

# Medical Biometry I

(Biostatistics 511)

Instructor:

David Yanez

Cartoons and images in these notes are from

Gonick L. Cartoon Guide to Statistics. HarperPerennial, New York, 1993.

Fisher L and vanBelle G. Biostatistics: A Methodology for the Health Sciences. Wiley, New York, 1993

Special thanks to past 511 instructors, James Hughes, and Lurdes Inoue for crafting these lecture slides.

# Lecture Outline

- Course Structure
- Overview
  - Scientific method
  - Classical Introduction

# Course Structure

- Instructors: David Yanez, Ph.D.
- TAs:
  - Lisa Brown
  - Michael Garcia
  - Phillip Keung
  - Sandrine Moutou
- Time and Place:
  - Lectures: 9:30 – 10:20 am MWF HSB T-625
  - Discussion Sessions:
    - 12:30-1:20 pm M HSB T-473 (AE)
    - 8:30-9:20 am W HSB K-439 (AB)
    - 8:30-9:20 am W HSB T-747 (AD)
    - 9:30-10:20 am Th HSB T-747 (AC)
    - 8:30-9:20 am F HSB T-473 (AA)

# Assumed Prior Knowledge

- Statistical coursework
  - None
- Mathematical coursework
  - High school algebra
    - Math pre-test and solutions: *See course webpage*

# Lectures

- Recording of lectures: Camtasia
  - Audio and computer video on course webpage
  - Posted approximately the evening after the lecture
- Technologic and human errors happen
  - Please attend the lectures!

# Textbook

- Baldi and Moore (2012, 2<sup>nd</sup> ed.): The Practice of Statistics in the Life Sciences
  - Classical organization
  - (Lectures organization will follow relatively closely)
  - Used primarily as a reference
  - Great for working additional practice problems/exercises

# Computer Software

- Used extensively for data analysis
- Students may use any program that will do what is required (Stata, SPSS, SAS, R, Excel, etc.), however
  - The course TA's are well versed in Stata
  - Stata is used heavily in Biost 512-513, 536, 537, 540
  - Computer labs/exercises will be performed in Stata
  - Stata commands will be provided for homeworks and lecture examples

# Stata

- Extremely flexible statistical package
  - Interactive
  - Excellent compliment of biostatistical methods
- Graphics and output are reasonable (or not unreasonable)
- Available in the microcomputer lab (HS Library)
- Plenty of supplemental information available online
- Can be obtained at a decent discount through the UW (gradplan)
  - See the course webpage under the “Data & Stata” link



# Homework Assignments

- Weekly homework assignments: conceptual problems, analysis of real data
- Handed out Mondays (generally) and due the following Monday (generally) by 9:30 am
  - To be handed in online at the Canvas system at: <https://canvas.uw.edu/>  
The homework “DropBox” link is also provided on the course webpage, <http://courses.washington.edu/b511/>
  - If you hand the homework in on time and make a good faith effort on each question, you will receive credit for the assignment
  - Approximately 8-9 homework assignments. Students are required to complete all but one to receive full credit for their homework grade

# Discussion Section

- Supplements lectures, forum for discussion of course topics/materials, question and answer, assistance with data analysis and statistical computing.
- Participation is strongly encouraged to required.
- Holidays (Veterans Day, Thanksgiving) – affected discussion sections should plan to sit in other discussion sections.
- First week's discussion sections will be held in the HS Library computer lab (3<sup>rd</sup> floor T-wing entrance). Please make it a habit to bring a PIN drive to these sessions to save files and data.

# Grading

- Homework (8-9 assignments): 20%
- Exams (3 exams, use best 2): 45%
- Final Exam (2 hours, Dec. 11): 35%
- Assignments
  - Encouraged you to work together, but please hand-in work that is solely your own.
- Exams
  - Closed note, closed book, no electronic devices except for a basic hand calculator. One hand-written crib sheet allowed. More later.

# Course Web Pages

- Address: <http://courses.washington.edu/b511/>
- Content:
  - Syllabus, Course Schedule/Outline
  - Class Lecture Notes (full size, four per page)
  - Recorded Lectures
  - Homework Assignments (and solutions)
  - Homework DropBox
  - Datasets and Stata information
  - Miscellaneous Handouts
  - Discussion Board
  - Current Announcements

# Office Hours

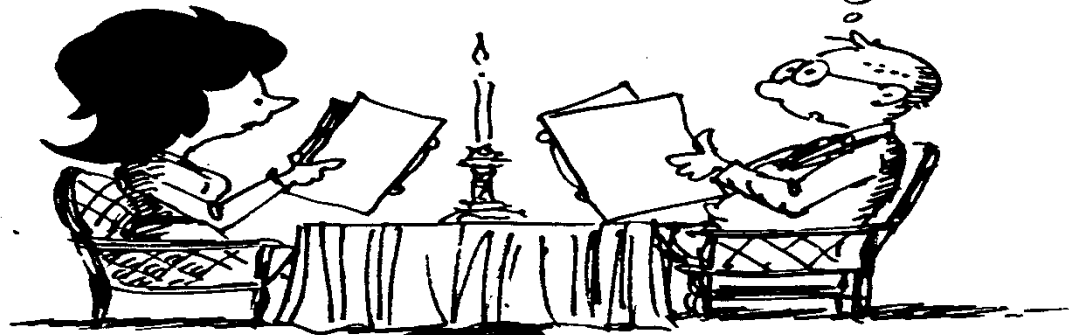
- Biost 511 students
  - Biost 511 instructors and TA's really like to see you during office hours.
  - Use of office hours (or lack thereof) may impact assignments and exams...

# What is Statistics?

WE MUDDLE THROUGH LIFE MAKING CHOICES  
BASED ON INCOMPLETE INFORMATION...

SHOULD I HAVE THE SOUP?  
EVERYTHING ELSE IS SO  
EXPENSIVE, AND I DON'T  
KNOW WHO'S PAYING... ARE  
STATISTICIANS STINGY? I'VE  
NEVER GONE OUT WITH  
ONE BEFORE... THOUGH I  
ONCE KNEW A VERY  
GENEROUS ACCOUNTANT...

SHOULD I HAVE THE SOUP?  
27 OUT OF THE 36 TIMES  
I'VE HAD IT, IT WAS PRETTY  
GOOD... BUT IS MONDAY THE  
REGULAR CHEF'S NIGHT  
OFF? AND WHAT IF ALL THE  
AIR MOLECULES IN THE  
ROOM SUDDENLY FLY UP TO  
THE CEILING?



# What is Statistics?

“Statistics is ... inference ... data ... variation...  
uncertainty”

GvB ...

1. What is the question?
2. Is it measurable?
3. Where/how will you get the data?
4. What do you think the data are telling you?

## Example – WWII Planes

What is the Question?

During WWII the British Air Force wanted to know what areas of it's fighter planes it should reinforce to prevent them from being shot down.

Is it Measureable?

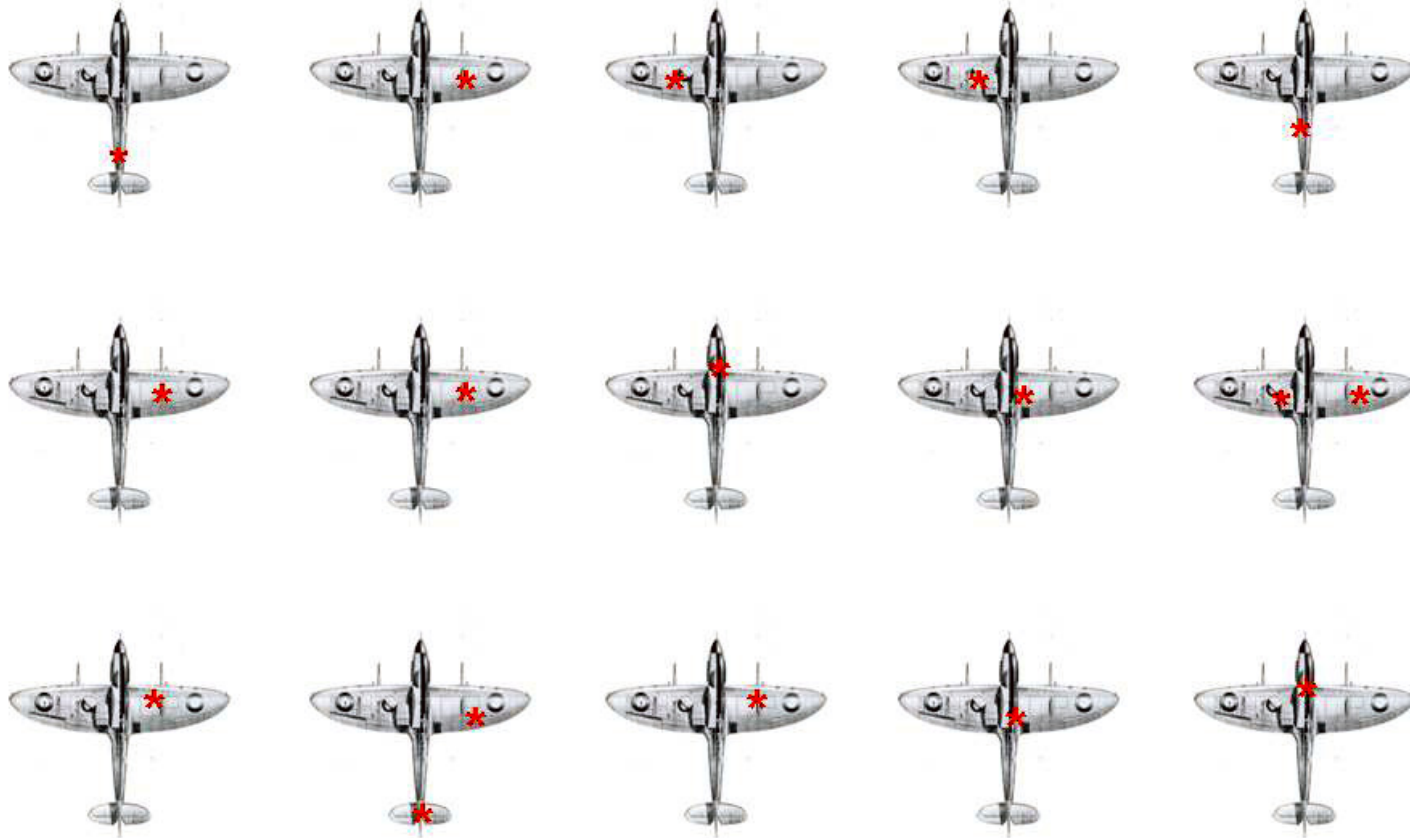


# Damaged Spitfire



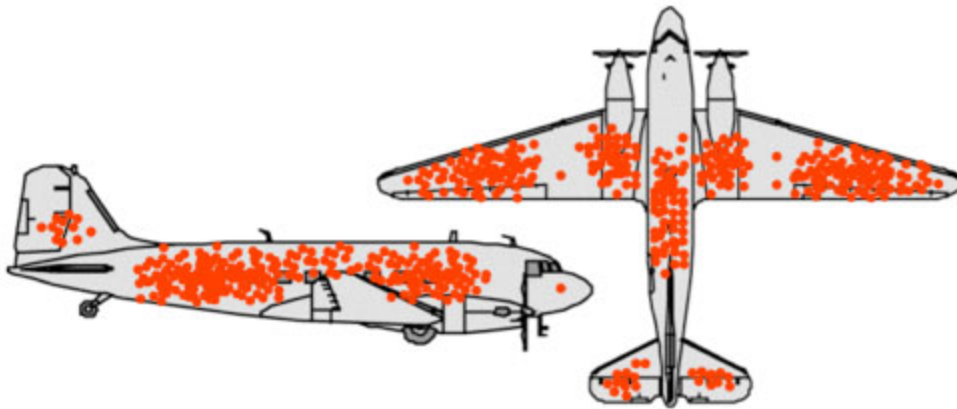
# The “Data”

- Vulnerability analysis of 400 Spitfires (15/400 shown)



# What do the data say?

- Specifically, where do you put the extra armor?
- Abraham Wald was a statistician charged with analyzing these data ...

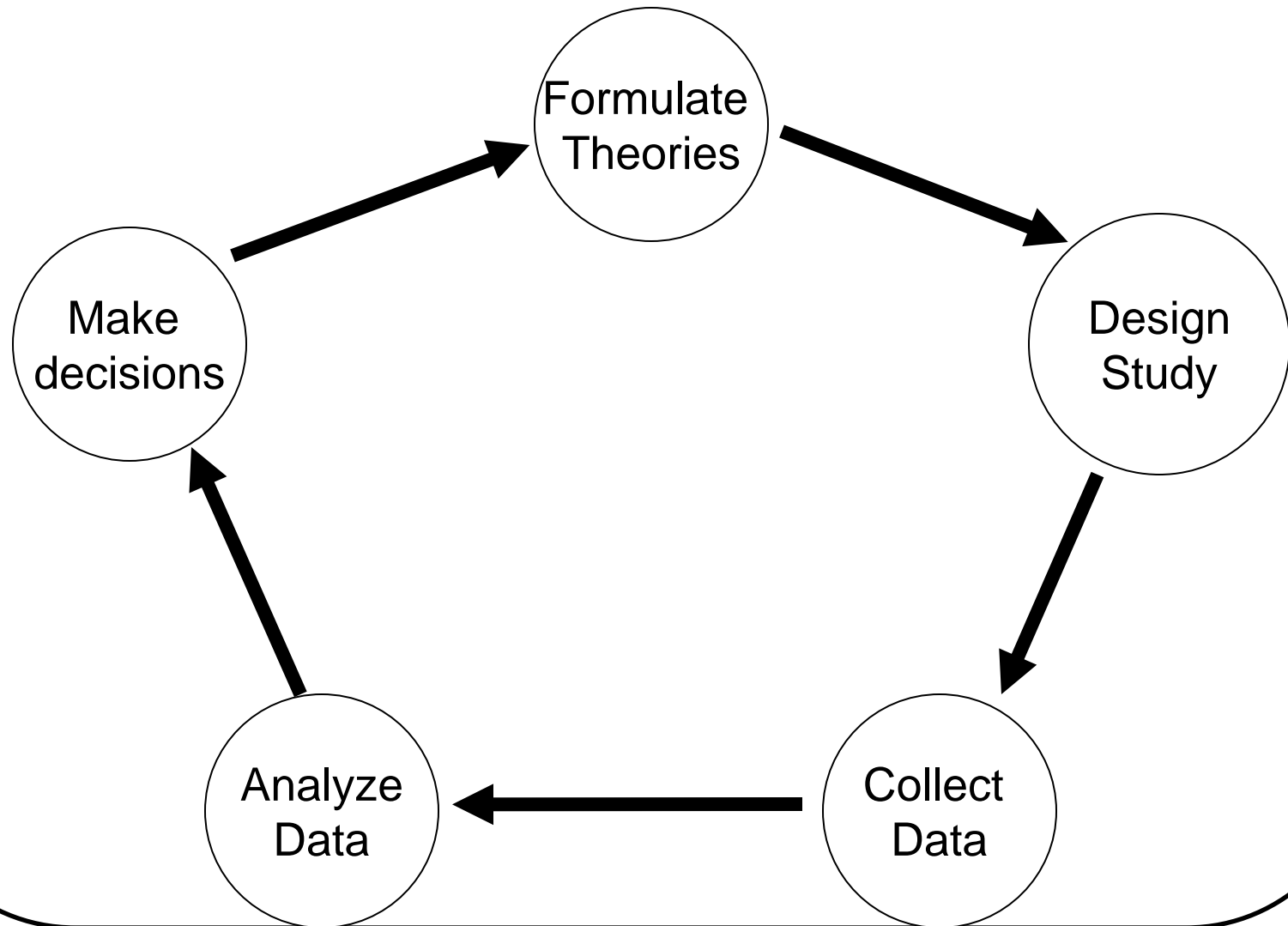


# Statistical Thinking

1. Conclusions should be based on data
2. All data have limitations
3. Variability is omnipresent
4. “All models are wrong but some are useful” (George Box)

Development of “Statistical Thinking” is as important as learning particular statistical methods.

# Scientific Method



# Producing Data

- Almost always, our data are incomplete ...
  - the data typically represent a sample from some larger population that we are really interested in
  - the data may not fully represent the population of interest
- Language and concepts associated with producing data
  - Random samples, convenience samples
  - Parameters and statistics
  - Observational vs experimental studies
  - Bias, variability, confounding

# Descriptive Statistics (EDA)

- Organization, summarization, and presentation of data
- “Exploratory data analysis is detective work - numerical detective work” – John Tukey
- Useful for generating hypotheses, finding unexpected patterns, forming new ideas (inductive reasoning)

## Tools:

- tables
- graphs
- numerical summaries

# Inferential Statistics

- Assess strength of evidence in data for/against a preconceived hypothesis (deductive reasoning)
- Make comparisons
- Make predictions
- Generalize findings from a sample to a larger population
- Powerful methods, but sensitive to assumptions

## Tools:

- Models
- Estimation and Confidence Intervals
- Hypothesis Testing



# Univariate Statistics

## Descriptive Statistics and EDA

---

- **Types of data**
  1. **Categorical**
  2. **Continuous**
- **Numerical Summaries**
  1. **Location - mean, median, mode.**
  2. **Spread - range, variance, standard deviation, IQR**
  3. **Shape - skewness**
- **Graphical Summaries**
  1. **Barplot**
  2. **Stem and Leaf plot**
  3. **Histogram**
  4. **Boxplot**
- **Mathematical Summaries**
  1. **Density curves**

# Purpose of Descriptive Analysis

- Identify missing data, errors in measurement, other data collection problems
- Assess validity of assumptions needed for formal (inferential) analyses
- Understand basic aspects of the data
  - Details of the “distribution” of each variable
  - Sizes of subgroups
  - Relationships between key variables

# Types of Data

**Data** - measurements or observations on “units of observation”

- Categorical (qualitative)
  - 1) Nominal scale - no natural order
    - gender, marital status, race
  - 2) Ordinal scale
    - severity scale, good/better/best
- Numerical (quantitative)
  - 1) Discrete - (few) integer values
    - number of children in a family
  - 2) Continuous - measure to arbitrary precision; sometimes “censored”
    - blood pressure, weight, time to event

Why bother? ⇒ PROPER DISPLAYS, PROPER ANALYSIS

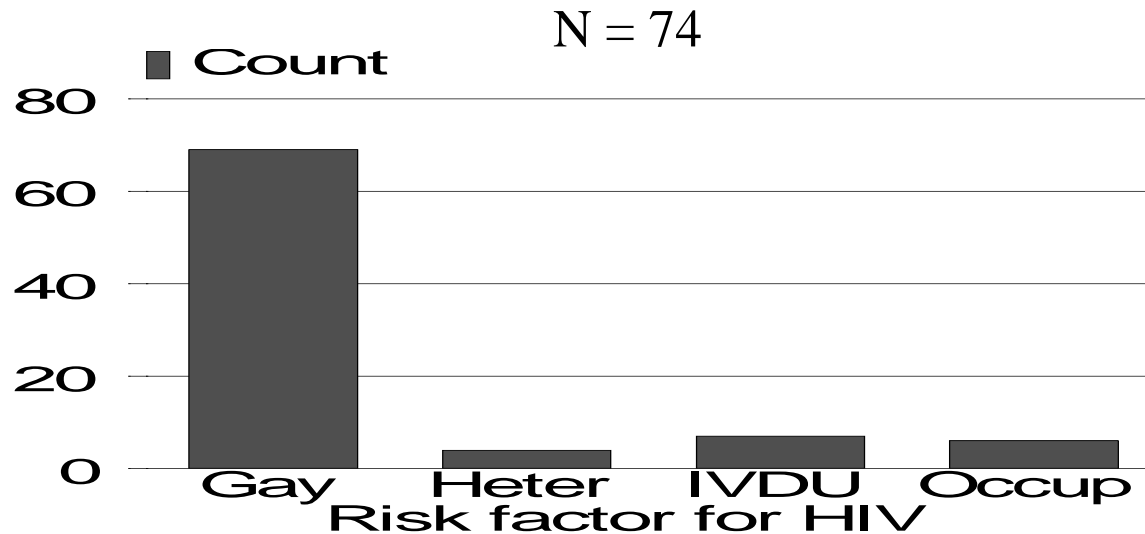
# Censored Data

Sometimes, a continuous variable may only be known to be greater than (or less than) a given amount. Such measures are said to be “censored”

- Right censored – it is only known that the true value is greater than some fixed number.
    - E.g. time to death following bypass surgery, but some people remain alive at the time of analysis.
  - Left censored - it is only known that the true value is less than some fixed number.
    - E.g. amount of selenium in a soil sample is below the detection limit of the machine
- Often, standard methods must be modified when for censored data
- “Survival analysis”, “failure time analysis”, “time to event” are terms you will hear ... see Biostat 513

# Categorical Data

Summarize categorical data with **counts** e.g. table or bar graph



Notes:

- vertical axis can be count or percent
- in the above example, counts do not add to 74 ... individuals can have multiple risk factors
- Presentation – use bar graph; paper – use table

# Continuous Data

- Barplot doesn't make sense for continuous data e.g. age
- We are more interested in the **distribution** of age:
  - where is the center of the age distribution (e.g. the average)?
  - how much does age vary?
  - are there some values far from the bulk of the data?
- What are some visual and numeric tools to help us answer these questions?

Consider the 11 ages:

21,32,34,34,42,44,46,48,52,56,64

# Stem and Leaf Diagram

We could group the data and tally the frequencies:

20: X  
30: XXX  
40: XXXX  
50: XX  
60: X

But why hide the details? Instead, we'll use the 10's place as "stems" and the units as "leaves":

2\* | 1  
3\* | 244  
4\* | 2468  
5\* | 26  
6\* | 4

The **stemplot** or **stem and leaf plot** is a quick, informative summary for small datasets.

## Stem and Leaf Diagram, construction

- All but the last digit form the stem.
- Stems are stacked vertically from the smallest to the largest.
- The leaf is the last digit in a value and is placed next to the appropriate stem (out from smallest to largest)
- Shows macro information - general shape, spread, range.
- Shows micro information - all values shown.
- Fast and easy to construct.
- Subjective decisions – rounding, splitting stems
- STATA – `stem age`
- (even better, download the `gr0028` package and use `stemplot`)



# Stem and Leaf - variations

To compare two sets of data, use a back-to-back stem and leaf diagram. Note, also, that we have “split” the stems.

Fig 1. Systolic blood pressure after 12 weeks treatment with daily calcium supplement or placebo

<u>Placebo</u>		<u>Calcium</u>
8	9*	
2	10*	2
9	10*	77
4220	11*	0122
97	11*	
3	12*	3
	12*	9
0	13*	
	13*	6

# Methods for Grouped Data

The stem and leaf effectively groups continuous data into intervals. Let's extend this idea. The following terms are useful for grouped data:

- frequency - the number of times the value occurs in the data.
- cumulative frequency - the number of observations that are equal to or smaller than the value.
- relative frequency - the % of the time that the value occurs (frequency/N).
- cumulative relative frequency - the % of the sample that is equal to or smaller than the value (cumulative frequency/N).

# Example - Birthweights

Sample of 100 birthweights in ounces. Complete the following table ...

Interval	Midpt	Freq.	Cum. Freq.	Rel. Freq.	Cum. Rel. Freq.
$29.5 \leq W < 69.5$	49.5	5			
$69.5 \leq W < 89.5$	79.5	10			
$89.5 \leq W < 99.5$	94.5	11			
$99.5 \leq W < 109.5$	104.5	19			
$109.5 \leq W < 119.5$	114.5	17			
$119.5 \leq W < 129.5$	124.5	20			
$129.5 \leq W < 139.5$	134.5	12			
$139.5 \leq W < 169.5$	154.5	6			

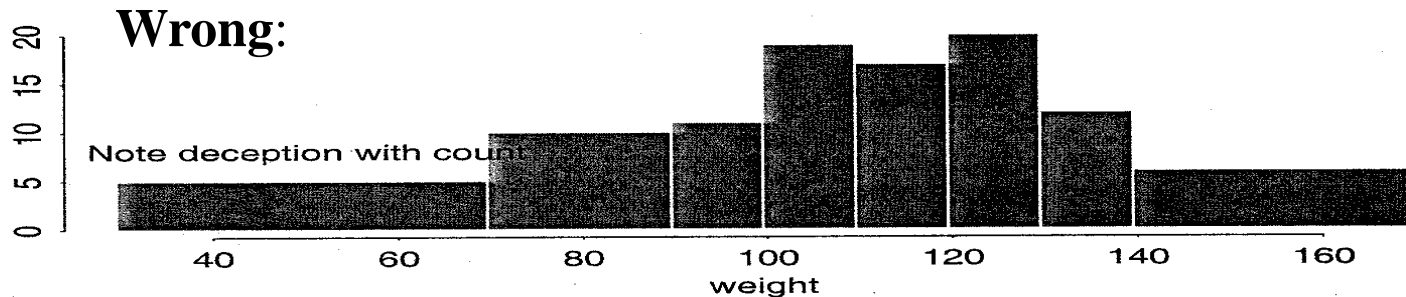
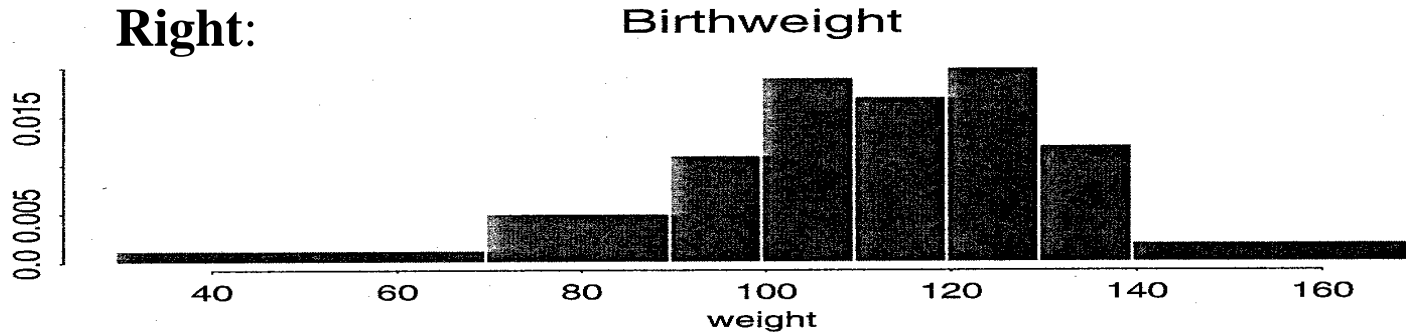
## Stata:

```
gen bwtcat = bwt
recode bwtcat min/69=1 69/89=2 89/99=3 99/109=4 109/119=5 119/129=6
129/139=7 139/max=8
tabulate bwtcat
```

# Histograms

- Similar to a barplot, but used for continuous data.
- Divide the data into intervals.
- A rectangle is constructed with the base being the interval end-points and the height chosen so the *area of the rectangle is proportional to the frequency* (if the width is one unit for all intervals, then  $\text{height} \propto \text{frequency}$ ).
- Shape can be sensitive to number and choice of intervals (rule of thumb: number of bins is smaller of  $\sqrt{n}$  or  $10 \cdot \log_{10} n$ )
- Histograms are more effective for moderate to large datasets.

# Example - Birthweights



Note: You can determine relative frequency ( $= \text{height} * \text{width}$ ) and cumulative relative frequency from a histogram.

# Characteristics of Distributions

## **Shape**

number of modes (peaks)

symmetry

## **Center**

where is the center?

## **Spread**

how much variation?

outliers?

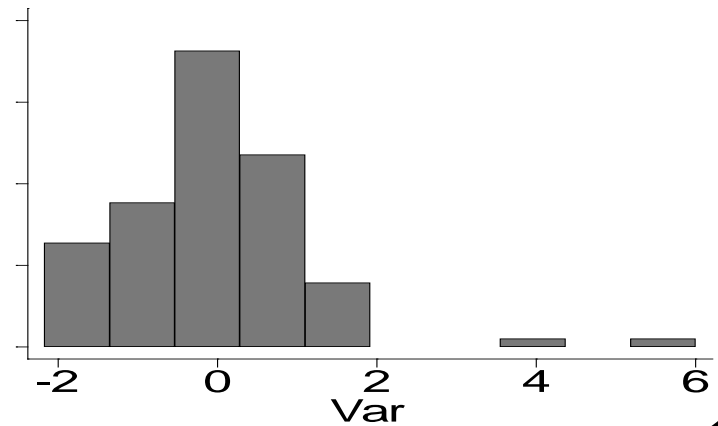
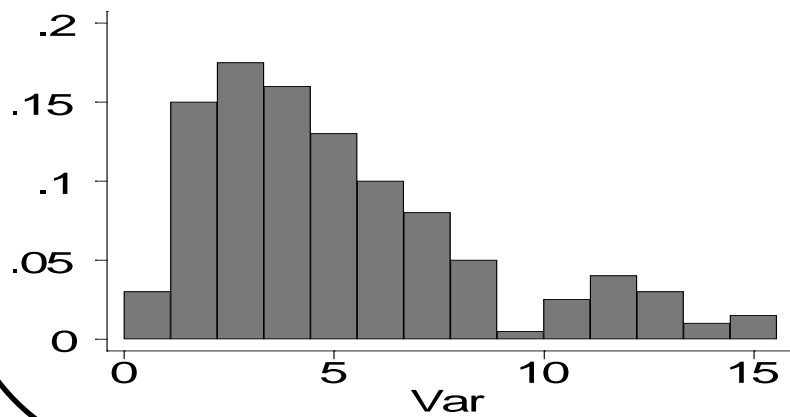
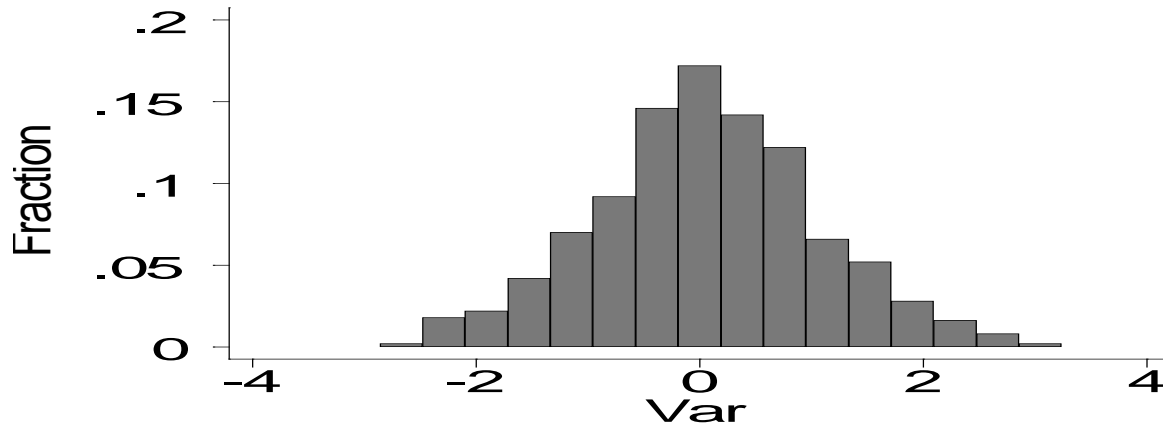
## **Other features**

boundaries

digit preference

...

# Example Distributions



# Notation

Suppose we have  $N$  measurements of a particular variable. We will denote these  $N$  measurements as:

$$X_1, X_2, X_3, \dots, X_N$$

where  $X_1$  is the first measurement,  $X_2$  is the second, etc.

Sometimes it is useful to order the measurements. We denote the ordered measurements as:

$$X_{(1)}, X_{(2)}, X_{(3)}, \dots, X_{(N)}$$

where  $X_{(1)}$  is the smallest value and  $X_{(N)}$  is the largest.



# Arithmetic Mean

The **arithmetic mean** is the most common measure of the **central location** of a sample. We use  $\bar{X}$  to refer to the mean and define it as:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

The symbol  $\Sigma$  is shorthand for “*sum*” over a specified range. For example:

$$\sum_{i=1}^4 X_i = (X_1 + X_2 + X_3 + X_4)$$

# Some Properties of the Mean

Often we wish to **transform** variables. Linear changes to variables (i.e.  $Y = a*X+b$ ) impact the mean in a predictable way:

(1) Adding (or subtracting) a constant to all values:

$$\begin{aligned} Y_i &= X_i + c \\ \bar{Y} &= \end{aligned}$$

(2) Multiplication (or division) by a constant:

$$\begin{aligned} Y_i &= cX_i \\ \bar{Y} &= \end{aligned}$$

Example: Convert mean 25°C to °F

Does this nice behavior happen for any change? NO! (show that  $\log \bar{X} \neq \overline{\log X}$  )

# Median

Another measure of central tendency is the **median** - the “middle one”. Half the values are below the median and half are above. Given the ordered sample,  $X_{(i)}$ , the median is:

N odd:            Median =  $X_{\left(\frac{N+1}{2}\right)}$

N even:            Median =  $\frac{1}{2}\left(X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}\right)$

# Mode

The **mode** is the most frequently occurring value in the sample.

## Example: Central Location

Suppose the ages in years of the first 10 subjects enrolled in your study are:

34,24,56,52,21,44,64,34,42,46

**Mean :**

$$\begin{aligned}\bar{X} &= (34 + 24 + 56 + 52 + 21 + 44 + 64 + 34 + 42 + 46) / 10 \\ &= 417 / 10 \\ &= 41.7 \text{ years}\end{aligned}$$

**Median:**

order the data: 21,24,34,34,42,44,46,52,56,64

$$\begin{aligned}\text{Median} &= \frac{1}{2} \left( X_{\left(\frac{10}{2}\right)} + X_{\left(\frac{10}{2}+1\right)} \right) \\ &= \frac{1}{2} (42 + 44) \\ &= 43 \text{ years}\end{aligned}$$

**Mode:** 34 years.

## Example (cont.)

Suppose the next patient enrolls and their age is 97 years.

How do the mean and median change?

$$\begin{aligned}\bar{X} &= (34 + 24 + 56 + 52 + 21 + 44 + 64 + 34 + 42 + 46 + 97) / 11 \\ &= 514 / 11 \\ &= 46.7 \text{ years}\end{aligned}$$

To get the median, order the data:

21, 24, 34, 34, 42, 44, 46, 52, 56, 64, 97

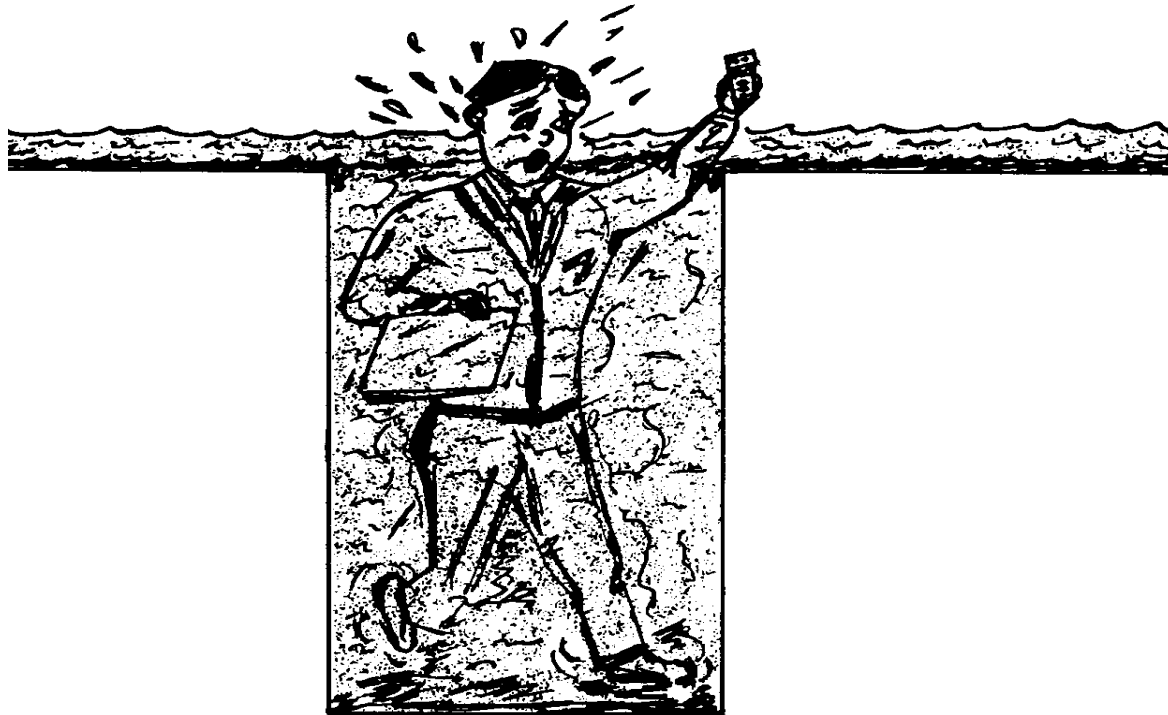
$$\begin{aligned}\text{Median} &= X_{(6)} \\ &= 44 \text{ years}\end{aligned}$$

If the new age was recorded incorrectly as 977, instead of 97, what would the new median be? What would the new mean be?

# Comparisons: Mean and Median

- Mean is sensitive to a few very large (or small) values - “outliers”
- Median is “resistant” to outliers
- Mean is attractive mathematically
- 50% of sample is above the median, 50% of sample is below the median.
- Note that a proportion is simply a mean of 0/1 data (e.g. 0 = no disease; 1 = disease)

# Variation is important!



# Descriptive Statistics and Exploratory Data analysis - Univariate

- **Types of data**
  1. Categorical
  2. Continuous
- **Numerical Summaries**
  1. Location - mean, median, mode.
  2. Spread - range, variance, standard deviation, IQR
  3. Shape - skewness
- **Graphical Summaries**
  1. Barplot
  2. Stem and Leaf plot
  3. Histogram
  4. **Boxplot**
- **Mathematical Summaries**
  1. Density curves



# Continuous Data

- Barplot doesn't make sense for continuous data e.g. age.
- We are more interested in the **distribution** of age:
  - where is the center of the age distribution (e.g. the average)?
  - how much does age vary?
  - are there some values far from the bulk of the data?
- What are some visual and numeric tools to help us answer these questions?

Consider the 11 ages:

21,32,34,34,42,44,46,48,52,56,64

# Measures of Spread: Range

The **range** is the difference between the largest and smallest observations:

$$\begin{aligned}\text{Range} &= \text{Maximum} - \text{Minimum} \\ &= X_{(N)} - X_{(1)}\end{aligned}$$

Alternatively, the range may be denoted as the pair of observations (more useful for quality control):

$$\begin{aligned}\text{Range} &= (\text{Minimum}, \text{Maximum}) \\ &= (X_{(1)}, X_{(N)})\end{aligned}$$

Disadvantage: the range typically increases with increasing sample size – hard to compare ranges from samples of different size

In the ages example, for the first 10 subjects, the range is

$$\begin{aligned}\text{Range} &= 64 - 21 = 43 \\ &\text{or } (21, 64)\end{aligned}$$

# Measures of Spread: Variance

Consider the following two samples:

20,23,34,26,30,22,40,38,37

30,29,30,31,32,30,28,30,30

These samples have the same mean and median, but the second is much less variable. The average “distance” from the center is quite small in the second. We use the **variance** to describe this feature:

$$\begin{aligned}s^2 &= \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \\ &= \frac{1}{N-1} \left( \sum_{i=1}^N X_i^2 - N\bar{X}^2 \right)\end{aligned}$$

The standard deviation is simply the square root of the variance:

$$\text{standard deviation} = s = \sqrt{s^2}$$

# Measures of Spread: Variance

For the first sample, we obtain:

2*		023	
2.		6	var = 59.25 yr <sup>2</sup>
3*		04	sd = 7.7 yr
3.		78	
4*		0	

For the second sample, we obtain:

2*			
2.		89	var = 1.25 yr <sup>2</sup>
3*		0000012	sd = 1.1 yr
3.			
4*			

# Properties of the variance and standard deviation

- Variance and standard deviation are **ALWAYS** greater than or equal to zero.
- Linear changes are a little trickier than they were for the mean:

(1) Add/subtract a constant:  $Y_i = X_i + c$

$$S_Y^2 = S_X^2$$

(2) Multiply/divide by a constant:  $Y_i = c \times X_i$

$$S_Y^2 = c^2 \times S_X^2$$

Example: Variance in °C is 25 degrees<sup>2</sup>; what is variance in °F?

- So what happens to the standard deviation?

# Measures of Spread: Quantiles & Percentiles

**Quartiles** are the (25,50,75) percentiles. The **interquartile range (IQR)** is  $Q_{.75} - Q_{.25}$  and is another useful measure of spread. The middle 50% of the data is found between  $Q_{.25}$  and  $Q_{.75}$ .

$Q_{.25}$  – median of the observations to the left (less than) the overall median.

$Q_{.75}$  – median of the observations to the right (less than) the overall median.

20,22,23,26,30,34,37,38,40

```
. centile age, centile(25 50 75)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
age	9	25	22.5	20	32.45832*
		50	30	22.07778	37.92222
		75	37.5	27.54168	40*

# Measures of Spread: Quantiles & Percentiles

More generally, define the **p'th percentile** as the value which has p% of the sample values less than or equal to it.

To find the p'th percentile, let  $k = p * N / 100$ .

(1) If k is an integer, pth percentile is the average of  $X_{(k)}$  and  $X_{(k+1)}$ .

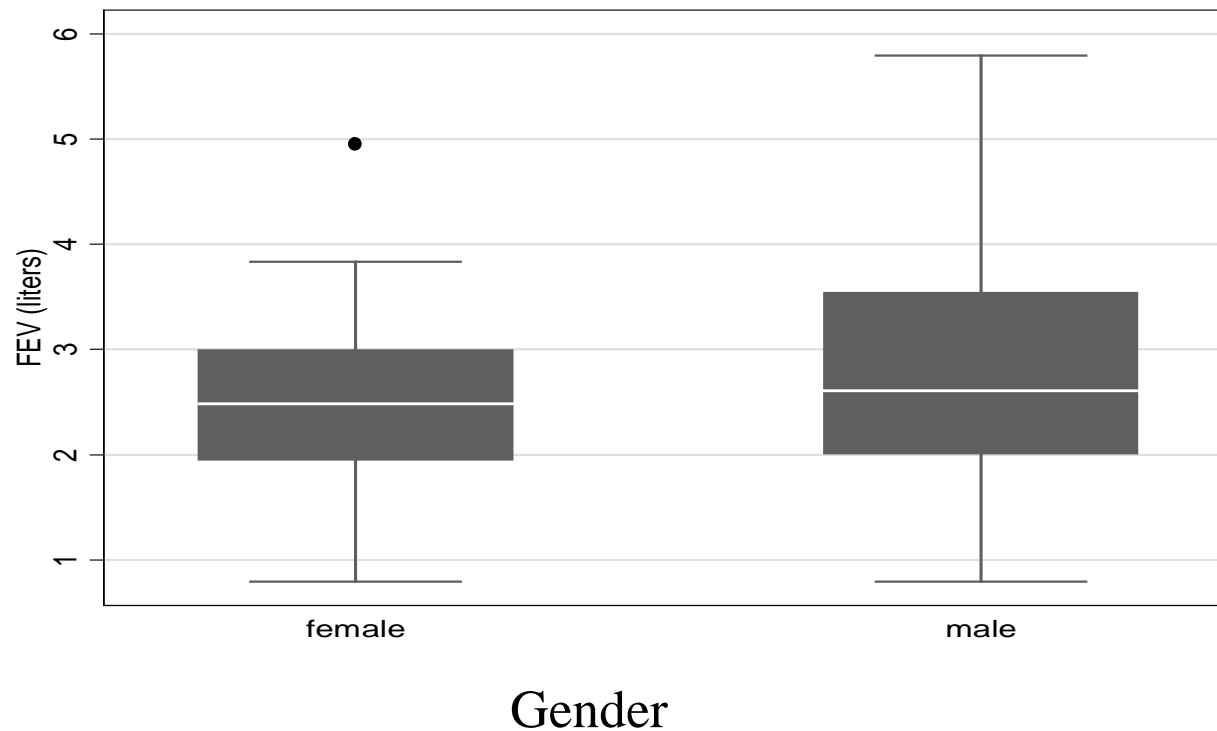
(2) If k is not an integer, pth percentile is  $X_{([k]+1)}$ .

[k] is the largest integer smaller than k (i.e. truncate the decimal).

Note: may not always agree with Stata result

# Boxplot

A graphical display of the quartiles of a dataset, as well as the range. Extremely large or small values are also identified.



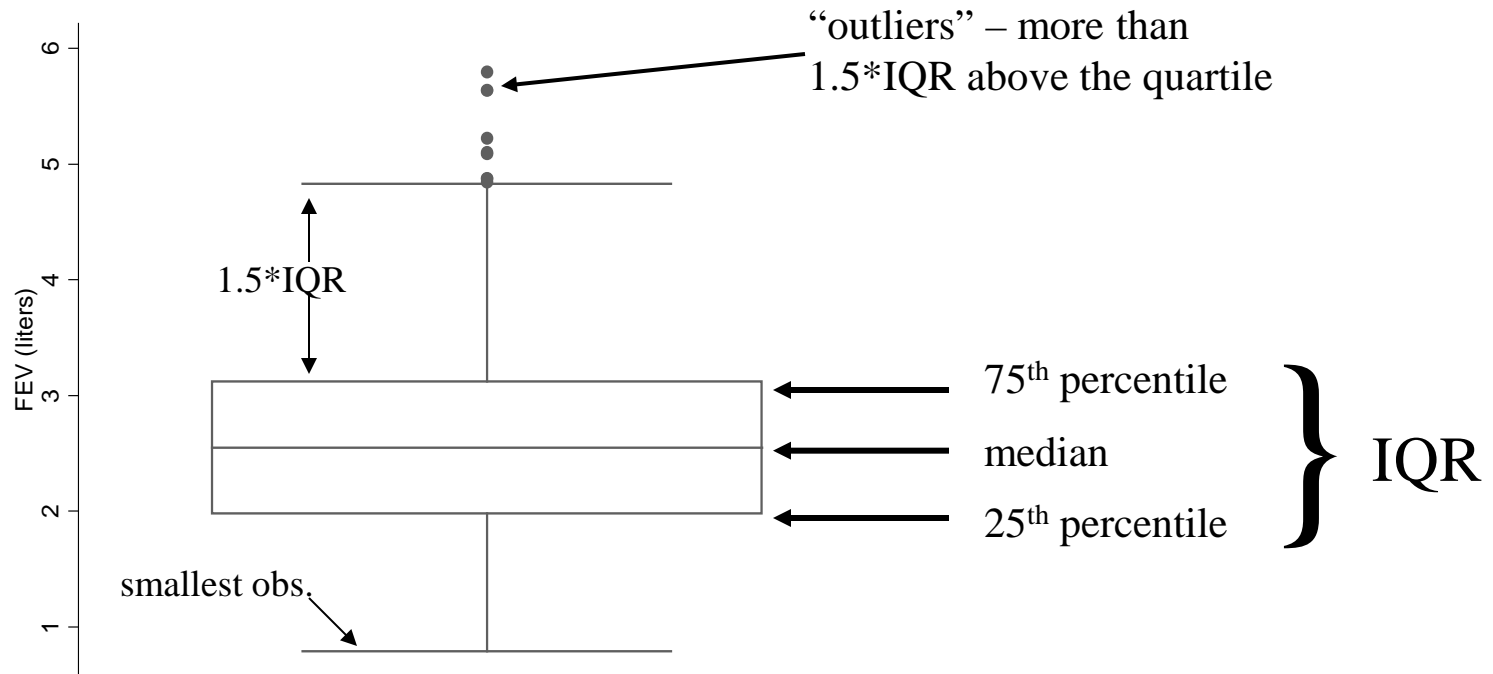
```
. graph box fev, over(sex)
```



# Boxplot: Construction

Define an **outlier** as any observation more than  $1.5 \times \text{IQR}$  above/below the quartile

The “whiskers” extend to the smallest/largest non-outlying observations.



# Boxplot: variations

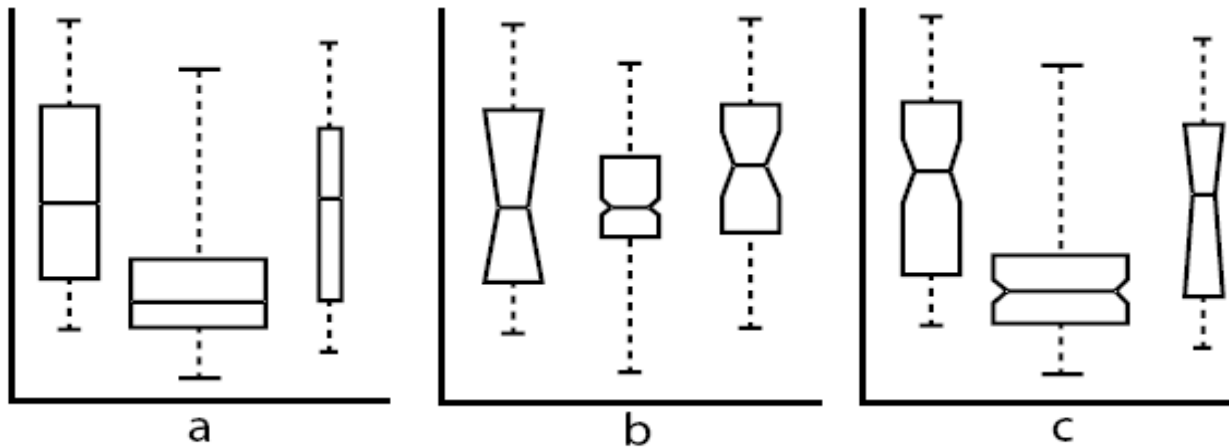
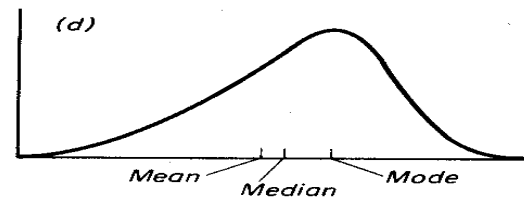
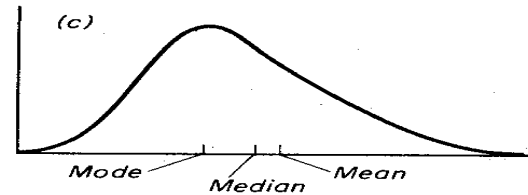
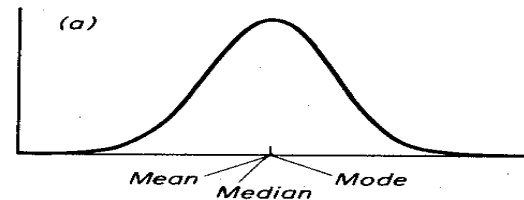


Figure 4: Variations of the box plot. a) Variable width box plot. b) Notched box plot. c) Variable width notched box plot.

# Skewness

Both histograms and boxplots can show us that a distribution is skewed. **Skewness** refers to the symmetry or lack of symmetry in the shape of the distribution. Neither the mean nor the variance tell us about symmetry.

1. “symmetric”;  
median = mean
2. “positive” or “right” skewed;  
median < mean
3. “negative” or “left” skewed;  
median > mean



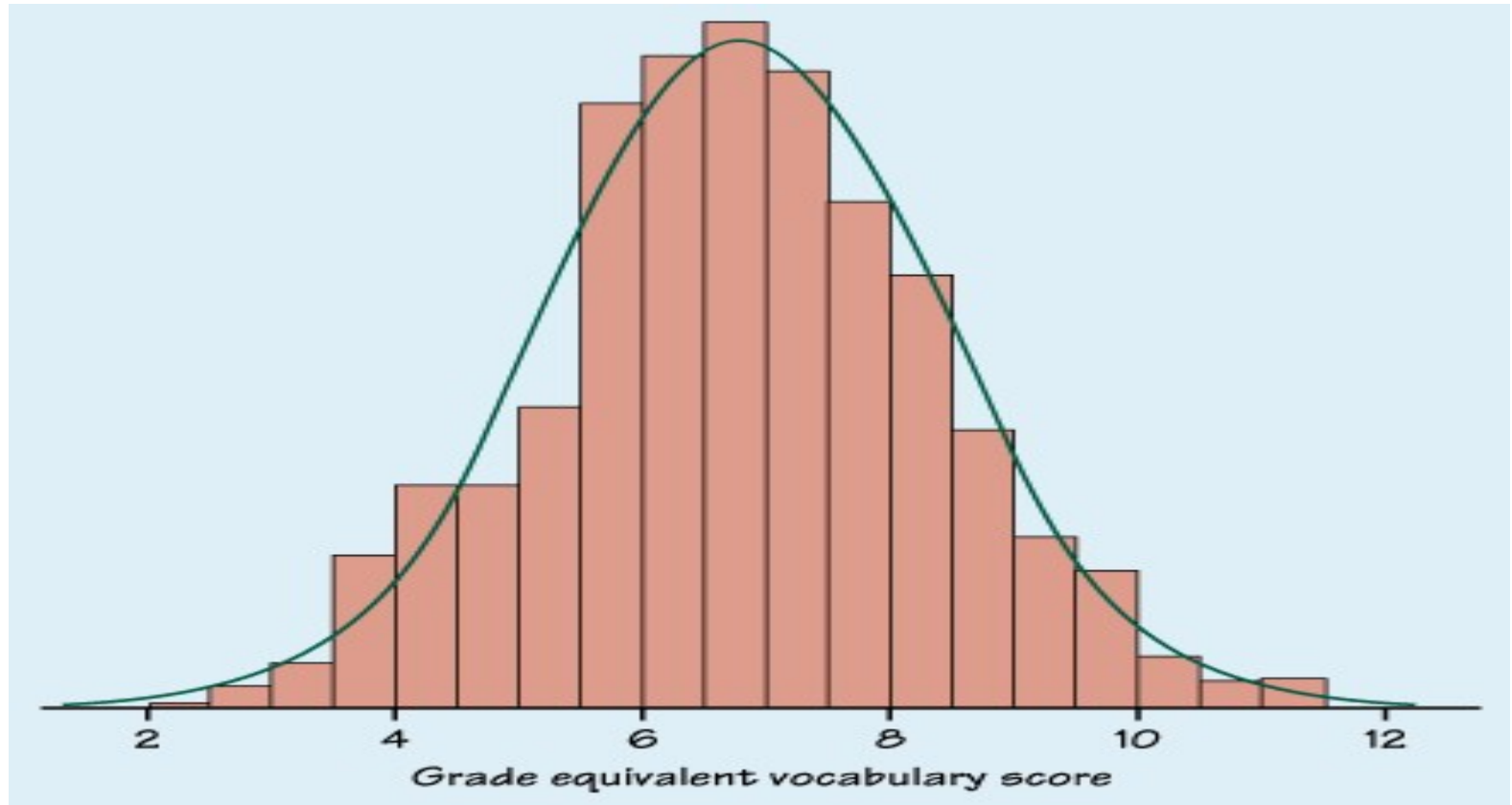
# Density Curves

We have seen how continuous data can be summarized with a **histogram**. Although histograms are summaries of the data, they still involve keeping track of a lot of numbers (i.e. the height and location of each bar). Also, histograms tend to be pretty jagged unless the dataset is reasonably large.

**Q:** Is there a way to smooth out the histogram and perhaps summarize the entire distribution of data with just a few numbers?

**A:** YES! We can use a type of mathematical model known as a **density curve**.

# Density Curves



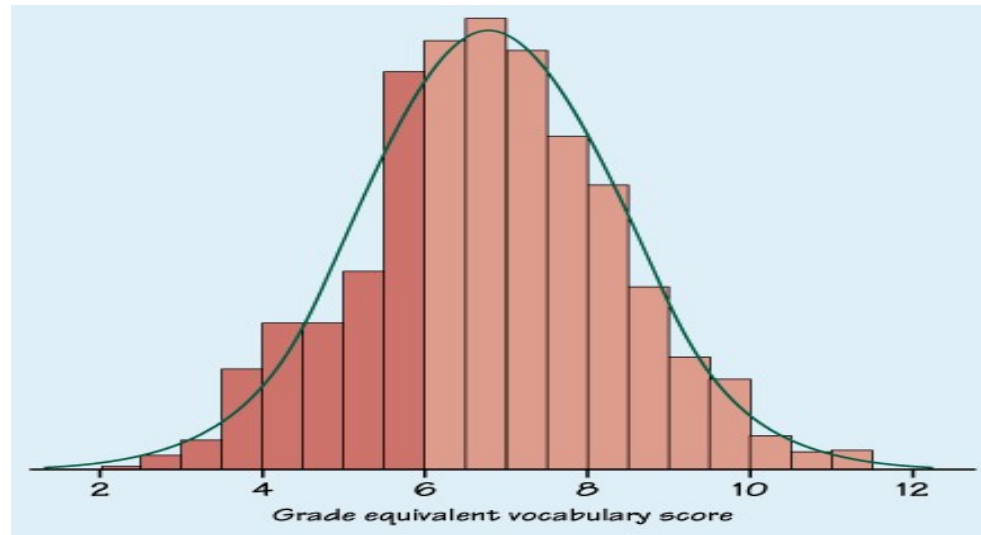
# Density Curves

We saw previously that we can use a histogram to determine the relative frequency (= proportion = **probability**) of obtaining observations in a particular interval.

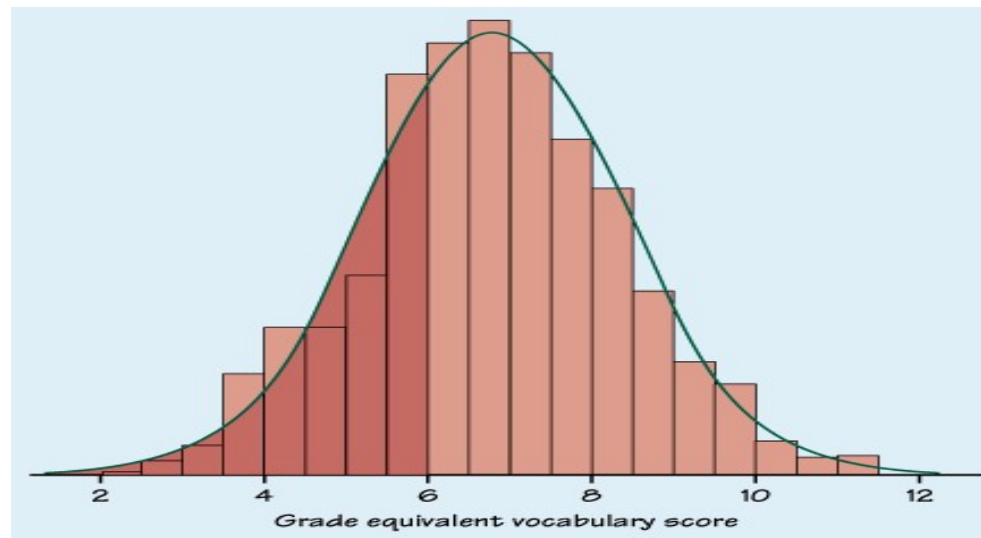
If a particular density curve provides a good fit to our data then we can use the density curve to approximate these probabilities. In particular, the probability of obtaining an observation in a particular interval is given by the **area** under the density curve.

Note: For continuous data, it does not make sense to talk about the probability of an individual value (i.e.  $P(X = 6) \approx 0.0$ )

Relative  
frequency of  
scores less than 6  
from histogram =  
.303



Probability of scores  
less than 6 from  
density curve = .293



# Probability Density Function (PDF)

1. A function, typically denoted  $f(x)$ , that gives probabilities based on the **area** under the curve.
2.  $f(x) \geq 0$
3. Total area under the function  $f(x)$  is 1.0.  $\int f(x)dx = 1.0$

# Cumulative Distribution Function (CDF)

The cumulative distribution function,  $F(t)$ , tells us the total probability less than some value  $t$ .

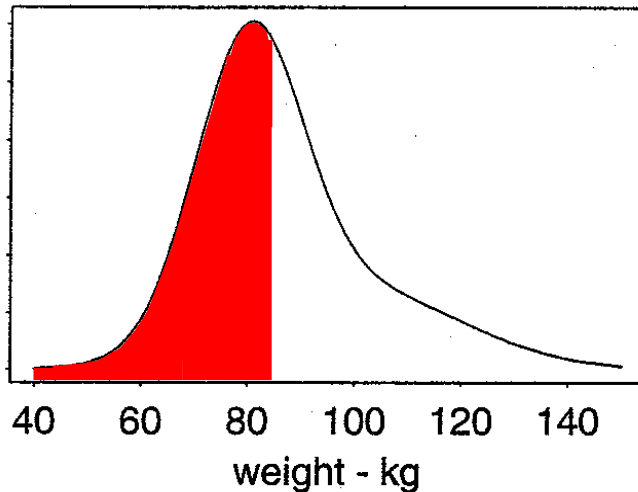
$$F(t) = P(X \leq t)$$

This is analogous to the cumulative relative frequency.

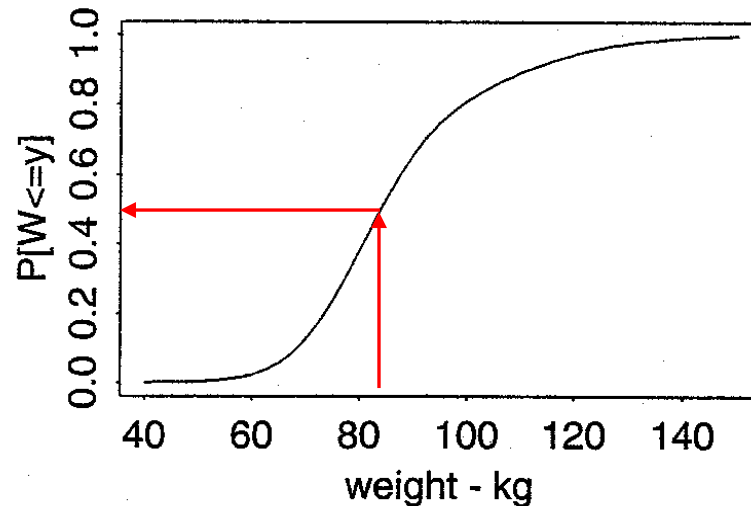


# Examples

Weight, males 30-40



Cumulative Dist Fn



The area under the curve on the PDF (left figure) represents the probability of being less than 82 lbs. In this example, the probability is approximately 0.5. This value corresponds to the probability of weighing less than 82 lbs. on the CDF (right figure).

# Gauss' Normal Distribution



**THE  
NORMAL  
LAW OF ERROR  
STANDS OUT IN THE  
EXPERIENCE OF MANKIND  
AS ONE OF THE BROADEST  
GENERALIZATIONS OF NATURAL  
PHILOSOPHY ♦ IT SERVES AS THE  
GUIDING INSTRUMENT IN RESEARCHES  
IN THE PHYSICAL AND SOCIAL SCIENCES AND  
IN MEDICINE AGRICULTURE AND ENGINEERING ♦  
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE  
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT**

# Normal Distribution

- A common model for continuous data

- Bell-shaped curve

⇒ takes values between  $-\infty$  and  $+\infty$

⇒ unimodal, symmetric about mean

⇒ mean=median=mode

- Examples

birthweights

blood pressure

CD4 cell counts (perhaps transformed)

# Normal Distribution

Specifying the mean and variance of a normal distribution completely determines the probability distribution function and, therefore, all probabilities (just 2 numbers!).

The **normal probability density function** is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$$

where  $\pi \approx 3.14$  (a constant)

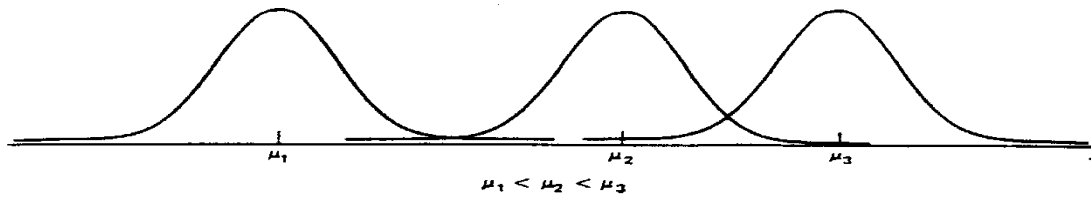
Notice that the normal distribution has two parameters:

$\mu$  = the mean of X

$\sigma$  = the standard deviation of X

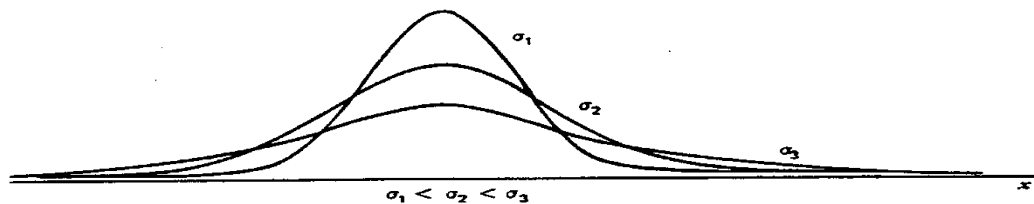
We write  $X \sim N(\mu, \sigma^2)$ . The **standard normal** distribution is a special case where  $\mu = 0$  and  $\sigma = 1$ .

# Examples



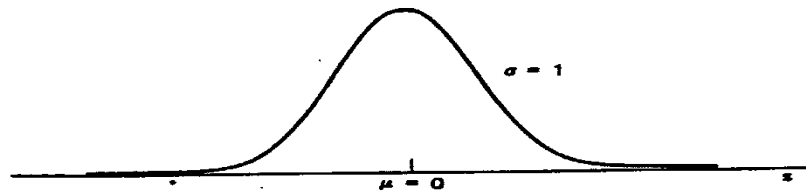
$\mu_1 < \mu_2 < \mu_3$

**FIGURE 3.6.3**  
**Three Normal Distributions with Different Means**



$\sigma_1 < \sigma_2 < \sigma_3$

**FIGURE 3.6.4**  
**Three Normal Distributions with Different Standard Deviations**

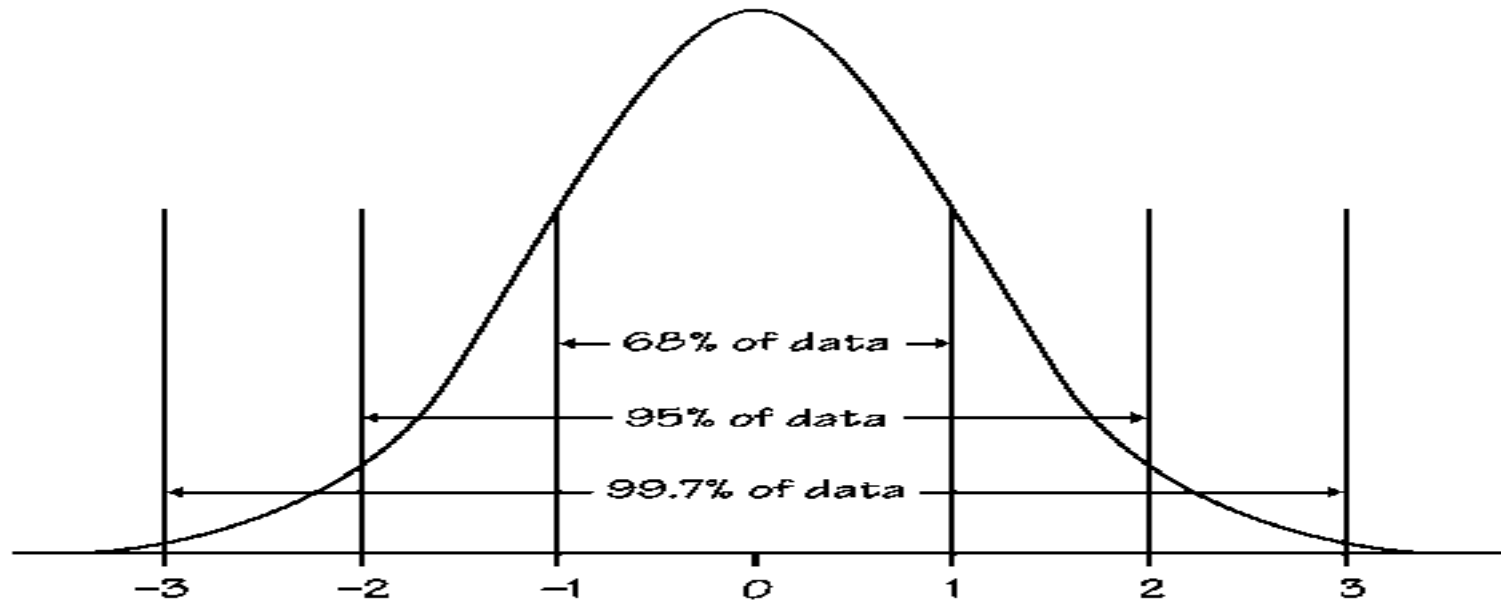


$\sigma = 1$

**FIGURE 3.6.5**  
**The Unit Normal Distribution**

4

# Standard Normal Dist<sup>n</sup>



In general,

~68% of data within  $\pm 1\sigma$  of  $\mu$

~95% of data within  $\pm 2\sigma$  of  $\mu$

~99.7% of data within  $\pm 3\sigma$  of  $\mu$

# Summary

- Types of data
  1. Categorical
  2. Continuous
- Numerical Summaries
  1. Location - mean, median, mode.
  2. Spread - range, variance, standard deviation, IQR
  3. Shape - skewness
- Graphical Summaries
  1. Barplot
  2. Stem and Leaf plot
  3. Histogram
  4. Boxplot
- Mathematical Summaries
  1. Density curves

# Descriptive Statistics and Exploratory Data Analysis – Bivariate/Multivariate

- **Quantitative Data**
  1. Scatterplots
  2. Starplot
  3. Correlation/Regression
- **Qualitative Data**
  4. Two-way (contingency) tables
- **Effect modification**

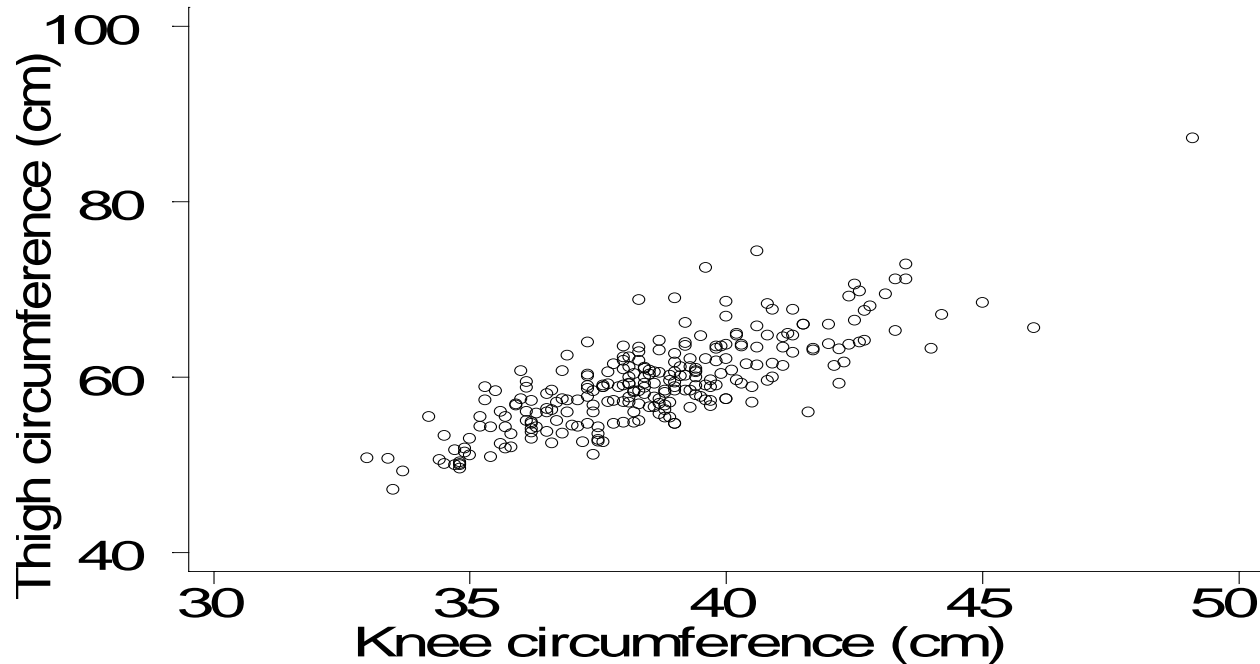


# Purpose of descriptive analysis

- Identify missing data, errors in measurement, other data collection problems
- Assess validity of assumptions needed for formal (inferential) analyses
- Understand basic aspects of the data
  - Details of the “distribution” of each variable within subgroups
  - **Relationships between key variables – associations, effect modification**

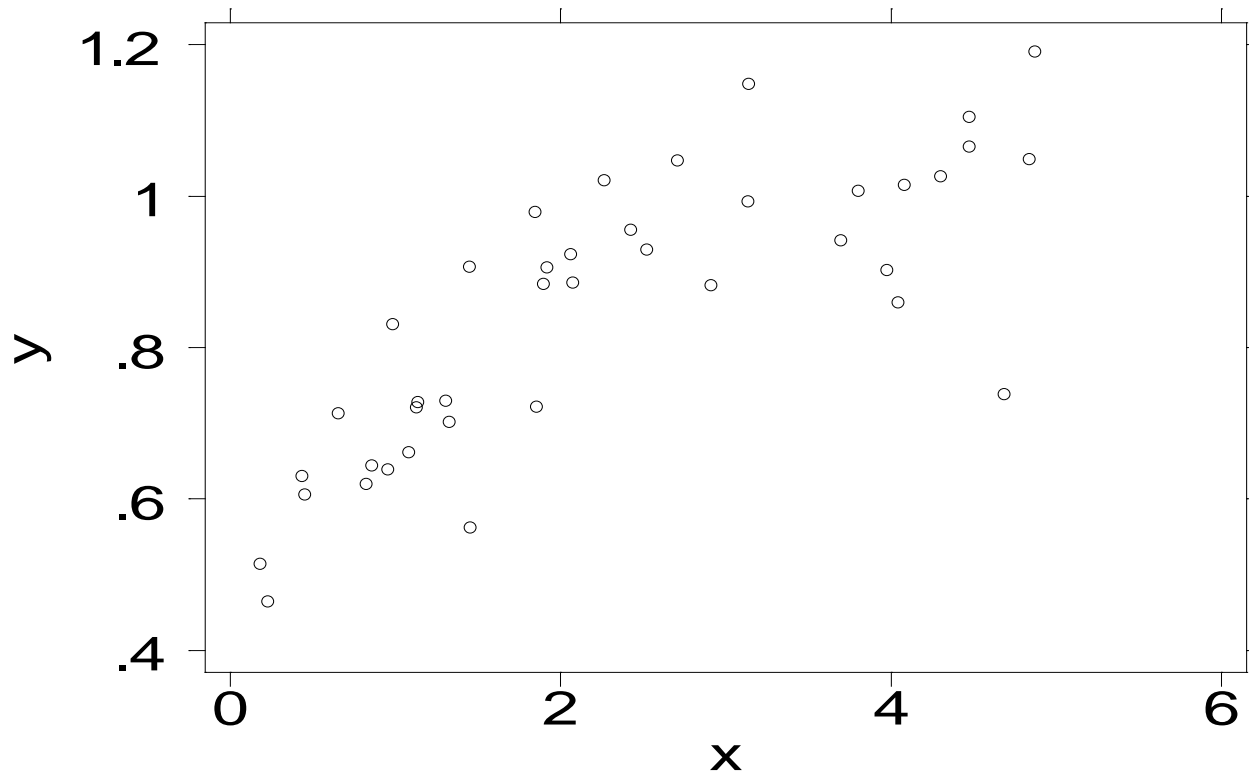
# Scatterplots

A **scatterplot** offers a convenient way of visualizing the relationship between pairs of quantitative variables.



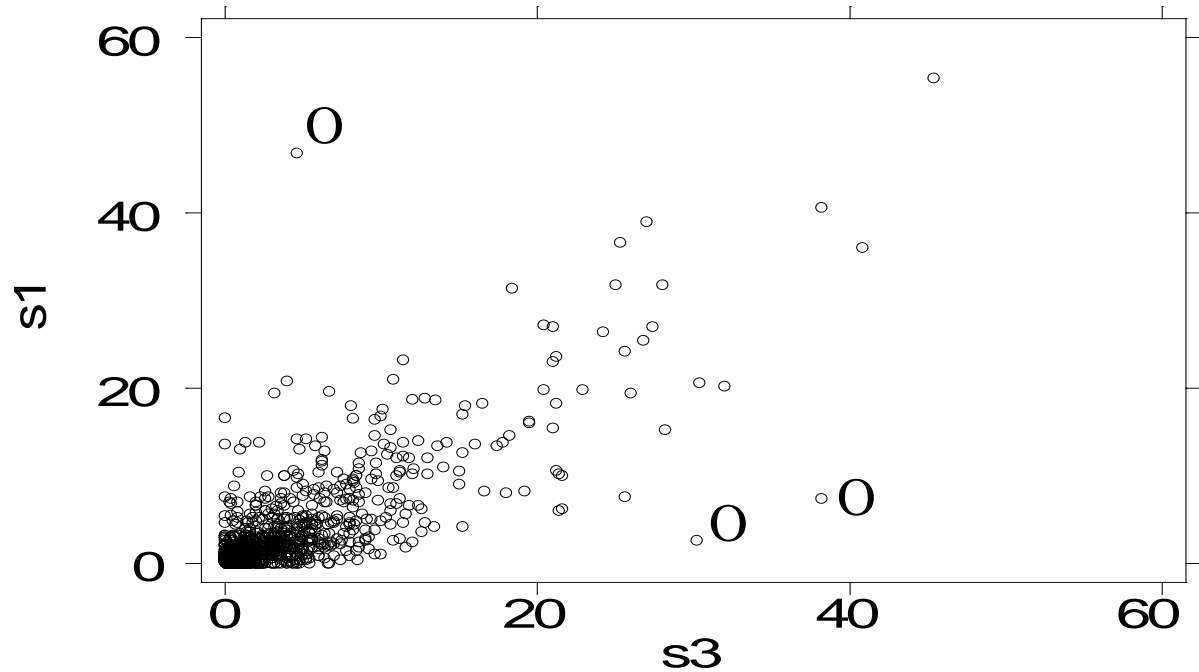
Many interesting features can be seen in a scatterplot including the overall pattern (e.g., linear, nonlinear, periodic), strength and direction of the relationship, and outliers.

# Examples



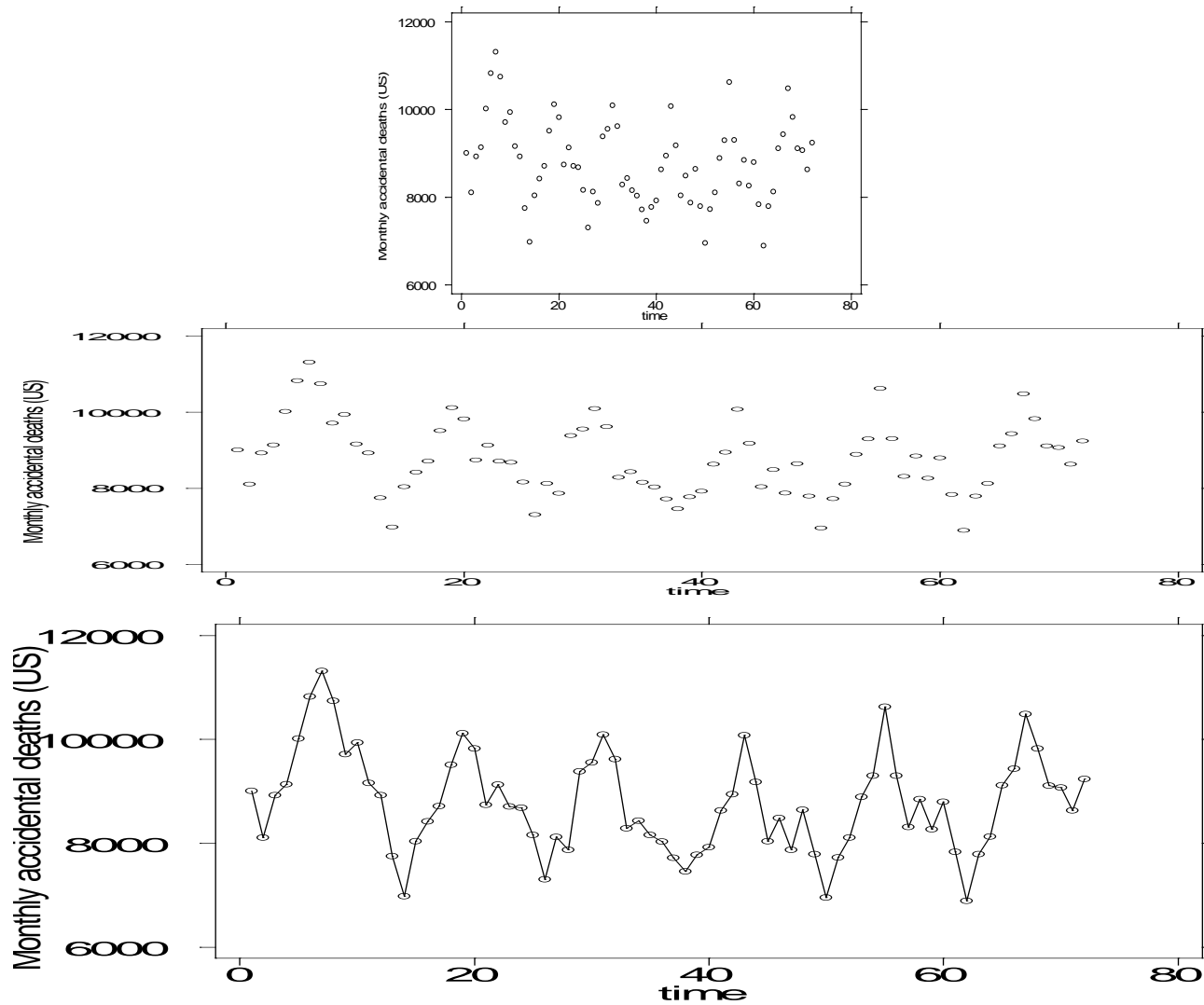
Scatterplot showing nonlinear relationship

# Examples

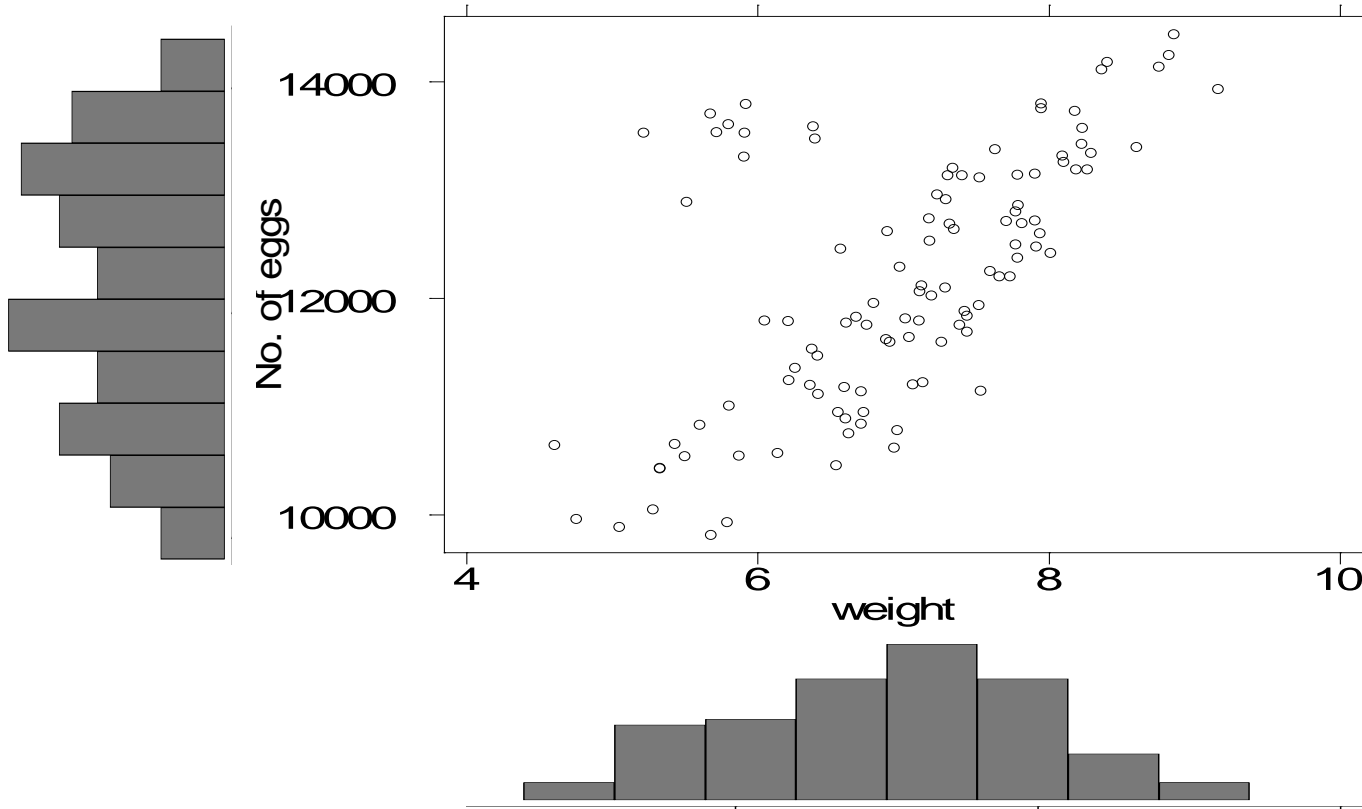


Scatterplot showing daily rainfall amount (mm) at nearby stations in SW Australia. Note outliers (O). Are they data errors ... or interesting science?!

# Presentation matters!

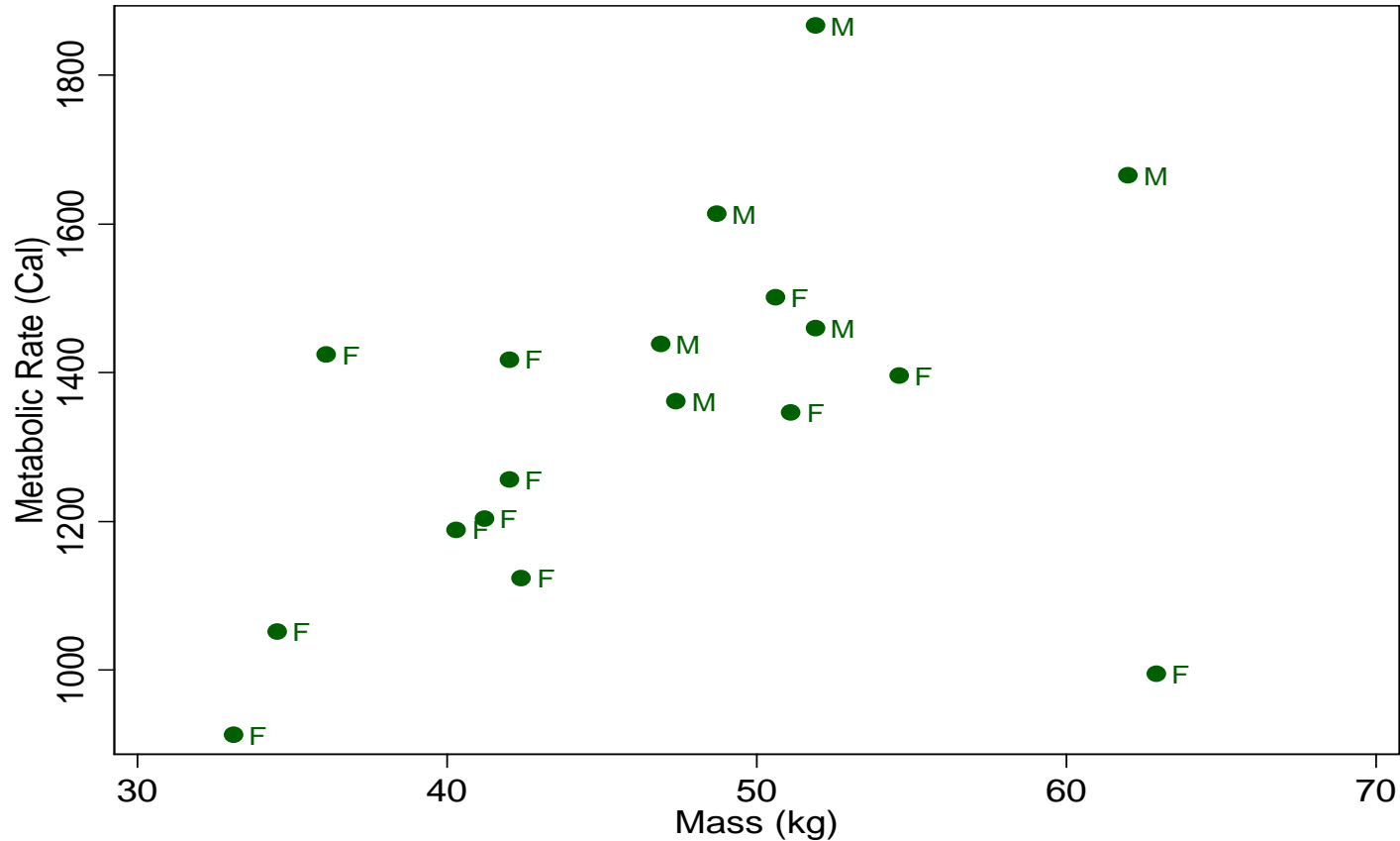


# One or two dimensions?



- Important information can be seen in two dimensions that isn't obvious in one dimension

# Adding dimensions...

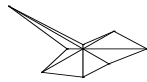


Use symbols or colors to add a third variable

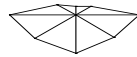
# Plotting Multivariate Data

Star plots are used to display multivariate data

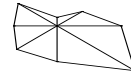
Concord



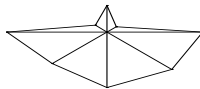
Pacer



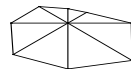
Century



Electra



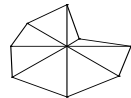
LeSabre



Regal



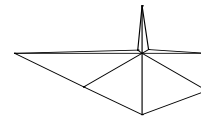
Riviera



Skylark



Deville



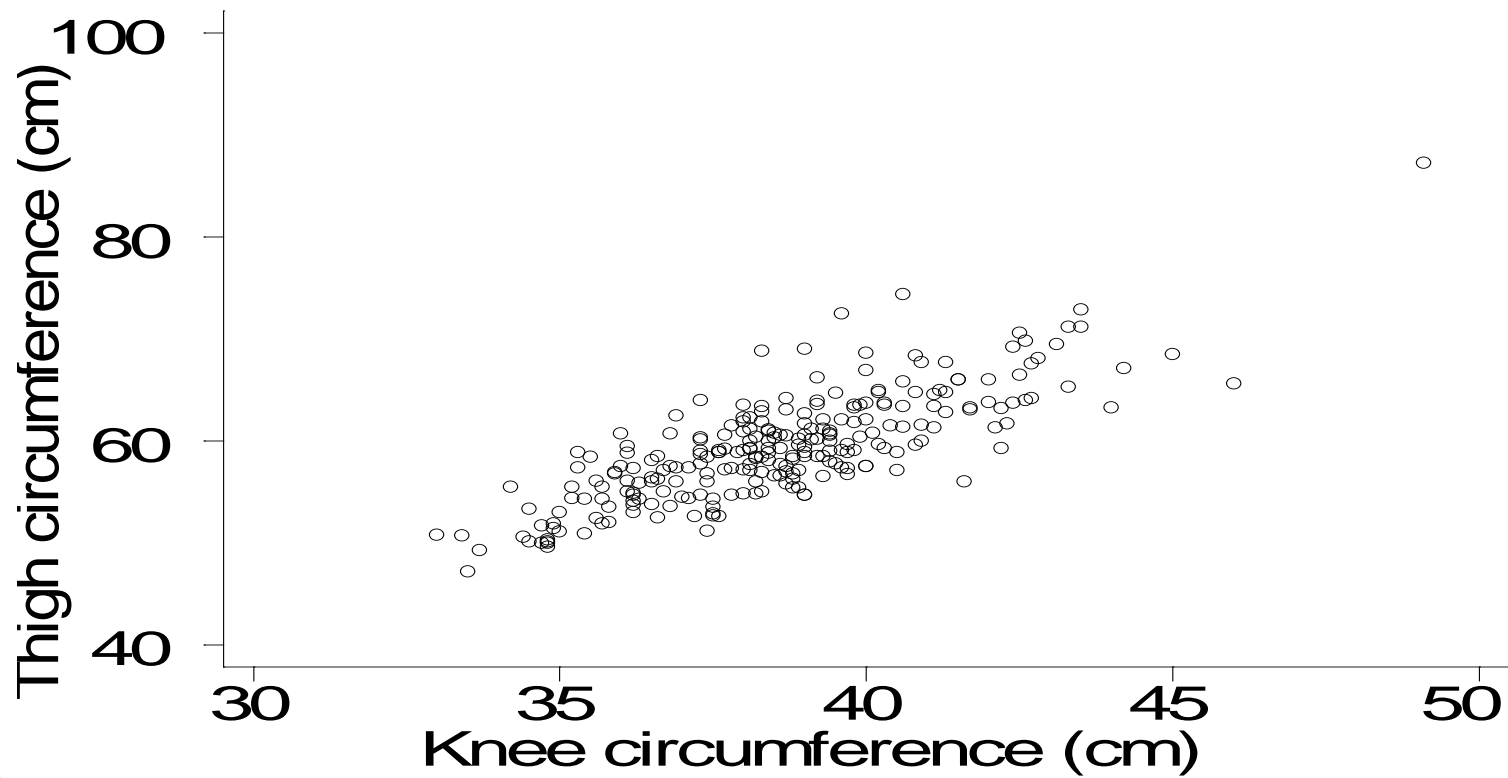
- ⊕ Price
- ⊙ Mileage (mpg)
- ⊖ Repair Record 19
- ⊙ Headroom (in.)
- ⊕ Weight (lbs.)
- ⊙ Turn Circle (ft.)
- ⊖ Displacement (cu
- ⊙ Gear Ratio

- Each ray corresponds to a variable
- Rays scaled from smallest to largest value in dataset



# Correlation

How can we summarize the “strength of association” between two variables in a scatterplot?



# Pearson's Correlation Coefficient

The **correlation** between two variables X and Y is:

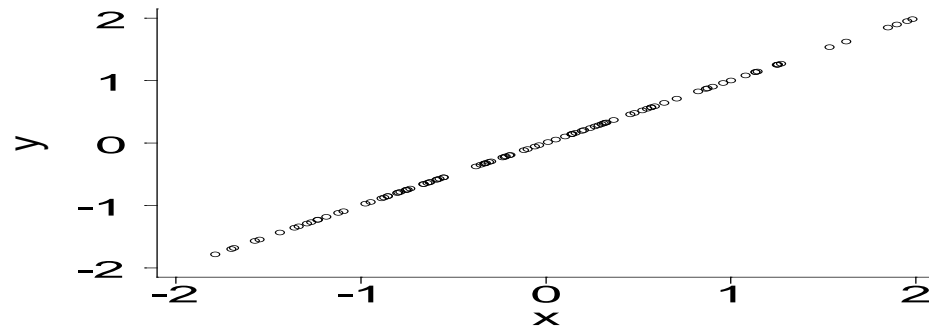
$$R = \frac{1}{N-1} \sum_{i=1}^N \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right)$$

## **Properties:**

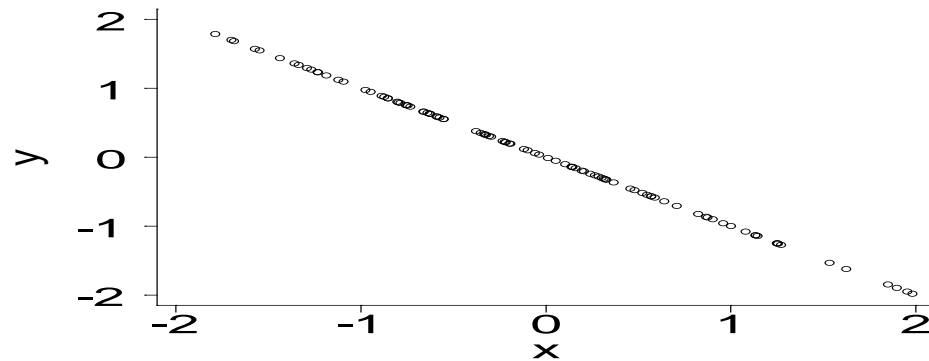
- No distinction between x and y.
- The correlation is constrained:  $-1 \leq R \leq +1$
- $|R| = 1$  means “perfect **linear** association”
- The correlation is a scale free measure (correlation doesn't change if there is a linear change in units).
- Pearson's correlation only measures strength of **linear** relationship.
- Pearson's correlation is sensitive to outliers.

# Examples

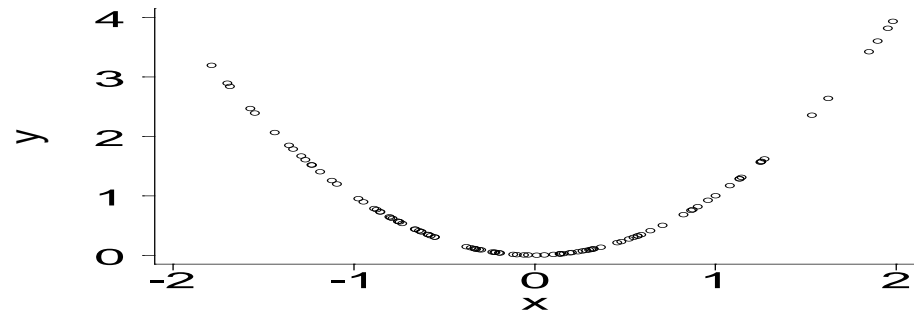
Perfect positive  
correlation ( $R = 1$ )



Perfect negative  
correlation ( $R = -1$ )

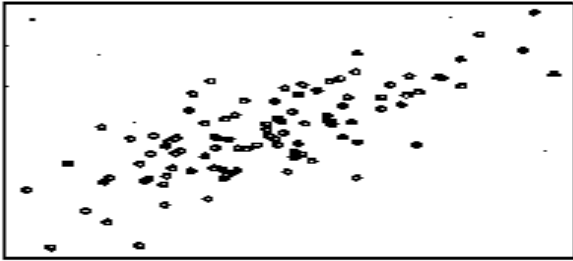


Uncorrelated ( $R = 0$ )  
but dependent

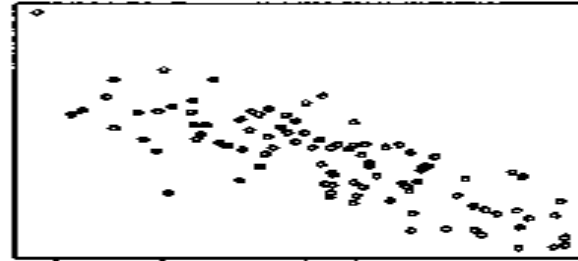


# More examples

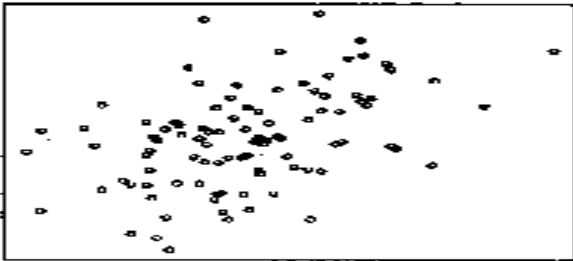
$r = 0.76$



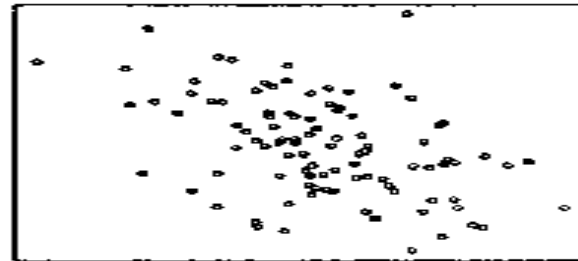
$r = -0.8$



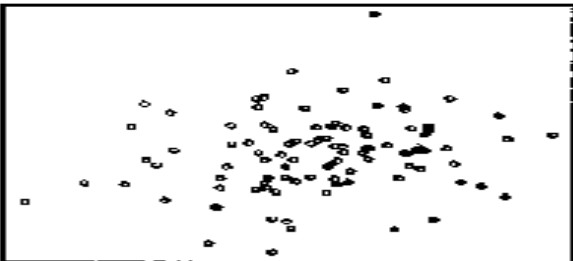
$r = 0.44$



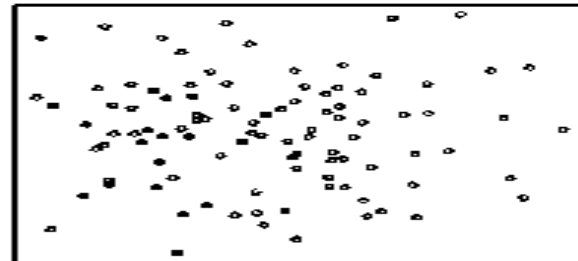
$r = -0.44$



$r = 0.19$



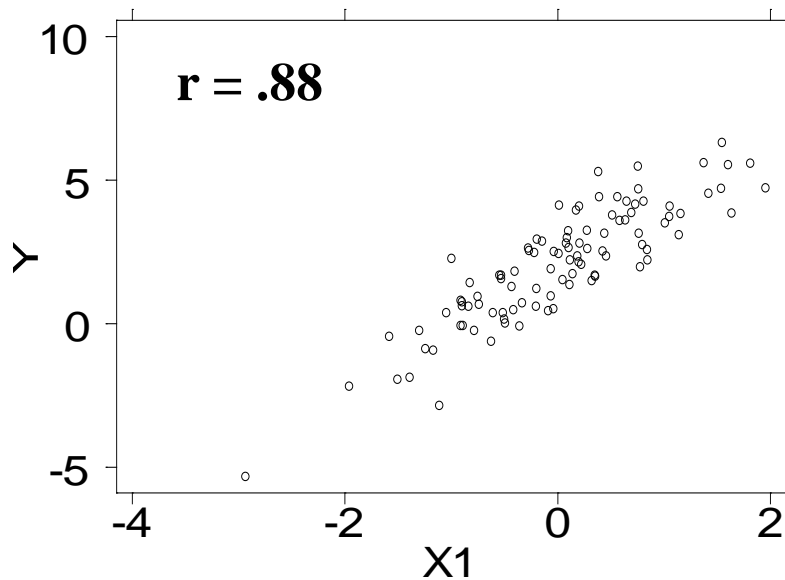
$r = -0.03$



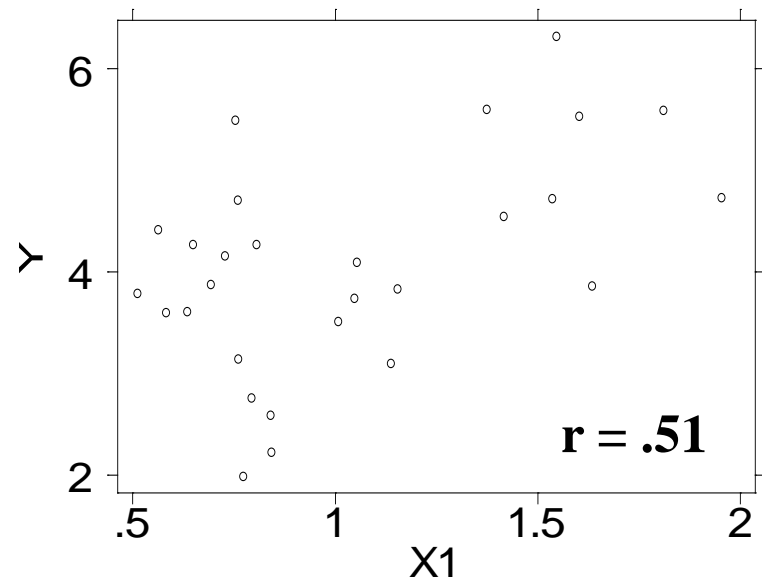
# Correlation Pitfalls

Correlation is attenuated if you restrict the range of X or Y ...

Plot of all data ...



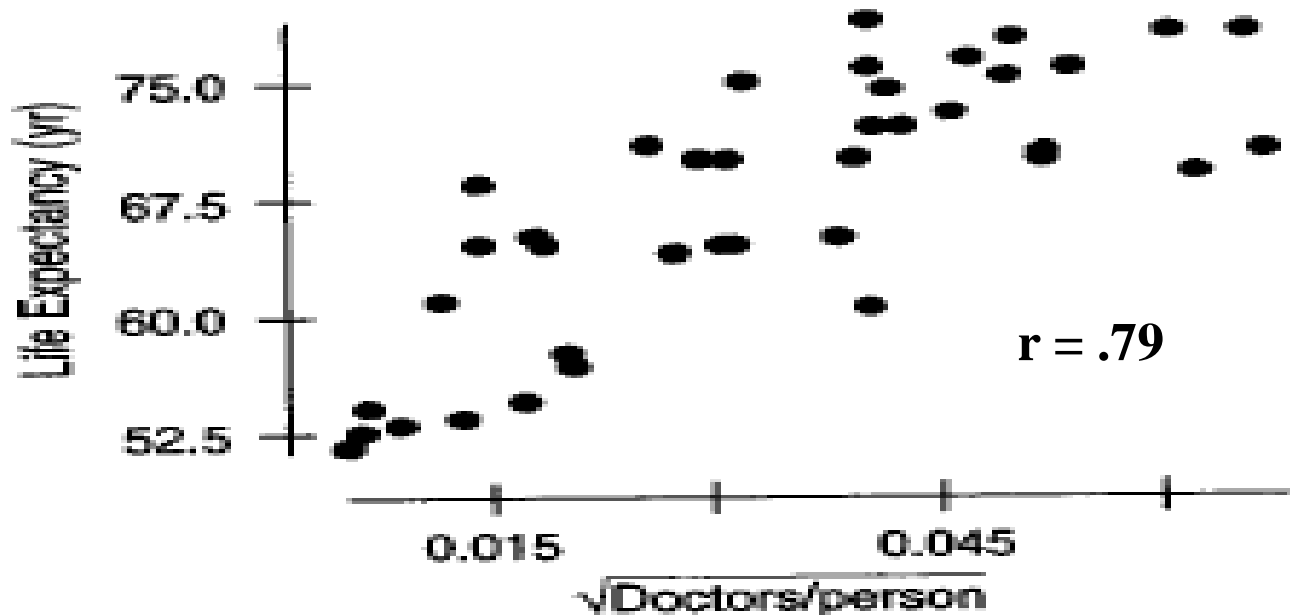
Now restrict the range of X ...



E.g. relationship between LSAT and GPA among law school students

# Correlation Pitfalls

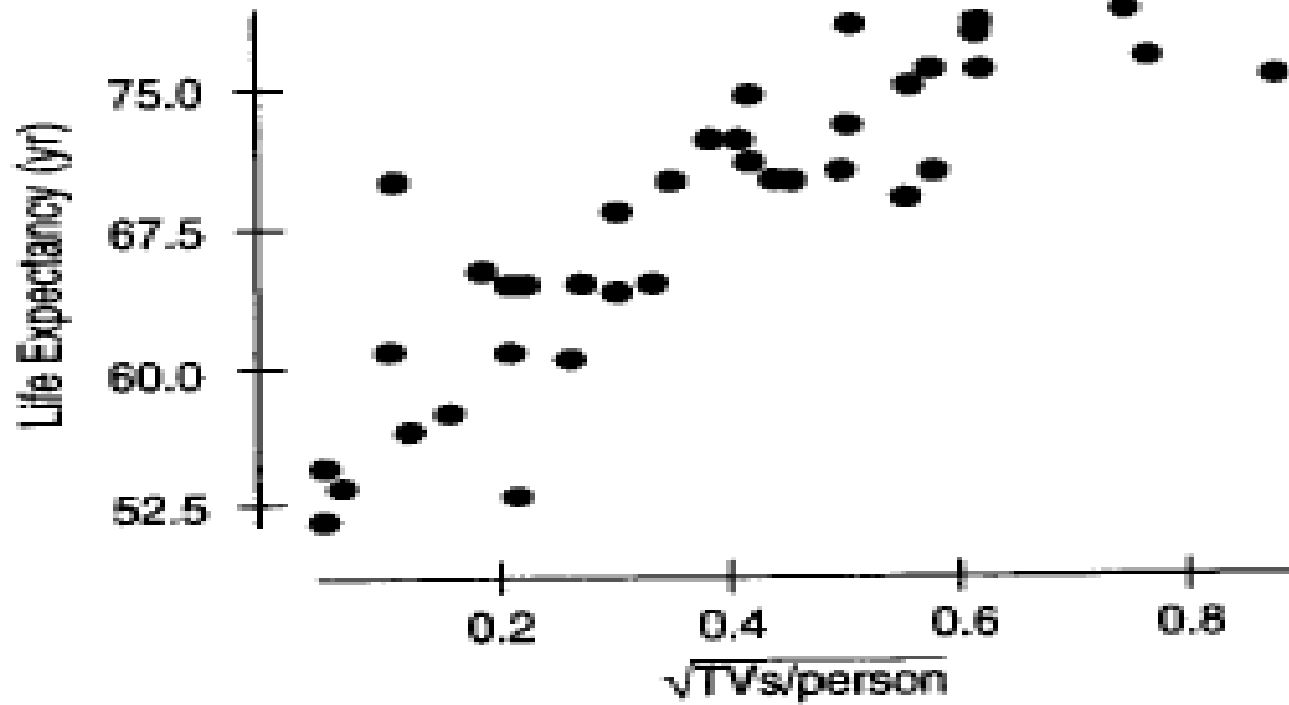
Careful about interpreting correlation causally ...



- There is an association between doctors/person and life expectancy.
- Can we increase life expectancy by providing more doctors?

# Correlation Pitfalls

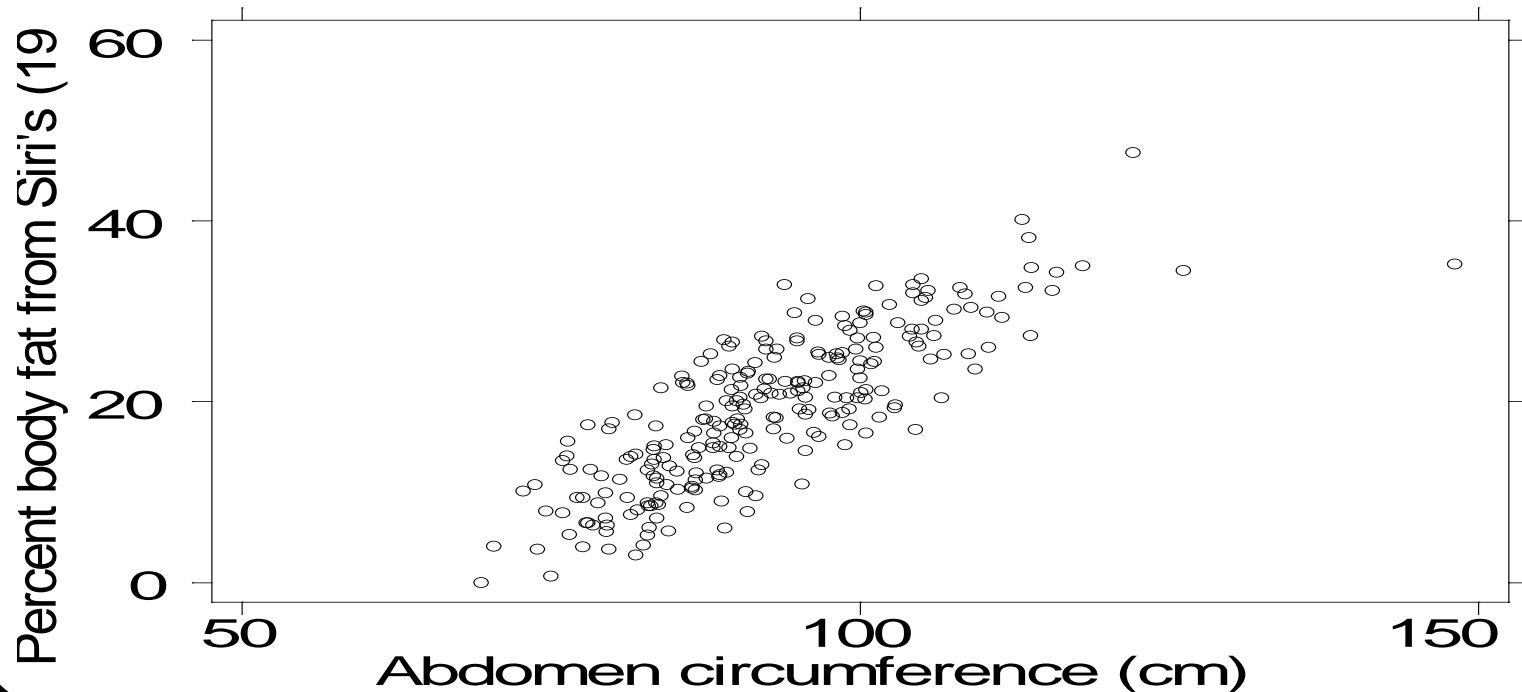
... or maybe we just sell more TV's?



# Linear Regression

The correlation coefficient was used to summarize the strength of the relationship between interchangeable X and Y.

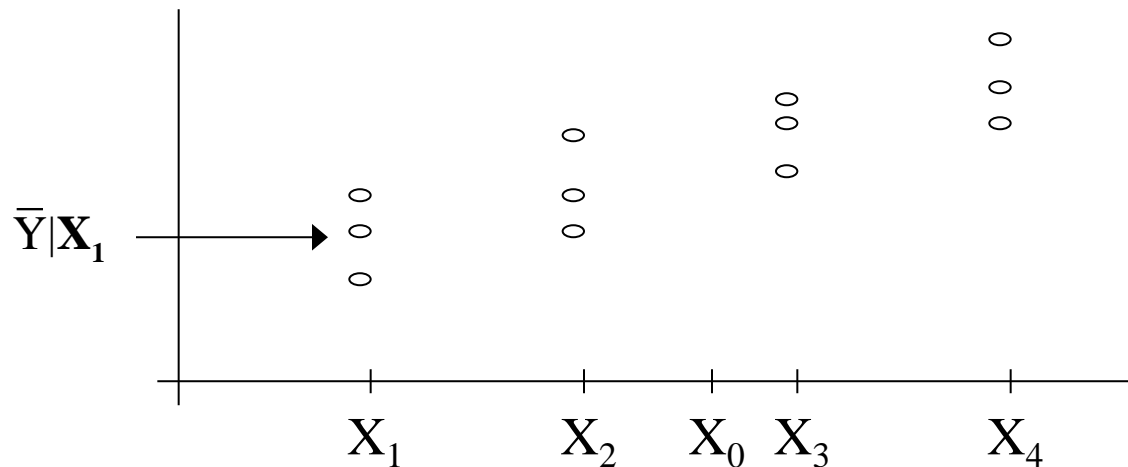
Sometimes, however, X and Y are not interchangeable. We may want to predict Y from X. What is a reasonable approach?





# Linear Regression

We could use the mean (or median) value of Y for whatever X we are interested in ...



But we would like something simpler (fewer numbers to keep track of) ...

... that will allow us to predict Y at  $X_0$ ...

... and make “fuller” use of the data.

# Linear Regression

- If a scatterplot suggests a linear relationship between X and Y we can draw a **linear regression line** to describe how the mean of Y **changes differs** when X **changes differs** or to predict the mean of Y for any given value of X:

$$Y = a + bX$$

- In **linear regression** one variable (X) is used to predict or explain another (Y) (the situation is asymmetric).
- X independent, predictor  $\Rightarrow$  Y dependent, response
- X and Y are both quantitative
- This is an example of a mathematical model

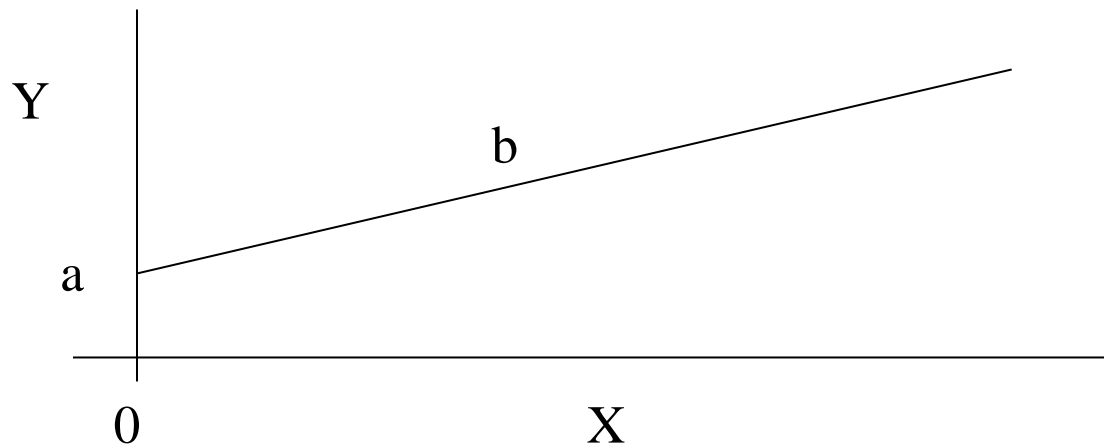
# Model Assumptions

## Straight line relationship

$$Y = a + bX$$

$a$  = intercept = value of mean of  $Y$  when  $X = 0$

$b$  = slope = expected **change difference** in the mean of  $Y$  for each 1 unit **change difference** in  $X$

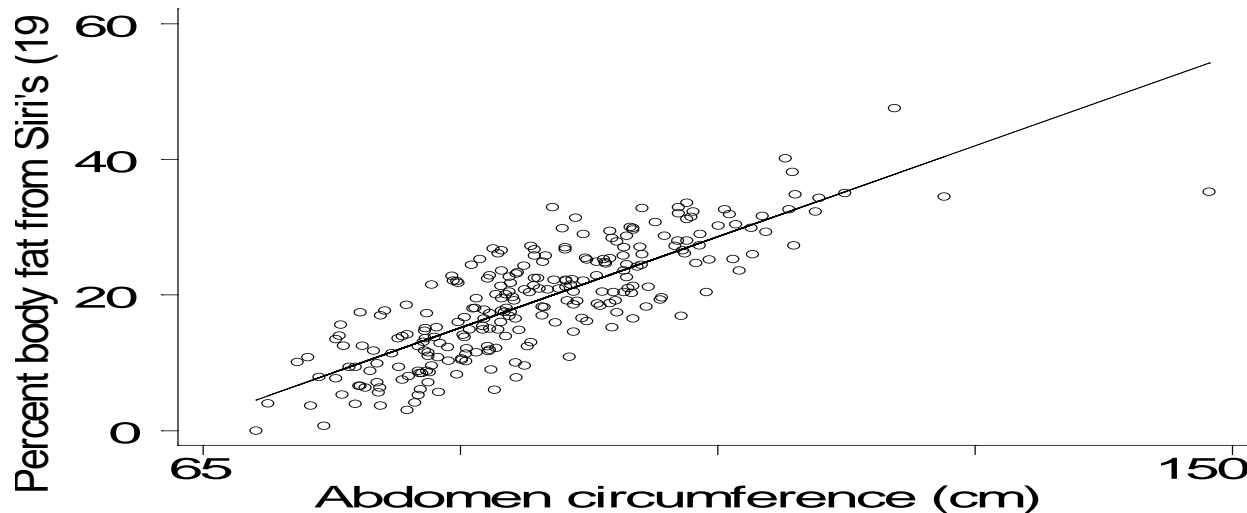


# Regression Slope

What is the interpretation of b, the slope in the regression equation?

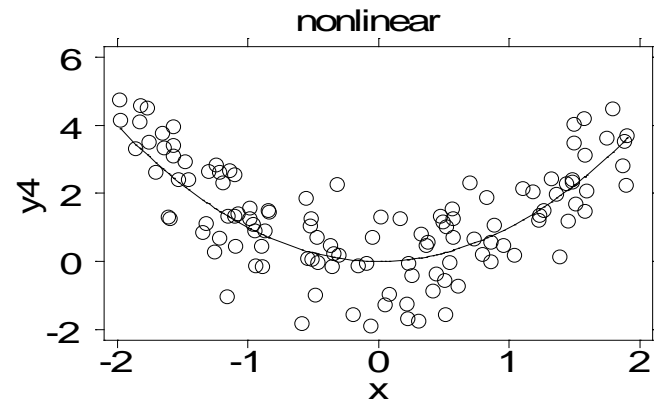
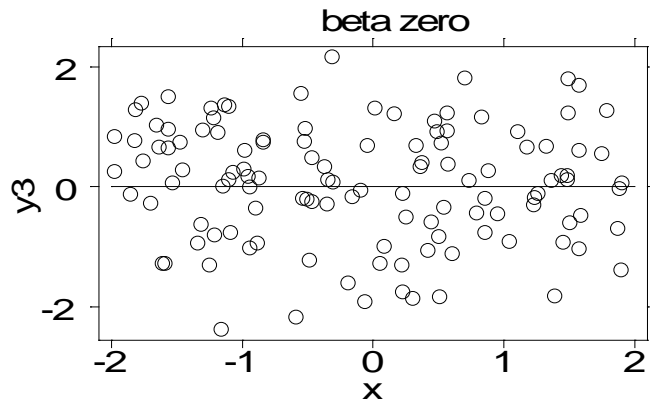
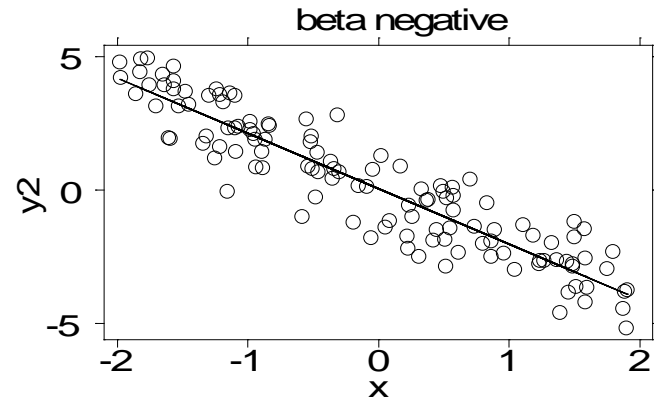
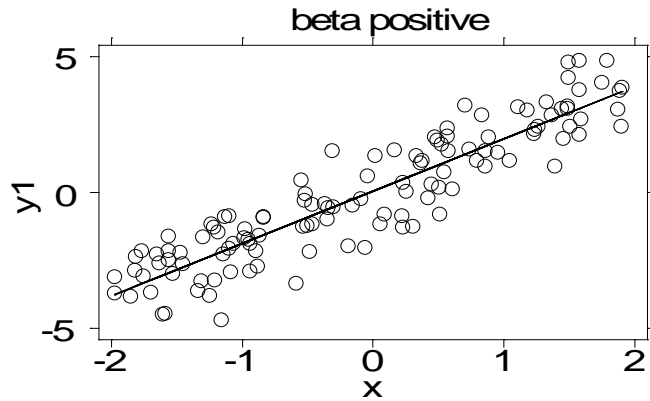
**b** describes the expected **change** difference in Y (**change** difference in the mean value of Y) for each 1 unit difference in X.

$$Y = -39.28 + .6312 X$$



For each 1 cm **increase** difference in abdominal circumference, the mean percent body fat **is increases by** 0.6312 points higher.

# Examples

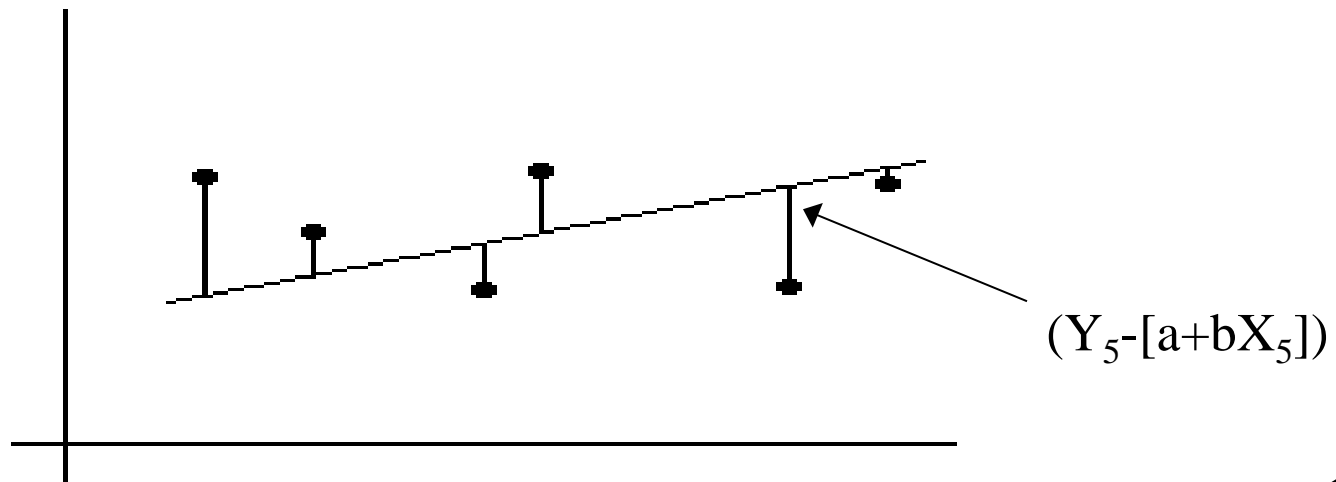


# Fitting Regression Lines

The method of **Least Squares** dates to at least Legendre (1805). We collect **pairs** of observations  $(X_i, Y_i)$  for  $i = 1, 2, \dots, n$ ; and choose a line,  $(a, b)$  where the total difference between the data and the fitted line is minimal - in the sense that the **sum of squared differences** between the **observed** and the **fitted** is minimized:

$$S = \sum_{i=1}^n (Y_i - [a + bX_i])^2$$

- Find the values  $(a, b)$  that make  $S$  as small as possible.



# Fitting Regression Lines

To obtain the minimum / maximum of a function we find the values that make the derivatives equal to zero. So to find (a, b) that minimizes  $S$  we solve the “**normal equations**”:

$$\frac{\partial}{\partial a} S = \sum_{i=1}^n -2(Y_i - [a + bX_i]) = 0$$

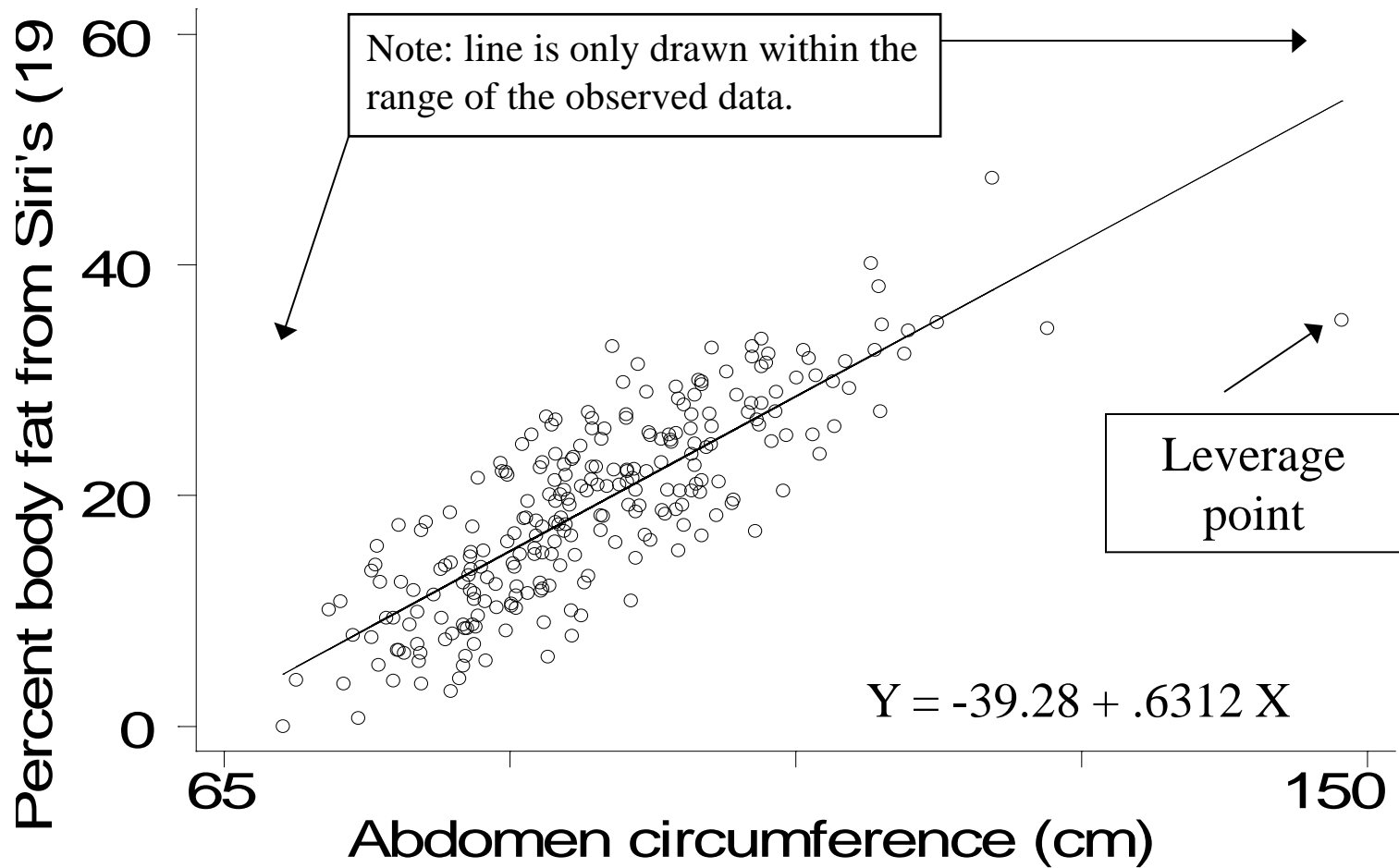
$$\frac{\partial}{\partial b} S = \sum_{i=1}^n -2X_i(Y_i - [a + bX_i]) = 0$$

Solving these equations yields the estimates:

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

# Example



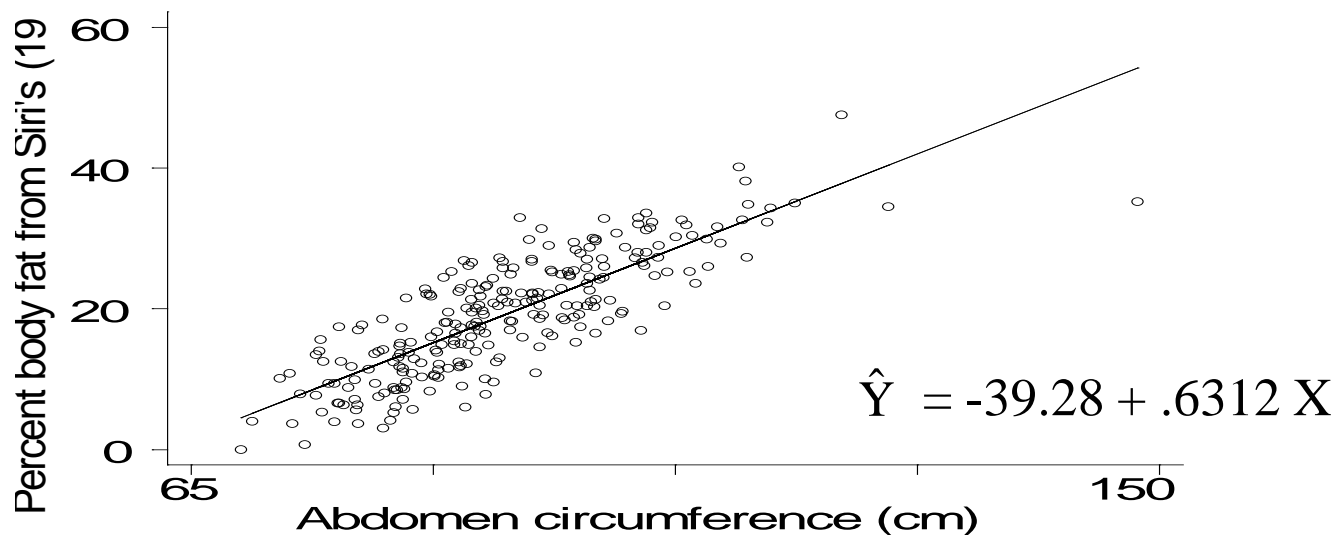


# Regression – Predicted Values

Given the estimates (a, b) we can find the **predicted value**,  $\hat{Y}$ , for any value of X.

$$\hat{Y} = \hat{a} + \hat{b}X$$

The interpretation of  $\hat{Y}$  is as the **estimated mean value of Y for a large sample of values taken at X**.



Predicted body fat when abdominal circumference is 90 cm  
 $= -39.28 + .6312 * 90 = 17.53$  percent

# Regression and Correlation

The least squares slope,  $b$ , and the correlation coefficient,  $r$ , are closely connected:

$$b = r \frac{s_y}{s_x}$$

- From this we see that slope = correlation  $\times$  scale change
- Unlike correlation, reducing the range or spread of  $x$  values (i.e. smaller  $s_x$ ) does not (systematically) attenuate  $b$  (because  $r$  decreases as well)

# Regression Pitfalls

“With great power comes great responsibility” - Spiderman

A computer program can always fit a linear regression model! That doesn't mean that the predictions or conclusions you draw from the model always make sense. Beware of

- Nonlinearity
- Outliers and leverage points
- Extrapolating outside the range of the data
- Mistaking association for causation (recall the life expectancy vs TVs example; more on this in a bit)

# Regression Pitfalls

## Caution

Predicted values assume the model is true. Presumably, we have checked that this is a reasonable assumption where we have data. It may not be true outside the range of our data!!

One disastrous example of extrapolation outside the range of the existing data was the decision to launch the space shuttle Challenger in 31 degree temperatures. Prior to that the coldest launch temperature was 53 degrees. Poor graphical presentation of data also played a role in this decision.

# Challenger

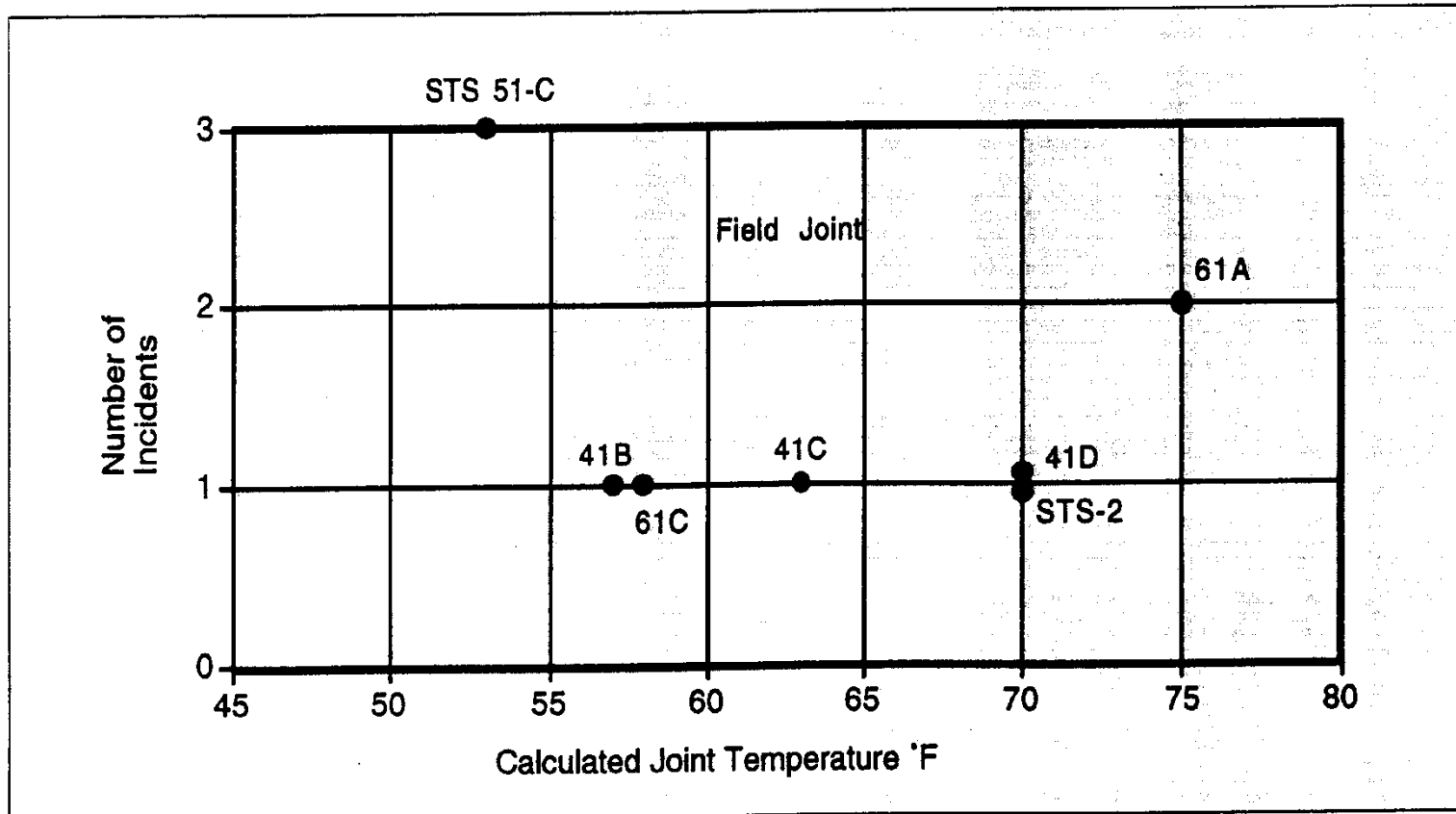


Figure 5. Joint temperature versus number of O-rings having some thermal distress identified by flight number; omits all flights in which there was no such problem. From Volume 1, p. 145, of the *Report of the Presidential Commission on the Space Shuttle Challenger Accident* (1986).

# Challenger

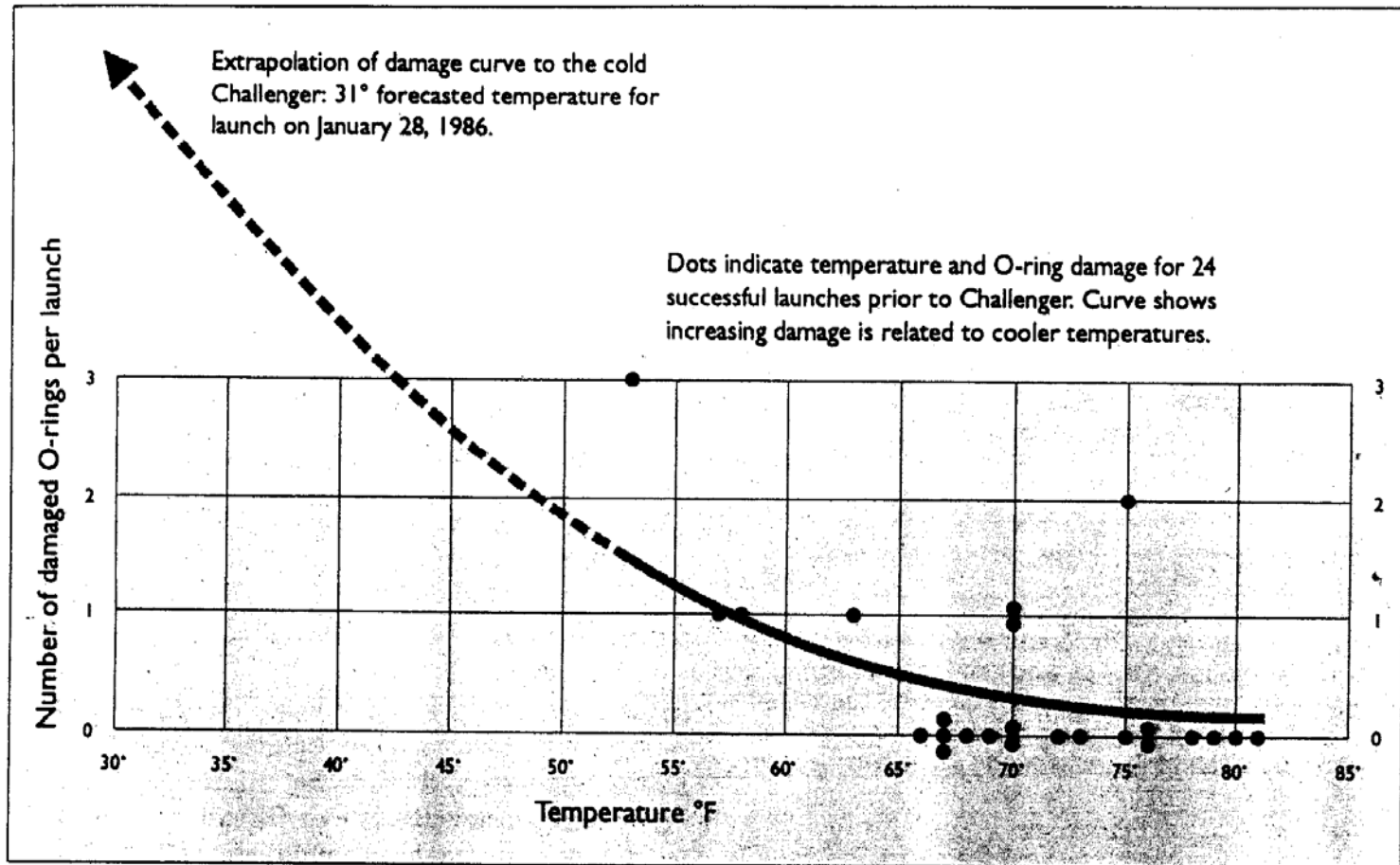


Figure 6. The data from Fig. 5 augmented with the “zero” entries from the 17 flights without O-ring incidents. A plausible nonlinear curve is fitted to the data describing the likely outcome at lower temperatures. Reproduced from Tufte (in press) with permission.

# Summary

**Scatterplots** provide a compact display of the relationship between two quantitative measures.

Colors or symbols can be used to add a third (categorical) dimension to a scatterplot.

**Starplots** can be used to display multivariate data.

The **correlation coefficient** summarizes the strength of the linear (Pearson's) or monotonic (Spearman's) relationship between two quantitative measures.

A **linear regression line** is a model summarizing how the mean value of one quantitative measure (Y) varies with another quantitative measure (X). A linear regression line can be used to **predict** Y from X. The **slope** of the regression line gives the expected **change difference** in Y for each 1 unit **change difference** in X.

# Descriptive Statistics and Exploratory Data Analysis – Bivariate/Multivariate

- **Quantitative Data**
  1. Scatterplots
  2. Starplot
  3. Correlation
  4. Regression
- **Qualitative Data**
  5. Two-way (contingency) tables
- **Effect modification**



# Two-way (contingency) Tables

Quantitative measures - correlation

Qualitative (categorical, discrete) measures – two-way tables

- Nominal or ordinal categories
- Cross-classify the observations according to the two factors
- **Two-way table = contingency table.**
- Measures of association

# Two-way Tables

**Example.** Education versus willingness to participate in a study of a vaccine to prevent HIV infection if the study was to start tomorrow. Counts, percents and row and column totals are given.

	definitely not	probably not	Probably	definitely	Total
< high school	52 1.1%	79 1.6%	342 7.0%	226 4.6%	699
high school	62 1.3%	153 3.2%	417 8.6%	262 5.4%	894
some college	53 1.1%	213 4.4%	629 13.0%	375 7.7%	1270
college	54 1.1%	231 4.8%	571 11.8%	244 5.0%	1100
some post college	18 0.4%	46 0.9%	139 2.9%	74 1.5%	277
graduate/ prof	25 0.5%	139 2.9%	330 6.8%	116 2.4%	610
Total	264	861	2428	1297	4850

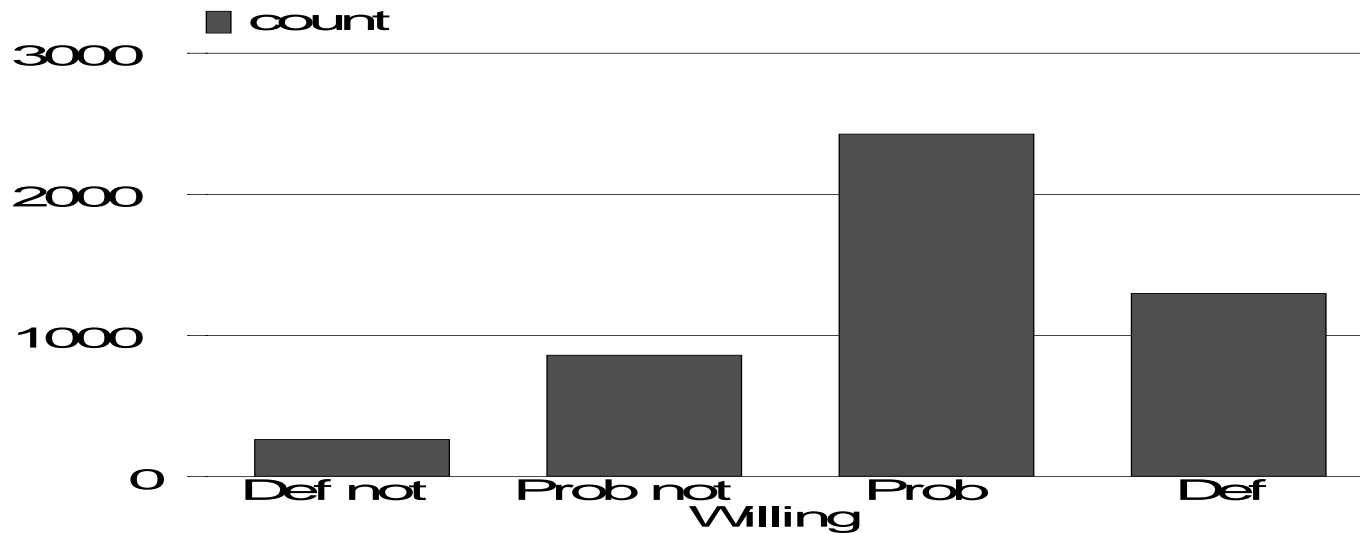
The table displays the **joint distribution** of education and willingness to participate.

# Two-way Tables

The **marginal distributions** of a two-way table are simply the distributions of each measure summed over the other.

E.g. Willingness to participate

Definitely not	Probably not	Probably	Definitely
264	861	2428	1297
5.4%	17.8%	50.1%	26.7%



# Two-way Tables

A **conditional distribution** is the distribution of one measure conditional on (given the) value of the other measure.

E.g. Willingness to participate among those with a college education.

Definitely not	Probably not	Probably	Definitely
54	231	571	244
4.9%	21.0%	51.9%	22.2%

# Two-way Tables

	definitely not	probably not	probably	definitely	Total
< high school	52	79	342	226	699
high school	62	153	417	262	894
some college	53	213	629	375	1270
college	54	231	571	244	1100
some post college	18	46	139	74	277
graduate/ prof	25	139	330	116	610
Total	264	861	2428	1297	4850

What proportion of individuals ...

- will definitely participate?
- have less than college education?
- will probably or definitely participate given less than college education?
- who will probably or definitely participate have less than college education?
- have a graduate/prof degree and will definitely not participate?

# Two-way Tables

Q: Is there an analogue to the correlation coefficient for quantifying association in two-way tables?

A: Yes. In fact, there are several. At this point we will only discuss two summary measures for  $2 \times 2$  tables – the *relative risk* and the *risk difference*.

# 2x2 Tables

**Example:** Pauling (1971)

Patients are randomized to either receive Vitamin C or placebo. Patients are followed-up to ascertain the development of a cold.

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

How can we summarize the association between treatment and disease?

## 2x2 Tables

Two measures are commonly used:

$$1) \text{ Relative Risk} = \frac{\text{Risk of disease among treated}}{\text{Risk of disease among placebo}}$$

RR < 1 – treatment associated with reduced risk of disease

RR = 1 – no association

RR > 1 – treatment associated with increased risk of disease

$$2) \text{ Risk Difference} = \text{Risk among treated} - \text{Risk among placebo}$$

RD < 0 – treatment associated with reduced risk of disease

RD = 0 – no association

RD > 0 – treatment associated with increased risk of disease

See article “Relative Risk vs Absolute Risk – Vastly Different”



## 2x2 Tables – Prospective Study

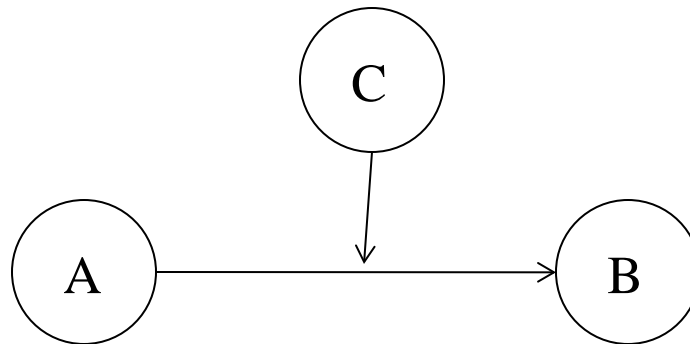
	Exposed	Unexposed	Total	
Cases	17	31	48	
Noncases	122	109	231	
Total	139	140	279	
Risk	.1223022	.2214286	.172043	
	Point estimate		[95% Conf. Interval]	
Risk difference	-.0991264		-.1868592	-.0113937
Risk ratio	.5523323		.3209178	.9506203
Prev. frac. ex.	.4476677		.0493797	.6790822
Prev. frac. pop	.2230316			
	chi2(1) =		4.81	Pr>chi2 = 0.0283

Note: RR and RD are appropriate when we sample exposure groups and measure disease; RR and RD should not be used when we sample disease groups and measure exposure (i.e. case-control study).

# Interaction (Effect Modification)

RR, RD, correlation and regression are all examples of measures of association between two variables. Complications in interpretation can arise when we involve a third variable.

**Interaction**, also known as **effect modification** in the epidemiology literature occurs when the degree of association between the two primary variables (A and B) depends on value of a third variable (C).



# Interaction (Effect Modification)

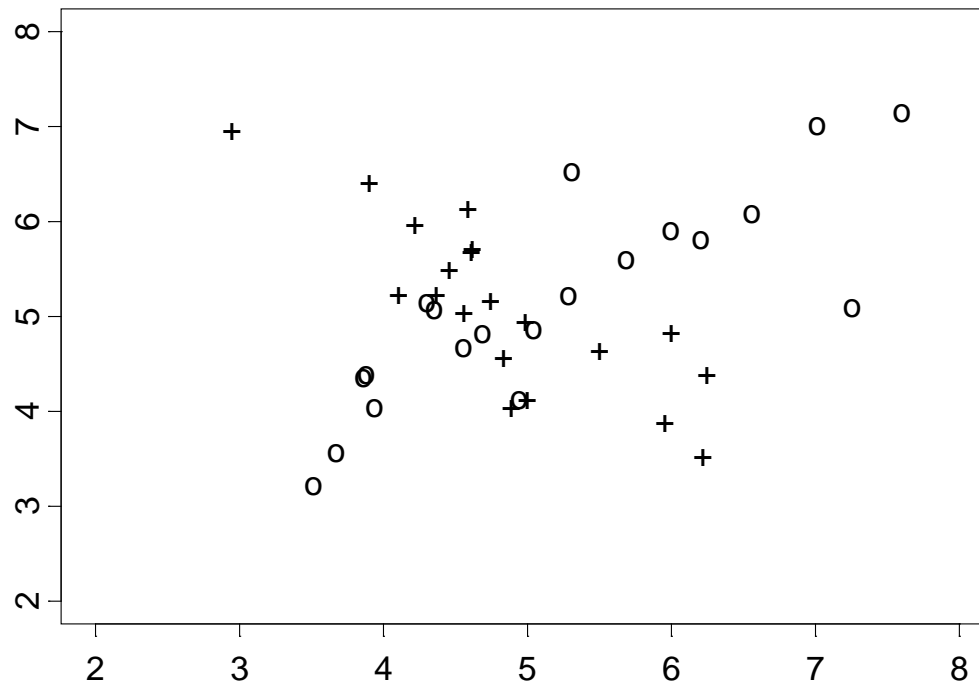
	Seat Belt	
Driver	Worn	Not worn
Dead	9	21
Alive	31	39
Total	40	60
Fatality Rate	22.5%	35%

	Impact Speed			
	$\leq 40$ mph		$> 40$ mph	
Driver	seat belt		seat belt	
	worn	not	worn	not
dead	2	3	7	18
alive	18	27	13	12
Total	20	30	20	30
Fatality Rate	10%	10%	35%	60%

- the effect of seat belt use on fatality rate depends on impact speed
- in general, the effect of A on B differs by levels of C; pooled table is intermediate

# Interaction (Effect Modification)

- The concept of effect modification applies to any measure of association. Here is an example with correlation. The “o” represent one subgroup (correlation = .7); the “+” represent a different subgroup (correlation = -.7). Overall correlation  $\approx 0$ .



# Summary

**Contingency tables** are used to study association between pairs of **categorical variables**.

The **joint distribution** of the two variables as well as the **marginal distributions** of each variable and the **conditional distribution** of one variable for a fixed level of the other variable can be obtained from the contingency table.

**Interaction (effect modification)** occurs if a third variable influences the association between the two variables of interest.

# Designing Studies

- Population vs. Sample
  1. Bias
  2. Variability
- Study Design
  1. Types of studies
    - a. Descriptive
    - b. Observational
    - c. Experimental
  2. Common Designs
    - a. Ecologic
    - b. Cross-sectional
    - c. Cohort
    - d. Case-control
    - e. Randomized trial
- Confounding

# Populations vs Samples

In almost all situations there is an implicit assumption that the conclusions we draw from our data apply to some larger group than just the individuals we measured.

## Population

- set of all “units”



## Parameter

- Numerical value that would be calculated using all units in the population

## Sample

- a subset of “units”



## Estimates/statistics

- Numerical value that is calculated using all units in the sample

**The objective of inferential statistics is to make valid inferences about the population from the sample.**



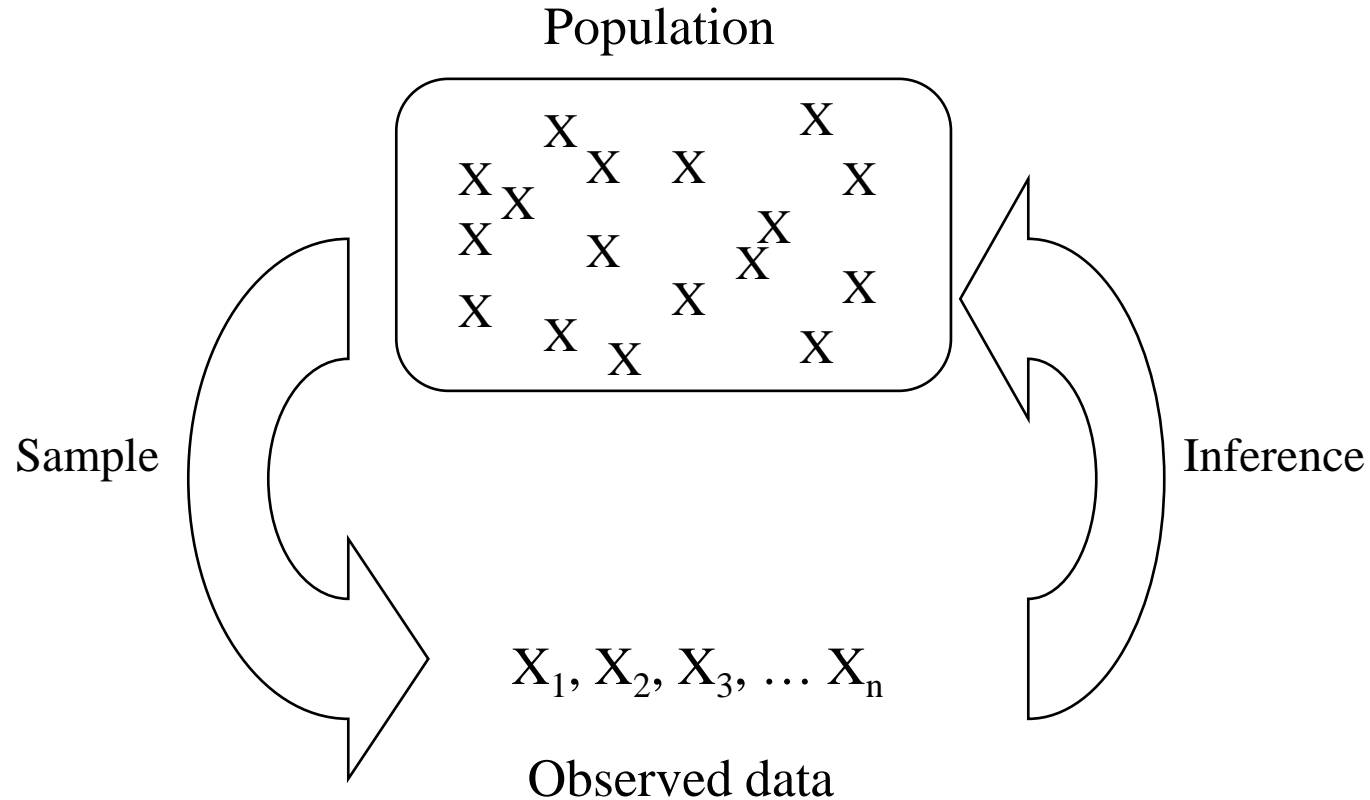
# Population vs Samples

Nine percent of the U.S. population has type B blood. In a sample of 400 individuals from the U.S. population, 12.5% were found to have type B blood.

- In this situation, the value of 9% is a (parameter, statistic).
- In this situation, the value of 12.5% is a (parameter, statistic).



# Basic Statistical Paradigm



# Language of Sampling

- Census
- Random (Probability) Sample
  - simple (,stratified, cluster, systematic, multistage)
  - sampling frame
- Voluntary sample
- Convenience Sample
- Sampling unit
- Bias
- Sampling variability

# Random (probability) Sampling

- Each unit in the population has a known non-zero chance of being included in the sample.
- Suppose our sample size is  $n$ . If all samples of size  $n$  have an equal chance of being drawn, the sample is a simple random sample.
- Probability sampling usually requires a sampling frame - a list of all units in the population e.g. census tracts/blocks, class list
- Random samples are very important when estimating absolute characteristics of a population e.g. percent who will vote, median income, seroprevalence of HBV in IDUs, mean mercury level in tuna
- Random samples are less important in comparative studies (implicit assumption that comparative effect is the same in all units) e.g. efficacy of behavioral intervention for reducing HBV infection in IDUs.
- Often, taking a truly random sample is impossible; we hope for a “representative” sample.

# Sampling Variability

- In making inferences about the population *from a random sample*, a key concern is sampling variability and its effect on our conclusions.
- If I repeat an experiment (draw a new sample), I don't expect to get exactly the same results i.e. mean, incidence rate, relative risk. These sample estimates are variable.
- How does that affect our inferences???
- The aim of experimental design and statistical analysis is to quantify/control/minimize effects of variability and to help us understand the effect of sampling variability on our inferences.



Key idea — the  
notion of  
repeated  
sampling

# Bias

- Quite often, we obtain data from a volunteer or convenience sample (note: random samples with high refusal rates are effectively convenience samples)
- Such samples are almost always subject to some sort of bias
- A sampling method is biased if it produces results that *systematically* differ from the population. Stated differently, do I expect that, on average, the estimate from my sample will equal the parameter of the population of interest? If so, the procedure is unbiased.
- E.g. Ann Landers survey, Pap smear study
- In general, statistical methods do not correct for bias

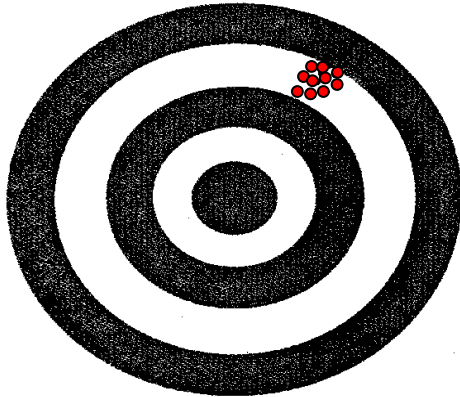
# Bias

- Bias can arise for a number of reasons:
  - Selection bias – sampling procedure systematically includes or excludes a portion of the population
  - Non-responses or refusals
  - Social desirability/response bias
  - Hawthorne effect, etc

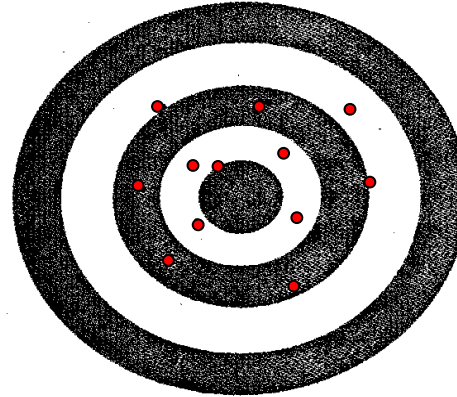
A study was conducted to estimate the average length of a prison sentence for prisoners at a correctional facility. A random sample of current prisoners was obtained on a particular day and they were monitored to the completion of their sentences. The information from this sample was used to estimate the average length of a prison sentence.

- (a) What is the population of interest?
- (b) What is the sample?
- (c) What is the variable of interest?
- (d) Why is the estimate obtained as explained above *almost certainly biased*? (which way?)

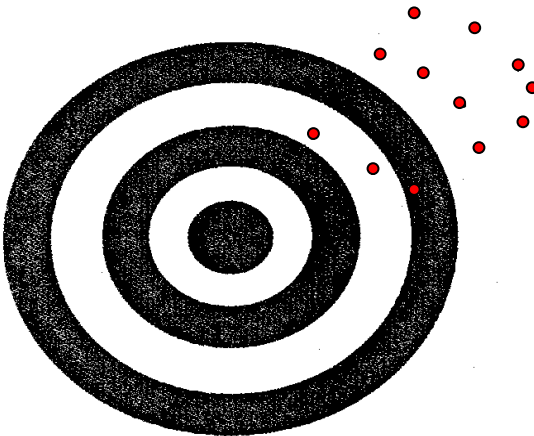
# Examples



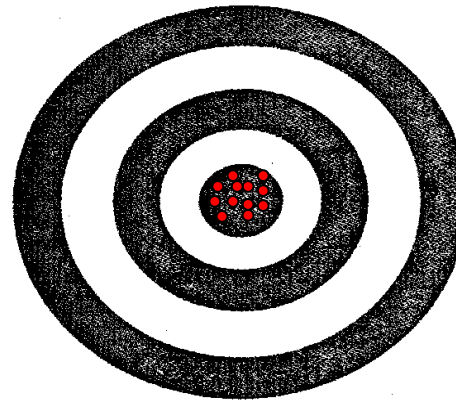
(a) High bias, high precision



(b) Low bias, low precision



(c) High bias, low precision



(d) Low bias, high precision

**Figure 1-4** Bias and lack of precision in sampling. The bull's-eye represents the truth about the population, and the bullet holes represent the results of repeated samples.

# Study Design

“Obtaining valid results from a test program calls for commitment to sound statistical design. In fact, proper experimental design is more important than sophisticated statistical analysis. Results of a well-planned experiment are often evident from simple graphical analyses. *However, the world’s best statistical analysis cannot rescue a poorly planned experiment.*”

*Gerald Hahn, Encyclopedia of Statistical Science, page 359, entry for **Design of Experiments***

“The plural of anecdote is not evidence”

*Dr. Stephen Straus who directs NCCAM*



# Study Types

Most scientific studies can be classified into one of these broad categories:

## 1) Descriptive Studies

Case reports, anecdotal evidence - typically arise serendipitously rather than as a result of a planned study.

## 2) Experimental Studies

The investigator deliberately sets one (or more) factors of interest to a specific level.

## 3) Observational Studies

The investigator collects data from an existing situation and does not (intentionally) interfere with the running of the system.

# Descriptive Studies

- Describe characteristics (case report/case series/anecdotes)
- First description of the disease or phenomenon
- Weak study design - cannot make causal inference
- First step to better-designed study

- Between October 1980-May 1981, 5 young men were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia (PCP) at 3 different hospitals in Los Angeles.
- Previously healthy, homosexual
- June 1981, MMWR
- Dec 1981 NEJM (*Gottlieb NEJM 1981; 305:1425-31*)
- Sept 1982 CDC “AIDS”

# Experimental Studies

- Sources of (major) variability are controlled by the researcher
- Randomization is often used to ensure that uncontrolled factors do not bias results
- The experiment is replicated on many subjects (units) to reduce the effect of chance variation
- Pairing or blocking can make the design more efficient (i.e. fewer units needed)
- Strong case for causation

## Examples

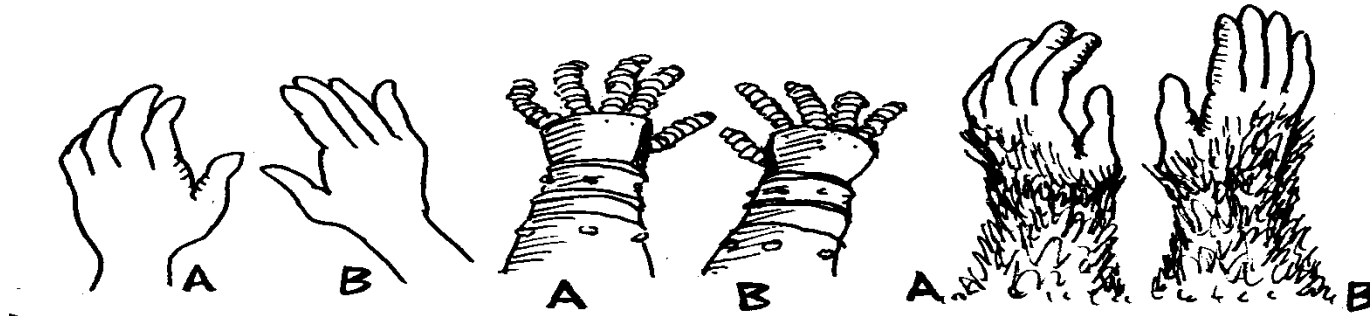
- effect of pesticide exposure on hatching of eggs
- RCT of two treatments for preventing perinatal transmission of HIV

## Basic principles for experimental studies

- Randomization
  - to ensure that uncontrolled factors do not bias the experimental results.
- Control/Placebo
  - group of subjects or experimental units that are treated identically in every way, except that they do not receive an actual treatment. Allows for the assessment of treatment effect.
- Blinding
  - neither subjects nor anyone working with subjects should know who is receiving the treatment and who is getting the placebo to avoid bias. Desirable but not always possible.
- Replication
  - same treatments are assigned to different sampling units to help assess variation in the responses.

# Blocking, Matching

A PAIRED COMPARISON EXPERIMENT IS ONE OF THE MOST EFFECTIVE WAYS TO REDUCE NATURAL VARIABILITY WHILE COMPARING TREATMENTS. FOR EXAMPLE, IN COMPARING HAND CREAMS, THE TWO BRANDS ARE RANDOMLY ASSIGNED TO EACH SUBJECT'S RIGHT OR LEFT HANDS. THIS ELIMINATES VARIABILITY DUE TO SKIN DIFFERENCES.



Hypothesis: Lotions A and B equally effective at softening skin

Unpaired analysis

- 20 possible ways of assigning hands to two groups of 3
- Lots of variation between groups even without treatment!

Paired Analysis:

- 8 possible ways of assigning hands to two groups of 3
- Little variation between groups in absence of treatment.

# Observational Studies

- Sources of variability (in the outcome) are not controlled by the researcher
- Adjustment for imbalances between groups, if possible, occurs at the analysis phase
- Randomization usually not an option; samples are assumed to be “representative”
- Can identify association, but usually difficult to infer causation

## Examples

- natural history of HIV infection
- Fiber intake and coronary heart disease
- association between chess playing and reading skill in elementary school children

# Common (Clinical) Study Designs

- Ecologic
- Cross-sectional survey
- Prospective cohort
- Case-Control
- Randomized Clinical trial

# Ecologic studies

- Units of study are populations, not individuals
- Correlate rates of exposure with rates of disease
- Fast, cheap, useful to generate hypotheses, but susceptible to the “ecologic paradox”; causal inferences highly suspect

## Circumcision and HIV: Ecologic survey

*Bongaarts, AIDS 1989;3:373-7*

- 409 African ethnic groups
- Capital city HIV seroprevalence
- 20 countries >90% circumcised: HIV seroprevalence 0.9%
- 5 countries <25% circumcised: HIV seroprevalence 16.4%
- Correlation non-circ/HIV 0.9



# Cross-sectional survey

- All measurements at one point in time; random or representative sample (i.e. not selected on the basis of one of the factors)
- Good for estimating prevalence of a disease
- Efficient for examining relationships between common factors or a common disease and risk factors
- Shows association, not causation
- Political polls, health surveys are examples

## Circumcision and HIV in Kenya

*Agot, Epidemiology 2004;15:157-163*

- Single ethnic community Western Kenya
- 845 men with HIV-1 test results
- Circumcised 398, uncircumcised 447
- Prevalence ratio (RR) 1.5 uncircumcised associated with HIV-1

# Cohort studies

- Groups defined by exposure (exposed versus unexposed)
- Usually prospective, longitudinal
- Measures disease incidence; compare incidence in exposed to unexposed (RR)
- Strengths
  - good when exposure is rare
  - can examine multiple effects of an exposure
  - if prospective, can minimize bias in exposure ascertainment
  - exposure known to precede disease
- Weakness
  - inefficient for rare diseases
  - if prospective, expensive and time-consuming
  - validity jeopardized by loss to follow-up
- Strongest observational design, but causal inferences still suspect

# Cohort Study

Dietary fat/fiber and breast cancer: Nurses Health Study

*Willett JAMA 1992;268:2037-44*

- 89,494 women in NHS
- Follow-up 8 years
- 1,439 incident cases breast cancer (1.6%)
- No association between fiber or fat intake and incident breast cancer

# Case-control study

- Cases have the disease; Controls do not have the disease
- Compare (past) exposure rates in the two groups
- Useful for evaluating exposures that cannot be randomized
- Strengths:
  - Less time, lower cost compared to cohort study
  - Good for rare diseases, diseases with long incubation period
- Weaknesses:
  - Selection of cases and controls may be difficult
  - Recall bias or misclassification in determination of exposure
  - Temporal ordering of exposure and disease may be uncertain
- Causal inferences suspect

# Case-control study

## Vaginal adenocarcinoma

*Herbst NEJM 1971;284:787-881*

- Vaginal adenocarcinoma in 8 young women
- 4 controls per case
- Interviewed mothers
- 7/8 cases DESB during pregnancy versus 0/32 controls

## Glioma and mobile phones

*Hepworth BMJ 2006; 332:883-887*

- 996 cases of glioma, aged 18 – 69
- 1716 controls
- Interviews to determine mobile phone use patterns
- No association between glioma and recent mobile phone use

# Randomized clinical trial

- Participants allocated randomly to intervention versus no intervention
- Identical enrollment, data collection, follow-up, defined outcomes
- Outcome compared between randomized groups
- Advantages
  - Controls bias/confounding (groups similar except for intervention)
  - Control on exposure/treatment assignment
  - Can examine multiple outcomes
- Disadvantages
  - Expensive, time-consuming
  - Depends on compliance, high followup rate
  - Needs ethical equipoise
  - Entry criteria/participation bias may limit external generalizability
- Causal conclusions possible

# Randomized clinical trial

## Circumcision and HIV

*Auvert PLoS 2006*

- 3,273 men in South Africa (18-24 yrs) randomized to immediate or delayed circumcision
- Median 18 month follow-up
- HIV infections: 49 in control, 20 intervention (RR 0.4)
- Controlling for sexual behavior, condom use - results unchanged

## Mwanza STD/HIV

*Grosskurth Lancet 1995;346:530-6*

- Community randomized clinical trial
- 6 pair-matched communities; 1,000 adults followed for 2 years each community
- Intervention: improved STD diagnosis and treatment infrastructure
- HIV incidence: 1.2% intervention, 1.9% control (RR 0.58)

# Confounding (aka Simpson's Paradox)

“Condom Use increases the risk of STD”

		STD rate	
Condom Use	Yes	55/95	(61%)
	No	45/105	(43%)

		STD rate	
<b># Partners &lt; 5</b>			
Condom Use	Yes	5/15	(33%)
	No	30/82	(37%)
<b># Partners ≥ 5</b>			
Condom Use	Yes	50/80	(62%)
	No	15/23	(65%)

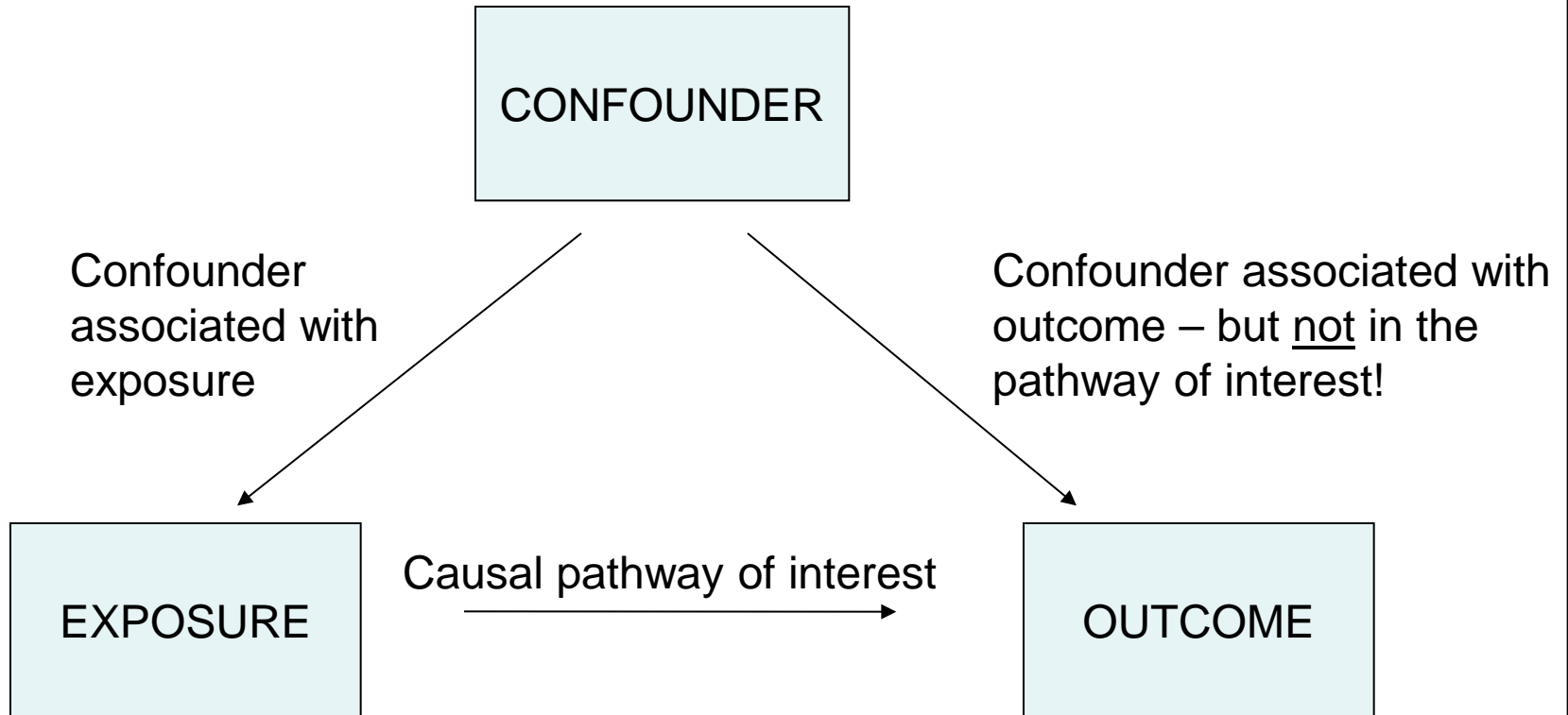
- Individuals with more partners are more likely to use condoms. But individuals with more partners are also more likely to get an STD.
- In general, effect of exposure on outcome is constant across levels of the confounder; however, pooled table is different (compare to EM)



# Confounding

- A critical and common problem in observational studies
- Confounding occurs when there is an imbalance between the exposure groups with respect to some other risk factor for the disease (think of it as a form of selection bias)
- Because of the confounding phenomenon, it's important that we don't automatically assume causation whenever we see an association.
- Statistical methods allow us to evaluate the association between two variables. As shown in the previous example, we can also use statistical methods to adjust for confounding provided we measured the confounder

# Confounding



# Confounding

USA Today: “Prayer lowers blood pressure”

“Attending religious services lowers blood pressure more than tuning into religious TV or radio, a new study says. People who attended a religious service once a week and prayed or studied Bible once a day were 40% less likely to have high blood pressure than those who didn’t go to church every week and prayed and studied the Bible less.”

# Confounding

## Conceptual Model

## Prayer causes lower BP?

Church & prayer → Low BP

Yes

healthy → Church & prayer  
          ↗ Low BP

No

social → Church & prayer  
          ↗ Low BP

No

From the information given you can't tell which model is correct!

# Confounding

- The difficulty with observational data is that “exposure” is not randomly assigned. Thus, the exposure groups (prayer/no prayer) may not be the same in all other important respects
- Additional examples:
  - CD4 cell count among those treated with AZT
  - Gender and college admission rates
  - and many more ...
- Confounding can also occur with other measures of association (e.g. regression)

**Q:** What can we do in these situations?

**A:** Control for imbalances via stratification (or other statistical methods)

**A:** Be cautious in our thinking and use of language

# Causal Inference Concepts

- What do we really mean when we say “A causes B”?
- How can we define the “causal effect” of prayer on blood pressure?
  - BP of subject  $i$  if he/she prays –  $Y_i(1)$
  - BP of subject  $i$  if he/she doesn't pray –  $Y_i(0)$
  - $Y_i(0)$  and  $Y_i(1)$  are called “counterfactual outcomes”
  - Define the **causal effect** as  $\Delta_i = Y_i(1) - Y_i(0)$
- In practice, we observe  $Y_i(0)$  or  $Y_i(1)$ , but not both, so ... *We can never observe  $\Delta_i$*

# Causal Inference Concepts

Define:

$\bar{Y}(1)$  is the average blood pressure **if everyone prayed**

$\bar{Y}(0)$  is the average blood pressure **if everyone did not pray**

- Although it is impossible to know  $\Delta_i$  it is sometimes possible to estimate the **average causal effect**:

$$\bar{\Delta} = \bar{Y}(1) - \bar{Y}(0)$$

- How? (Hint: If you can't know  $Y_i(1)$  on everybody, what's the next best thing?)

# Causal Inference Concepts

- Take a sample!

BP w/ prayer

BP w/o prayer

$Y_1(1)$

$Y_1(0)$

$Y_2(1)$

$Y_2(0)$

$Y_3(1)$

$Y_3(0)$

$Y_4(1)$

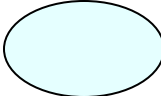
$Y_4(0)$

$Y_5(1)$

$Y_5(0)$

:

:

 = observed data

Estimate  $\bar{Y}(1)$  by  $Y_1(1) + Y_2(1) + Y_5(1) + \dots / n_1$

Estimate  $\bar{Y}(0)$  by  $Y_3(0) + Y_4(0) + \dots / n_0$

- What key assumption did you have to make?



# Causal Inference Concepts

- We can estimate the average causal effect when there is nothing (other than exposure) that systematically differs between exposed and unexposed groups
- Randomization guarantees this – “no unmeasured confounding”
- With observational data the average outcome among those actually exposed **may not be equal** to the average outcome that would be observed if everyone was exposed.
- Sometimes, we can control for imbalances via stratification (or other statistical methods) *but only if you have measured the confounding factors.*
- Different populations (i.e. young, old) may have different average causal effects (effect modification)

# Confounding

**Example:** Does consumption of fish oil reduce risk of a heart attack?

Consider the following two study designs:

1. Individuals with and without a recent MI are asked about their consumption of fish and/or fish oil over the past 5 years.
  2. Individuals at risk for MI are randomized to daily fish oil capsules or placebo and followed for 2 years.
- Which design is less likely to be affected by confounding?

# Problems in Design/Data Collection

## **Example:**

33% reduction in blood pressure after treatment with medication in a sample of 60 hypertensive men.

Problem:

## **Example:**

Daytime telephone interview of voting preferences

Problem:

## **Example:**

Higher proportion of “abnormal” values on tests performed in 1990 than a comparable sample taken in 1980.

Problem:

# Summary

1. Statistics plays a role from study conception to study reporting.
2. Statistics is concerned with making valid inferences about populations from samples that are subject to various sources of variability.
3. Different studies designs have different strengths and weaknesses and may require different statistical approaches.
4. The potential for confounding means we must be careful in making causal interpretations from observational studies
5. **You must understand the study design and sampling procedures before you can hope to interpret the data!!**



# Probability

- Probability - meaning
  - 1) classical
  - 2) frequentist
  - 3) subjective (personal)
- Sample space, events
- Mutually exclusive, independence
- and, or, complement
- Joint, marginal, conditional probability
- Probability - rules
  - 1) Addition
  - 2) Multiplication
  - 3) Total probability
  - 4) Bayes
- Screening
  - sensitivity
  - specificity
  - predictive values

# Probability

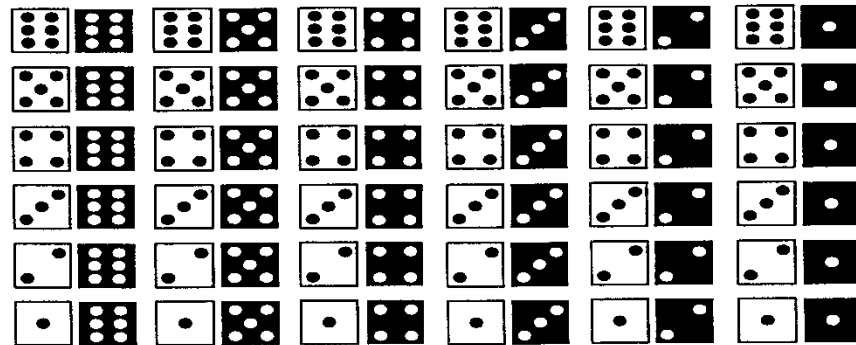
Probability provides a measure of uncertainty associated with the occurrence of events or outcomes

Definitions:

1. **Classical:**  $P(E) = m/N$

If an event can occur in  $N$  mutually exclusive, equally likely ways, and if  $m$  of these possess characteristic  $E$ , then the probability is equal to  $m/N$ .

Eg What is the probability of rolling a total of 7 on two dice?

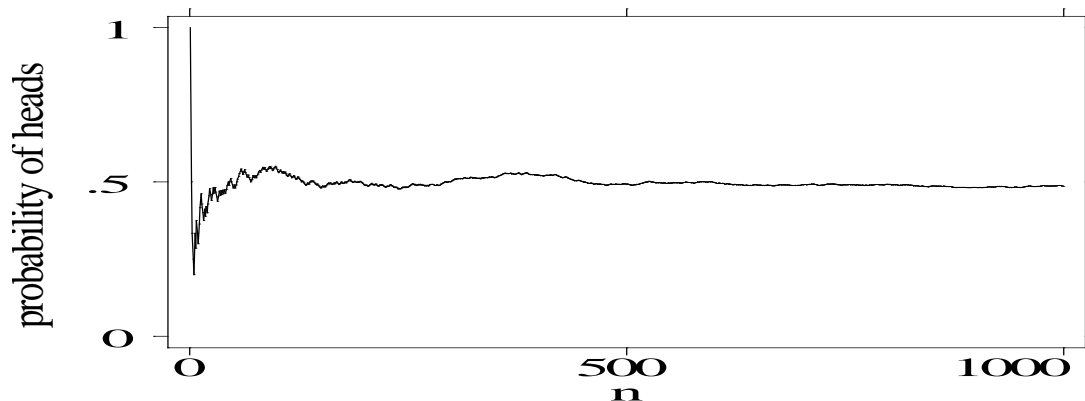


# Relative Frequency

2. **Relative frequency:**  $P(E) \approx m/n$

If a process or an experiment is repeated a large number of times,  $n$ , and if the characteristic,  $E$ , occurs  $m$  times, then the relative frequency,  $m/n$ , of  $E$  will be approximately equal to the probability of  $E$ .

E.g. Around 1900, the English statistician Karl Pearson heroically tossed a coin 24,000 times and recorded 12,012 heads, giving a proportion of 0.5005.



# Personal Probability

## 3. **Personal probability (subjective)**

- Your personal degree of uncertainty.

E.g. What is the probability of life on Mars?

What is probability Huskies will win Pac-12 in basketball in 2012-2013 season?



# Basic Terminology and Properties

- The **sample space** consists of the possible outcomes of an experiment (could be infinite). An **event** is an outcome or set of outcomes. e.g. roll a die; sample space is (1,2,3,4,5,6), an event is (roll a 3 or a 5)
- Two events, A and B, are said to be **mutually exclusive** (disjoint) if only one or the other, but not both, can occur in a particular experiment. e.g. roll a die; events (roll a 3) and (roll an even number) are mutually exclusive
- Probability of an event A, denoted  $P(A)$ , must be between 0 and 1 (inclusive)
- Probabilities of any exhaustive collection (i.e. at least one must occur) of mutually exclusive events is 1
- The probability of all events other than an event A is denoted by  $P(A^c)$  [“A complement”] or  $P(\bar{A})$  [“A bar”], and  $P(A^c) = 1 - P(A)$

# Basic Properties of Probability

**Example:** Roll a single die and consider the following events:

$E_1 =$  roll a 1

$E_2 =$  roll an even number

$E_3 =$  roll a 4, 5 or 6

$E_4 =$  roll a 3 or 5

- 1) What is  $\Pr(E_4)$ ?
- 2) Are  $E_2$  and  $E_3$  mutually exclusive?  $E_2$  and  $E_4$ ?
- 3) Find a mutually exclusive, exhaustive collection of events. Do the probabilities add to 1?
- 4) What is  $\Pr(E_4^c)$ ?

# Combining events

- If A and B are any two events then we write

$$P(A \text{ or } B)$$

to indicate the probability that event A or event B (or both) occurred.

- If A and B are any two events then we write

$$P(A \text{ and } B) \text{ or } P(AB)$$

to indicate the probability that both A and B occurred.

- If A and B are any two events then we write

$$P(A \text{ given } B) \text{ or } P(A|B)$$

to indicate the probability of A among the subset of cases in which B is known to have occurred.

# Probability

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	9	80	89
	Neg.	1	9910	9,911
		10	9990	10,000

What is  $P(\text{test positive})$ ?

What is  $P(\text{test positive or disease positive})$ ?

What is  $P(\text{test positive and disease positive})$ ?

What is  $P(\text{test positive} \mid \text{disease positive})$ ?

What is  $P(\text{disease positive} \mid \text{test positive})$ ?

2.6.2. The following table shows the first 1000 patients admitted to a clinic for retarded children by diagnostic classification and level of intelligence. For this group find:

- 9/1000 (a)  $P(A_3 \cap B_4)$ .
- 114/1000 (b) The probability that a patient picked at random is severely retarded.
- 376/1000 (c) The probability that a patient picked at random is either not retarded or is borderline.
- 4/1000 (d) The probability that a patient picked at random is profoundly retarded and has Down's syndrome.
- 4/160 (e) The probability that a patient is profoundly retarded, given that he has Down's syndrome.

Major Diagnostic Classification	Level of Retardation						Total
	$A_1$ Not Retarded	$A_2$ Pro-found	$A_3$ Severe	$A_4$ Moderate	$A_5$ Mild	$A_6$ Border-line	
$B_1$ Encephalopathies	33	38	57	114	103	55	400
$B_2$ Down's syndrome	2	4	34	88	27	5	160
$B_3$ Congenital cerebral defect	10	2	6	6	6	0	30
$B_4$ Mental retardation of unknown cause	0	0	9	36	62	35	142
$B_5$ Other	161	0	8	16	8	75	268
<b>Total</b>	<b>206</b>	<b>44</b>	<b>114</b>	<b>260</b>	<b>206</b>	<b>170</b>	<b>1000</b>

# General Rules

## 1) Addition (“or”) rule

If two events A and B are not mutually exclusive, then the probability that event A or event B occurs is:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

**E.g.** Of the students at Anytown High school, 40% have had the mumps, 70% have had measles and 32% have had both. What is the probability that a randomly chosen student has had at least one of the above diseases?

		Measles		
		Yes	No	Total
Mumps	Yes	32		40
	No			
Total		70		100

# General Rules

## 2) Multiplication (“and”) rule (special case – independence)

If two events, A and B, are “independent” (probability of one does not depend on outcome of the other) then

$$P(AB) = P(A)P(B)$$

E.g. From the data on the previous page, does it appear that mumps and measles are independent?

Easy to extend for independent events A,B,C,...

$$P(ABC\dots) = P(A)P(B)P(C)\dots$$

# Independence

To check for independence, you can check any of the following ...

$$\begin{aligned} P(A|B) &= P(A) \text{ or} \\ P(B|A) &= P(B) \text{ or} \\ P(AB) &= P(A)P(B). \end{aligned}$$

If any one holds, then all three hold; if any one is violated, then all are violated

		Measles		
		Yes	No	Total
Mumps	Yes	32	8	40
	No	38	22	60
Total		70	30	100

The notion of independent events is pervasive throughout statistics ...





# General Rules

## 2) Multiplication (“and”) rule – general case

The general formula for the probability that both A and B will occur is

$$P(AB) = P(A | B)P(B) = P(B | A)P(A)$$

E.g.  $P(\text{mumps}) = 0.40$  and  $P(\text{both}) = 0.32$ . Find  $P(\text{measles}|\text{mumps})$ .

$$P(\text{measles} | \text{mumps}) P(\text{mumps}) = P(\text{both})$$

$$P(\text{measles} | \text{mumps}) * 0.40 = 0.32$$

$$P(\text{measles} | \text{mumps}) = 0.80$$

### 3) Total probability rule

If  $A_1, \dots, A_n$  are mutually exclusive, exhaustive events, then

$$P(B) = \sum_{i=1}^n P(B | A_i)P(A_i)$$

**E.g.** The following table gives the estimated proportion of individuals with Alzheimer's disease by age group. It also gives the proportion of the general population that are expected to fall in the age group in 2030. What proportion of the population in 2030 will have Alzheimer's disease?

		Proportion population	Proportion with AD	Hypoth. population	Number affected
Age group	< 65	.80	.00	80,000	0
	65 – 75	.11	.01	11,000	110
	75 – 85	.07	.07	7,000	490
	> 85	.02	.25	2,000	500
				100,000	1100

$$P(\text{AD}) = 0*.8 + .01*.11 + .07*.07 + .25*.02 = .011$$

# Bayes Rule (Theorem)

Bayes rule - combines multiplication rule with total probability rule

$$\begin{aligned}P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{\sum_{i=1}^n P(B|A_i)P(A_i)}\end{aligned}$$

We will only apply this to the situation where A and B have two levels each, say, A and A<sup>c</sup>, B and B<sup>c</sup>. The formula becomes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B | A^c)P(A^c)}$$

# Screening – application of Bayes Rule

Suppose we have a random sample of a population...

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	30	120
	Neg.	10	970	980
		100	1000	1100

A = disease pos.      B = test pos.

$$\text{Prevalence} = P(A) = 100/1100 = .091$$

$$\text{Sensitivity} = P(B | A) = 90/100 = .9$$

$$\text{Specificity} = P(B^c | A^c) = 970/1000 = .97$$

$$\text{PVP} = P(A | B) = 90/120 = .75$$

$$\text{PVN} = P(A^c | B^c) = 970/980 = .99$$

# Screening – application of Bayes Rule

Now suppose we have taken a sample of 100 disease positive and 100 disease negative individuals (e.g. case-control design)

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	3	93
	Neg.	10	97	107
		100	100	200

A = disease pos.      B = test pos.

Prevalence = ??? (not .5!)

Sensitivity =  $P(B | A) = 90/100 = .9$

Specificity =  $P(B^c | A^c) = 97/100 = .97$

PVP =  $P(A | B) = 90/93$  **NO!**

PVN =  $P(A^c | B^c) = 97/107$  **NO!**

# Screening – application of Bayes Rule

A = disease pos.

B = test pos.

Assume we know, from external sources, that  $P(A) = 100/1100$ .  
Then for every 100 disease positives we should have 1000  
disease negatives .... 1:10.

Make a mock table ...

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	$3 \times 10$	120
	Neg.	10	$97 \times 10$	980
		100	$100 \times 10$	1100

$$PVP = \frac{90}{90 + 3 \times 10} = .75$$

# Screening – application of Bayes Rule

Now, use Bayes rule ...

$$\begin{aligned} \text{PVP} = P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{.9 \times \frac{100}{1100}}{.9 \times \frac{100}{1100} + .03 \times \frac{1000}{1100}} \\ &= \frac{.9 \times 100}{.9 \times 100 + .03 \times 1000} = .75 \end{aligned}$$

Another way of thinking about this – the case control design has given us a biased sample (too many cases). Bayes formula is just giving the appropriate weights to remove the bias.

# Summary

- Probability - meaning
  - 1) classical
  - 2) frequentist
  - 3) subjective (personal)
- Sample space, events, complement
- Mutually exclusive, independence
- and, or, given
- Joint, marginal, conditional probability
- Probability - rules
  - 1) Addition
  - 2) Multiplication
  - 3) Total probability
  - 4) Bayes
- Screening
  - sensitivity
  - specificity
  - predictive values