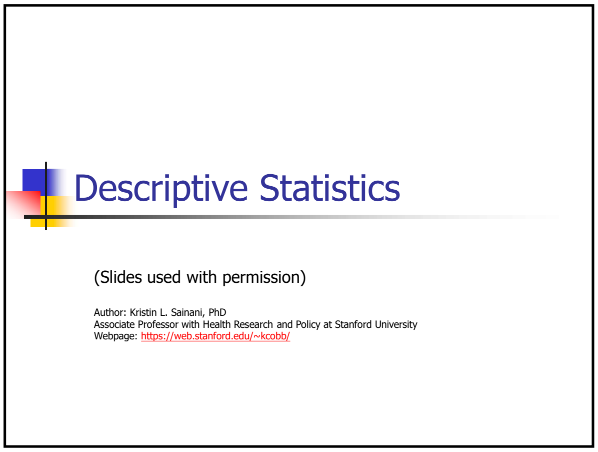
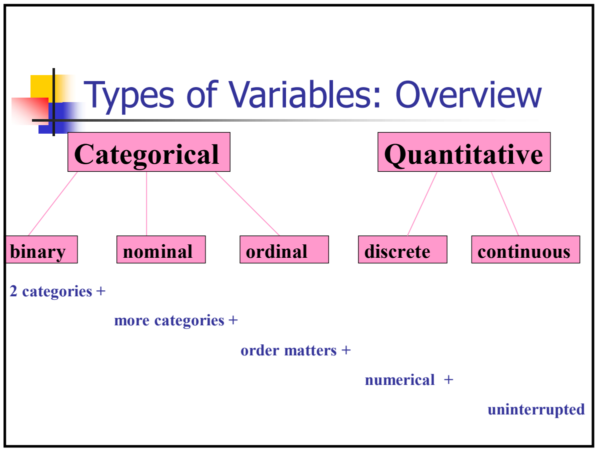


Descriptive Statistics

(Slides used with permission)

Author: Kristin L. Sainani, PhD
Associate Professor with Health Research and Policy at Stanford University
Webpage: https://web.stanford.edu/~kcobb/

1



Types of Variables: Overview

Categorical

Quantitative

binary

nominal

ordinal

discrete

continuous

2 categories +

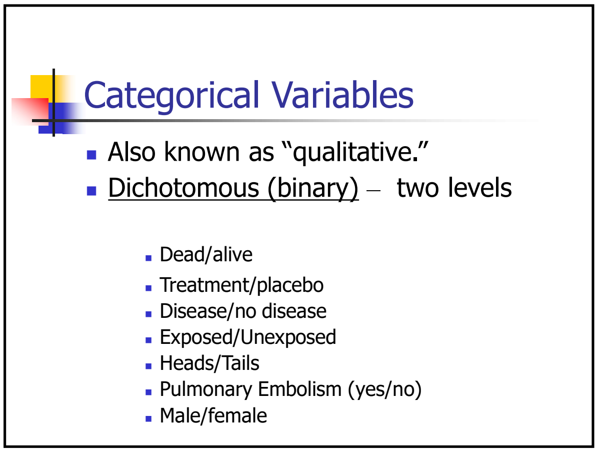
more categories +

order matters +

numerical +

uninterrupted

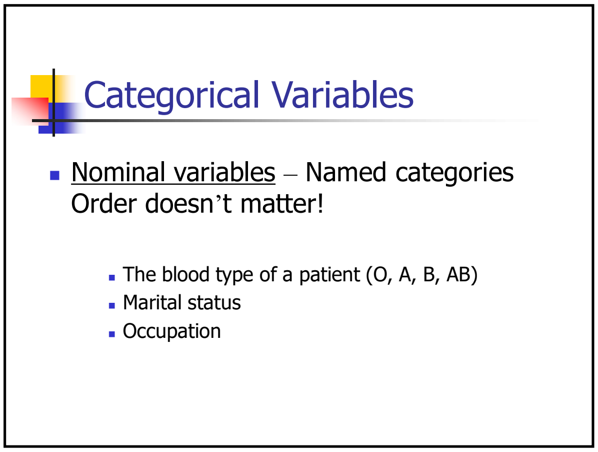
2



Categorical Variables

- Also known as "qualitative."
- Dichotomous (binary) – two levels
  - Dead/alive
  - Treatment/placebo
  - Disease/no disease
  - Exposed/Unexposed
  - Heads/Tails
  - Pulmonary Embolism (yes/no)
  - Male/female

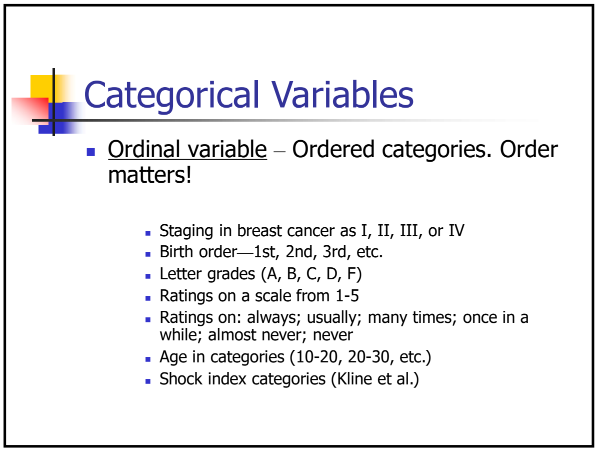
3



Categorical Variables

- Nominal variables – Named categories Order doesn't matter!
  - The blood type of a patient (O, A, B, AB)
  - Marital status
  - Occupation

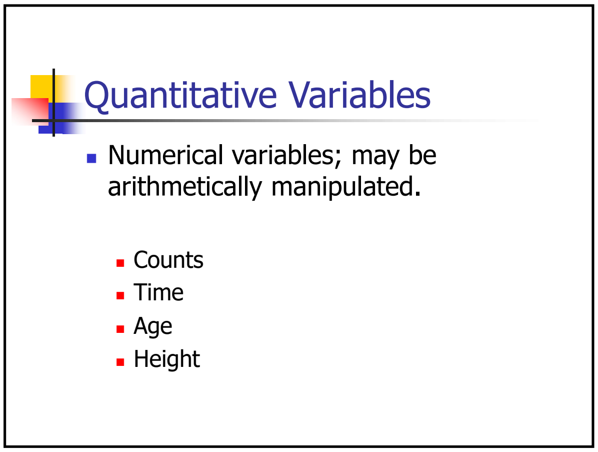
4



Categorical Variables

- Ordinal variable – Ordered categories. Order matters!
  - Staging in breast cancer as I, II, III, or IV
  - Birth order—1st, 2nd, 3rd, etc.
  - Letter grades (A, B, C, D, F)
  - Ratings on a scale from 1-5
  - Ratings on: always; usually; many times; once in a while; almost never; never
  - Age in categories (10-20, 20-30, etc.)
  - Shock index categories (Kline et al.)

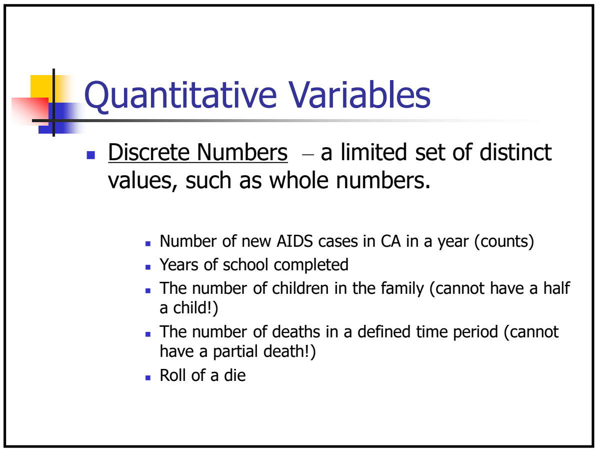
5



Quantitative Variables

- Numerical variables; may be arithmetically manipulated.
  - Counts
  - Time
  - Age
  - Height

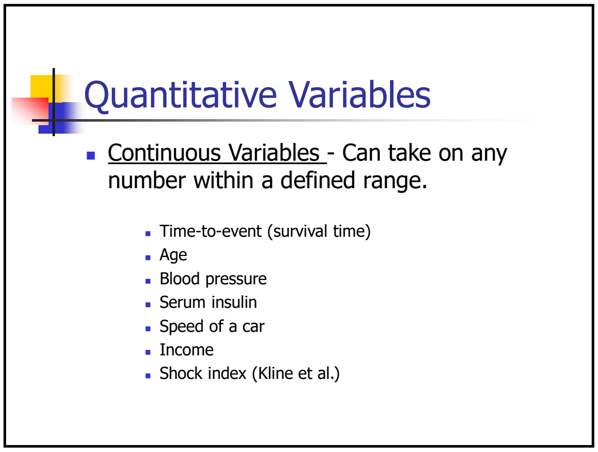
6



Quantitative Variables

- Discrete Numbers – a limited set of distinct values, such as whole numbers.
  - Number of new AIDS cases in CA in a year (counts)
  - Years of school completed
  - The number of children in the family (cannot have a half a child!)
  - The number of deaths in a defined time period (cannot have a partial death!)
  - Roll of a die

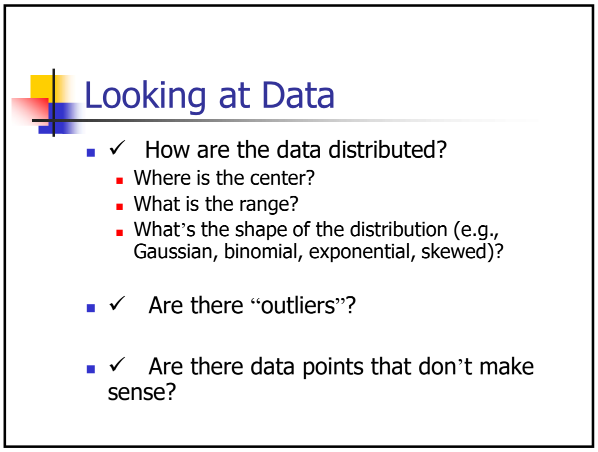
7



Quantitative Variables

- Continuous Variables - Can take on any number within a defined range.
  - Time-to-event (survival time)
  - Age
  - Blood pressure
  - Serum insulin
  - Speed of a car
  - Income
  - Shock index (Kline et al.)

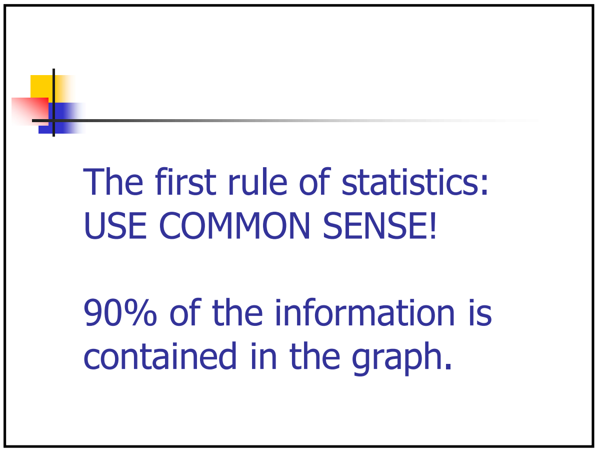
8



Looking at Data

- ✓ How are the data distributed?
  - Where is the center?
  - What is the range?
  - What's the shape of the distribution (e.g., Gaussian, binomial, exponential, skewed)?
- ✓ Are there "outliers"?
- ✓ Are there data points that don't make sense?

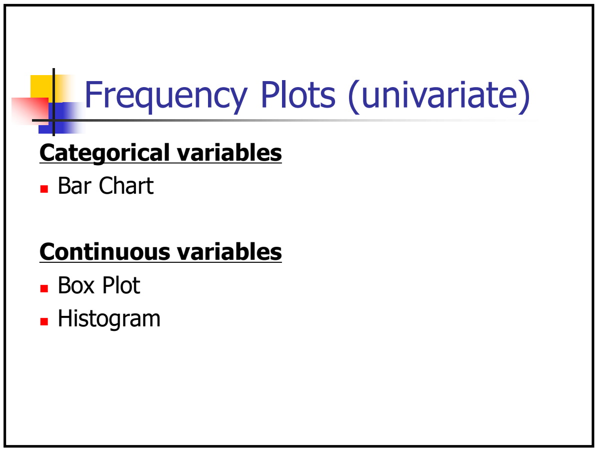
9



The first rule of statistics: USE COMMON SENSE!

90% of the information is contained in the graph.

10



Frequency Plots (univariate)

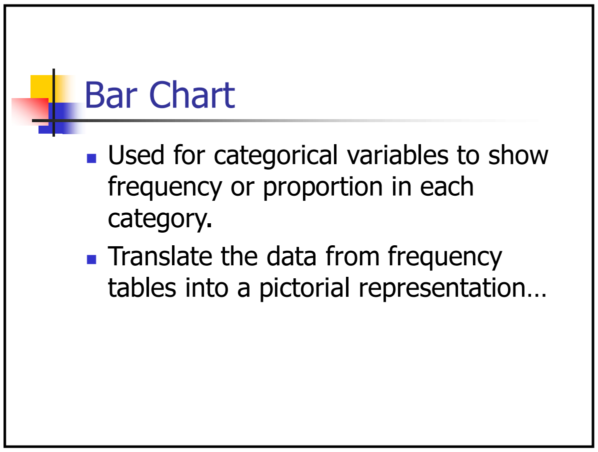
Categorical variables

- Bar Chart

Continuous variables

- Box Plot
- Histogram

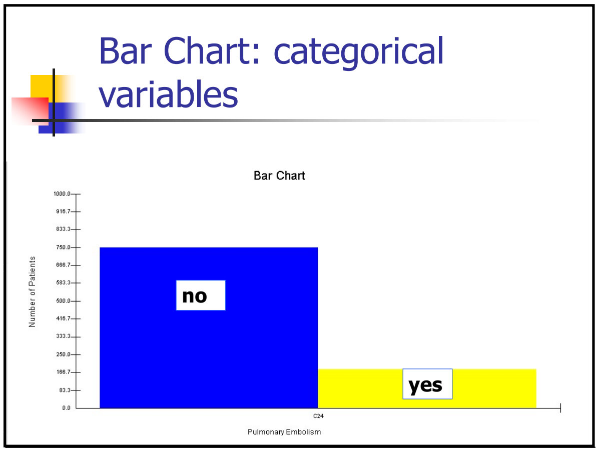
11



Bar Chart

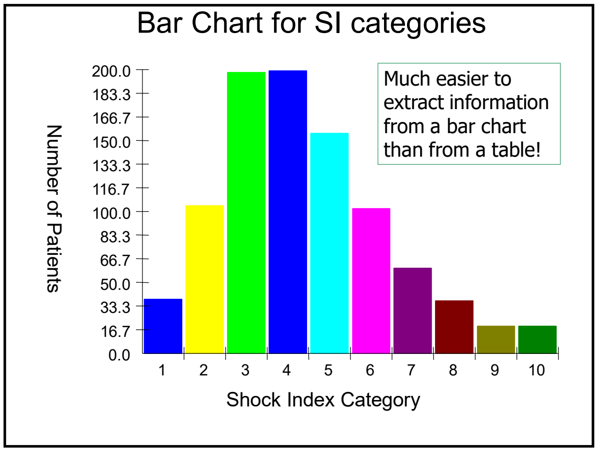
- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables into a pictorial representation…

12



Bar Chart: categorical variables

13



Bar Chart for SI categories

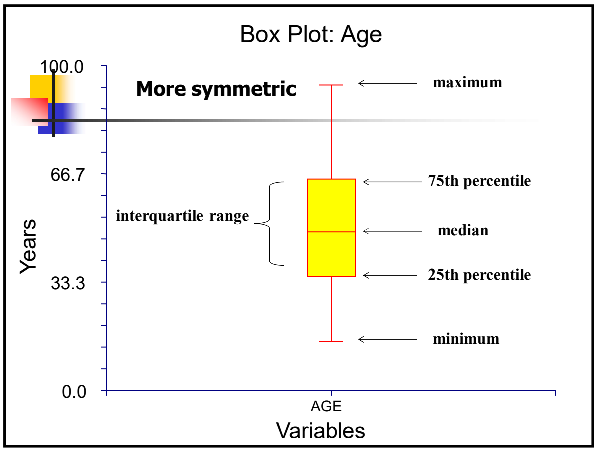
Much easier to extract information from a bar chart than from a table!

14

Box plot and histograms: for continuous variables

- To show the distribution (shape, center, range, variation) of continuous variables.

15



Box Plot: Age

More symmetric

maximum

75th percentile

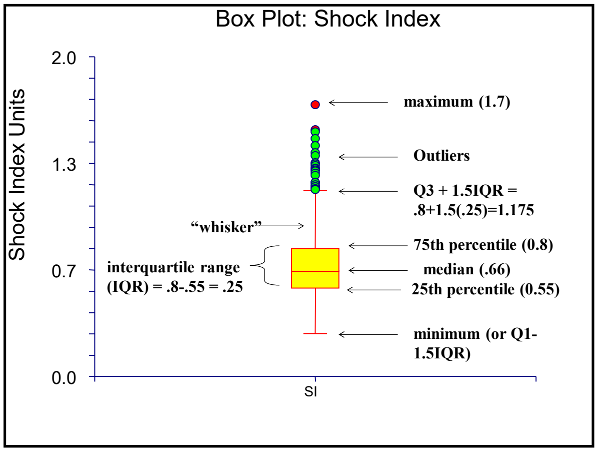
median

25th percentile

minimum

interquartile range

16



Box Plot: Shock Index

maximum (1.7)

Outliers

Q3 + 1.5IQR = .8+1.5(.25)=1.175

“whisker”

75th percentile (0.8)

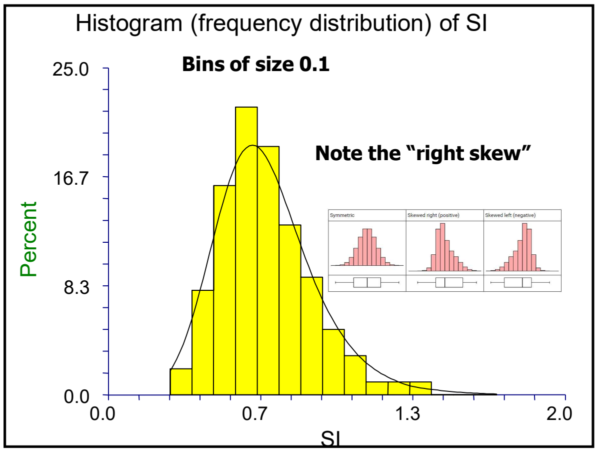
median (.66)

25th percentile (0.55)

interquartile range (IQR) = .8-.55 = .25

minimum (or Q1-1.5IQR)

17

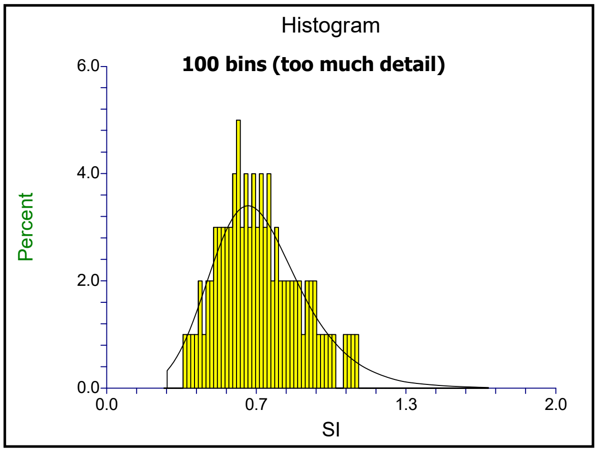


Histogram (frequency distribution) of SI

Bins of size 0.1

Note the “right skew”

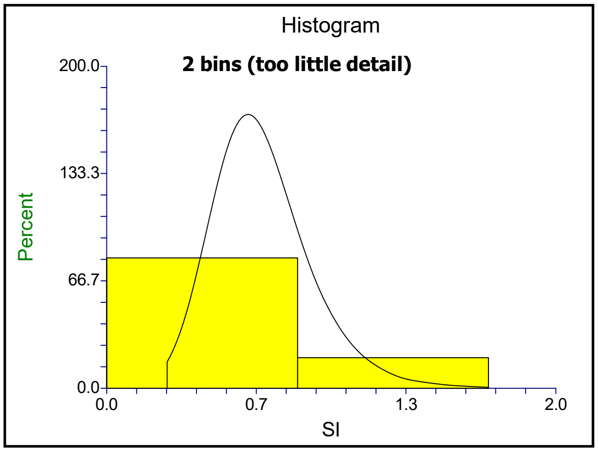
18



Histogram

100 bins (too much detail)

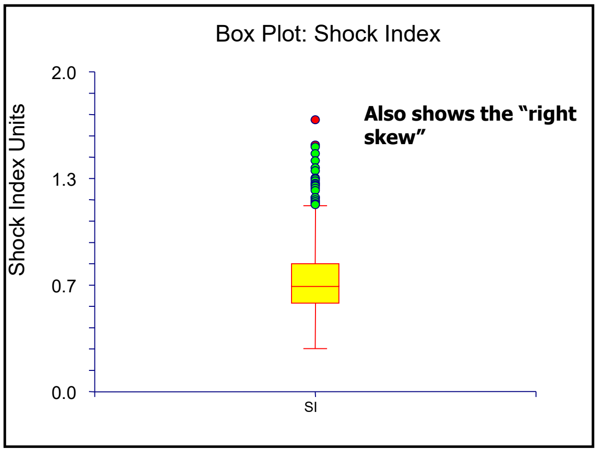
19



Histogram

2 bins (too little detail)

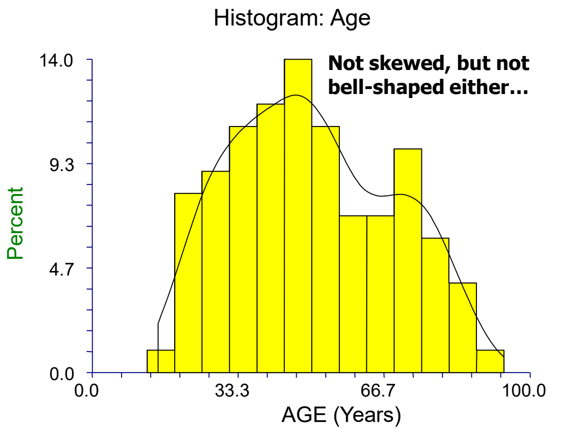
20



Box Plot: Shock Index

Also shows the "right skew"

21



Not skewed, but not bell-shaped either…

22

Measures of central tendency

- Mean
- Median
- Mode

23

Central Tendency

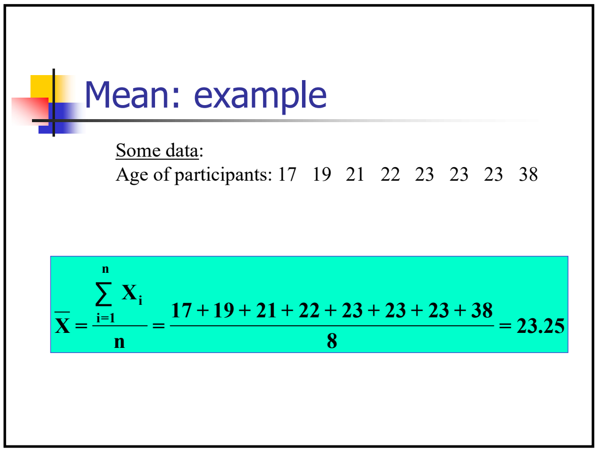
- Mean – the average; the balancing point

calculation: the sum of values divided by the sample size

In math shorthand:

$$\bar{X} = \frac{\sum_{i=1}^{n} x}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

24

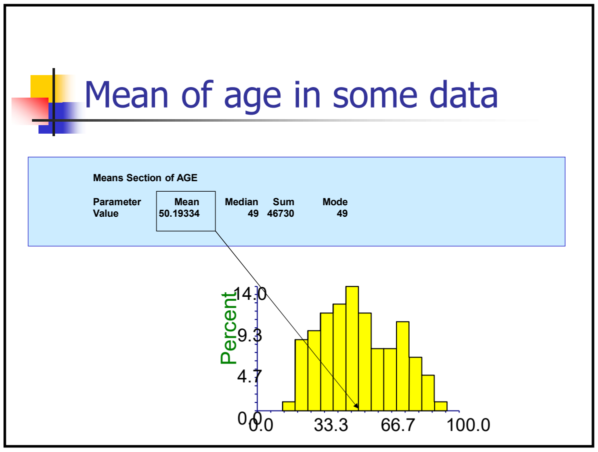


Mean: example

Some data:
Age of participants: 17 19 21 22 23 23 23 38

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{17+19+21+22+23+23+23+38}{8} = 23.25$$

25

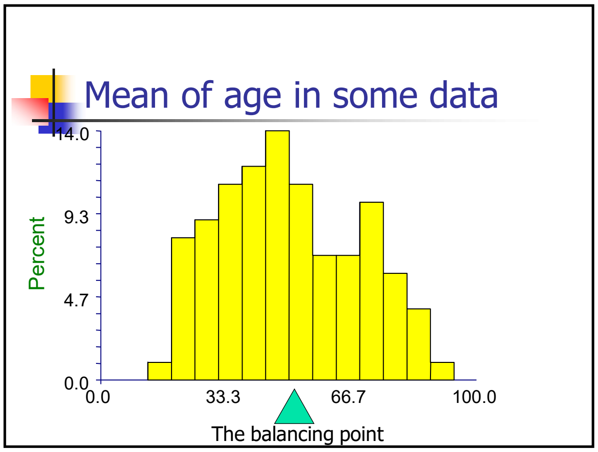


Mean of age in some data

Means Section of AGE

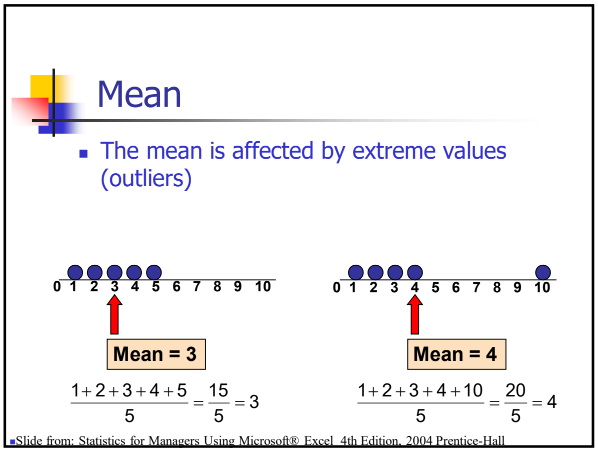
Parameter	Mean	Median	Sum	Mode
Value	50.19334	49	46730	49

26



Mean of age in some data

27



Mean

- The mean is affected by extreme values (outliers)

0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10

Mean = 3

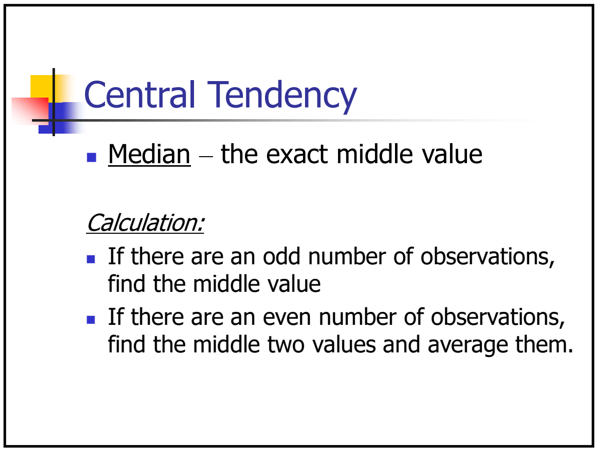
$$\frac{1+2+3+4+5}{5}=\frac{15}{5}=3$$

Mean = 4

$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

28



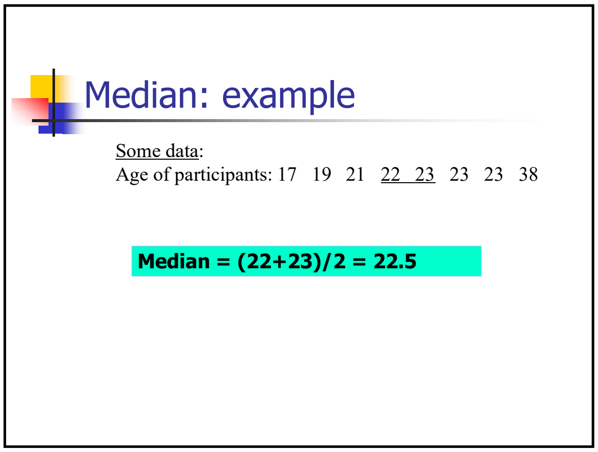
Central Tendency

- Median – the exact middle value

Calculation:

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them.

29

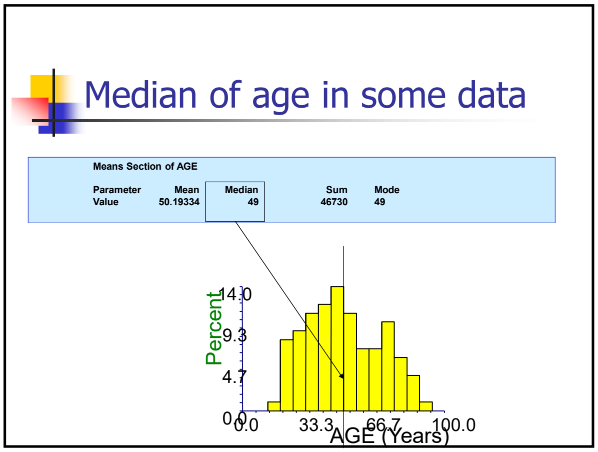


Median: example

Some data:
Age of participants: 17 19 21 22 23 23 23 38

Median = (22+23)/2 = 22.5

30

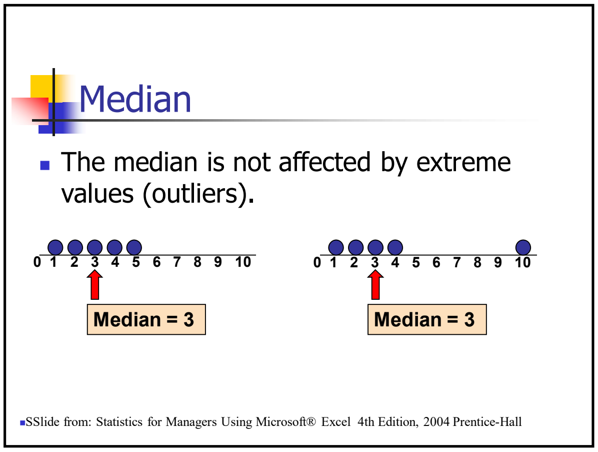


Median of age in some data

Means Section of AGE

Parameter	Mean	Median	Sum	Mode
Value	50.19334	49	46730	49

31

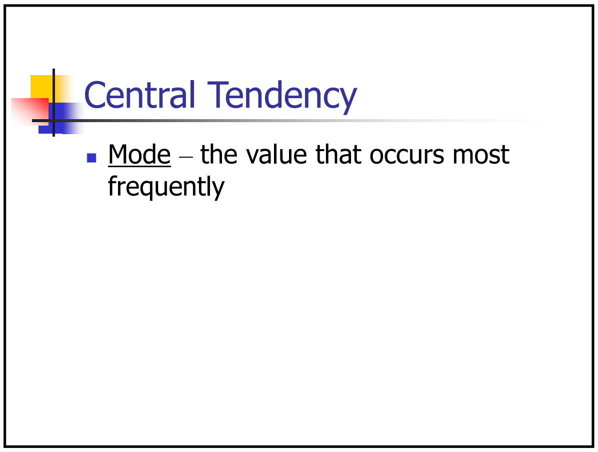


Median

- The median is not affected by extreme values (outliers).

SSlide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

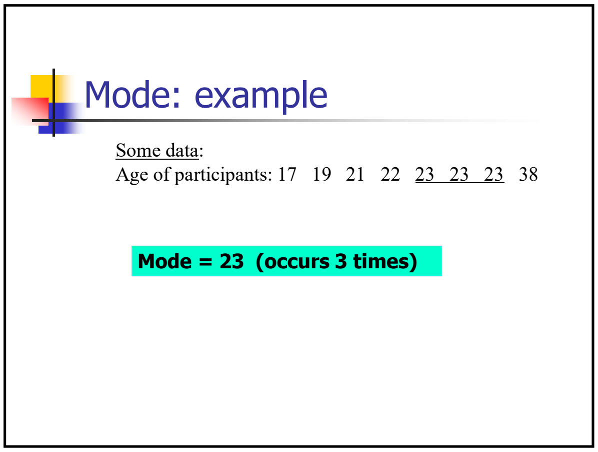
32



Central Tendency

- Mode – the value that occurs most frequently

33

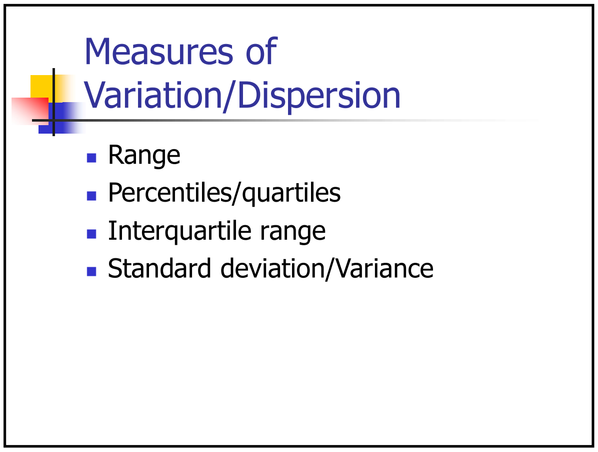


Mode: example

Some data:
Age of participants: 17 19 21 22 23 23 23 38

Mode = 23 (occurs 3 times)

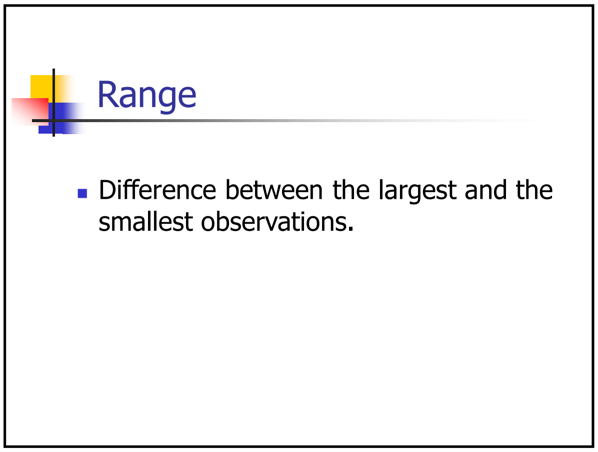
34



Measures of Variation/Dispersion

- Range
- Percentiles/quartiles
- Interquartile range
- Standard deviation/Variance

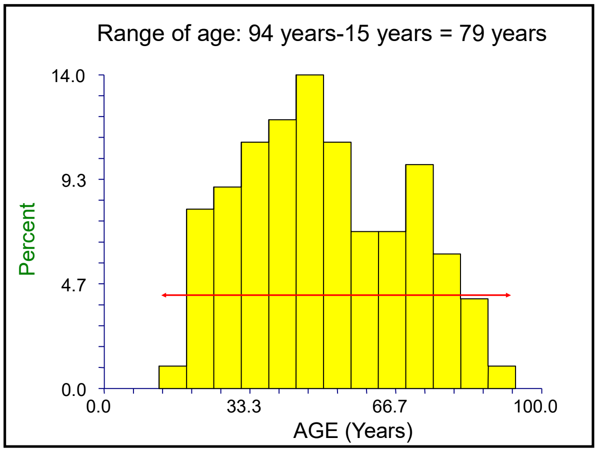
35



Range

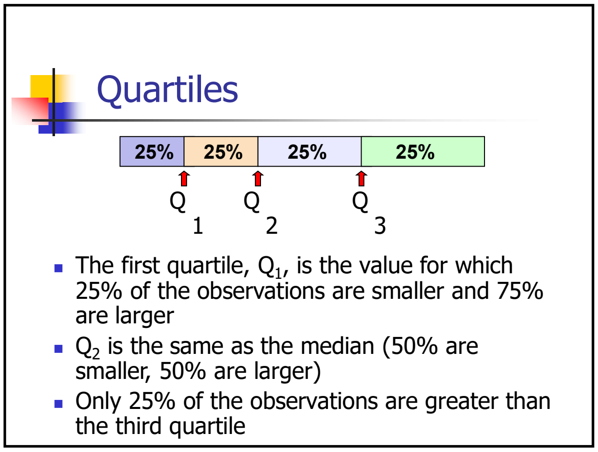
- Difference between the largest and the smallest observations.

36



Range of age: 94 years-15 years = 79 years

37

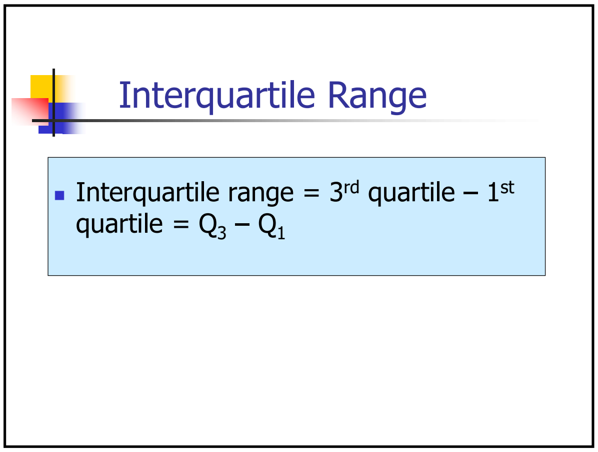


Quartiles

25%
25%
25%
25%
Q
1
Q
2
Q
3

- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

38

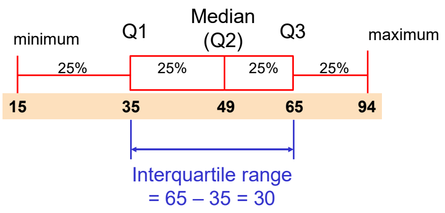


Interquartile Range

- Interquartile range = 3rd quartile – 1st quartile = $Q_3 - Q_1$

39

Interquartile Range: age



40

Variance

- Average (roughly) of squared deviations of values from the mean

$$S^2 = \frac{\sum_{i}^{n} (x_i - \overline{X})^2}{n-1}$$

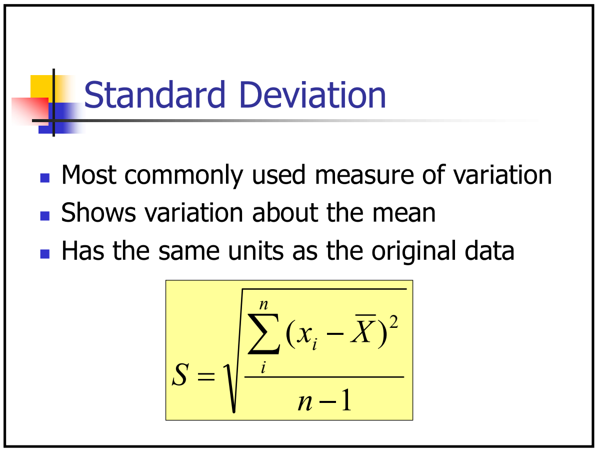
41

Why squared deviations?

- Adding deviations will yield a sum of 0.
- Absolute values are tricky!
- Squares eliminate the negatives.

- Result:
  - Increasing contribution to the variance as you go farther from the mean.

42

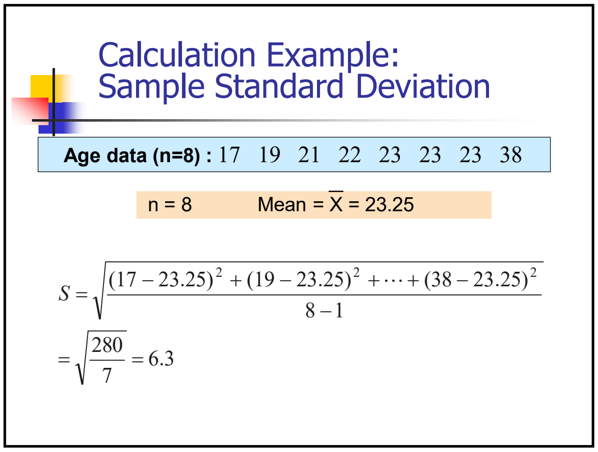


Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

$$S = \sqrt{\frac{\sum_{i}^{n}(x_i - \overline{X})^2}{n-1}}$$

43



Calculation Example: Sample Standard Deviation

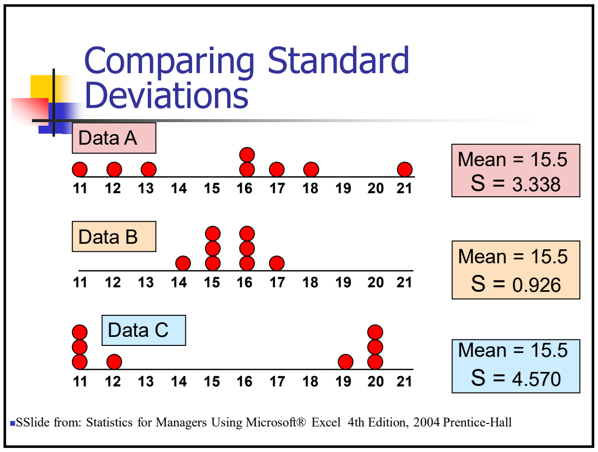
Age data (n=8) : 17 19 21 22 23 23 23 38

n = 8 Mean = $\overline{X}$ = 23.25

$$S = \sqrt{\frac{(17-23.25)^2 + (19-23.25)^2 + \cdots + (38-23.25)^2}{8-1}}$$

$$= \sqrt{\frac{280}{7}} = 6.3$$

44



Comparing Standard Deviations

Data A

Mean = 15.5
S = 3.338

Data B

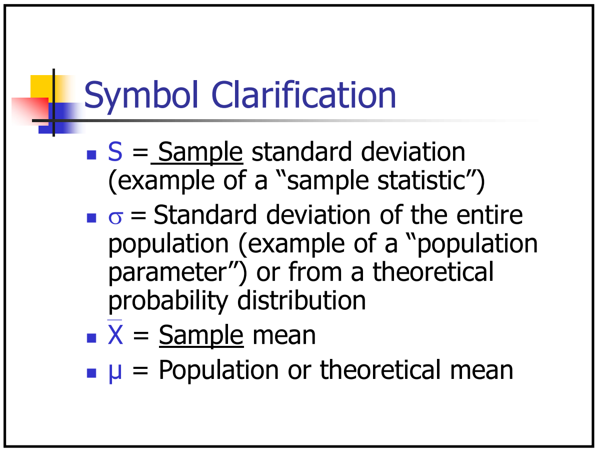
Mean = 15.5
S = 0.926

Data C

Mean = 15.5
S = 4.570

SSlide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

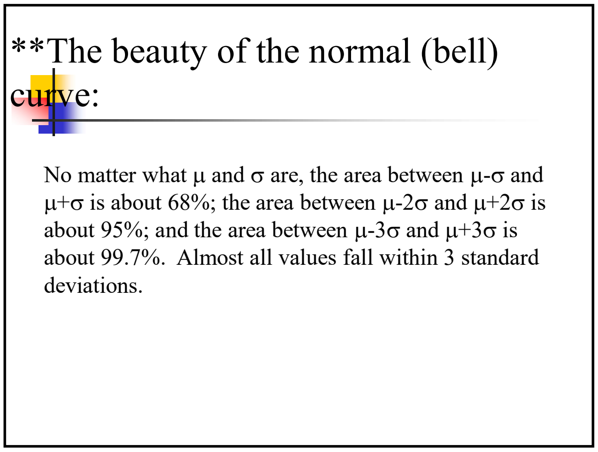
45



Symbol Clarification

- S = Sample standard deviation (example of a "sample statistic")
- σ = Standard deviation of the entire population (example of a "population parameter") or from a theoretical probability distribution
- $\bar{X}$ = Sample mean
- μ = Population or theoretical mean

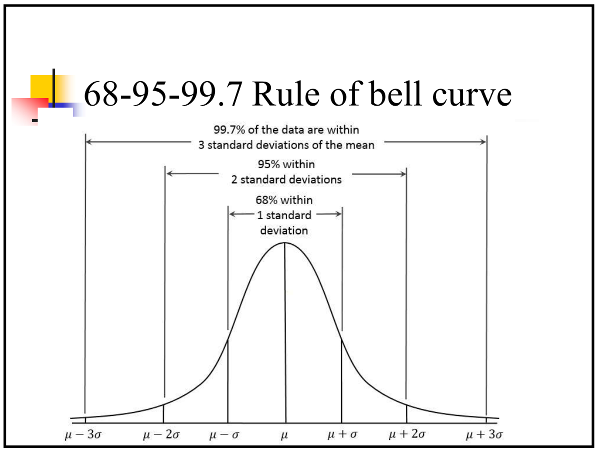
46



**The beauty of the normal (bell) curve:

No matter what μ and σ are, the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%; and the area between μ-3σ and μ+3σ is about 99.7%. Almost all values fall within 3 standard deviations.

47



68-95-99.7 Rule of bell curve

48

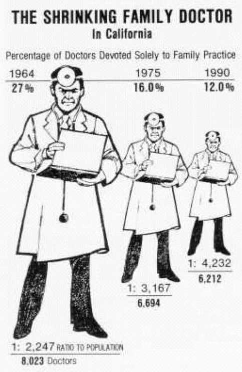
Summary of Symbols

- S^2= Sample variance
- S = Sample standard dev
- σ^2 = Population (true or theoretical) variance
- σ = Population standard dev.
- $\overline{X}$ = Sample mean
- μ = Population mean
- IQR = interquartile range (middle 50%)

49

Examples of bad graphics

50

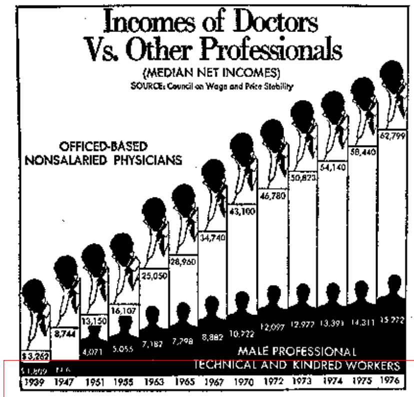


What's wrong with this graph?

Los Angeles Times, August 5, 1979, p. 3.

from: ER Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983, p.69

51

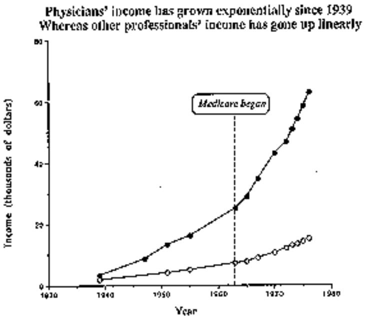


Notice the X-axis

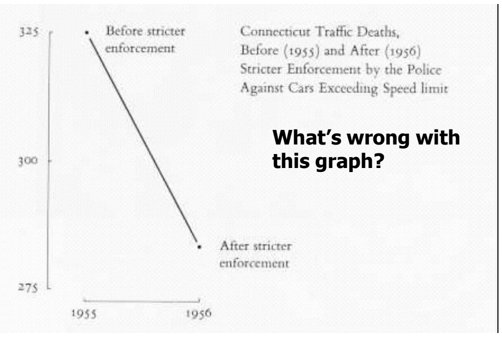
From: Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot Wainer, H. 1997, p.29.

52

Correctly scaled X-axis…



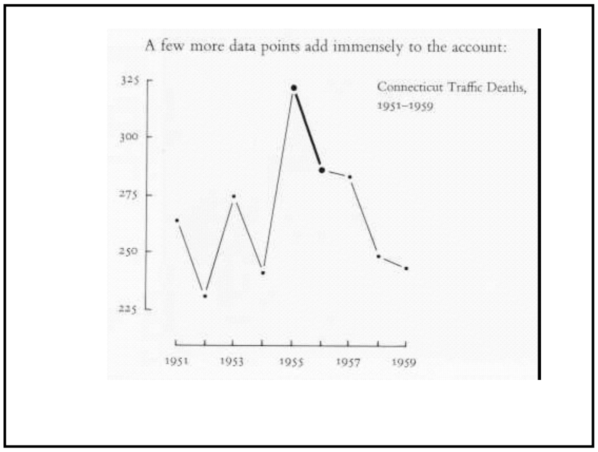
53



What's wrong with this graph?

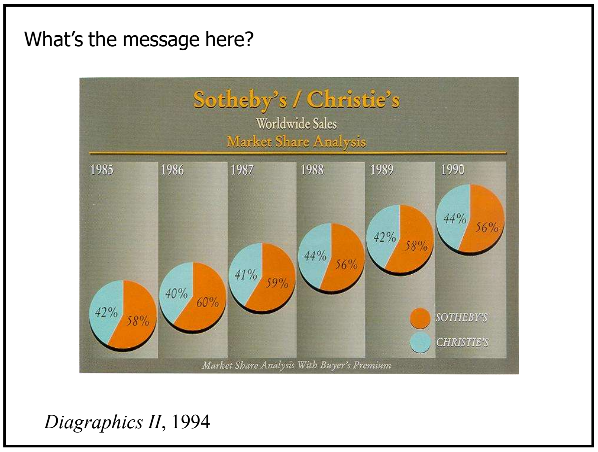
from: ER Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983, p.74

54



A few more data points add immensely to the account:

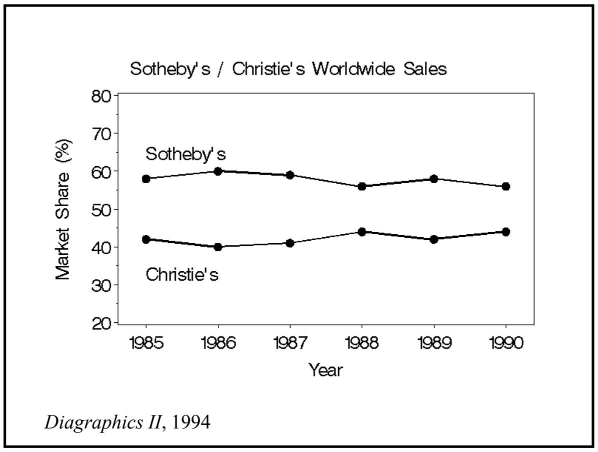
55



What's the message here?

Diagraphics II, 1994

56



Diagraphics II, 1994

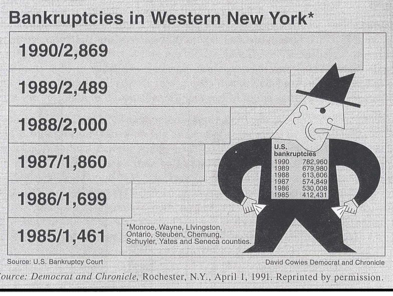
57

MORE PEOPLE FILE: BLAME RECESSION AND EASY CREDIT

By Janet Lively

Last Wednesday, a substance abuse counselor, a single mother on disability and the owner of a foreign automotive repair business gave up any hope of paying their bills.

They filed for bankruptcy, joining the more than 800 people and businesses who have asked for relief this year from the Western New York District of the U.S. Bankruptcy Court in Rochester. If filings continue at the same rate, 1991 will easily be another record year for the court.



Source: U.S. Bankruptcy Court

Source: Democrat and Chronicle, Rochester, N.Y., April 1, 1991. Reprinted by permission.

From: Johnson R. *Just the Essentials of Statistics.* Duxbury Press, 1995.

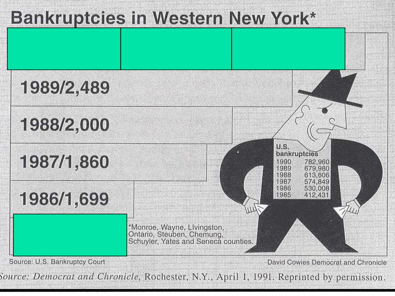
58

MORE PEOPLE FILE: BLAME RECESSION AND EASY CREDIT

By Janet Lively

Last Wednesday, a substance abuse counselor, a single mother on disability and the owner of a foreign automotive repair business gave up any hope of paying their bills.

They filed for bankruptcy, joining the more than 800 people and businesses who have asked for relief this year from the Western New York District of the U.S. Bankruptcy Court in Rochester. If filings continue at the same rate, 1991 will easily be another record year for the court.



Source: U.S. Bankruptcy Court

Source: Democrat and Chronicle, Rochester, N.Y., April 1, 1991. Reprinted by permission.

From: Johnson R. *Just the Essentials of Statistics.* Duxbury Press, 1995.

59

References

- http://www.math.yorku.ca/SCS/Gallery/
- Kline et al. *Annals of Emergency Medicine* 2002; 39: 144-152.
- Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall
- Tappin, L. (1994). "Analyzing data relating to the Challenger disaster". *Mathematics Teacher*, 87, 423-426
- Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983.
- Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot Wainer, H. 1997.

60