UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: Ivo Dinov,

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

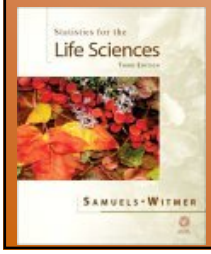
Fred Phoa; Anwer Khan & Jason Shen

University of California, Los Angeles, Fall 2005

http://www.stat.ucla.edu/~dinov/courses_students.html

Administrative

- The book for this course -- Statistics for the Life Sciences
  - Homework will be primarily assigned from the text
  - You are responsible for keeping up with reading
  - Some chapters will be covered by reading only



Survey – Must Complete Anonymously

- http://moodle.stat.ucla.edu/mod/questionnaire/view.php?id=501
- http://moodle.stat.ucla.edu/ → Resources → SOCR → **socrsurvey05** (password) → SOCR Intro Survey

Please complete this anonymous survey ONCE and provide your candid responses

1. Logon using any browser: http://moodle.stat.ucla.edu/
2. Go to: Resources
3. Click on: Statistics Online Computational Resource (SOCR)
4. Enter password: socrsurvey05
5. Click on: SOCR Intro Survey
6. Complete the Survey (Thank you!)
7. Close Browser

UCLA STAT 13

to just hear is to forget

to see is to remember

to do it yourself is to understand …

(… to go to class is to … comprehend …)

What is Statistics? A practical example

Michael Benton & Francisco Ayala, *Dating the Tree of Life*, Science 2003 300: 1698-1700

Molecular vs. **Paleontological** dating of major branching points in the tree of life are debated

Molecular date estimates are up to twice as old (due to statistical bias) as **Paleontological** dates (missing fossils).

Goals: Same as that set out by Darwin: to understand *where life came from*, the *shape of evolution*, the *place of humans in nature* and to determine the *extent of modern biodiversity* and where *it is threatened*.

What is Statistics? A practical example

Plants: The first vascular land plants are found as **fossils** in the Silurian, and earlier evidence from possible vascular plant spores may extend the range back to the Ordovician, **475 Ma** considerably < a molecular estimate of **700 Ma**.

Birds: Molecular estimates place the split of basal clades and modern orders at 70 to 120 Ma. The oldest uncontroversial fossils of modern bird orders date from the Paleocene (60 Ma), much younger.

Mammals: Molecular dates split of modern placentals in the mid- to Late Cretaceous (80 to 100 Ma). The oldest fossil representatives of modern mammals dated from the Paleocene and Eocene (50 to 65 Ma).

What is Statistics? Topics!

It is proposed that molecular dates are correct (with **confidence intervals**) and that methods exist to correct for that **error**. However, critics have pointed out several **pervasive biases** that make molecular dates too old.

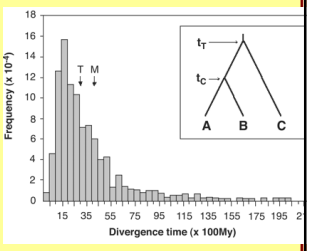
First, if calibration dates are too old, then all other dates **estimated** from them will also be too old.

A second biasing factor is that undetected fast-evolving genes could **bias** estimates of timing. **Empirical and statistical studies** of vertebrate sequences suggest that such non-clock-like genes may be detected and that they do not affect **estimates** of dating. However, **statistical tests** may have low **power** and could produce consistently > dates.

What is Statistics? A practical example

A 3rd source of **bias** relates to polymorphism. Two species often become fixed for alternative alleles that existed as a polymorphism in their ancestral species.

A 4th biasing **factor** is that molecular time estimates show (**skewed**) **asymmetric distributions**, with a constrained (large numbers) younger left-end and an unconstrained (smaller numbers) older right-end.



What is Statistics? Estimate Variation!

Data Source	Metazoa (Animals) In MYA	Bilateria (metazoans except sponges, e.g., anemones)	Deuterostomia (backboned animals)
Gene (8 G)		1200 ± 100	1001 ± 100
Protein (64 E)	930 ± 115	790 ± 60	590
Gene (4 G)	940 ± 80	700 ± 80	
Gene (18 G)		670 ± 60	600 ± 60
Gene (22 G)		830 ± 55	
Gene (50 G)	1350 ± 150 (est.)	993 ± 46	
Gene (22 G)		659 ± 131	
Protein (10 E)		627 ± 51	
Gene (MtDNA 18S rRNA)		588 min.	586/589 min.

molecular estimates are that – basal splits among major animal clades happened about 1000 MYA

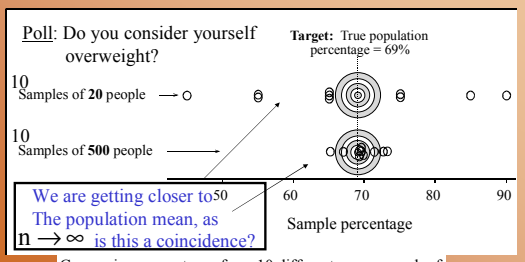
Statistics Example

- What do you think of when you hear "statistics"?
- **Definition**: *Statistics* is the science of understanding data and making decisions in the face of variability and uncertainty.
- To utilize statistics we need to understand:
  - how the data was collected
  - why it was collected
  - how to analyze and interpret the data APPROPRIATELY!

Newtonian science vs. chaotic science

- Article by Robert May, Nature, vol. 411, June 21, 2001
  - Science we encounter at schools deals with crisp certainties (e.g., prediction of planetary orbits, the periodic table as a descriptor of all elements, equations describing area, volume, velocity, position, etc.)
  - As soon as uncertainty comes in the picture it shakes the foundation of the deterministic science, because only probabilistic statements can be made in describing a phenomenon (e.g., roulette wheels, chaotic dynamic weather predictions, Geiger counter, earthquakes, etc.)
  - What is then science all about – describing absolutely certain events and laws alone, or describing more general phenomena in terms of their behavior and chance of occurring? Or may be both!

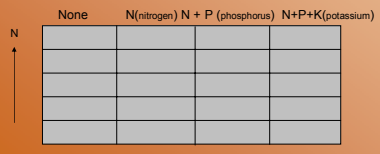
Variation in sample percentages



Comparing percentages from 10 different surveys each of 20 people with those from 10 surveys each of 500 people (all surveys from same population).

Statistics Example

Example: A plant ecologist measured the growth response of cotton grass (cm) to four different fertilizer treatments in Northern Alaska. For each treatment, five small 4 ft^2 plots were selected, all within the particular field of interest.



What points seem important from this description?

Statistics Example

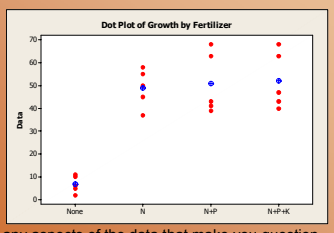
Example (cont'): The data for the experiment were:

	Fertilizer			
	None	Nitrogen	Nitrogen + Phosphorous	Nitrogen + Phosphorous + Potassium
	10	58	63	68
	6	45	43	47
	11	55	68	63
	2	50	41	43
	5	37	39	40
mean	6.8	49	50.8	52.2

- What are the important features of this data?
- Can we say that one treatment is definitively better?

Statistics Example

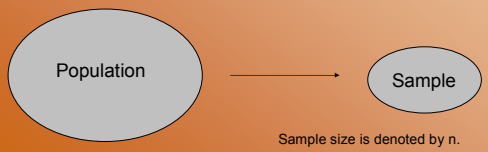
Example (cont'): Another look at the data from a visual standpoint:



- Are there any aspects of the data that make you question whether a true difference in the treatment groups exists?

Statistical Jargon

- **Definition**: A *population* is an entire group of which we want to characterize.
- **Definition**: A *sample* is a collection of observations on which we measure one or more characteristics.



Sample size is denoted by n.

Statistical Jargon

- **Definition**: A variable is a characteristic of an observation that can be assigned a number or a category.
  - For example the year in college (variable) of a student (observational unit).
- There are two types of variables:
  1. categorical and
  2. quantitative
  - these types of variables can be split further into two types…

Categorical Variables

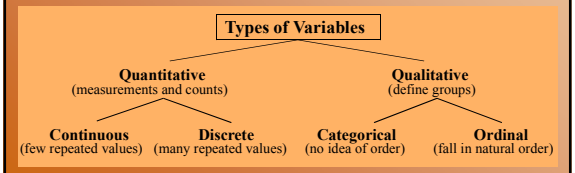
- Categorical (qualitative) variables are variables that are classified into groups.
- There are two types of categorical variables:
  - Ordinal (arranged in a meaningful order)
  - Not ordinal (no meaningful order)
- What type of categorical variable are following:
  - gender (M/F)?
  - size of soda (small, medium, large)?
  - political affiliation (democrat, republican, independent, green party, other)?

Quantitative Variables

- Quantitative variables are variables that have a meaningful numerical value.
- There are two types of quantitative variables:
  - Continuous (lies on an interval scale with infinite possible values)
  - Discrete (space between each value, countable)
- What type of quantitative variable are following:
  - weight (lbs.)?
  - height (in.)?
  - number of cars in the library parking lot?

Notation

- Y is used to denote a random variable
- y is used to denote the observations
  - subscripts, such as y_1, can be used to denote a particular observation
- What is the difference?



Using Statistical Jargon

Example: Most breast cancer patients (>80%) are over the age of 50 at diagnosis. A researcher at a particular New York cancer center believes that his patients are even older than the norm, typically older than 65 years at diagnosis. To investigate he reviews the ages of a random sample of 100 of his female patients diagnosed with breast cancer.

Using Statistical Jargon

- Identify the following:
  - Population
  - Sample
  - Sample size
  - Variable of interest
    - quantitative or qualitative?
  - Other variables
    - quantitative or qualitative?
  - Observational unit

Describing Data

- There are two ways to describe a data set:
  - Graphs and tables
  - Numbers
- Both are important for analyzing data

Graphs and Tables

- **Definition**: A *frequency distribution* is a display of the number (frequency) of occurrences of each value in a data set.
- **Definition**: A *relative frequency distribution* is a display of the percent (frequency/n) of occurrences of each value in a data set.

Graphs and Tables

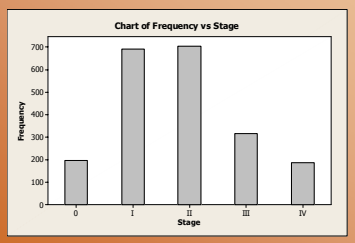
- Categorical variables
  - Easier to deal with than quantitative variables

Example: Stage of disease at diagnosis of breast cancer in a random sample of US women.

Stage	Frequency	Relative Frequency
0	197	0.09
I	691	0.33
II	703	0.34
III	314	0.15
IV	187	0.09
Total	2092	1.00

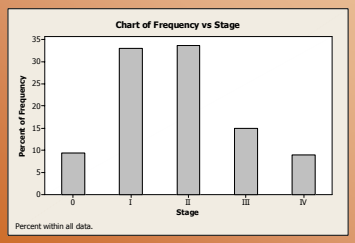
Graphs and Tables – frequency histogram

- Example: Stage of disease (cont'):



Graphs and Tables – relative histogram

- Example: Stage of disease (cont'):



Graphs and Tables

- Quantitative variables
  - need to make classes (meaningful intervals) first
  - some work needs to be done to get quantitative data into classes. One common rule of thumb is that the number of classes should be close to $\sqrt{n}$
  - important that classes are of equal width for accurate interpretation of data
- Once we have our classes we can create a frequency/relative frequency table or histogram.

Graphs and Tables

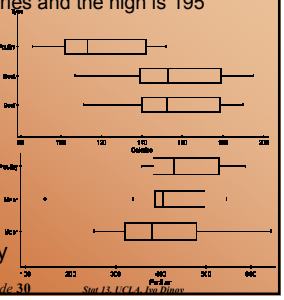
Example: People who are concerned about their health may prefer hot dogs that are low in salt and calories. The "Hot dogs" datafile (http://www.stat.ucla.edu/~dinov/courses_students.dir/05/Fall/DataFiles/) contains data on the sodium and calories contained in each of 54 major hot dog brands. The hot dogs are also classified by type: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). For now we will focus on the calories of these sampled hotdogs.

Graphs and Tables

- Example: Hotdogs (cont') Make a frequency table.
  - Overall, the low is 86 calories and the high is 195 calories

$\sqrt{n} = \sqrt{54} = 7.35 \approx 7$

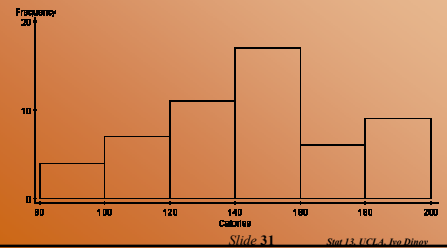
Calories	Frequency	Relative Frequency
70 - <90	2	0.04
90 - <110	7	0.13
110 - <130	3	0.06
130 - <150	21	0.39
150 - <170	6	0.11
170 - <190	10	0.18
190 - <210	5	0.09
Total	54	1.00



- Seems slightly arbitrary

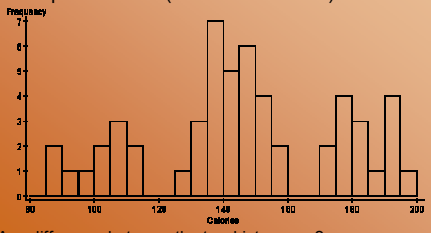
Graphs and Tables – bin-size effect

- Example: Hotdogs (cont') Histogram using previously defined classes.



Graphs and Tables – bin-size effect

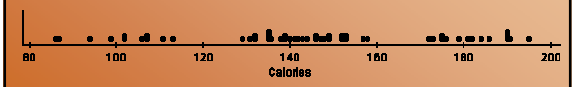
- Example: Hotdogs (cont')
  - Most of the time it is easiest to just let the computer decide (ie. use the default)



- Any difference between the two histograms?

Graphs and Tables – Dot plot on calories

- Another widely used graphical display of data is called a dot plot.
  - Looks just like it's name



Graphs and Tables

- The next graphical display we will review is called a stem and leaf display.
  - Each observation is split into a stem and a leaf
  - A good place to start is to use the last digit of the observation as the leaf and the rest as the stem

Character Stem-and-Leaf Display

```
Stem-and-leaf of Calories  N  = 54
Leaf Unit = 1.0

  2     8 67
  4     9 49
  9    10 22677
 11    11 13
 12    12 9
 22    13 1225556899
(11)   14 01234667899
 21    15 223378
 15    16
 15    17 235569
  9    18 1246
  5    19 00015
```


Graphs and Tables

- Suppose you got a stem and leaf that looked like the following.

Character Stem-and-Leaf Display

```
Stem-and-leaf of Calories  N  = 54
Leaf Unit = 1.0
  2     8 67
  3     9 4
  4     9 9
  6    10 22
  9    10 677
 11    11 13
 11    11
*** part of display removed to fit on slide
 15    17 23
 13    17 5569
  9    18 124
  6    18 6
  5    19 0001
  1    19 5
```


Graphs and Tables - Summary

- Advantages:
  - histogram: can handle large data sets
  - dot plot: can get a better picture of data values
  - stem and leaf: can see actual data values
- Disadvantages:
  - histogram:can't tell exact data values; need to set-up classes
  - dot plot: can't handle large data sets
  - stem and leaf:can't handle large data sets

The BIG Three

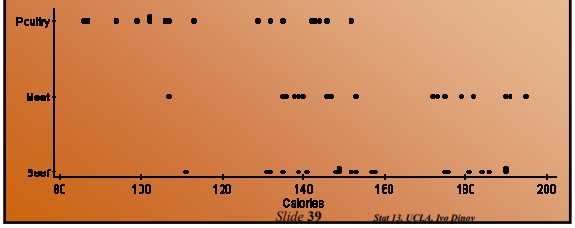
- There are three main features of data that should *always* be addressed in an analysis
  - Shape
  - Center
  - Spread

Shapes of Distributions

- The shape of a distribution can usually be determined by just looking at it as a histogram, dot plot or stem and leaf display.
- **Definition:** A distribution is *unimodal* if it has one mode
  - Unimodal distributions include:
    - Bell (symmetric, *Normal*)
    - Skewed right
    - Skewed Left
  - Other examples of distributions are:
    - Bimodal
    - Multimodal
    - Exponential

Shapes of Distributions

- What seems like a logical reason for the shape of the hot dog calorie data?
- Dot Plot for Hot-dogs: Calories vs. Type of meat:



Shapes of Distributions

- Classify and draw a sketch each of the following scenarios with respect to mode. Also, if unimodal, classify symmetry (symmetric, skewed right or skewed left).
  - Data collected on height of randomly sampled college students.
  - Data collected on height of randomly sampled female college students.
  - The salaries of all persons employed by a large university.
  - The amount of time spent by students on a difficult exam.
  - The grade distribution on a difficult exam.

Shapes of Distributions