Introduction to Biostatistics Math 4310-Biol6317

August 23, 2011

Syllabus

Department of Mathematics — University of Houston

Biostatistics

Math4310/Biol6317
Fall 2011

Class: TuTh 1pm-2:20pm, PGH 348

Instructor: Bernhard Bodmann, bgb@math.uh.edu

Office: PGH 604; Tu 11:30-12:20pm, Wed 2:00-2:50pm,

Objectives: This course covers applications of statistics in biology and medicine, motivated by typical case studies. The students will learn a variety of uses, and abuses, of statistical methods. The material will be interspersed with simple programming projects, which allows the students to become familiar with R, the open-source software package used in this course.

Contents: The first part of the course is a rapid review of essentials in probability and statistics. The main part of the material focuses on typical estimation problems and hypothesis testing applied to data from medicine as well as population, molecular and physiological biology.

Topic	*Approximate Time*
Probability and statistics essentials	2 weeks
Inferences for one sample	2 weeks
Summarizing and describing data	1 week
The two sample problem	2 weeks
Contingency tables	2 weeks
Case-control and cross-sectional studies	2 weeks
Introduction to non-parametric methods	2 weeks
Large datasets	1 week

Topics include: Independence, Bayes rule, sensitivity and specficity of a test, likelihood ratio; normal and chi-squared distribution, condence intervals; students t-distribution; empirical quantiles, boxplot, qauntile-quantile plot; kernel density estimates, stem and leaf plots, histograms; bootstrap principle; binomial confidence intervals; group comparisons; Pearsons chi-squared test; retrospective case/control studies; multiplicity: Bonferroni adjustment for family-wise error, false-discovery rate; stratified tables; matched pairs; Poisson processes and rate estimate.

Prerequisites: MATH 1432 and MATH 2311 or equivalent.

Text: The lectures will be as self-contained as possible. The course material follows the book: Bernard Rosner, Fundamentals of Biostatistics, 6th edition, Thomson Brooks/Cole, 2006. Due to its price, this is recommended, not mandatory. An alternative text, which covers most of the material in the course, is the book: Michael Whitlock and Dolph Schluter, The analysis of biological data, Roberts and Company, 2009

Website:
www.math.uh.edu/~bgb/Courses

Office hours:
Tu 11:30-12:20pm,
We 2-2:50pm
Q.: Can everyone make it to at least one day?

Book:
Rosner is helpful, but $$$.
Recommended, not mandatory

Syllabus, continued

Department of Mathematics University of Houston

Biostatistics

Math4310/Biol6317
Fall 2011

Class: TuTh 1pm-2:20pm, PGH 348

Instructor: Bernhard Bodmann, bgb@math.uh.edu

Office: PGH 604; Tu 11:30-12:20pm, Wed 2:00-2:50pm,

Objectives: This course covers applications of statistics in biology and medicine, motivated by typical case studies. The students will learn a variety of uses, and abuses, of statistical methods. The material will be interspersed with simple programming projects, which allows the students to become familiar with R, the open-source software package used in this course.

Contents: The first part of the course is a rapid review of essentials in probability and statistics. The main part of the material focuses on typical estimation problems and hypothesis testing applied to data from medicine as well as population, molecular and physiological biology.

Topic	*Approximate Time*
Probability and statistics essentials	2 weeks
Inferences for one sample	2 weeks
Summarizing and describing data	1 week
The two sample problem	2 weeks
Contingency tables	2 weeks
Case-control and cross-sectional studies	2 weeks
Introduction to non-parametric methods	2 weeks
Large datasets	1 week

Topics include: Independence, Bayes rule, sensitivity and specficity of a test, likelihood ratio; normal and chi-squared distribution, condence intervals; students t-distribution; empirical quantiles, boxplot, qauntile-quantile plot; kernel density estimates, stem and leaf plots, histograms; bootstrap principle; binomial confidence intervals; group comparisons; Pearsons chi-squared test; retrospective case/control studies; multiplicity: Bonferroni adjustment for family-wise error, false-discovery rate; stratified tables; matched pairs; Poisson processes and rate estimate.

Prerequisites: MATH 1432 and MATH 2311 or equivalent.

Text: The lectures will be as self-contained as possible. The course material follows the book: Bernard Rosner, Fundamentals of Biostatistics, 6th edition, Thomson Brooks/Cole, 2006. Due to its price, this is recommended, not mandatory. An alternative text, which covers most of the material in the course, is the book: Michael Whitlock and Dolph Schluter, The analysis of biological data, Roberts and Company, 2009

Homework: 10 sets, short statistics problems, some exercises with R (freely available stats package)

Biology graduate students: Apart from 10 regular homework sets, 3 or 4 projects with data analysis provided by Biology faculty (Azevedo, Frankino, Ziburkus).

Syllabus, continued

Software: R, freely available at www.cran.r-project.org

Midterm Exam: Tuesday, October 18, 2011, in class.

Assignments: You will be asked to hand in approximately ten assignments, which will be due on Thursdays in the lecture. To obtain full credit for the course, graduate students will need to complete 4 additional projects on biological datasets.

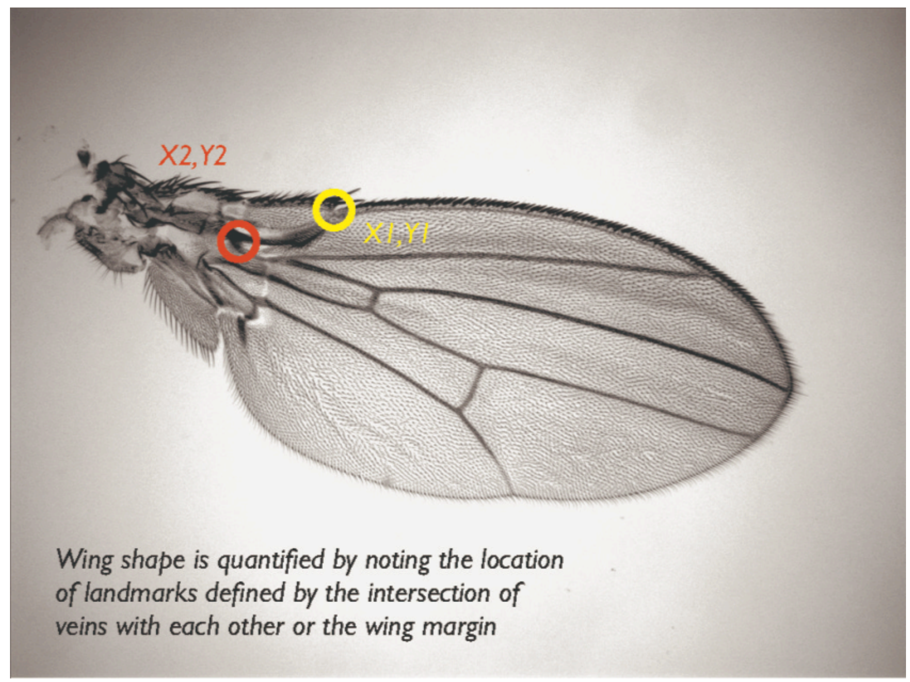
Final Grade: Final exam contributes 40%, midterm 30%, assignments 30%. All grades are summed and divided by the total number of points you can collect in the course. A percentage of 46% or more is D- , 54% or more is D, 62% or more is C, 70% is B-, 77% is B, 85% or more is A- , of 90% or more is A.

Midterm: October 18, in class.

Project Example: Wing shapes

Worlds between theory and experiment

- Learning objective: Exposure to realistic conditions of research in a laboratory
- Method: **Case studies**
- Example: Statistical analysis of wing shape measurements to **distinguish genotypes**

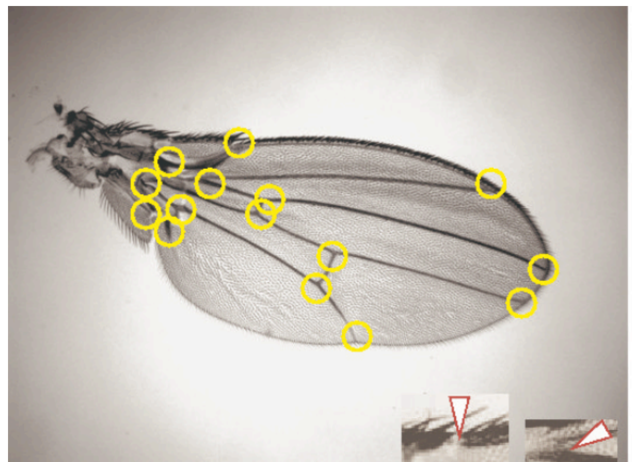


Data provided by Tony Frankino, Biology

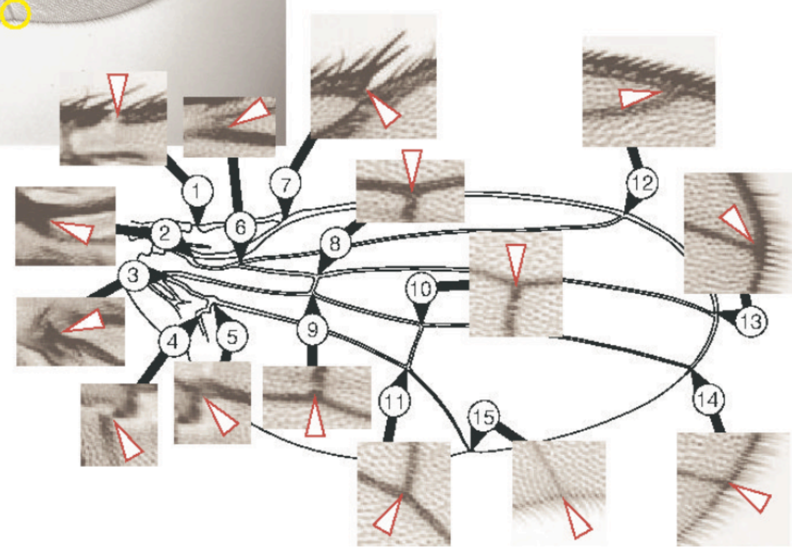
Project Example:From Data to Textbook Method

Modern research: facing a flood of data

- Challenge: Too much data to apply standard recipes
- Strategy: Extract relevant quantities



Total of 15 landmarks used.




```
comp1<-wing_eigen$vectors[,1]
```

Project Example: Result

- Solution: Combine data reduction and hypothesis testing to distinguish two different genotypes

```
> t.test(Z[1:42],Z[43:90],var.equal=TRUE)

Two Sample t-test

data:  Z[1:42] and Z[43:90]
t = -4.6546, df = 88, p-value = 1.140e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3428719 -0.1376926
sample estimates:
   mean of x    mean of y
0.0001783903 0.2404606482
```

What is Biostatistics?

From the Wikipedia entry on biostatistics:
Biostatistics (a combination of the words biology and statistics; sometimes referred to as biometry or biometrics) is the application of statistics to a wide range of topics in biology and medicine. The science of biostatistics encompasses

- the design of biological experiments, especially in medicine and agriculture;

- the collection, summarization, and analysis of data from those experiments; and

- the interpretation of, and inference from, the results.

Example: Mendel and Pea Counts

Gregor Mendel was an Augustinian monk who lived in the late 19th century and, through studying peas, developed the basis for today's genetics.

Expt. 1. — ***AB***, seed parents ***ab***, pollen parents
A, form round ***a***, form wrinkled
B, albumen yellow ***b***, albumen green

The fertilized seeds appeared round and yellow like those of the seed parents. The plants raised therefrom yielded seeds of four sorts, which frequently presented themselves in one pod. In all, 556 seeds were yielded by 15 plants, and of these there were:

315	round and yellow,
101	wrinkled and yellow,
108	round and green,
32	wrinkled and green.

101+32 =133 wrinkled
of 556 total,

fraction: 133/556=24%.

Why 24%?

Example: Mendel and Pea Counts

Gregor Mendel was an Augustinian monk who lived in the late 19th century and, through studying peas, developed the basis for today's genetics.

R=round
r=wrinkled

Pollen/Egg combined
R dominant

Eggs \ Pollen	1/2 R	1/2 r
1/2 R	1/4 RR	1/4 Rr
1/2 r	1/4 rR	1/4 rr

Example: Mendel and Pea Counts

Gregor Mendel was an Augustinian monk who lived in the late 19th century and, through studying peas, developed the basis for today's genetics.

Expt. 1. — *AB*, seed parents *ab*, pollen parent
A, form round *a*, form wrinkle
B, albumen yellow *b*, albumen gree

The fertilized seeds appeared round and yellow like t seed parents. The plants raised therefrom yielded seeds of which frequently presented themselves in one pod. In all were yielded by 15 plants, and of these there were:

315 round and yellow,
101 wrinkled and yellow,
108 round and green,
32 wrinkled and green.

Eggs

Pollen	**1/4 R Y**	**1/4 R y**	**1/4 r Y**	**1/4 r y**
1/4R Y	**RR YY**	**RR Yy**	**Ry YY**	
1/4 R y		**RR yy**		
1/4 r Y				**rr Yy**
1/4 r y				**rr yy 1/16**

R: round r: wrinkled
Y yellow y: green

rryy fraction: 32/556=5.7%.

Example: Smoking and Longevity

Raymond Pearl

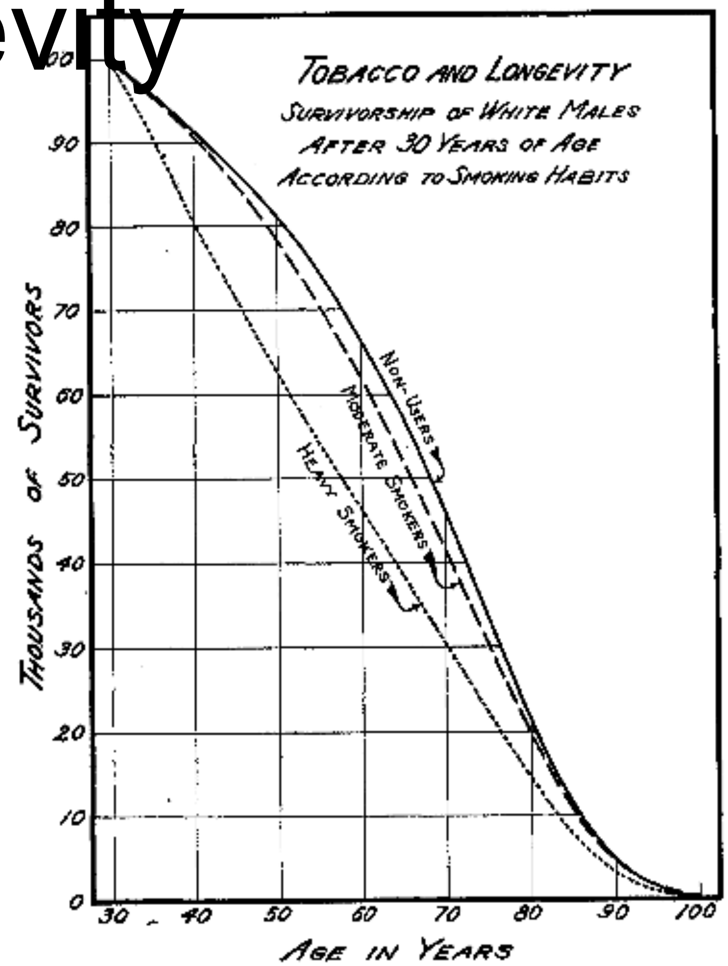


FIG. 1. The survivorship lines of life tables for white males falling into three categories relative to the usage of tobacco. A. Non-users (solid line); B. Moderate smokers (dash line); C. Heavy smokers (dot line).

Example: Smoking and Longevity

1938: Raymond Pearl publishes **Smoking and Longevity**
1964: Advisory Committee to the Surgeon General publishes **Smoking and Health**, holding cigarette smoking responsible for a 70 percent increase in the mortality rate of smokers over non-smokers. The report estimates that average smokers had a nine to ten-fold risk of developing lung cancer compared to non-smokers: heavy smokers had at least a twenty-fold risk. The report also named smoking as the most important cause of chronic bronchitis and pointed to a correlation between smoking and emphysema, and smoking and coronary heart dise

Q.: Why more than 25 years in between?

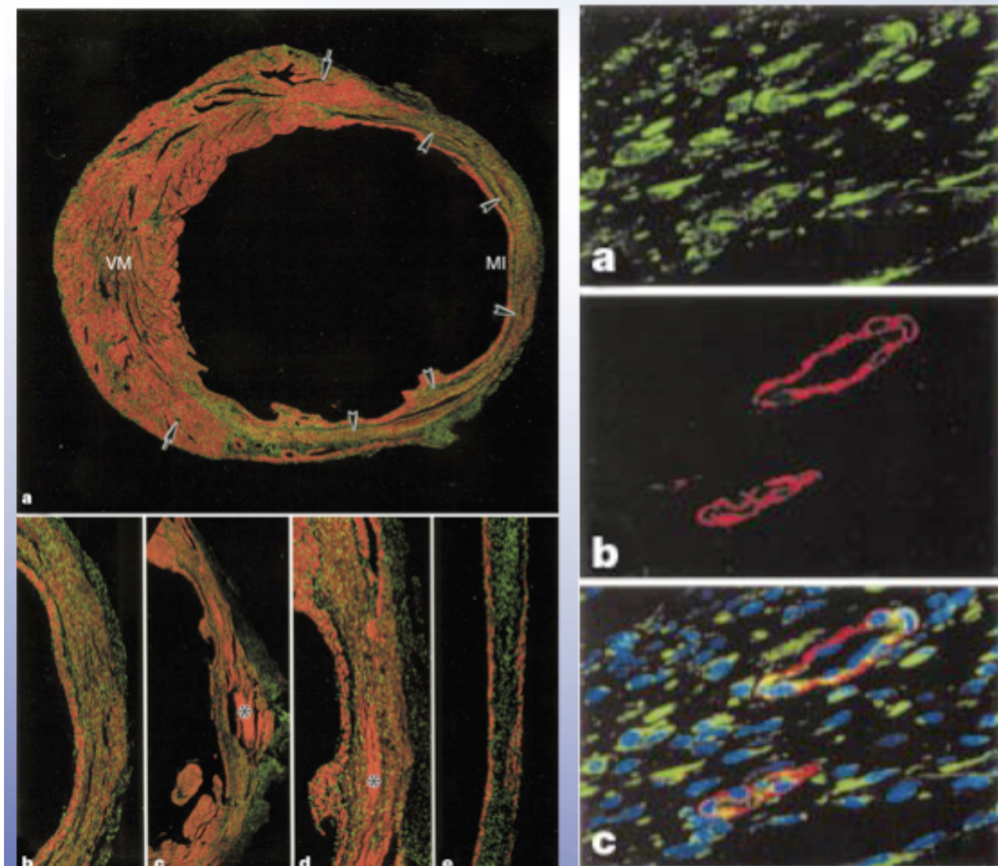
Example: Smoking and Longevity

1938: Raymond Pearl publishes **Smoking and Longevity**
1964: Advisory Committee to the Surgeon General publishes **Smoking and Health**, holding cigarette smoking responsible for a 70 percent increase in the mortality rate of smokers over non-smokers. The report estimates that average smokers had a nine to ten-fold risk of developing lung cancer compared to non-smokers: heavy smokers had at least a twenty-fold risk. The report also named smoking as the most important cause of chronic bronchitis and pointed to a correlation between smoking and emphysema, and smoking and coronary heart dise

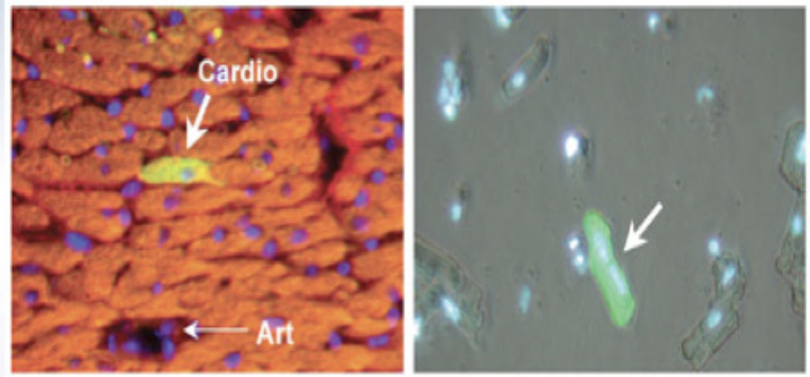
Q.: Why more than 25 years in between? Pearl's methods and interpretation (not outcomes!) were disputed.

Example: Stem cells

Orlic et al, Nature, 2001

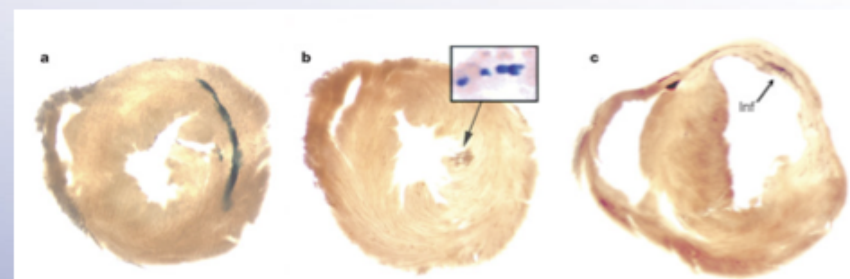


Murry et al, Nature, 2004



Bone marrow transplant (MHC/EGFP)

Bone marrow transplant (EGFP)



University of Waterloo

Injected HaemSCs regenerate 53% of CardioMyo's in MI

Injected HaemSCs regenerate <0.0001% of CardioMyo's in MI

Experiments: From reality to mathematical description

The outcomes of a statistical experiment could be:

…?

Experiments: From reality to mathematical description

The outcomes of a statistical experiment could be:

- an election
- fragments from DNA nucleotide sequences
- the result of a clinical trial
- the output of a computer simulation
- information gathered from hospital records
- ...

Mathematical description of experiments

The **sample space**, Ω, is the collection of possible **outcomes** of an experiment.

Example: die roll $\Omega = \{1,2,3,4,5,6\}$.

An **event**, say E, is a subset of Ω.

Example: die roll is even $E = \{2,4,6\}$.

The set ϕ is called the null event or the empty set.

Set theoretic notation and interpretation

$\omega \in E$ means that if ω occurs then E occurs, too.
$E \subset F$ means that the occurrence of E implies the occurrence of F.
$E \cap F$ means the event that both E and F occur.
$E \cup F$ means the event that at least one of E or F occur.
$E \cap F = \phi$ means that E and F are **mutually exclusive**, or cannot both occur.
E^c or $\bar{E}$ is the event that E does not occur.

Probability measures

A **probability measure** P is a real valued function from the collection of possible events so that the following hold

1. For each event $E \subset \Omega$, $0 \leq P(E) \leq 1, P(\Omega) = 1$.
2. If $\{E_j\}_{j=1}^{\infty}$ is a sequence of mutually exclusive (disjoint) events, then $P(\cup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} P(E_j)$.