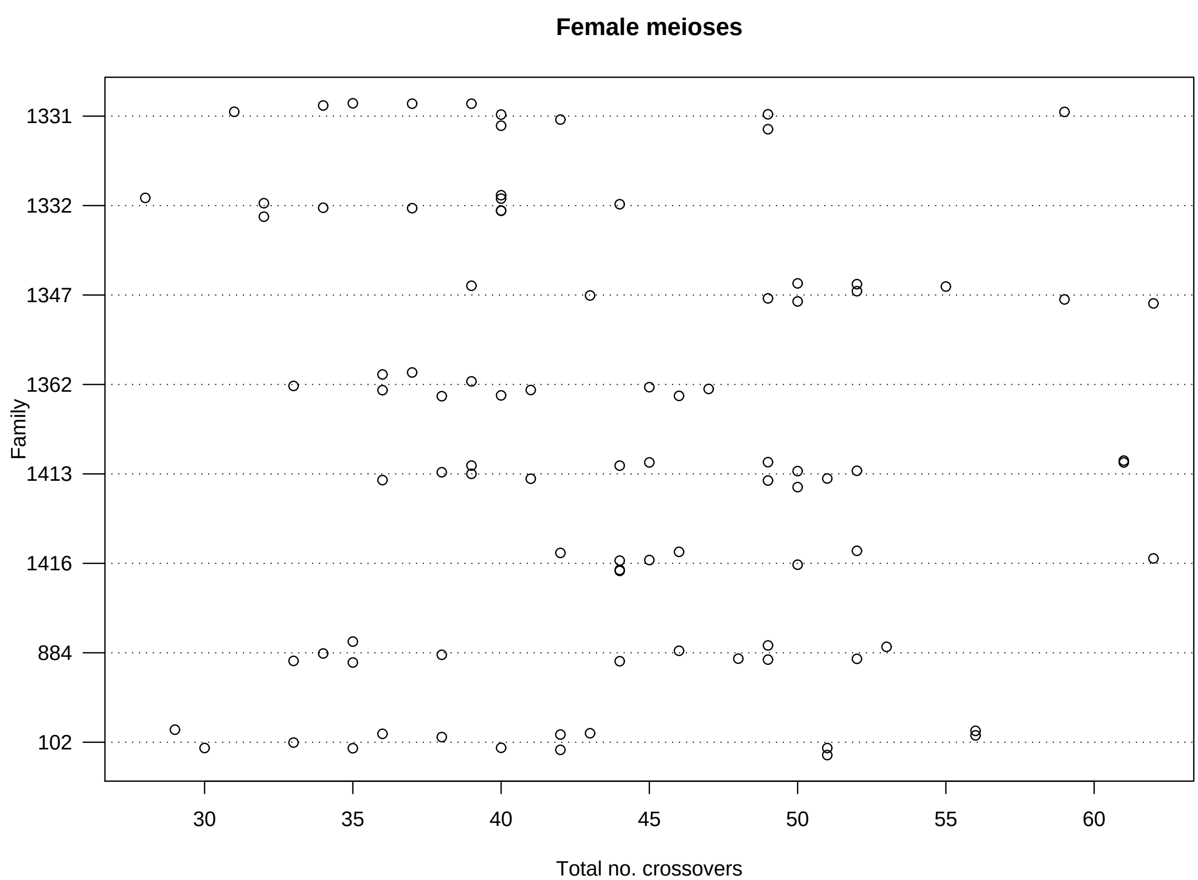
Example 4

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?


Male meioses
1331
1332
1347
1362
1413
1416
884
102
Family
20
25
30
35
Total no. crossovers


Female meioses
1331
1332
1347
1362
1413
1416
884
102
Family
30
35
40
45
50
55
60
Total no. crossovers

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?

Fundamental idea # 1

Separate the underlying population from the sample/data.

Separate features of the population (called parameters) from estimates based on the sample/data.

Fundamental idea # 2

Imagine repeating the whole process again.

What other data might we have obtained?

If, in truth, the world is boring, would these data be a surprise?

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

Why is statistics difficult?

- Strange way of thinking
- Semantics
- Probability eludes intuition
- Computing
- A bag of tricks (apparently)