Biostatistics – Course Notes

Bernhard Bodmann

September 2, 2009

(Week 1 only)

Contents

1 Overview

Syllabus

Department of Mathematics — University of Houston

Biostatistics

Math4397/6397
Fall 2009

Class:	TuTh 2:30pm-3:50pm, AH 301
Instructor:	Bernhard Bodmann, bgb@math.uh.edu
Office:	PGH 604; Wed 2:00-2:50pm, Th 11:00-11:50am
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.

General 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

Detailed topics include: Independence, Bayes rule, sensitivity and specificity of a test, likelihood ratio; normal and chi-squared distribution, confidence intervals; students t-distribution; empirical quantiles, boxplot, quantile-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:	Bernard Rosner, Fundamentals of Biostatistics, 6th edition, Thomson Brooks/Cole, 2006.

1.1 What is biostatistics?

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.

All were sown the following year. Eleven of the round yellow seeds did not yield plants, and three plants did not form seeds. Among the rest:

38	had round yellow seeds	*AB*
65	round yellow and green seeds	*ABb*
60	round yellow and wrinkled yellow seeds	*AaB*
138	round yellow and green, wrinkled yellow and green seeds	*AaBb*

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: Smoking and cancer across 26 years

1938: Raymond Pearl publishes "Smoking and Longevity"

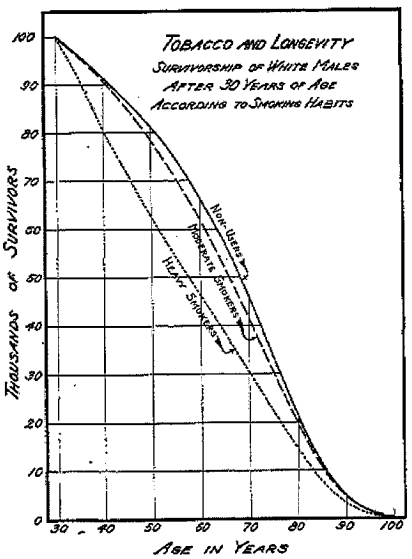
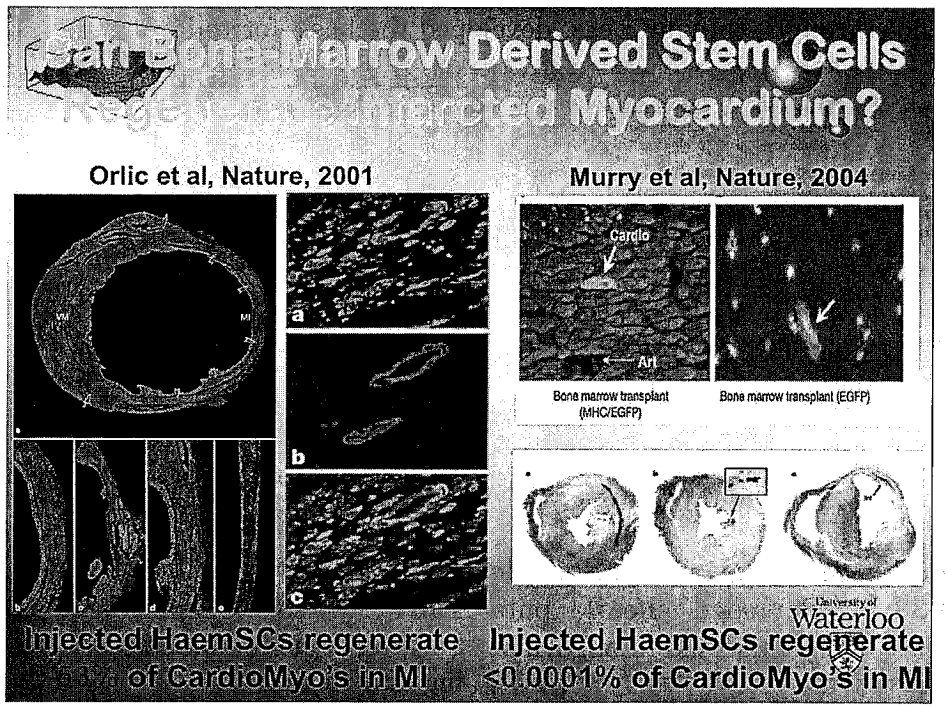


FIG. 1. The survivorship lines of life tables for white males falling into three categories relative to the usage of tobacco. A. Non-users (solid line); B. Moderate smokers (dash line); C. Heavy smokers (dot line).

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 estimated 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 disease.

Example: stem cells and cardiovascular regeneration



1.2 Experiments: From reality to mathematical description

Statistical experiments: From reality...

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
- ...

Experiments: ...to a mathematical description

- 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 $\emptyset$ is called the **null event** or the **empty set**.

Set theoretic notation and interpretation

Set operations have particular interpretations for events.

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

2 Probability essentials [Ch. 3.1-3.5, 4.1-4.3, 5.1-5.2]

Probability measures

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

1. For an event $E \subset \Omega$, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1$

3. If $\{E_j\}_{j=1}^{\infty}$ is a sequence of mutually exclusive (disjoint) events, then $P(\cup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} P(E_j)$.

Discrete vs. continuous outcomes

- P is defined on $\mathcal{F}$ a collection of subsets of Ω
- Example $\Omega = \{1, 2, 3\}$ then

$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

- When Ω is a continuous set such as $\mathbb{R}$, we always assume that $\mathcal{F}$ contains all (bounded or semi-bounded) intervals, their complements, and all countable intersections and unions thereof.

Rules for computing probabilities

Based on the axioms, we can prove all of the following:

- $P(\emptyset) = 0$,
- $P(E) = 1 - P(E^c)$,
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,
- if $A \subset B$ then $P(A) \leq P(B)$,
- $P(A \cup B) = 1 - P(A^c \cap B^c)$,
- $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$,
- $P(\cup_{i=1}^{n} E_i) \geq \max_i P(E_i)$.

Proving a rule

Prove that $P(E) = 1 - P(E^c)$.

We have $E \cup E^c = \Omega$ and $E \cap E^c = \emptyset$. Thus,

$$\begin{aligned} 1 &= P(\Omega) \\ &= P(E \cup E^c) \\ &= P(E) + P(E^c) \end{aligned}$$

■

Proving another rule

Prove that $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$

We proceed by induction, starting with the case $\boldsymbol{n = 2}$.

We have $E_1 = (E_1 \cap E_2) \cup (E_1 \cap E_2^c)$. Similarly, $E_2 = (E_2 \cap E_1) \cup (E_2 \cap E_1^c)$ and $E_1 \cup E_2 = (E_1 \cap E_2) \cup (E_1 \cap E_2^c) \cup (E_2 \cap E_1^c)$, with the events in parentheses being mutually exclusive.

Additivity gives

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &\leq P(E_1) + P(E_2) \end{aligned}$$

Now consider $\boldsymbol{n}$, assuming the rule holds for $\boldsymbol{n-1}$ and $\boldsymbol{n = 2}$. Using the induction assumption for $n = 2$ and the sets E_n, $\cup_{i=1}^{n-1} E_i$ and for $n - 1$ and the sets $E_1, E_2, \ldots E_{n-1}$ gives

$$\begin{aligned} P(\cup_{i=1}^{n} E_i) &\overset{n=2}{\leq} P(E_n) + P(\cup_{i=1}^{n-1} E_i) \\ &\overset{n-1}{\leq} P(E_n) + \sum_{i=1}^{n-1} P(E_i) \\ &= \sum_{i=1}^{n} P(E_i) \end{aligned}$$

■

Using rules with a study on middle ear infections

In vol. 166 of the the European Journal of Pediatrics, Bulut et al. write that from 120 children having an acute middle ear infection, "respiratory viruses were isolated in 39 patients (32.5%). In total 69 bacterial species were isolated from 65 (54.8%) of 120 patients."

Question: Does this imply that 87.3% of patients tested positive for a virus or for bacteria?

Answer: No, because

$$P(\text{viruses or bacteria}) = P(\text{viruses}) + P(\text{bacteria}) - P(\text{viruses and bacteria})$$

so if some patients have both, then

$$P(\text{viruses or bacteria}) < P(\text{viruses}) + P(\text{bacteria}) = 0.873\,.$$

Rules and drosophila mutation rates

Probability of spontaneous, lethal mutation in X chromosome (Crow and Temin, 1964)

$$P(\text{mutation}) = 0.0025$$

Question: Since $P(\dots) = 1/400$, we need 400 fruit flies to observe a mutant with certainty?

Rebuttal: The probability of having no mutation among 400 is easier to compute, so we use complements.

$$P(\text{no mutation for 1 fly}) = 1 - \frac{1}{400}.$$

If the mutation for each fly is independent (see further below), then

$$\begin{aligned} P(\text{no mutation among 400}) &= P(\{\text{none for 1st }\} \cap \{\text{none for 2nd }\} \cap \cdots \cap \{\text{none for 400th}\}) \\ &= P(\{\text{none for 1st }\})P(\{\text{none for 2nd }\}) \dots P(\{\text{none for 400th}\}) \\ &= (1 - \frac{1}{400})^{400}. \end{aligned}$$

Thus,

$$\begin{aligned} P(\text{at least 1 mutation among 400}) &= 1 - P(\text{no mutation among 400}) \\ &= 1 - (1 - 0.0025)^{400} \end{aligned}$$

We compute $P(\text{at least 1 mutation among 400}) \approx 0.63$.

2.1 Random variables

Random variables

- A **random variable** is a map from outcomes of an experiment to numbers.
- The random variables that we study will come in two varieties, **discrete** or **continuous**.
- Discrete random variable are random variables that take on only a countable number of possibilities, e.g. $\{0, 1, 2, 3, \dots\}$.
- A continuous random variable can take any value on the real line or some subset of the real line.

Examples of random variables

- The fortune of a casino player at some time.
- The value $\{0, 1\}$ associated with the outcome of a coin flip.
- The systolic blood pressure of a person randomly drawn from a population.
- The level of gene expression in some cell.

2.2 Probability Mass Functions and Probability Distribution Functions

Probability Mass Function (PMF)

A **probability mass function** evaluated at an outcome corresponds to the probability that a random variable takes that value. To be a valid pmf, a function p must satisfy

1. $p(x) \geq 0$ for all x
2. $\sum_x p(x) = 1$

The sum is taken over all of the possible values for x.

Probability Density Funtion

A **probability density function (pdf)** is an integrable function associated with a continuous random variable.

> Areas under the graph of a pdf correspond to probabilities for that random variable.

To be a valid pdf, a function f must satisfy

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Example: Density for goldfish life

Assume that the natural life time of a goldfish in years follows a density like

$$f(x) = \begin{cases} \frac{e^{-x/5}}{5} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

More compactly written: $f(x) = \frac{1}{5}e^{-x/5}$ for $x > 0$. Is this a valid density? We check

1. The number e raised to any power is always positive, thus f is.
2. $$\int_0^\infty f(x)dx = \int_0^\infty e^{-x/5}/5dx = -e^{-x/5}\Big|_0^\infty = 1$$

Example continued

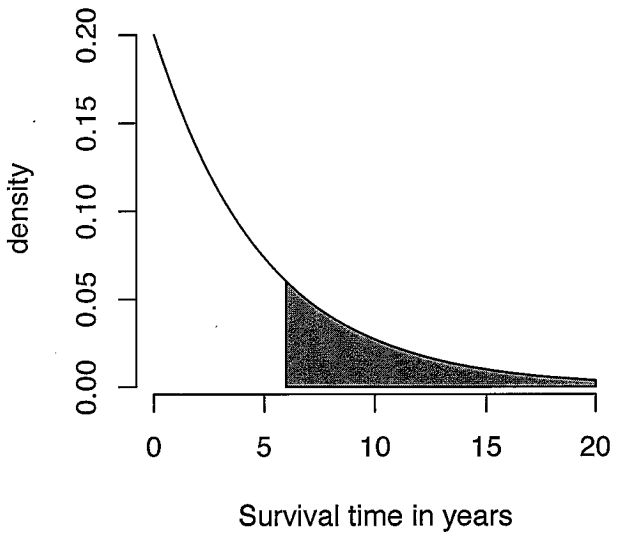
What is the probability that a randomly selected goldfish from this distribution survives more than 6 years?

$$P(X \geq 6) = \int_6^\infty \frac{e^{-t/5}}{5}dt = -e^{-t/5}\Big|_6^\infty = e^{-6/5} \approx .301.$$

Approximation in R

```
pexp(6, 1/5, lower.tail = FALSE)
```

Example continued



2.3 Cumulative Distribution Functions, survival functions and quantiles

Cumulative Distribution Functions and the survival function

- The **cumulative distribution function** (CDF) of a random variable X is defined as the function

$$F(x) = P(X \leq x)$$

- This definition applies, whether X is discrete or continuous.
- The **survival function** of a random variable X is defined as

$$S(x) = P(X > x)$$

- Notice that $S(x) = 1 - F(x)$
- If a continuous random variables has a PDF, then it is the derivative of the CDF.

Example: Survival function and CDF for exponential density

What are the survival function and CDF from the exponential density considered before?

$$S(x) = \int_x^\infty \frac{e^{-t/5}}{5} dt = -e^{-t/5}\Big|_x^\infty = e^{-x/5}$$

hence we know that

$$F(x) = 1 - S(x) = 1 - e^{-x/5}$$

Notice that we can recover the PDF by

$$f(x) = F'(x) = \frac{d}{dx}(1 - e^{-x/5}) = e^{-x/5}/5$$

Quantiles

- The α^{th}**quantile** of a distribution with distribution function F is the point x_α so that
$$F(x_\alpha) = \alpha$$
- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

Example: Quantiles for exponential distribution

- What is the 25^{th} percentile of the exponential survival distribution considered before?
- We want to solve (for x)
$$\begin{aligned} .25 &= F(x) \\ &= 1 - e^{-x/5} \end{aligned}$$
resulting in $x = -\log(.75) \times 5 \approx 1.44$
- Therefore, 25% of the goldfish from this population live less than 1.44 years
- R can approximate exponential quantiles: `qexp(.25, 1/5)`