# PBIO 504

# Introductory Biostatistics

Goal: Learn principles and basic methods to reason with good statistical insight, understand how to read and interpret statistical methods when reviewing articles, and learn how to perform basic statistical analysis

# Lecture 1: Introduction, Graphs and Descriptive Statistics

# What is Statistics?

*Statistics: Your chance for happiness (or misery)* - XL Meng

- *“I keep saying the sexy job in the next 10 years will be statisticians.” Hal Varian, Google*
- *"AP Statistics was the most boring course I took in high school!“*
- *“You are in pain, which treatment should you get”*
- *Statistics is everywhere. We are drowning in data (Big Data) but starving for information.*

# Which treatment is better?

- Simpson's paradox → Personalized medicine

- Charig et. al. (BMJ. 292 (6524): 879. 1986) evaluated two treatments for kidney stones. Rx A has a success rate of 78% (273/350) and Rx B, 83% (289/350). Which one should you choose? *Surely B?*

# Which treatment should I get

- When Rx A and Rx B are applied to those who suffer **small** stones, the success rates are, respectively, 93% (81/87) and 87% (234/270)

- When they are applied to those with **large** stones, the success rate for Rx A is 73% (192/263) and for Rx B it is 69% (55/80).

- That is, regardless of the size of the kidney stone, treatment A has a higher success rate.

# Simpson's Paradox

- Edward Simpson first described this paradox in his 1951 paper "The Interpretation of Interaction in Contingency Tables."
- Actually Pearson and Yule independently observed a similar paradox half a century earlier.

- *Often when we look at aggregated data we may encounter this paradox.*

# What is Statistics?

- The **OBJECTIVE** of **STATISTICS** is to make an **INFERENCE** about a **POPULATION** based on **INFORMATION** in a **SAMPLE** taken from that population.
- A **POPULATION** is the set of all measurements on all patients of interest.
- A **SAMPLE** is a subset of measurements selected from a population
- A discipline to study uncertainty for decision making

# Understanding Biostatistics

- Interpreting vital statistics
- Understanding epidemiologic data
- Interpreting information about testing new drugs or medical equipment
- Understanding diagnostic tests and procedures
- Staying informed
- Appraising guidelines
- Evaluating study protocols and articles
- Participating in or directing research projects

WSJ December 2012: Reasoning with statistical insights

THE NUMBERS GUY | By Carl Bialik

# Statistical Habits to Add, or Subtract, in 2013

In the year ending Monday, we saw some gains in statistical savvy: Data-crunching pollsters accurately forecast the outcome of the presidential election; the Memphis Grizzlies hired a vice president of basketball operations for his statistical expertise; and folks grew comfortable with the phrase "big data," to describe the billions of billions of bytes generated daily by information technology.

The growing importance of statistical analysis is set to be a theme of next year, too, with more than 150 professional organizations worldwide, including the American Statistical Association, designating 2013 as the International Year of Statistics.

All of which should increase the pressure on writers, public speakers, researchers and corporate officials to improve how they present data to the public. I asked a half-dozen professional statisticians for their pet peeves about how numbers are presented and what resolutions they would suggest statistically-minded people make for the new year. I also solicited ideas from readers on my blog.

1. A major concern running through the responses is that data tend to be fuzzy—much fuzzier than they can seem when stated with neither margins of error nor qualification. "The most important numerical fallacy is that people tend to think of numbers as known, constant and having no variability," said Donald Berry, a biostatistician at the University of Texas MD Anderson Cancer Center in Houston.


Associated Press

Revelers in New York ring in 2012. Some New Year's resolutions could be useful in the statistical world.

Many readers agreed. Richard Hoffbeck, a retired research data analyst in Minneapolis, asked for reports on job numbers and economic forecasts to include estimates of uncertainty—the statistical margin of error that is common in poll reporting.

"I never see them mentioned," he said.

Statisticians and readers also suggested ways to avoid some common traps next year:

2. **Be patient:** Don't rush to anoint the next big stock, or slugger or pill. A small sample size can yield extreme results just by chance. A company or athlete on a hot streak may be really good, but likely isn't as good as, say, two record weeks would suggest.

Statisticians call this phenomenon regression to the mean, and taking it into account involves using prior knowledge. In the case of a previously untouted stock or category of drug, the prior knowledge is that most stock prices don't skyrocket and most tested drugs don't work. So treat short-term gains with caution.

With a baseball player, a hot start is likely to fizzle. Prof. Berry predicts that whoever is leading the major leagues in batting average on May 15 next season won't be able to sustain the hot start.

"His batting average is going to drop," he said. "You can bank on it."

3. **Provide context:** One reader complained that too many commentators report the daily movement in stock indexes in terms of points gained or lost, rather than percentages. "These are meaningless numbers without knowing what the baseline is," the reader wrote.

Context is also useful in medical studies.

A seemingly large effect from a drug discovered in an observational study carries much less weight than one found in an experimental, double-blind study, said Robert N. Rodriguez, president of the American Statistical Association. "Experimental studies are much more reliable for deciding whether a factor has a causal effect."

An experimental study would randomly assign people to two groups that receive different treatments, while an observational one would compare people in the real world who already get one of the two treatments.

Conversely, the fact that a single study doesn't find a meaningful effect doesn't mean there is none—the study might be flawed, or the effect might be too small to be picked up. "The crux of the problem is that people interpret an absence of evidence as evidence for absence," said Carlisle Rainey, a doctoral student in political science and statistics at Florida State University.

Gregory Taylor, a math teacher in Ottawa, said that too often the context of a percentage or fraction is left out. "A diet that is effective on eight out of 10 people won't necessarily be useful to me if it was only tested on teenagers," he said. "Too often people focus on the number, and miss what it's talking about."

Context can also help temper excitement about apparent record breakers. That box-office mark? Try adjusting it for ticket-price inflation.

4. **Believe in miracles:** Seemingly improbable events do happen. People win the lottery twice, have three children years apart with the same birth date, or bump into old acquaintances on the other side of the world.

To the people involved, such occurrences may seem providential, but it is worth remembering that with seven billion people in the world, the same thing could happen to a lot of people.

The probability of a seemingly surprising coincidence, like winning the lottery twice, "is actually quite high, if you mean anyone, anytime" winning for a second time, said Jessica Utts, who heads the statistics department at the University of California, Irvine.

*Learn more about this topic at WSJ.com/NumbersGuy. Email numbersguy@wsj.com.*

# *Flying Is Safest Since Dawn of Jet Age*

**Safer Skies**

While the number of fatal crashes involving smaller turboprops

# Multivitamins in the Prevention of Cardiovascular Disease in Men

## The Physicians' Health Study II Randomized Controlled Trial

Howard D. Sesso, ScD, MPH
William G. Christen, ScD
Vadim Bubes, PhD
Joanne P. Smith, BA
Jean MacFadyen, BA
Miriam Schvartz, MD
JoAnn E. Manson, MD, DrPH
Robert J. Glynn, ScD
Julie E. Buring, ScD
J. Michael Gaziano, MD, MPH

**Context** Although multivitamins are used to prevent vitamin and mineral deficiency, there is a perception that multivitamins may prevent cardiovascular disease (CVD). Observational studies have shown inconsistent associations between regular multivitamin use and CVD, with no long-term clinical trials of multivitamin use.

**Objective** To determine whether long-term multivitamin supplementation decreases the risk of major cardiovascular events among men.

**Design, Setting, and Participants** The Physicians' Health Study II, a randomized, double-blind, placebo-controlled trial of a common daily multivitamin, began in 1997 with continued treatment and follow-up through June 1, 2011. A total of 14 641 male US physicians initially aged 50 years or older (mean, 64.3 [SD, 9.2] years), including 754 men with a history of CVD at randomization, were enrolled.

**Intervention** Daily multivitamin or placebo.

**Main Outcome Measures** Composite end point of major cardiovascular events, including nonfatal myocardial infarction (MI), nonfatal stroke, and CVD mortality. Secondary outcomes included MI and stroke individually.

**Results** During a median follow-up of 11.2 (interquartile range, 10.7-13.3) years, there were 1732 confirmed major cardiovascular events. Compared with placebo, there was no significant effect of a daily multivitamin on major cardiovascular events (11.0 and 10.8 events per 1000 person-years for multivitamin vs placebo, respectively; hazard ratio [HR], 1.01; 95% CI, 0.91-1.10; $P=.91$). Further, a daily multivitamin had no effect on total MI (3.9 and 4.2 events per 1000 person-years; HR, 0.93; 95% CI, 0.80-1.09; $P=.39$), total stroke (4.1 and 3.9 events per 1000 person-years; HR, 1.06; 95% CI, 0.91-1.23; $P=.48$), or CVD mortality (5.0 and 5.1 events per 1000 person-years; HR, 0.95; 95% CI, 0.83-1.09; $P=.47$). A daily multivitamin was also not significantly associated with total mortality (HR, 0.94; 95% CI, 0.88-1.02; $P=.13$). The effect of a daily multivitamin on major cardiovascular events did not differ between men with or without a baseline history of CVD ($P=.62$ for interaction).

**Conclusion** Among this population of US male physicians, taking a daily multivitamin did not reduce major cardiovascular events, MI, stroke, and CVD mortality after more than a decade of treatment and follow-up.

**Trial Registration** clinicaltrials.gov Identifier: NCT00270647

*JAMA. 2012;308(17):1751-1760* www.jama.com

DESPITE UNCERTAINTY REGARDING the long-term health benefits of vitamins, many US adults take vitamin supplements[1] to prevent chronic diseases[2] or for general health and well-being.[3] Because multivitamins are the most common supplement taken by US adults,[4,5] there are broad public health implications regarding their everyday use. Individuals who believe they are deriving benefits from supplements may be less likely to engage in other preventive health behaviors, and chronic use of daily supplements poses a financial burden, with annual vitamin supplement sales in the billions of US dollars.[6]

A daily multivitamin, with its combination of essential vitamins and minerals that meet minimum recommended

**For editorial comment see p 1802.**

**Author Video Interview available at www.jama.com.**

**Author Affiliations:** Divisions of Preventive Medicine (Drs Sesso, Christen, Bubes, Schvartz, Manson, Glynn, Buring, and Gaziano and Mss Smith and MacFadyen), Aging (Drs Sesso, Buring, and Gaziano), and Cardiovascular Disease (Dr Gaziano), Department of Medicine, Brigham and Women's Hospital and Harvard Medical School; Departments of Epidemiology (Drs Sesso, Manson, and Buring) and Biostatistics (Dr Glynn), Harvard School of Public Health; Department of Ambulatory Care and Prevention, Harvard Medical School (Dr Buring), and Harvard Medical School and VA Boston Healthcare Center (Dr Gaziano), Boston, Massachusetts. Dr Gaziano is also Contributing Editor, *JAMA*.
**Corresponding Author:** Howard D. Sesso, ScD, MPH, Brigham and Women's Hospital, 900 Commonwealth Ave E, Third Floor, Boston, MA 02215 (hsesso@hsph.harvard.edu).

# Multivitamins in the Prevention of Cancer in Men

## The Physicians' Health Study II Randomized Controlled Trial

J. Michael Gaziano, MD, MPH
Howard D. Sesso, ScD, MPH
William G. Christen, ScD
Vadim Bubes, PhD
Joanne P. Smith, BA
Jean MacFadyen, BA
Miriam Schvartz, MD
JoAnn E. Manson, MD, DrPH
Robert J. Glynn, ScD
Julie E. Buring, ScD

**Context** Multivitamin preparations are the most common dietary supplement, taken by at least one-third of all US adults. Observational studies have not provided evidence regarding associations of multivitamin use with total and site-specific cancer incidence or mortality.

**Objective** To determine whether long-term multivitamin supplementation decreases the risk of total and site-specific cancer events among men.

**Design, Setting, and Participants** A large-scale, randomized, double-blind, placebo-controlled trial (Physicians' Health Study II) of 14 641 male US physicians initially aged 50 years or older (mean [SD] age, 64.3 [9.2] years), including 1312 men with a history of cancer at randomization, enrolled in a common multivitamin study that began in 1997 with treatment and follow-up through June 1, 2011.

**Intervention** Daily multivitamin or placebo.

**Main Outcome Measures** Total cancer (excluding nonmelanoma skin cancer), with prostate, colorectal, and other site-specific cancers among the secondary end points.

**Results** During a median (interquartile range) follow-up of 11.2 (10.7-13.3) years, there were 2669 men with confirmed cancer, including 1373 cases of prostate cancer and 210 cases of colorectal cancer. Compared with placebo, men taking a daily multivitamin had a statistically significant reduction in the incidence of total cancer (multivitamin and placebo groups, 17.0 and 18.3 events, respectively, per 1000 person-years; hazard ratio [HR], 0.92; 95% CI, 0.86-0.998; $P$=.04). There was no significant effect of a daily multivitamin on prostate cancer (multivitamin and placebo groups, 9.1 and 9.2 events, respectively, per 1000 person-years; HR, 0.98; 95% CI, 0.88-1.09; $P$=.76), colorectal cancer (multivitamin and placebo groups, 1.2 and 1.4 events, respectively, per 1000 person-years; HR, 0.89; 95% CI, 0.68-1.17; $P$=.39), or other site-specific cancers. There was no significant difference in the risk of cancer mortality (multivitamin and placebo groups, 4.9 and 5.6 events, respectively, per 1000 person-years; HR, 0.88; 95% CI, 0.77-1.01; $P$=.07). Daily multivitamin use was associated with a reduction in total cancer among 1312 men with a baseline history of cancer (HR, 0.73; 95% CI, 0.56-0.96; $P$=.02), but this did not differ significantly from that among 13 329 men initially without cancer (HR, 0.94; 95% CI, 0.87-1.02; $P$=.15; $P$ for interaction=.07).

**Conclusion** In this large prevention trial of male physicians, daily multivitamin supplementation modestly but significantly reduced the risk of total cancer.

**Trial Registration** clinicaltrials.gov Identifier: NCT00270647

*JAMA. 2012;308(18):doi:10.1001/jama.2012.14641* www.jama.com

MULTIVITAMINS ARE THE most common dietary supplement, regularly taken by at least one-third of US adults.[1,2] The traditional role of a daily multivitamin is to prevent nutritional deficiency. The combination of essential vitamins and minerals contained in multivitamins may mirror healthier dietary patterns such as fruit and vegetable intake, which have been modestly and inversely associated with cancer risk in some,[3] but not all,[4,5] epidemiologic studies. Observational studies of long-term multivitamin use and cancer end points have been inconsistent.[6-12] To date, large-scale randomized trials testing single or small numbers of higher-dose individual vitamins and minerals for cancer have generally found a lack of effect.[13-18]

According to the 2010 Dietary Guidelines for Americans, "For the general, healthy population, there is no evidence to support a recommendation for the use of multivitamin/mineral supplements in the primary prevention of

**Author Affiliations:** Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Drs Gaziano, Sesso, Christen, Bubes, Schvartz, Manson, Glynn, and Buring and Mss Smith and MacFadyen); VA Boston Healthcare System, Boston, Massachusetts (Dr Gaziano); and Departments of Epidemiology (Drs Sesso, Manson, and Buring) and Biostatistics (Dr Glynn), Harvard School of Public Health, Boston, Massachusetts. Dr Gaziano is also Contributing Editor, *JAMA*.
**Corresponding Author:** J. Michael Gaziano, MD, MPH, Department of Medicine, Brigham and Women's Hospital, 1620 Tremont St, Boston, MA 02120 (jmgaziano@partners.org).

Annals of Internal Medicine

# Effectiveness of Acupuncture as Adjunctive Therapy in Osteoarthritis of the Knee

## A Randomized, Controlled Trial

Brian M. Berman, MD; Lixing Lao, PhD; Patricia Langenberg, PhD; Wen Lin Lee, PhD; Adele M.K. Gilpin, PhD; and Marc C. Hochberg, MD

**Background:** Evidence on the efficacy of acupuncture for reducing the pain and dysfunction of osteoarthritis is equivocal.

**Objective:** To determine whether acupuncture provides greater pain relief and improved function compared with sham acupuncture or education in patients with osteoarthritis of the knee.

**Design:** Randomized, controlled trial.

**Setting:** Two outpatient clinics (an integrative medicine facility and a rheumatology facility) located in academic teaching hospitals and 1 clinical trials facility.

**Patients:** 570 patients with osteoarthritis of the knee (mean age [±SD], 65.5 ± 8.4 years).

**Intervention:** 23 true acupuncture sessions over 26 weeks. Controls received 6 two-hour sessions over 12 weeks or 23 sham acupuncture sessions over 26 weeks.

**Measurements:** Primary outcomes were changes in the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain and function scores at 8 and 26 weeks. Secondary outcomes were patient global assessment, 6-minute walk distance, and physical health scores of the 36-Item Short-Form Health Survey (SF-36).

**Results:** Participants in the true acupuncture group experienced greater improvement in WOMAC function scores than the sham acupuncture group at 8 weeks (mean difference, −2.9 [95% CI, −5.0 to −0.8]; $P = 0.01$) but not in WOMAC pain score (mean difference, −0.5 [CI, −1.2 to 0.2]; $P = 0.18$) or the patient global assessment (mean difference, 0.16 [CI, −0.02 to 0.34]; $P > 0.2$). At 26 weeks, the true acupuncture group experienced significantly greater improvement than the sham group in the WOMAC function score (mean difference, −2.5 [CI, −4.7 to −0.4]; $P = 0.01$), WOMAC pain score (mean difference, −0.87 [CI, −1.58 to −0.16]; $P = 0.003$), and patient global assessment (mean difference, 0.26 [CI, 0.07 to 0.45]; $P = 0.02$).

**Limitations:** At 26 weeks, 43% of the participants in the education group and 25% in each of the true and sham acupuncture groups were not available for analysis.

**Conclusions:** Acupuncture seems to provide improvement in function and pain relief as an adjunctive therapy for osteoarthritis of the knee when compared with credible sham acupuncture and education control groups.

Ann Intern Med 2004;141:901–910. www.annals.org

For author affiliations, see end of text.

See related articles on pp 911–919 and pp 920–928.

# Reviewing the literature

- Sampling of the study population
- Unit of observation (person or thing) upon which data is collected
- Study hypotheses
- Data collection
  - Sample size; selection bias; observation bias; confounding
- Data analyses
  - Methods to control confounding; measures of association; confidence intervals; statistical tests
- Interpretation of data - Generalizability

# Population and Samples

- Population: a set of units (usually people, objects, or events).
- Sample: a subset of the units of a population.
- Random sample: a subset in which every element in the population had an equal chance of being selected in the sample.

# Data Analyses

- Understand the study design
- Determine appropriate use of data and adequate statistical methods to test hypotheses
- Evaluate inconsistencies of data and identify problems with data collection
- Select and define relevant variables
- Make inferences from data

# DATA

- **Data** consists of observations or measurements of characteristics or certain properties of the study population

  - **Quantitative:** observations or measurements that are numerical

  - **Qualitative:** a general description of properties that cannot be recorded numerically

# Types of Data

- **Categorical**: only a finite number of values exist.
  - *ordinal* (where the values are ordered)
  - *nominal* (where the values are in no particular order)

- **Continuous**: a large or infinite number of values exist.

Review the textbook: pages 7-11

**Variables**: Any particular characteristic may “vary” among the units in a population

- Gender: male or female
- Pain severity: mild, moderate, high
- Number of children in a family is a variable that is limited to integer values (0, 1, 2, etc.)
- Temperature
- Height, Weight, Cholesterol level, etc.

# Variables

- **Independent** variables are those independent of any effects of other variables

- **Dependent** variable is the variable we are interested in. It depends on the level or presence or amount of some other variable.

# Presenting Nominal and Ordinal Data

**Frequency distribution** (or frequency table) is a tabular summary of a set of data points representing the frequency (or number) of points in each group or category.

**Relative frequency** of a category is the frequency of that category divided by the total number of observations in the sample.

# Proportion

- The number of observations with the characteristic of interest divided by the total number of observations

- ***Example*: at an animal shelter today there are 10 poodles, 5 labradors, and 5 beagles**

Proportion of Beagles = 5 / (10+5+5) = 0.25

(or 25%)

# Ratio

- The number of observations with the characteristic of interest dived by the number without the characteristic of interest.

- ***Example*: at an animal shelter today there are 10 poodles, 5 labradors, and 5 beagles**

Ratio of Beagles to non-Beagles = 5 / (10+5) = 1/3 (or 0.33)

# Rate

- The number of those with a particular outcome divided by the size of the study population in a certain time period, multiplied by a base (e.g. 100, 1,000, 10,000, or 100,000)
- Usually it involves a dimension of time
- Examples:
  - Mortality rate
  - Attack rate
  - Person-time rate

# Examples from the Literature

Review the textbook: chapter 2

# Use of Graphics to Represent Data and Predictive Modeling

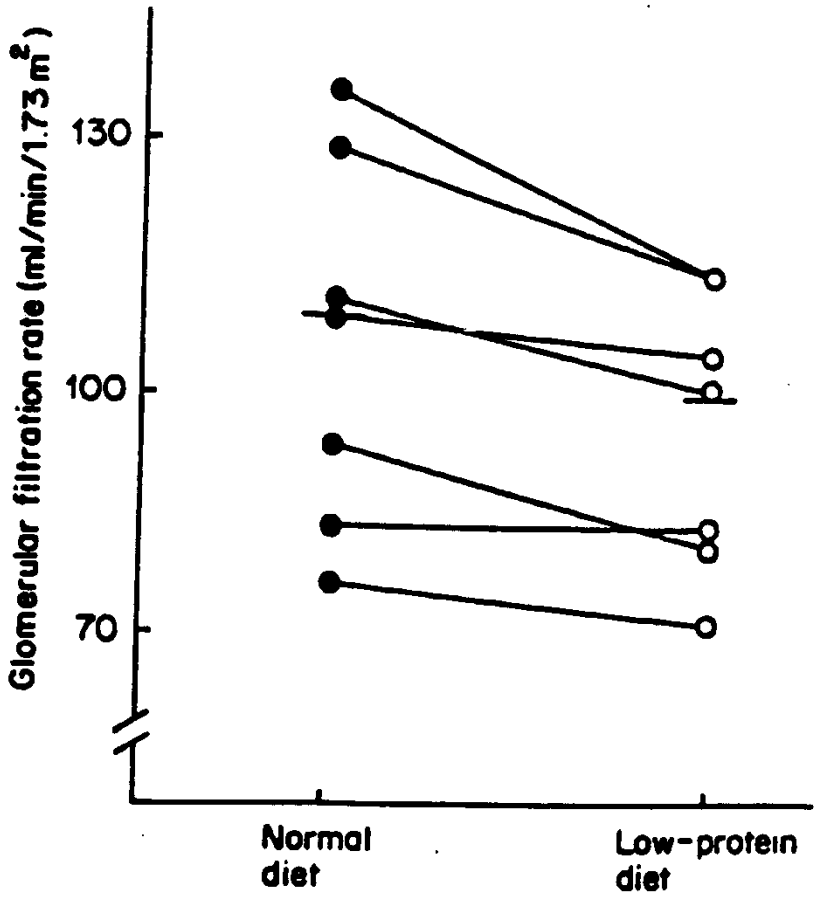
Perot Chart: National Debt and Deficit 1992

**USA Today Oct 1, 2012: Perot's economic stance resonates 20 years later**

**Nate Silver: Prediction of 2012 presidential election**

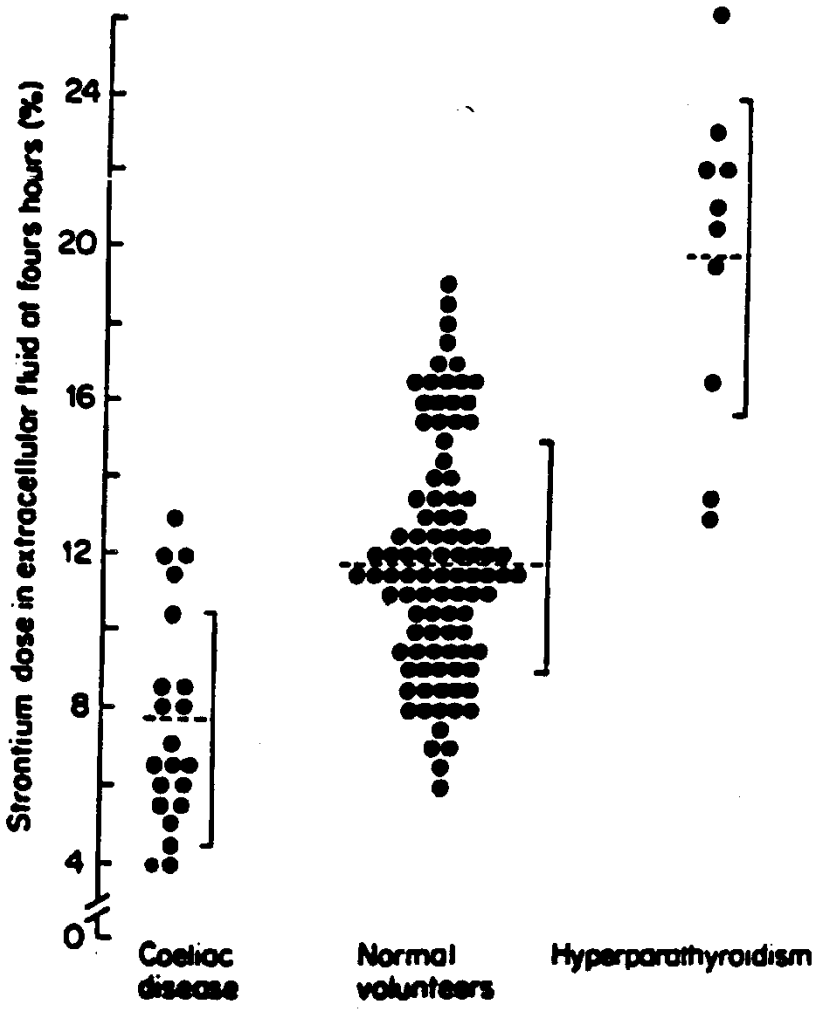
# Example of a Line Graph

Glomerular filtration rate during normal and low protein diets in diabetic patients



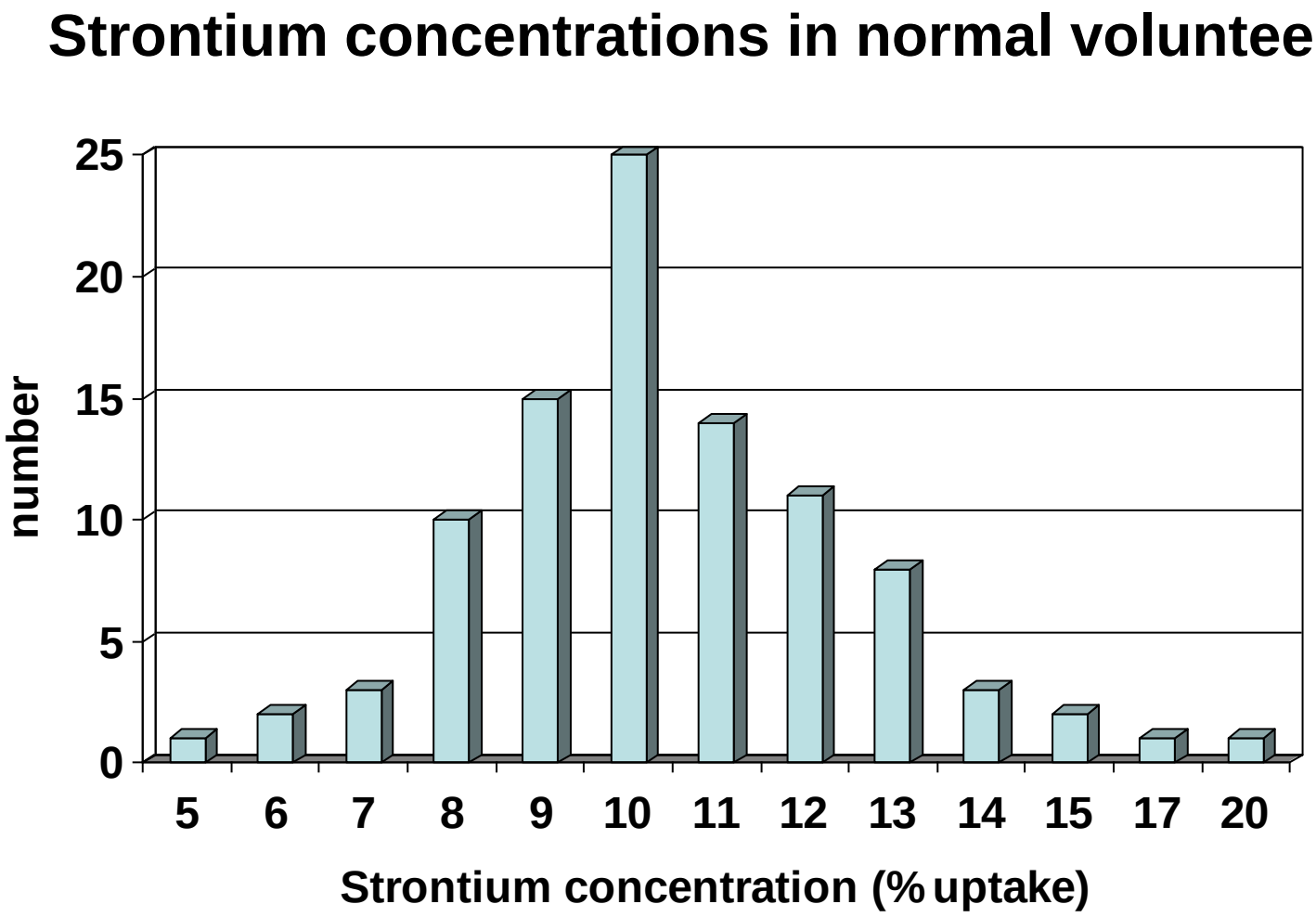

- This is a type of line graph from Coehn et al. (1987) showing the changes in kidney function of patients who were first fed a normal diet and then switched to a low protein diet.
- A line connects the pair of data points for each of the 7 subjects in the study.
- Note the elements of a good graph:
  - ***a clear title***
  - ***X and Y axes are labeled***
  - ***the unit of measurement is specified***.

# A Dot Plot Example

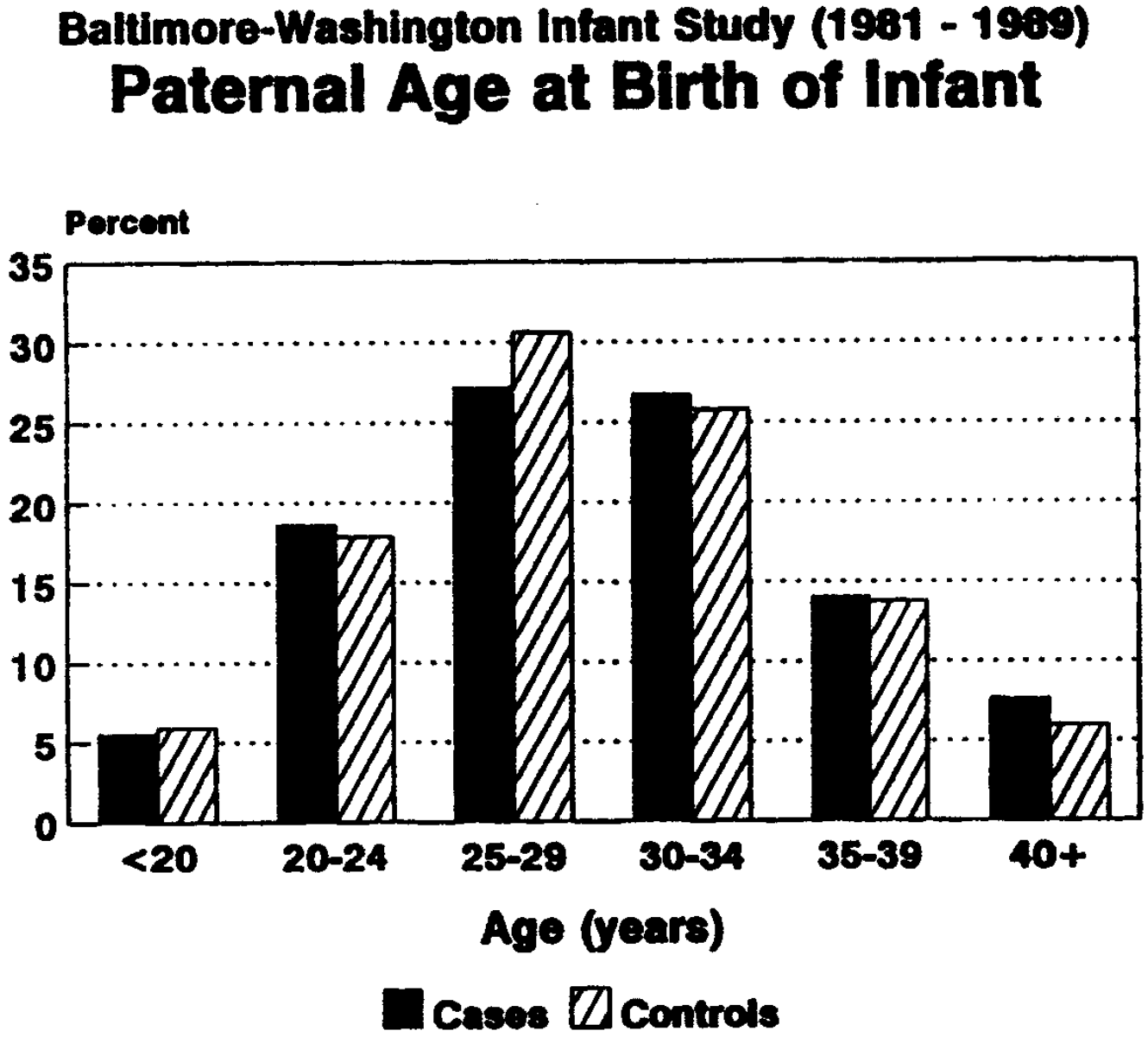



- This is a “dot plot” from Milsom et al. (1987) showing the distribution of strontium in the extracellular fluid of 3 groups of subjects.
- The three groups are: persons with coeliac disease, normal volunteers, and patients with hyperparathyroid disease.
- Each dot represents the strontium reading for one person.
- The dashed line represents the mean level of strontium for the group (Celiac, Normal, Hyperparathyroidism).

# A Bar Graph of the Same Data



# A Bar Graph Comparing 2 Groups



Data: 2.3, 1.2, 1.4, 6.3, 8.3, 7.7, …

## Constructing a Simple Stem-and-Leaf Diagram

1. Choose some convenient numbers to serve as stems. To be useful in ascertaining shape at least five stems are needed. The stems chosen are usually the first one or two digits of the numbers in the data set.
2. Label the rows via the chosen stems.
3. Reproduce the data graphically by recording the digit following the stem as a leaf on the appropriate stem.
4. Turn the graph on its side to see how the numbers are distributed. In particular, try to answer such questions as
   *a.* Do the data tend to cluster near a particular stem or stems or do they spread rather evenly across the diagram?
   *b.* Do the data tend to taper toward one end or the other of the diagram?
   *c.* If a smooth curve is sketched across the top of the diagram, does it form a rough bell? Is it flat? Is it symmetric?

An example should clarify the idea.

# Stem and Leaf Example

Note: order the leaves to get the final figure (e)

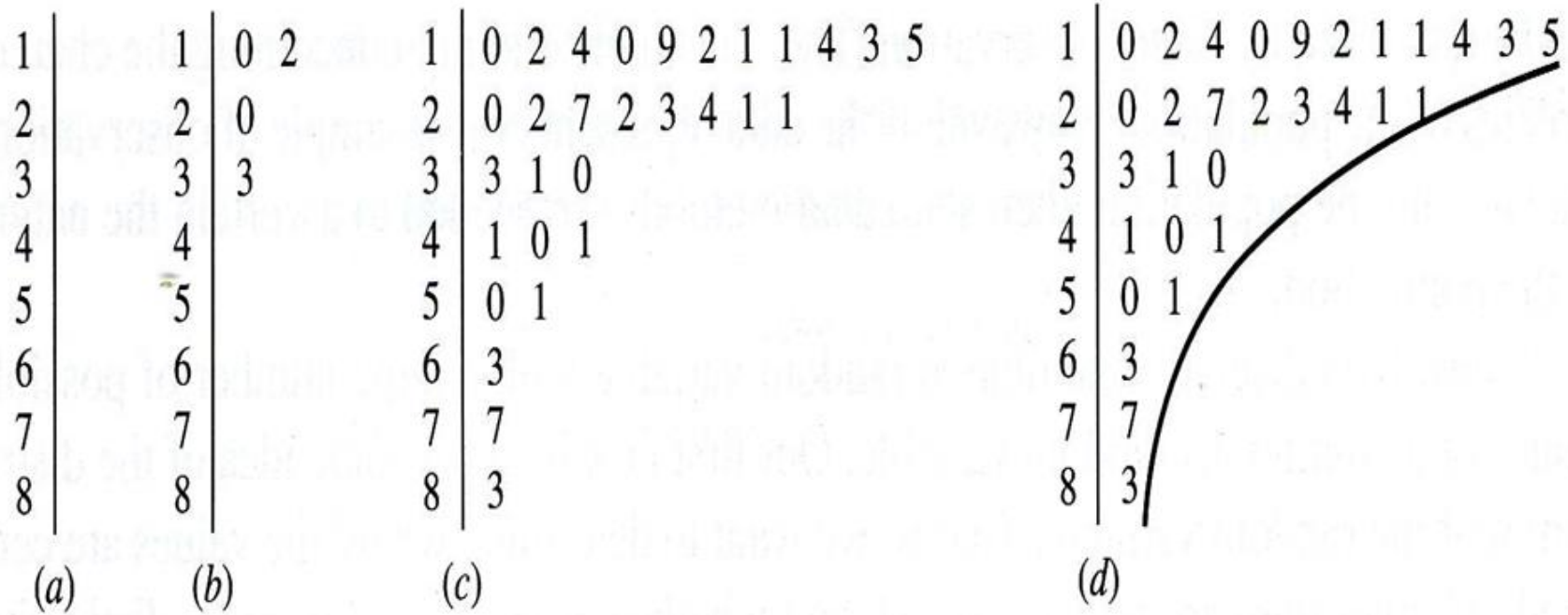


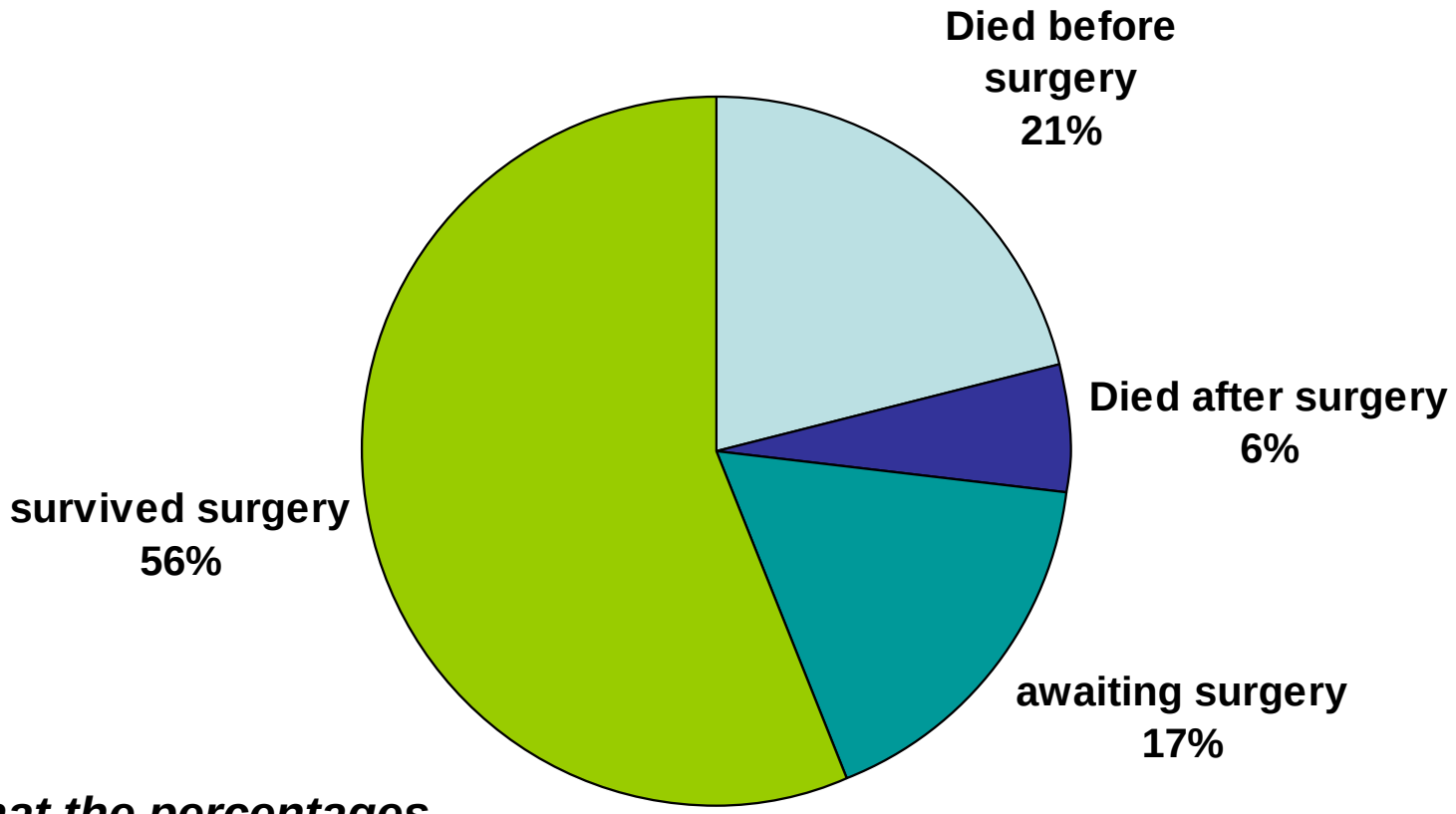

**FIGURE 1.4**

Data: 2.3, 1.2, 1.4, 6.3, 8.3, 7.7, 5.0, 4.1, 3.3, …

Stem-and-leaf display for the magnitude of a sample of California earthquakes as measured on the Richter scale: (*a*) choosing stems, (*b*) recording the first four data points, (*c*) the entire data set displayed and (*d*) looking for shape.

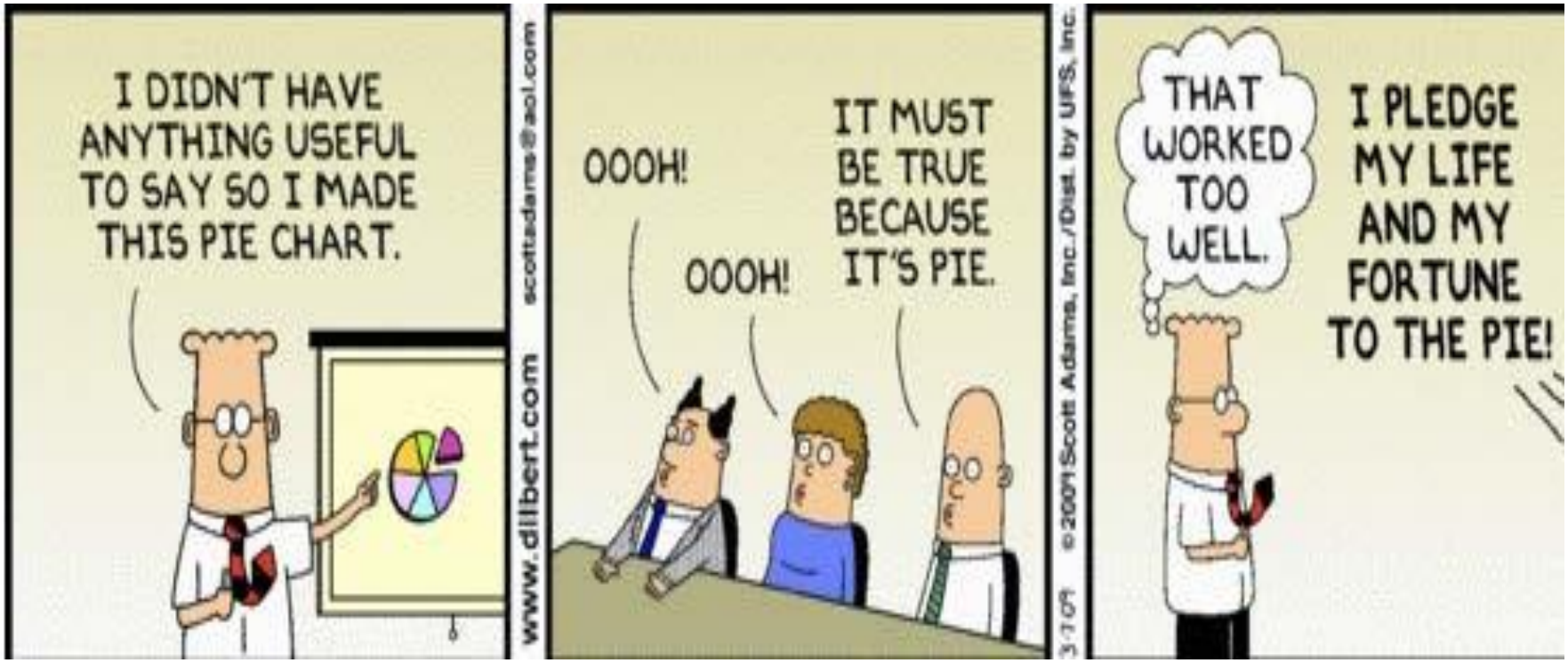
# Pie Chart

**Surgical history of 123 patients with Tetralogy of Fallot**




***Note that the percentages add to 100%***

The pie chart is useful for showing relative proportions of a few categories. The more categories, the greater the number of “slices”, the more difficult the chart is to read. Consider using a bar chart or histogram instead.

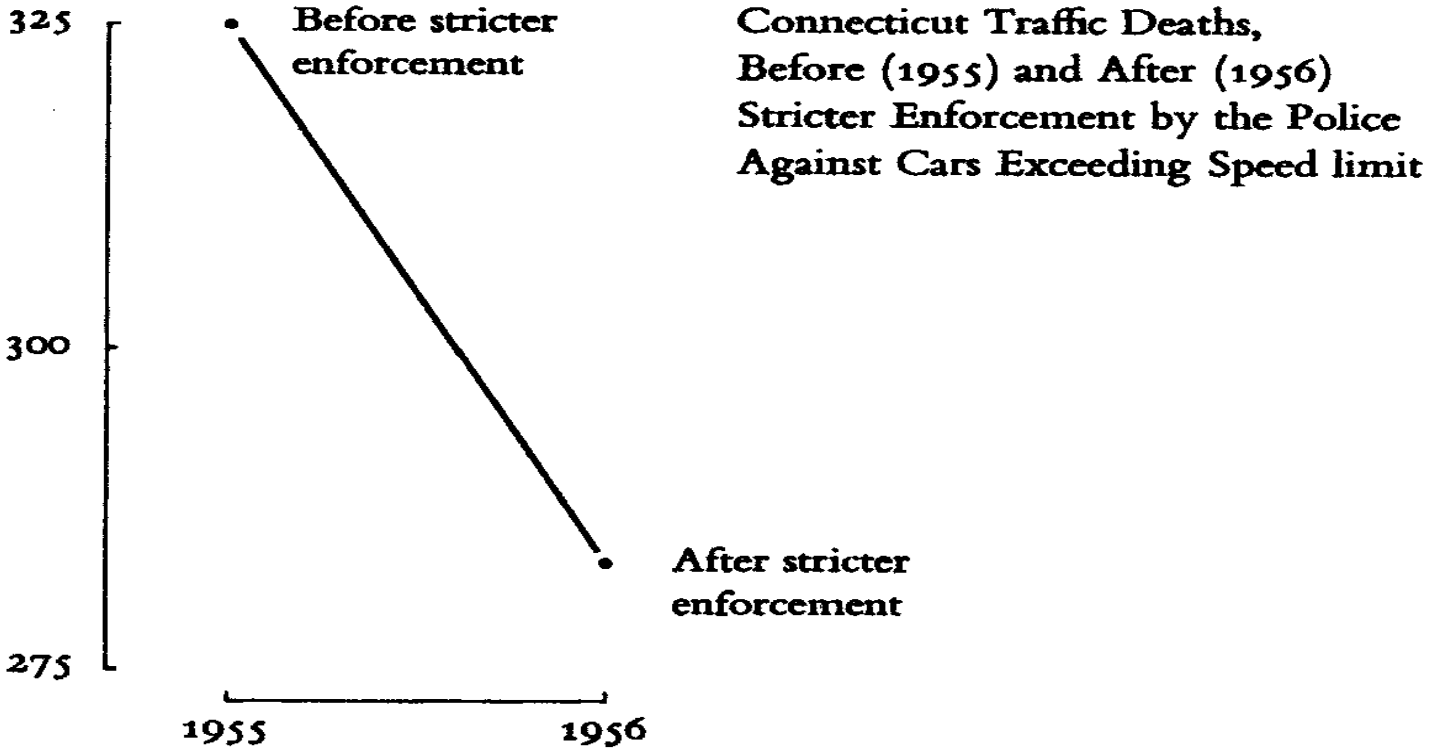


## Context is Essential for Graphical Integrity

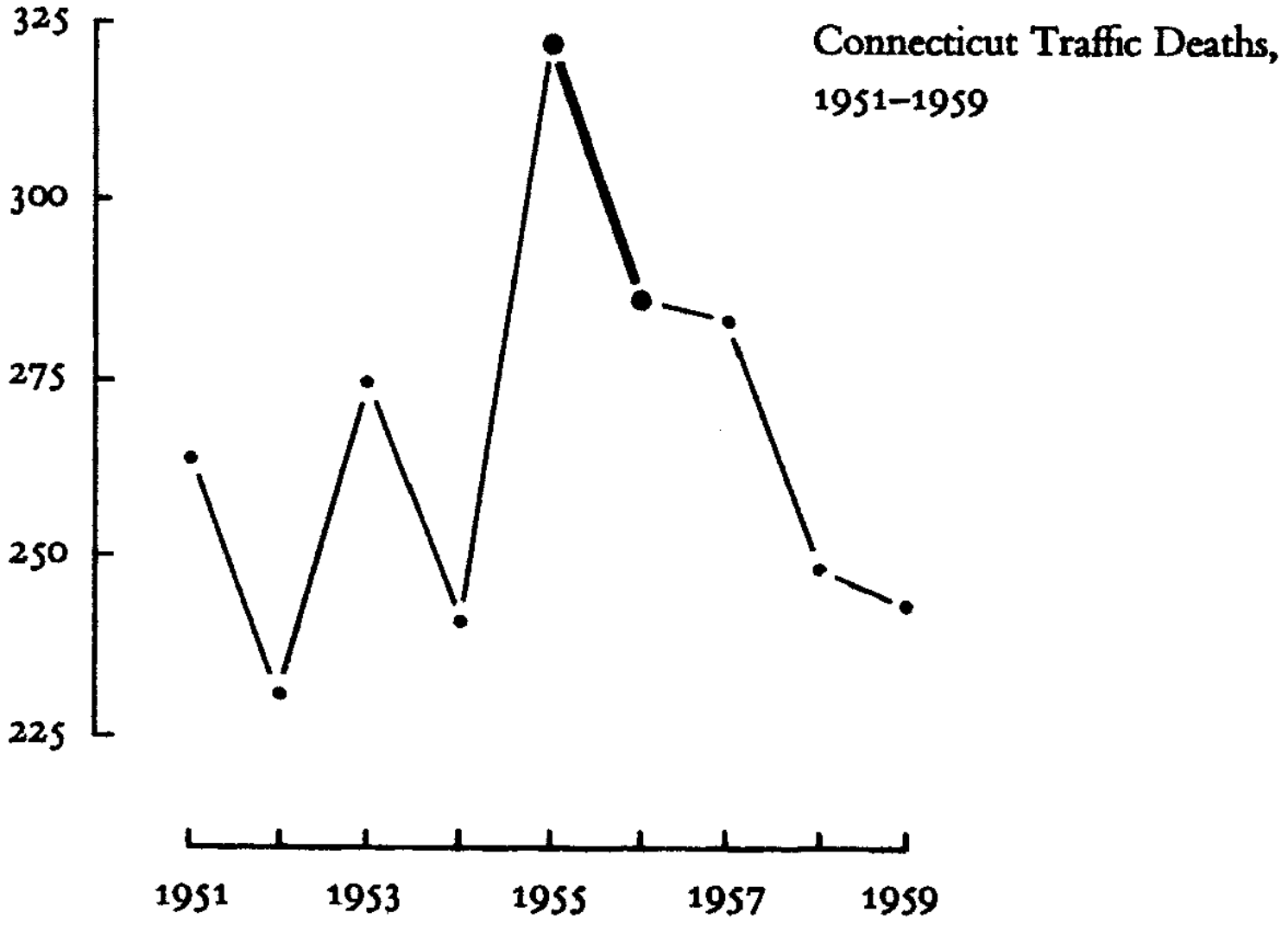
To be truthful and revealing, data graphics must bear on the question at the heart of quantitative thinking: "Compared to what?" The emaciated, data-thin design should always provoke suspicion, for graphics often lie by omission, leaving out data sufficient for comparisons. The principle:

Graphics must not quote data out of context.

Nearly all the important questions are left unanswered by this display:



A few more data points add immensely to the account:




Connecticut Traffic Deaths,
1951–1959

# Presenting Numerical Data

## Descriptive Statistics
## Measures of Central Tendency

Terminology:

*Statistic* – Sample parameter

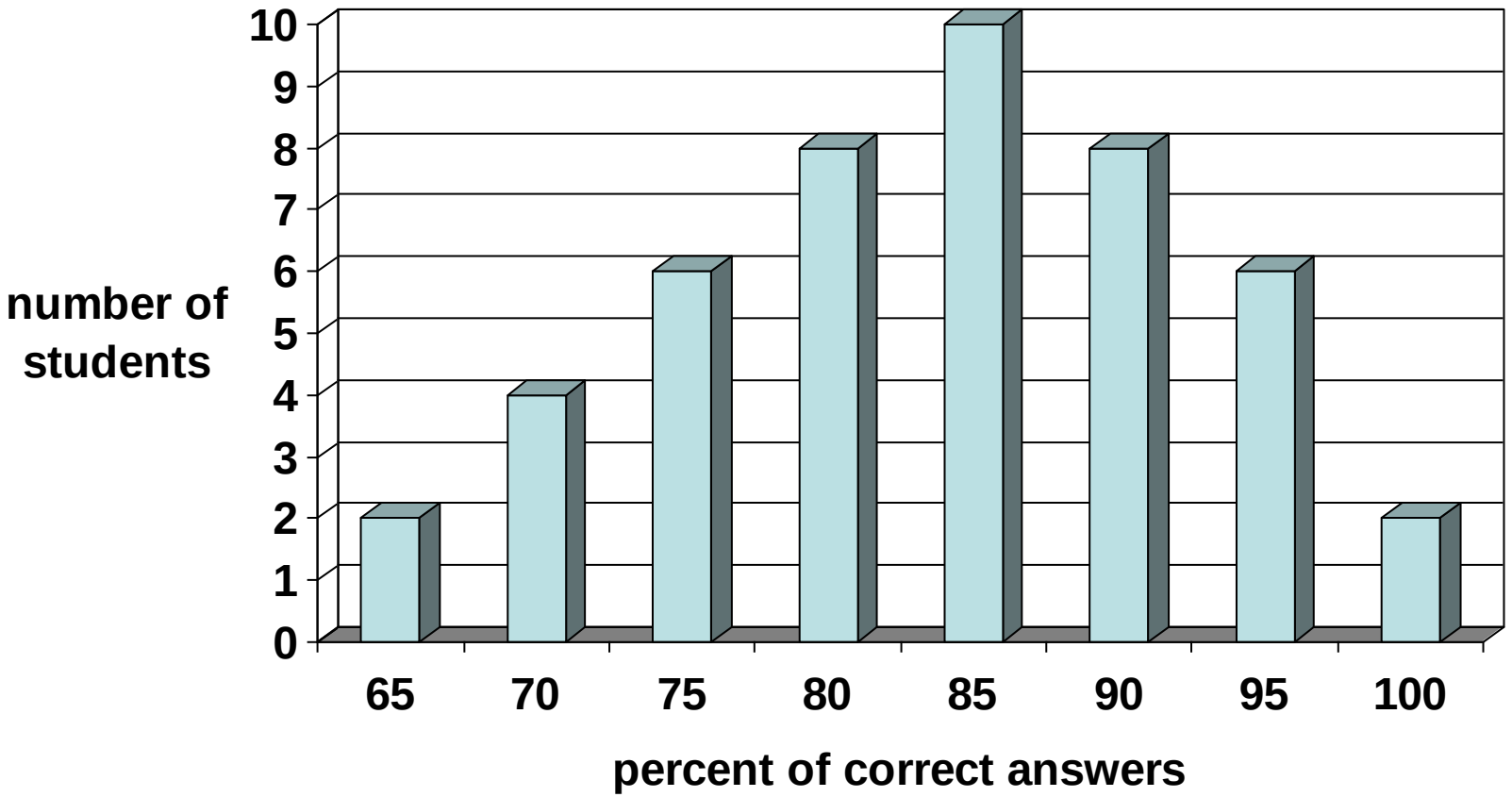
*Parameter* – Population parameter

# What is Central Tendency?

- Do the data resemble a bell shaped curve? look for the central value of the data and the amount of dispersion around this point. *{see grades example}*
- If not, is there a pattern or not? is there any evidence of bimodal or multimodal centering? Do the concepts of central value and dispersion have any meaning if the data are like this?

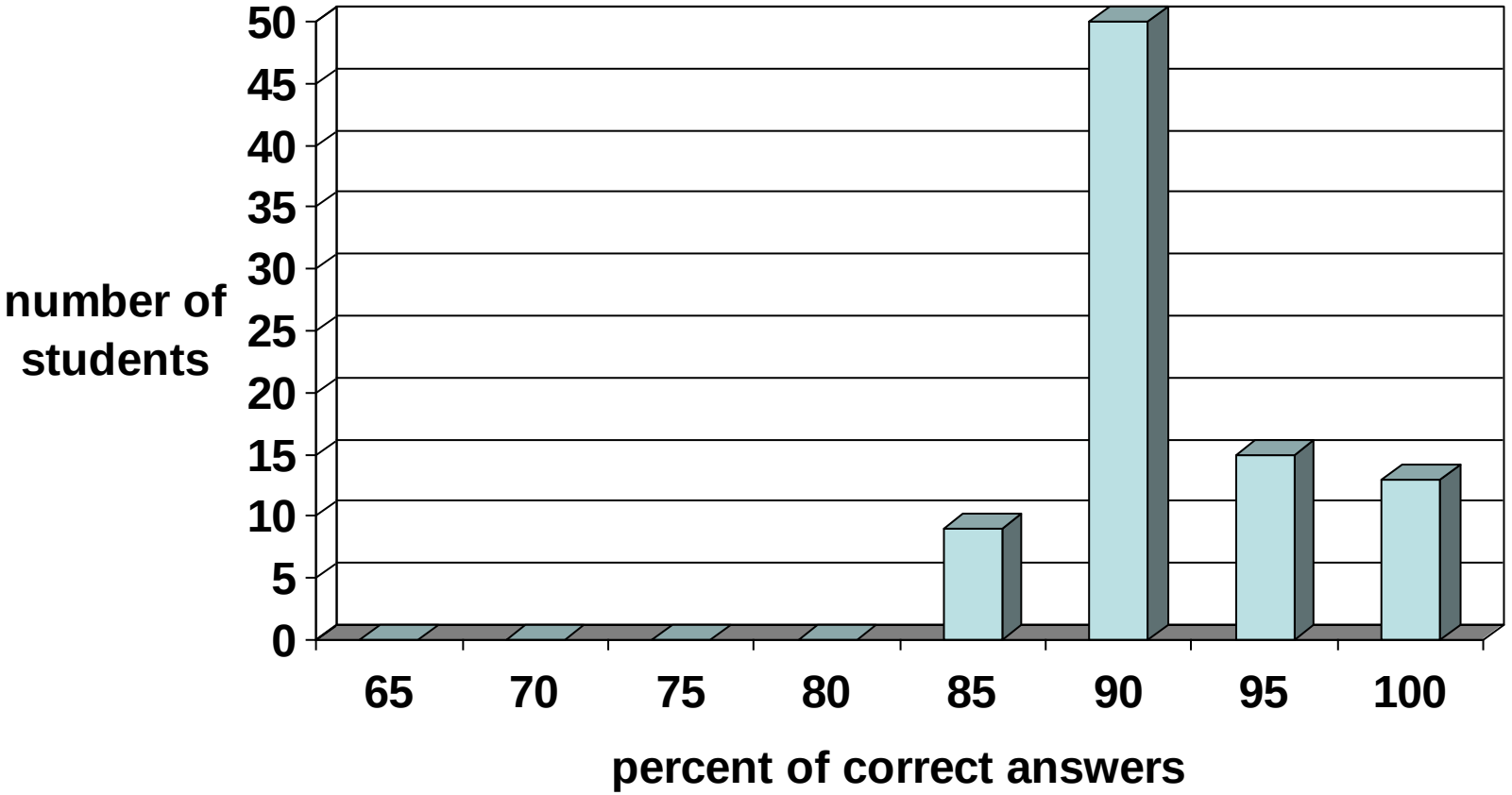
  *{see beak length example}*

# Grades in Stat 101 Class Exam

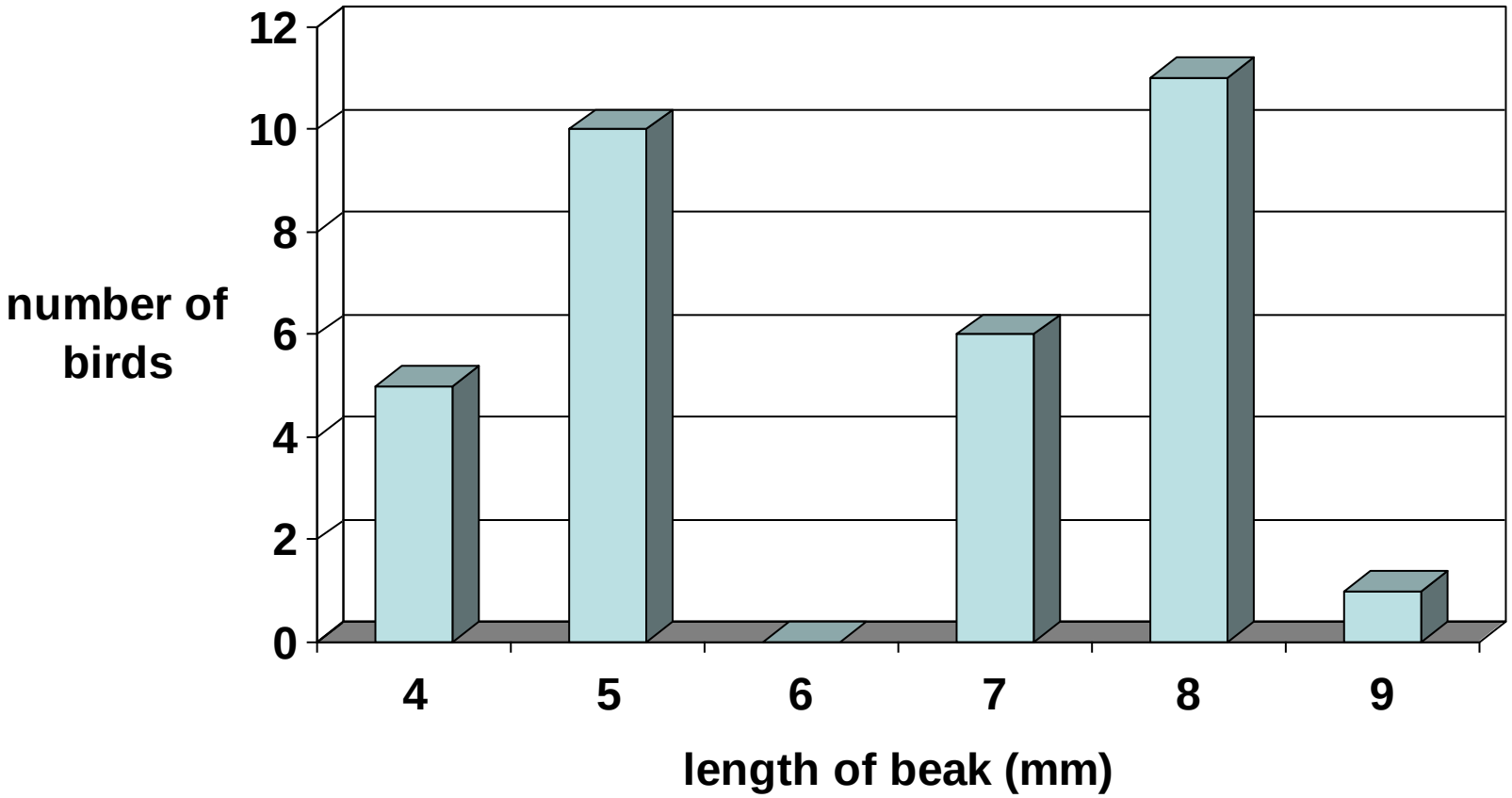


# Grades on Drivers License Exam

(among those who completed Drivers Ed class)



# Beak Length in Darwin's Finches



# Important Descriptive Statistics that Help Us Look for Central Tendency

- Sample Mean (arithmetic average)
- Geometric mean is the **n**$^{th}$ root of the product of **n** observations. It is used when data is measured on a logarithmic scale
- Sample Median (the middle observation)
- Sample Mode (most observations have this value)

Review the textbook: pages 11-15, 38-48

# Measures of Spread (or Dispersion)

- Range (max value minus min value)
- Variance
- Standard deviation (always positive)
- Coefficient of variation
- Percentiles
- Inter-quartile range (IQR)

# Example

- To find the $25^{th}$, $50^{th}$ (median) and $75^{th}$ percentiles in your data, you have to identify the position of those values using the formula below
- $Q_1$ position 1/4*(n+1)
- $Q_2$ position 2/4*(n+1)
- $Q_3$ position 3/4*(n+1)

These positions refer to the ordered data (low-high)

- Ex: 55, 57, 58, 61, 66, 71, 78, 79, 83, 85, 99
- Range = 99-55=44
- **IQR** = 83-58= 25

# Standard Deviation

- Subtract the mean from every value in the data set and square the difference; then take the sum of the squared differences, divide it by n-1, and take its square root.
- ***Example*: 10, 20, 30, 40 miles per hour**
- Recall that the mean is 25, and so:
- $10-25=-15$ and $(-15)^2=225$
- $20-25=-5$ and $(-5)^2=25$
- $30-25=5$ and $5^2=25$
- $40=25=15$ and $(-15)^2=225$

# Standard Deviation (continued)

- The sum of the squared differences = 225+25+25+225=500 $mph^2$
- Now divide this sum by n-1 (where n is the total number of observations)to calculate the *sample variance*:

  $s^2=500/(4-1)=500/3=166.666\ mph^2$

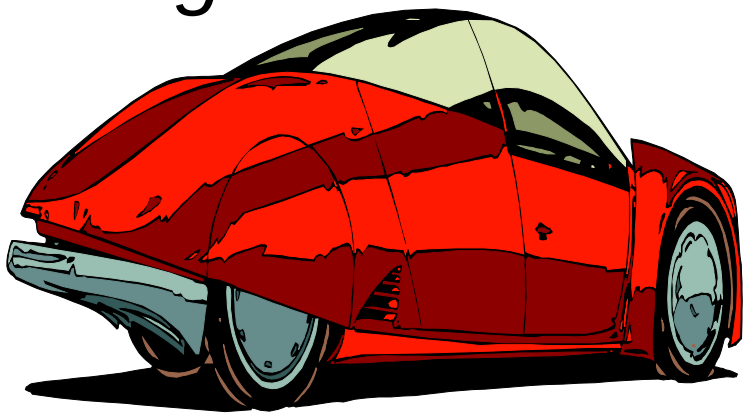
- Standard deviation s = sqrt(sample variance) sqrt (166.666 $mph^2$) = 12.90 mph

# Median

- Locate the observation in the exact center of the data, after listing them in order from low to high
- ***Example*: 1,2,4,7,11,19,21 years old**
- Median=7 years (note three observations below it and three above it)
- This is easy when there is an odd number of observations; in case of an even number, take the average of the two middle observations to get the median.
- ***Example*: 10, 20, 30, 40 mph**
- Median = (20+30) / 2 = 25 mph

# Range

- Subtract the lowest from the highest value in the data set.
- ***Example*: 10, 20, 30, 40 mph**
- Range = 40-10 = 30 mph
- It is also acceptable to report both the lowest and highest values when describing the range, e.g. "*The values ranged from 10 to 40 miles per hour.*"



# Add mean to the plot: Zink and Citrate Levels in mM Concentration

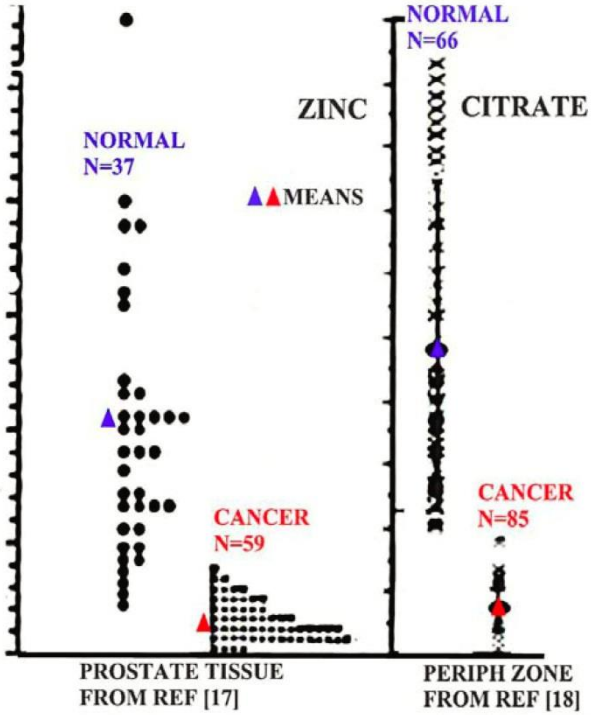
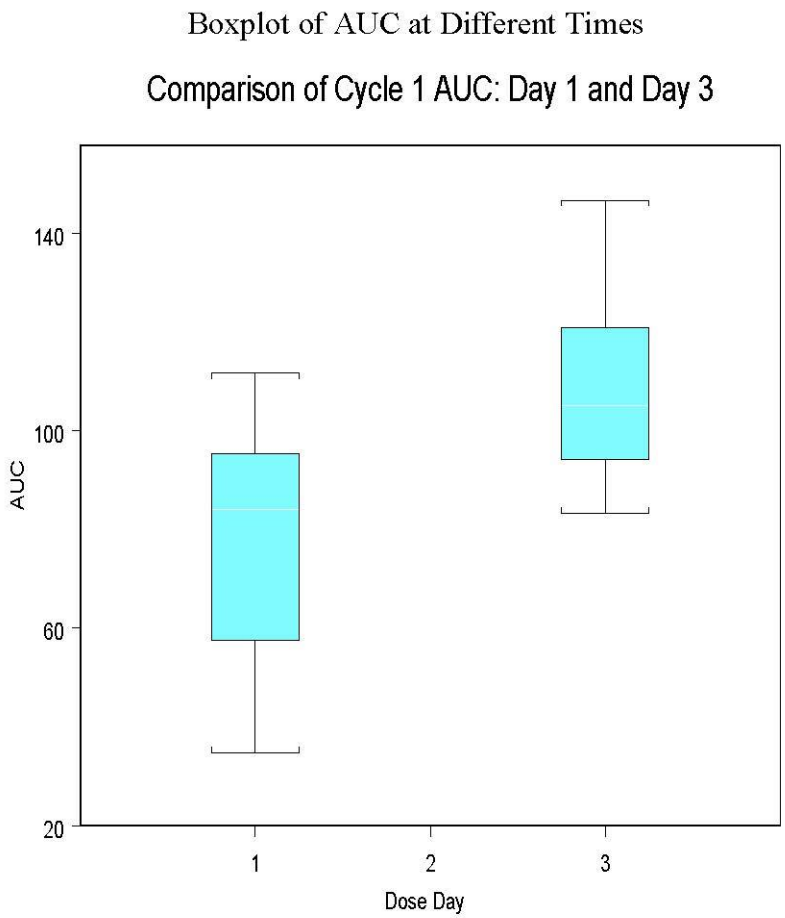



**Fig. (1).** The comparison of the zinc and citrate levels in normal prostate tissue versus prostate cancer.

***The Open Urology & Nephrology Journal*, 2008, *1:*26-35**

# Five-number summary statistics

- **Min, $Q_1$, Median ($Q_2$), $Q_3$, Max**



# The Mean may be misleading

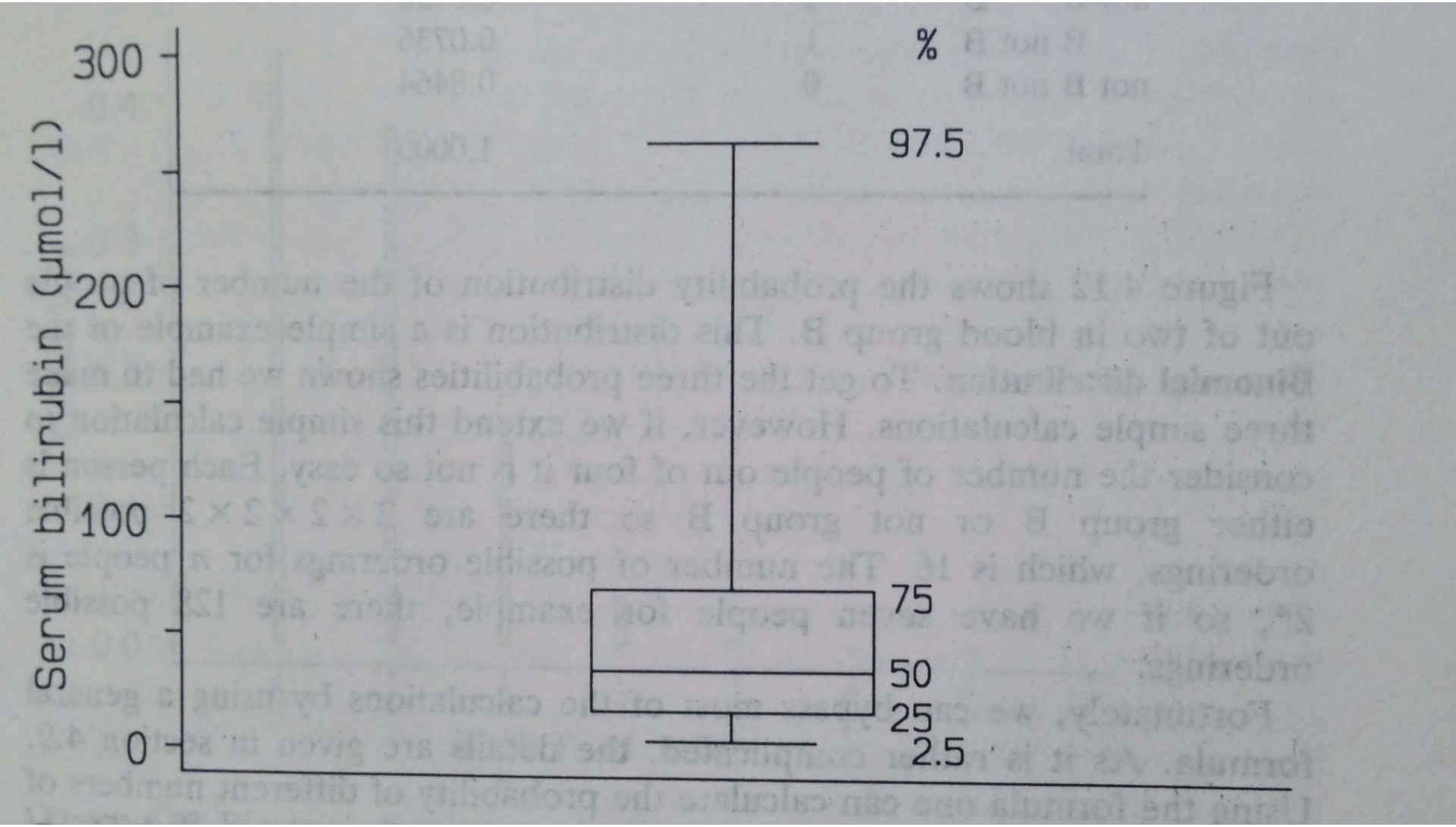
Group Practice A (5 physicians) mean=$240K

Group Practice B (5 Physicians) mean=$200K

In which group would you prefer to work (data points represent thousands)?

Salary Data Group A: 150, 150, 150, 150, 600

Salary Data Group B: 200, 200, 200, 200, 200


Serum bilirubin (μmol/l)
300
200
100
0
%
97.5
75
50
25
2.5

# Role of data transformation, log-scale

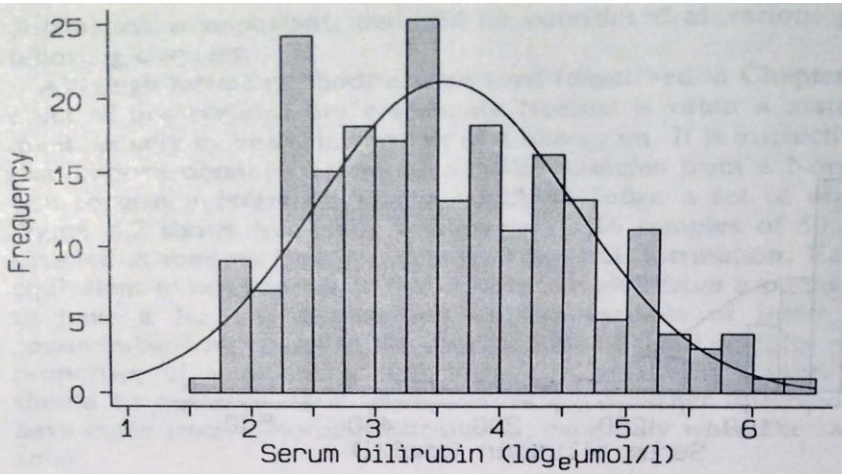



**Figure 4.9** Histogram of log serum bilirubin with fitted Normal distribution (logarithms to base e).

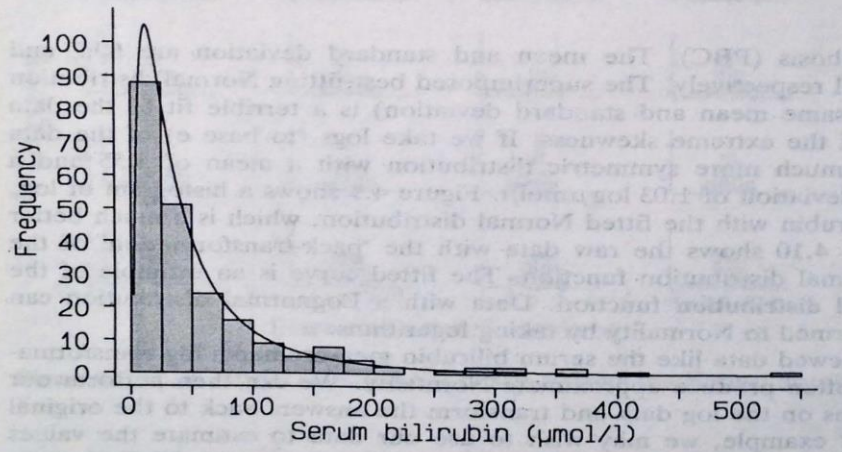



**Figure 4.10** Histogram of serum bilirubin with fitted Lognormal distribution.

# Is my lab test normal?

- Be careful about normal limits when the distribution of the lab result is highly skewed. You could have the typical value that may be near or beyond the normal limits

# Assignment

Download STATA (IC) on your computer.

- See STATA folder for tutorial and Exercise
- Note: Do not turn the STATA Exercise in. Try to work on it independently. We will go over it during lab session next class.

- Reading: textbook chapters, see Syllabus