Introduction to Biostatistics
Math 4310-Biol6317

August 23, 2011
Syllabus

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

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).
Software: R, freely available at www.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 Grades: 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 96% or more is A+, 84% or more is A, 72% or more is B+, 77% is B, 80% or more is A-, 68% or more is C.
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

Wing shape is quantified by noting the location of landmarks defined by the intersection of veins with each other or the wing margin

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

```r
comp1<-wing_eigen$vectors[,1]
```
• Solution: Combine data reduction and hypothesis testing to distinguish two different genotypes

```r
> 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. —  

\[ AB, \text{seed parents} \]
\[ A, \text{form round} \]
\[ B, \text{albumen yellow} \]
\[ ab, \text{pollen parents} \]
\[ a, \text{form wrinkled} \]
\[ b, \text{albumen green} \]

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:

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>round and yellow</td>
<td>315</td>
</tr>
<tr>
<td>wrinkled and yellow</td>
<td>101</td>
</tr>
<tr>
<td>round and green</td>
<td>108</td>
</tr>
<tr>
<td>wrinkled and green</td>
<td>32</td>
</tr>
</tbody>
</table>

\[ 101+32 = 133 \text{ wrinkled of 556 total,} \]

fraction: \( \frac{133}{556} = 24\% \).

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

<table>
<thead>
<tr>
<th>Pollen</th>
<th>Eggs</th>
<th>1/2 R</th>
<th>1/2 r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2 R</td>
<td>1/4 RR</td>
<td>1/4 Rr</td>
<td></td>
</tr>
<tr>
<td>1/2 r</td>
<td>1/4 rR</td>
<td>1/4 rr</td>
<td></td>
</tr>
</tbody>
</table>
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
Y yellow  y: green

<table>
<thead>
<tr>
<th>Eggs</th>
<th>1/4 R Y</th>
<th>1/4 r Y</th>
<th>1/4 R Y</th>
<th>1/4 r Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/4 R Y</td>
<td>RR YY</td>
<td>RR Yy</td>
<td>Ry YY</td>
<td></td>
</tr>
<tr>
<td>1/4 R y</td>
<td></td>
<td></td>
<td></td>
<td>RR yy</td>
</tr>
<tr>
<td>1/4 r Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/4 r y</td>
<td></td>
<td></td>
<td></td>
<td>rr Yy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rr yy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/16</td>
</tr>
</tbody>
</table>

The fertilized seeds appeared round and yellow like the 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.

rerry  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 (dashed 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 disease.

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 disease.

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


Murry et al, Nature, 2004

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**, $\Omega$, 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 $\Omega$.

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

The set $\phi$ is called the null event or the empty set.
Set theoretic notation and interpretation

\(\omega \in E\) means that if \(\omega\) 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.
\(E \cap F = \emptyset\) means that \(E\) and \(F\) are mutually exclusive, or cannot both occur.
\(E^c\) or \(\overline{E}\) 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 $P(\bigcup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} P(E_j)$. 