UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

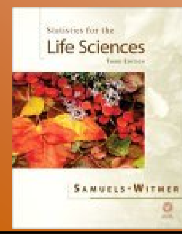
Instructor: **Ivo Dinov,**
Asst. Prof. of Statistics and Neurology

Teaching Assistants:
Jacquelina Dacosta & Chris Barr

University of California, Los Angeles, Fall 2006
http://www.stat.ucla.edu/~dinov/courses_students.html

Administrative

- The book for this course -- Statistics for the Life Sciences, by Samuels & Witmer, 3rd edition.
  - Homework will be primarily assigned from the text
  - You are responsible for keeping up with reading
  - Some chapters will be covered by reading only



Consent Forms and Surveys

- What is Statistics Research on Education?
- What is SOCR (www.SOCR.ucla.edu)?
- Fill in the Consent Form – NOW!
- **In Discussion with TA on Tuesday 10/03/06!**
  - **http://moodle.stat.ucla.edu/course/view.php?id=72**
    - **http://lab.stat.ucla.edu/accounts/MoodleLookup.php** - Stats Lab Access
    - **http://moodle.stat.ucla.edu/** → Fall 2006 → SOCR Course Resources → **Stat Lab Access (above)** → SOCR Fall 2006 FS LSI Survey
  - Complete **STATISTICAL REASONING ASSESSMENT (PRE) survey**
  - Complete **ATTITUDE SURVEY (PRE) survey**

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

Modeling the Spread of the Flu Virus

Goals: Quantify long-range dissemination of infectious diseases (e.g., flu virus)

Methods: Use influenza-related mortality data to analyze the between-state progression of interpandemic influenza in the United States over the past 30 years.

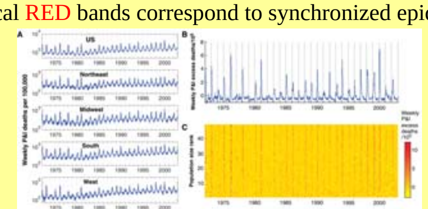
Results: Outbreaks show hierarchical spatial spread evidenced by higher pairwise synchrony between more populous states. Seasons with higher influenza mortality are associated with higher disease transmission and more rapid spread than are mild ones.

Cécile Viboud, **Ottar Bjørnstad, David Smith, Lone Simonsen, Mark Miller, Bryan Grenfell**
Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza
Science 21 April 2006: Vol. 312. no. 5772, pp. 447 – 451 DOI: 10.1126/science.1125237

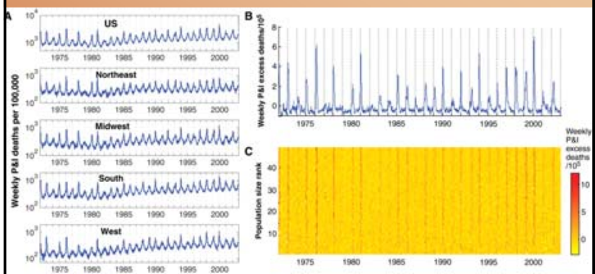
What is Statistics? A practical example

Modeling the Spread of the Flu Virus

Weekly epidemics: (**A**) Death rates from pneumonia and influenza per 100,000 population on a log10 scale. (**B** and **C**) Death rates in excess attributed to influenza in the United States (B) and by state as a color intensity plot (C). Vertical RED bands correspond to synchronized epidemics



What is Statistics? A practical example

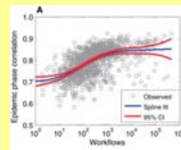


http://www.sciencemag.org/cgi/content/full/312/5772/447?maxtoshow=&HITS=20&hits=20&RESULTFORMAT=&andorexacttitle=or&andorexacttitleabs=or&fulltext=confidence+interval&andorexactfulltext=and&searchid=1&FIRSTINDEX=0&sortspec=relevance&fdate=7/1/2005&tdate=9/30/2006&resourcetype=HWCIT

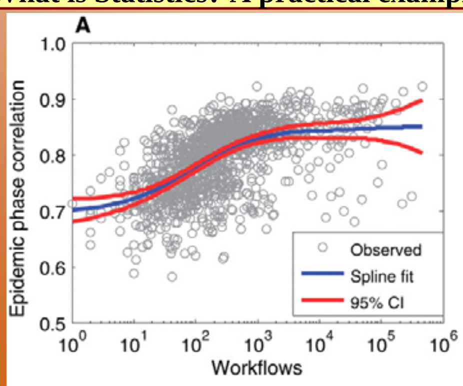
What is Statistics? A practical example

Modeling the Spread of the Flu Virus

Influenza spread and workflows. (**A**) **Gray dots** represent the observed phase synchrony in influenza epidemics (***y* axis**) plotted against the total number of individuals commuting between each pair of states (***x* axis**, log10 scale). Superimposed is the ***best fit statistical model*** (spline, red curve) and **95% confidence intervals** (CI).



What is Statistics? A practical example



What is Statistics? A practical example

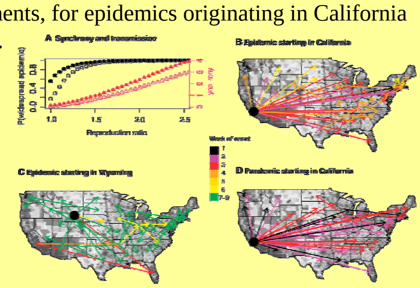
Parameter estimates for the piecewise gravity model fitted to U.S. workflow data by county. Models are fitted separately for distances above and below 119 km. *P* is the county population size; ***d* is the Euclidian distance** between the population centers of two counties; t_1, t_2, and ρ represent dependence of dispersal workflows on the population size of the donor (resident county) and recipient (work county) and the distance between them, respectively. A total of 3,109 counties in 49 continental U.S. states are used, yielding 161,710 pairs of counties with nonzero flow of workers.

Parameter	Point Estimates (Standard Error)	
	d=Distance < 119 km	d=Distances > 119 km
population of residence county (donor), t_1	0.30 ± (0.004)	0.24 ± (0.001)
population of work county (recipient), t_2	0.64 ± (0.004)	0.14 ± (0.001)
ρ distance (km)	3.05 ± (0.012)	0.29 ± (0.003)

What is Statistics? A practical example

Modeling the Spread of the Flu Virus

Simulated spread of influenza by a gravity model based on work movements, for epidemics originating in California or Wyoming.



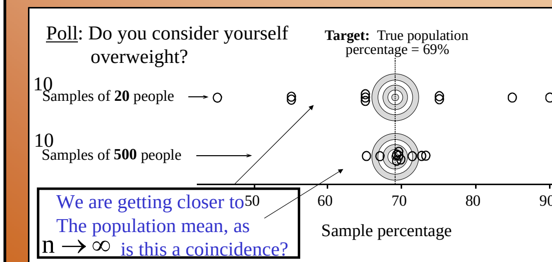
Statistics Example

- 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!

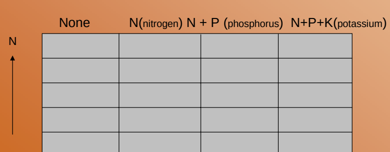
Variation in sample percentages



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

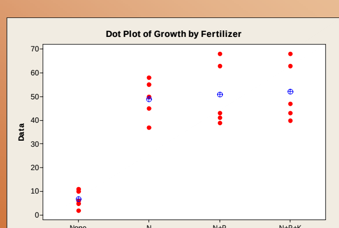
Example (cont'): The data for the experiment were:

	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

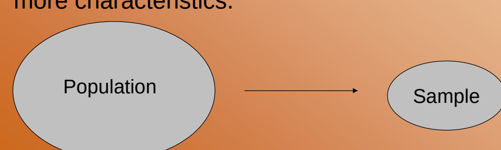
Example (cont'): Another look at the data from a visual standpoint:



- 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

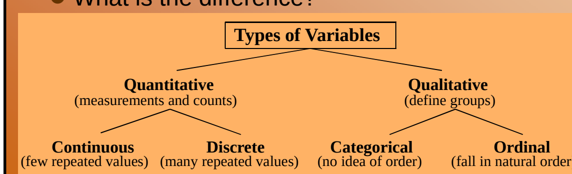
- Categorical (qualitative) variables are variables that are classified into groups.
- There are two types of categorical variables:
  - Ordinal (arranged in a meaningful order)
  - Not ordinal (no meaningful order)
- What type of categorical variable are following:
  - gender (M/F)?
  - size of soda (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

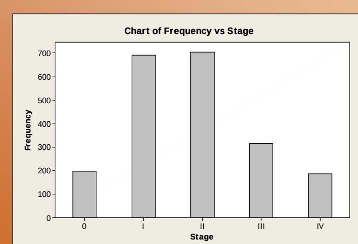
- Categorical variables
  - Easier to deal with than quantitative variables

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

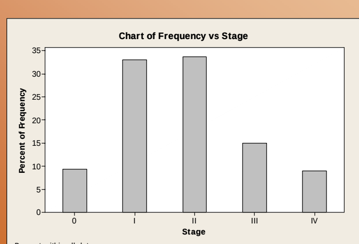
Graphs and Tables – frequency histogram

- 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

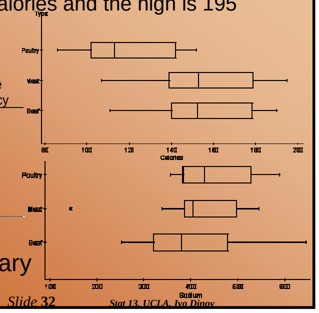
Example: People who are concerned about their health may prefer hot dogs that are low in salt and calories. The "Hot dogs" datafile (http://www.stat.ucla.edu/~dinov/courses_students.dir/05/Fall/DataFiles/) contains data on the sodium and calories contained in each of 54 major hot dog brands. The hot 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

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

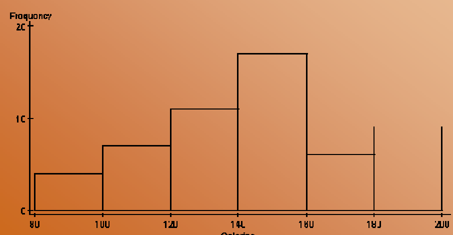
Calories	Frequency	Relative Frequency
70 - <90	2	0.04
90 - <110	7	0.13
110 - <130	3	0.06
130 - <150	21	0.39
150 - <170	6	0.11
170 - <190	10	0.18
190 - <210	5	0.09
Total	54	1.00



- Seems slightly arbitrary

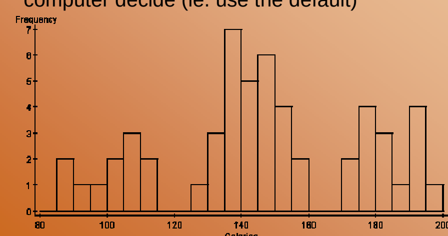
Graphs and Tables – bin-size effect

- 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

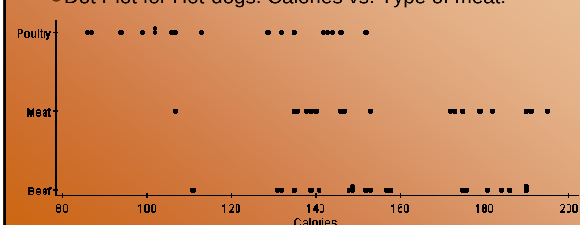
- There are three main features of data that should *always* be addressed in an analysis
  - Shape
  - Center
  - Spread

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

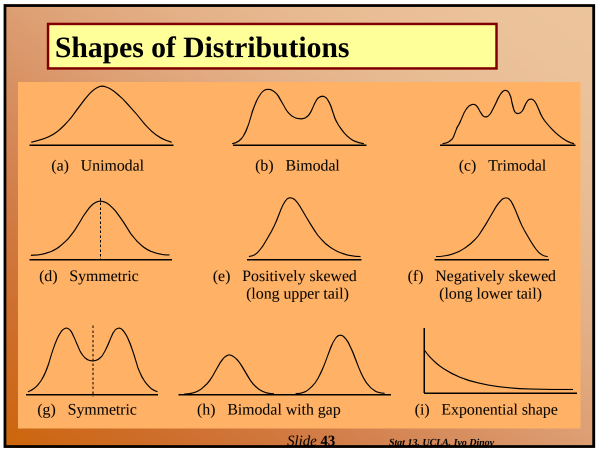
Shapes of Distributions

- 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 <u>all</u> 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
(a) Unimodal
(b) Bimodal
(c) Trimodal
(d) Symmetric
(e) Positively skewed (long upper tail)
(f) Negatively skewed (long lower tail)
(g) Symmetric
(h) Bimodal with gap
(i) Exponential shape
Slide 43
Stat 13, UCLA, Ivo Dinov