

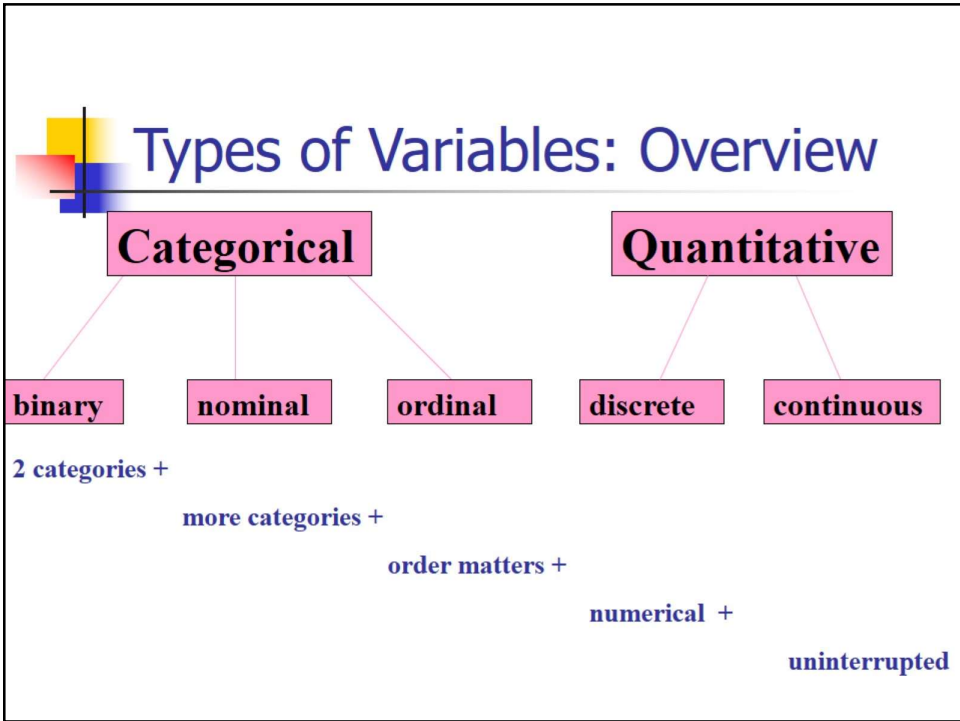


Descriptive Statistics


(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




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

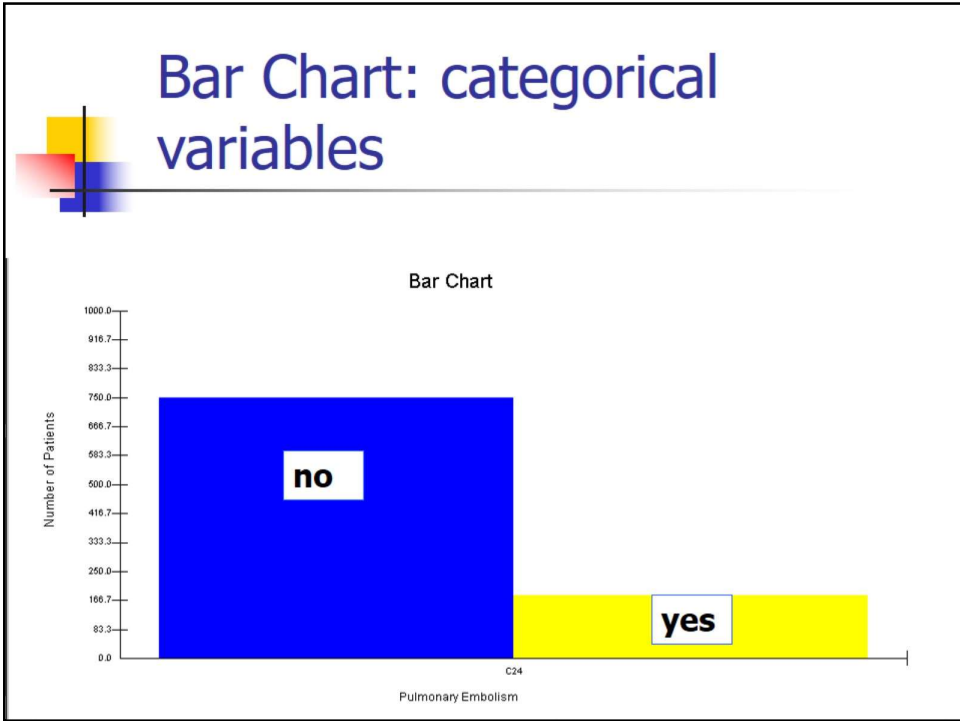
11



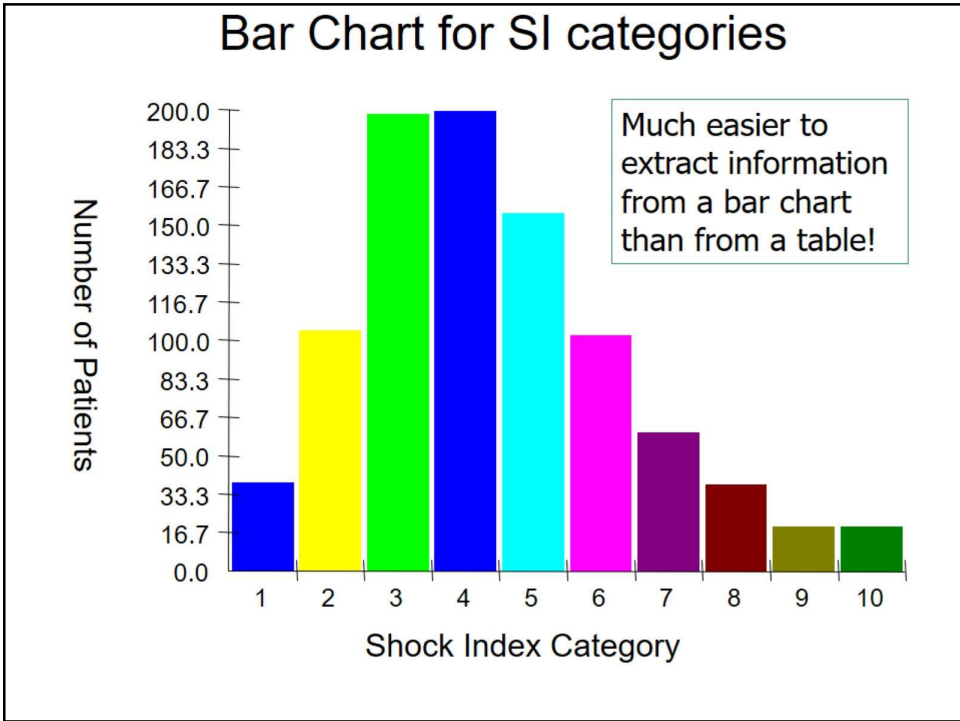
Bar Chart

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

12



13

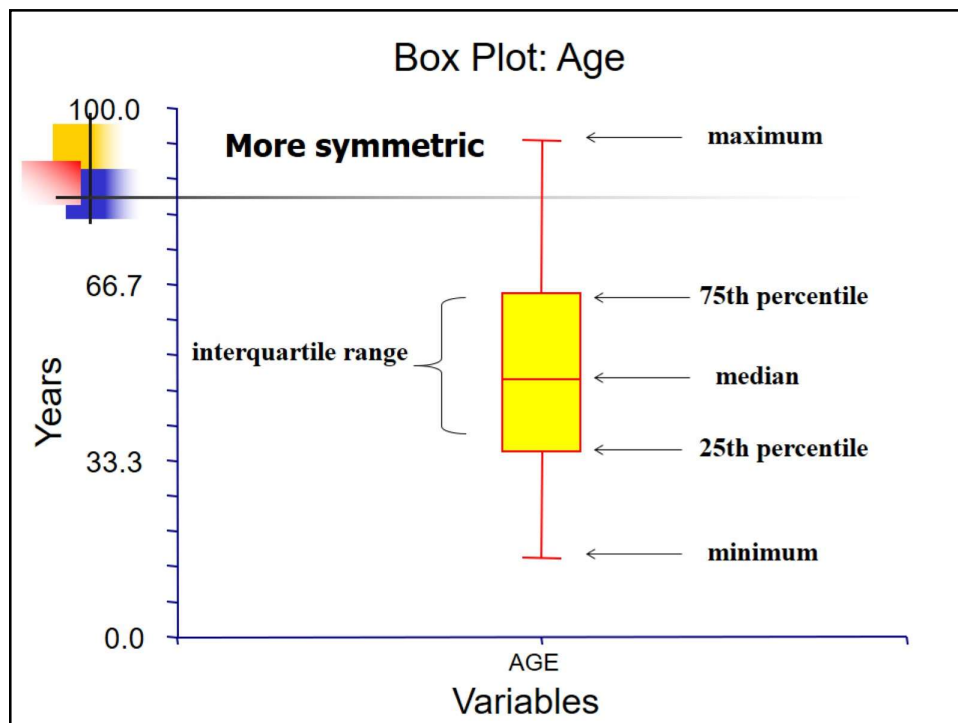


14

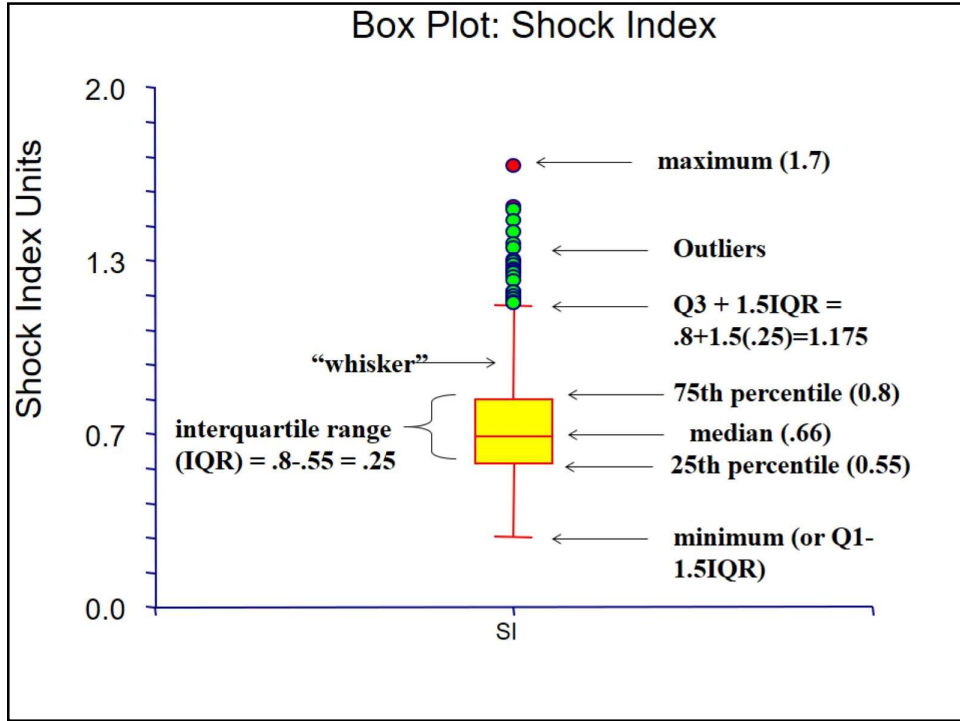
Box plot and histograms: for continuous variables

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

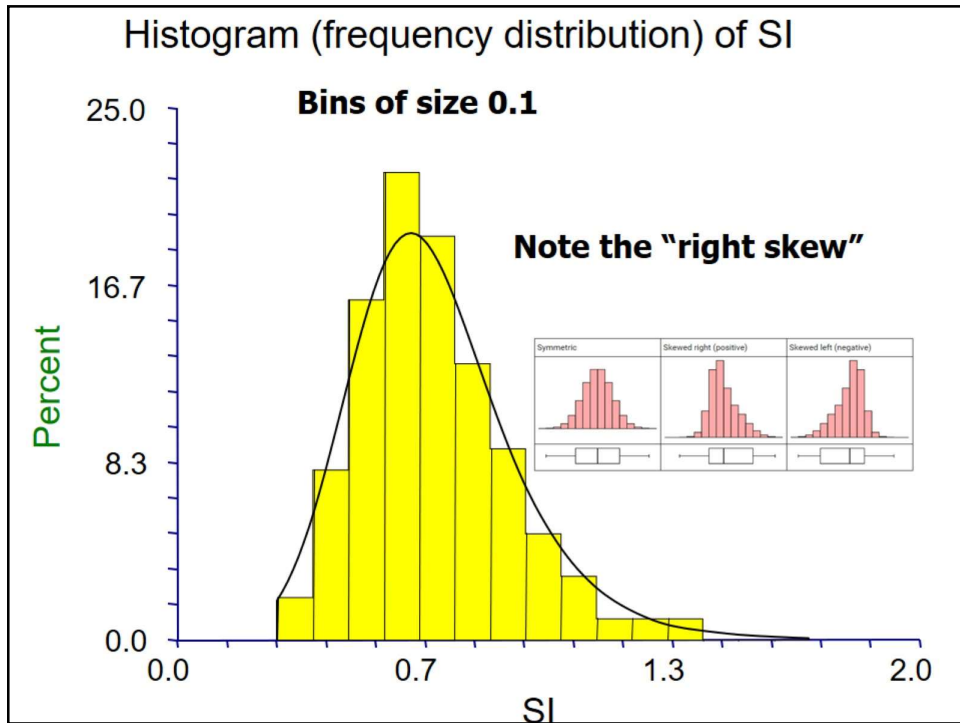
15



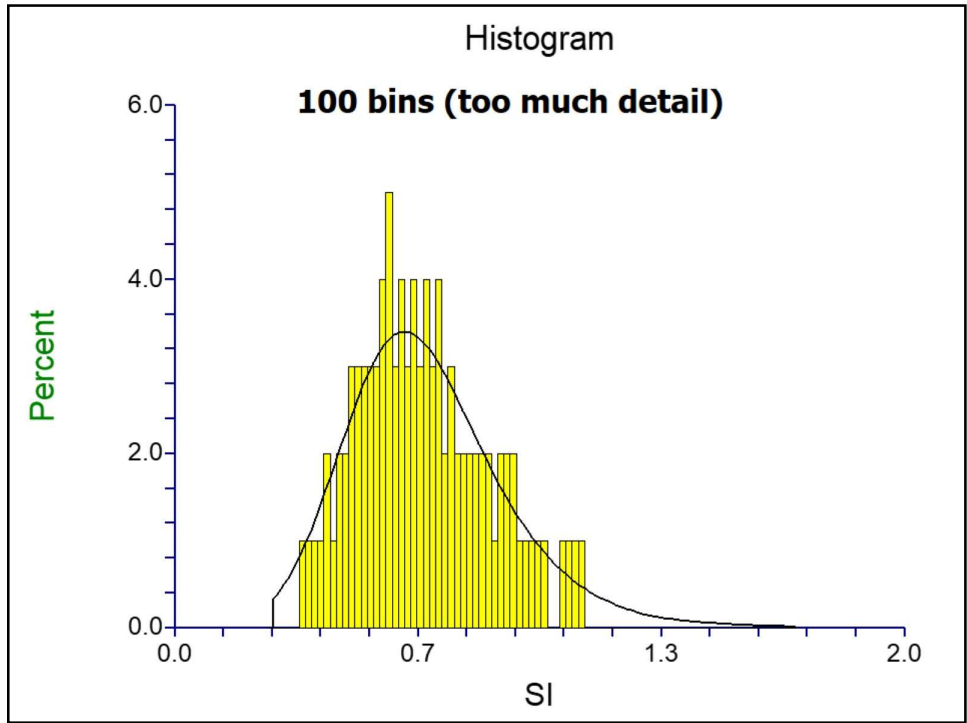
16



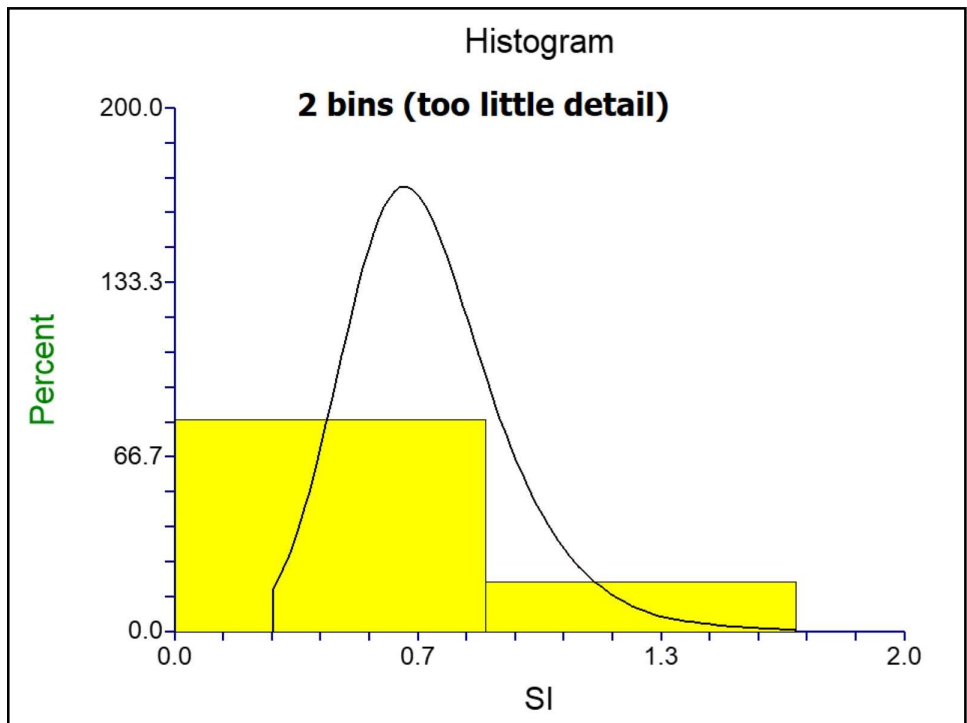
17



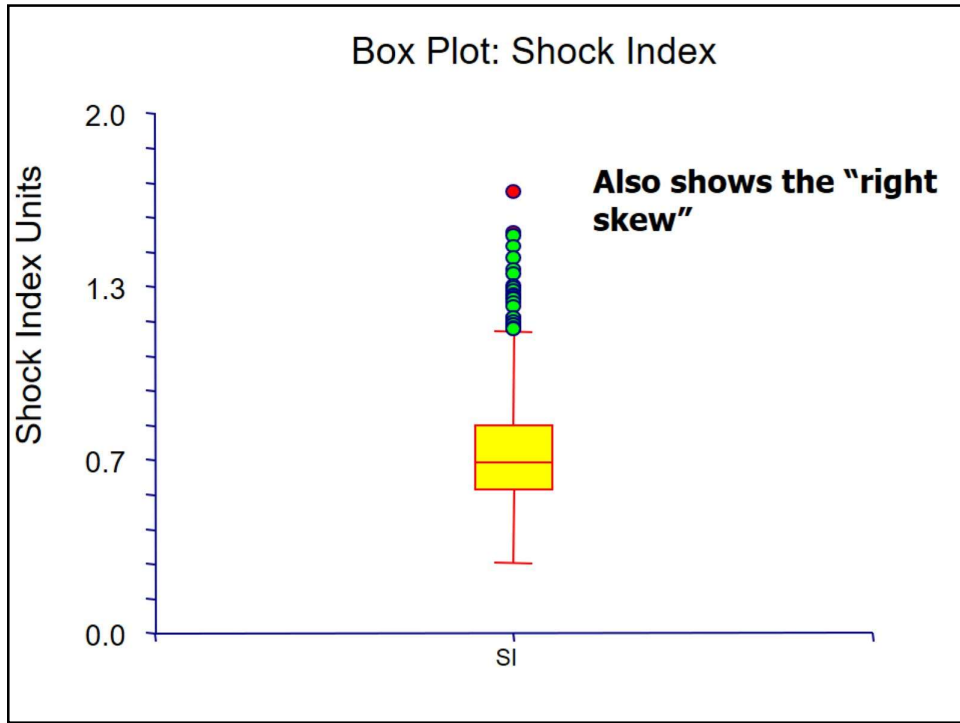
18



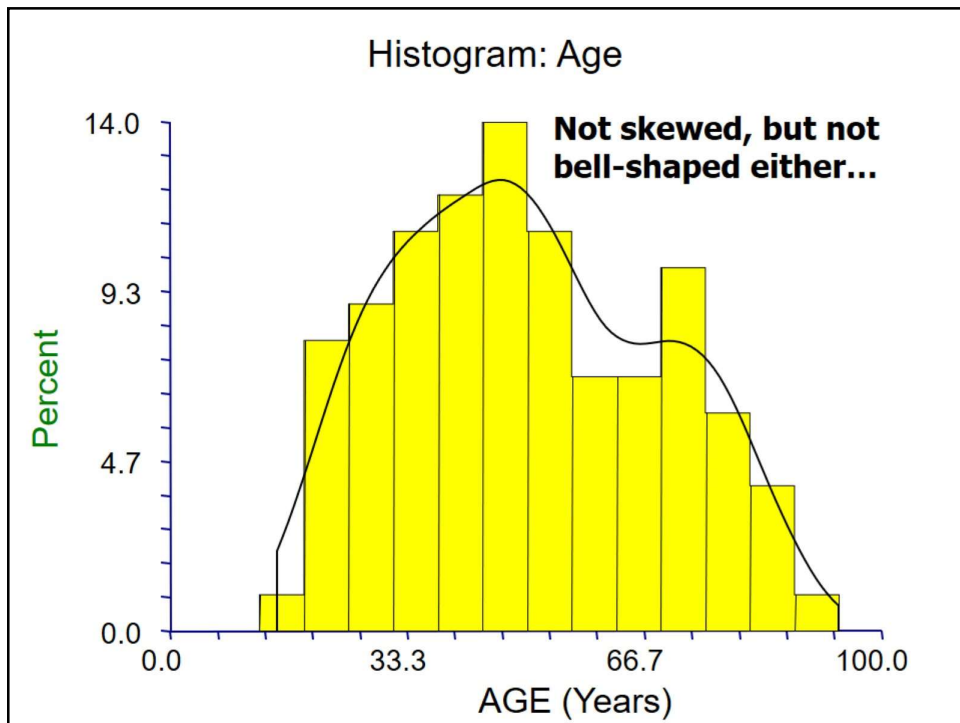
19



20



21



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:

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

24

Mean: example

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

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

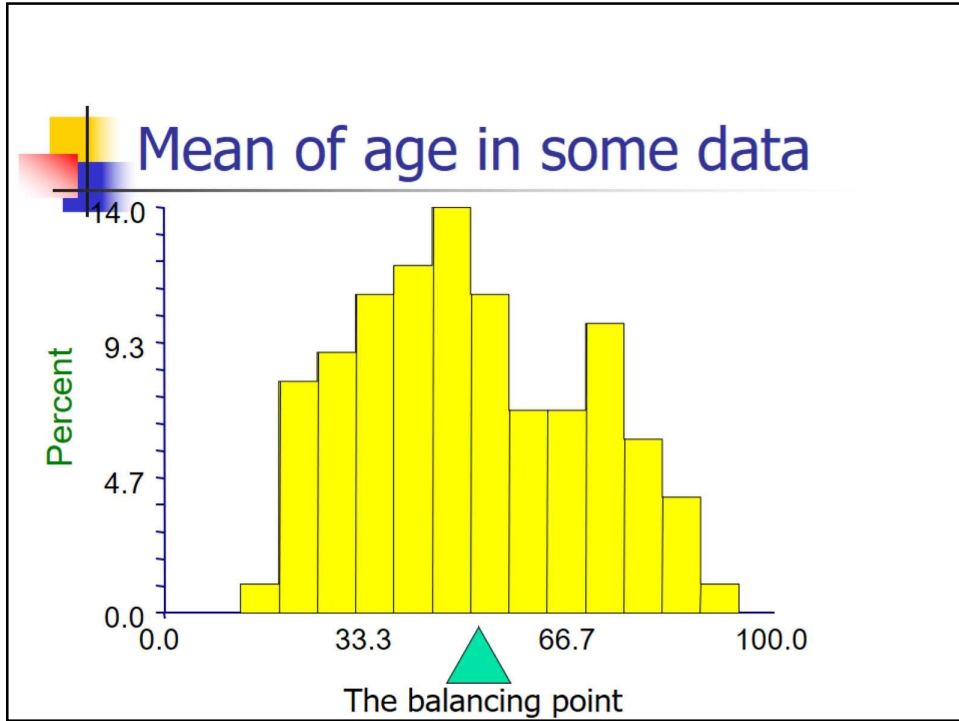
25

Mean of age in some data

Means Section of AGE				
Parameter Value	Mean	Median	Sum	Mode
	50.19334	49	46730	49

A histogram showing the distribution of age data. The x-axis is labeled from 0.0 to 100.0 with major ticks at 0.0, 33.3, 66.7, and 100.0. The y-axis is labeled 'Percent' and ranges from 0.0 to 14.0 with major ticks at 0.0, 4.7, 9.3, and 14.0. The histogram consists of yellow bars. A callout box from the table above points to the mean value of 50.19334, which is located on the x-axis between 33.3 and 66.7.

26



27


Mean

- The mean is affected by extreme values (outliers)

The diagram shows two number lines from 0 to 10. The first number line has blue dots at 1, 2, 3, 4, and 5. A red arrow points to the value 3, with a box below it stating 'Mean = 3'. Below this is the calculation: $\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$. The second number line has blue dots at 1, 2, 3, 4, and 10. A red arrow points to the value 4, with a box below it stating 'Mean = 4'. Below this is the calculation: $\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$.

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

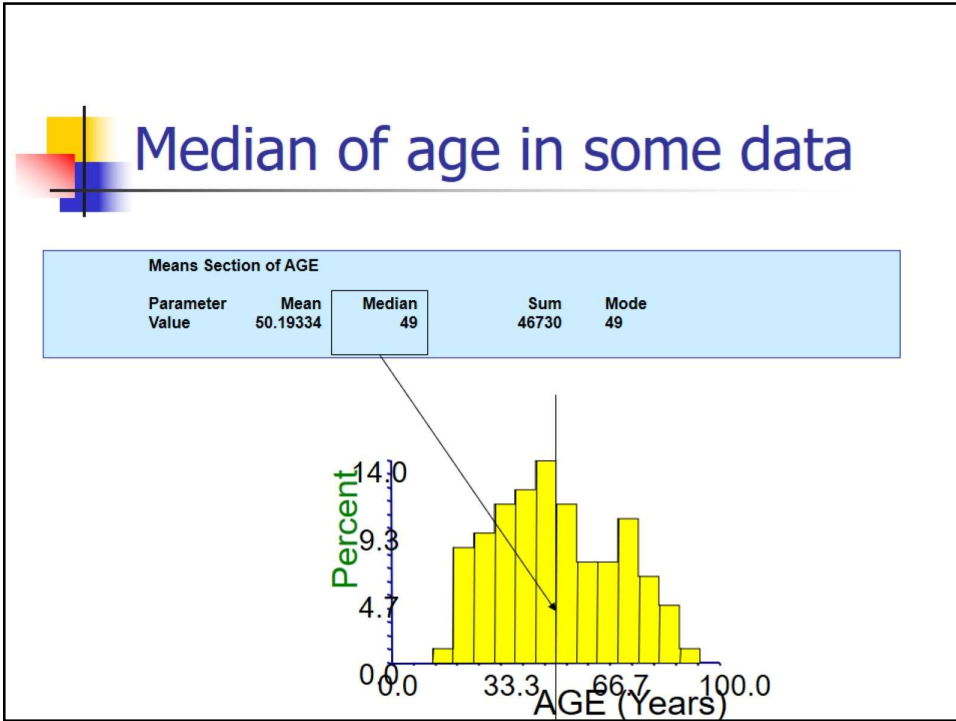


Median: example

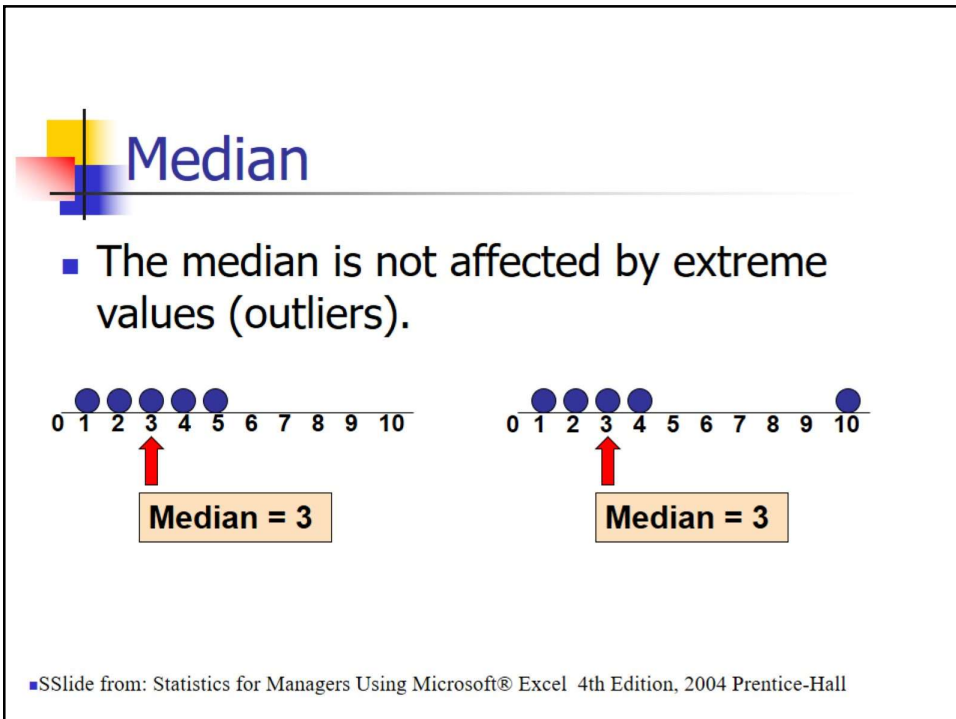
Some data:
Age of participants: 17 19 21 22 23 23 23 38

Median = $(22+23)/2 = 22.5$


30



31




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

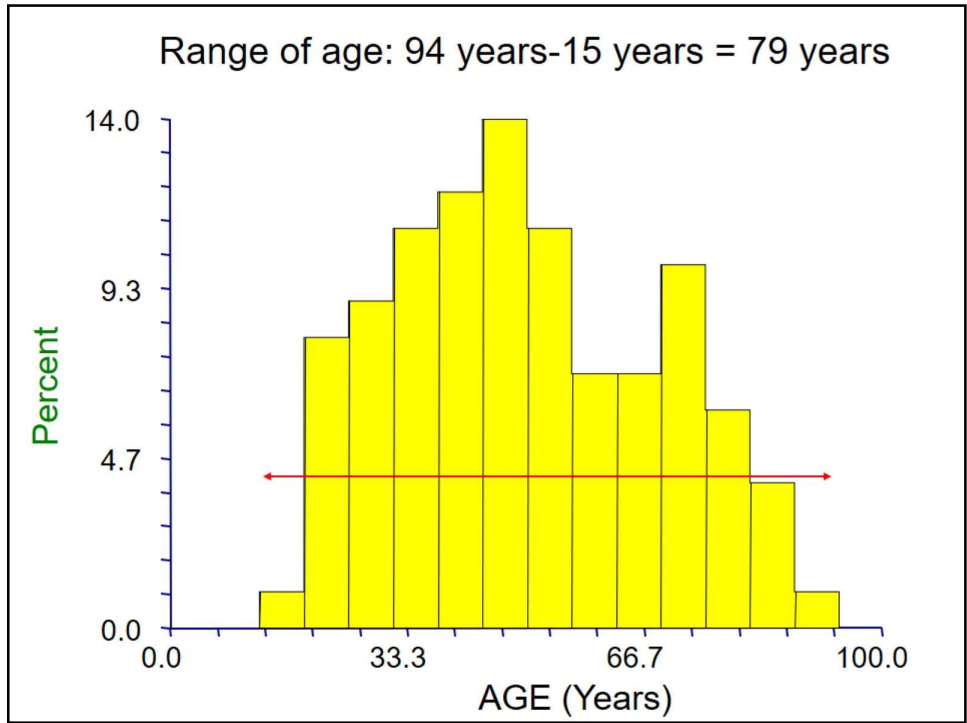
35



Range

- Difference between the largest and the smallest observations.

36




37

Quartiles

Quartile	Percentage
Q ₁	25%
Q ₂	25%
Q ₃	25%
Q ₄	25%

- 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

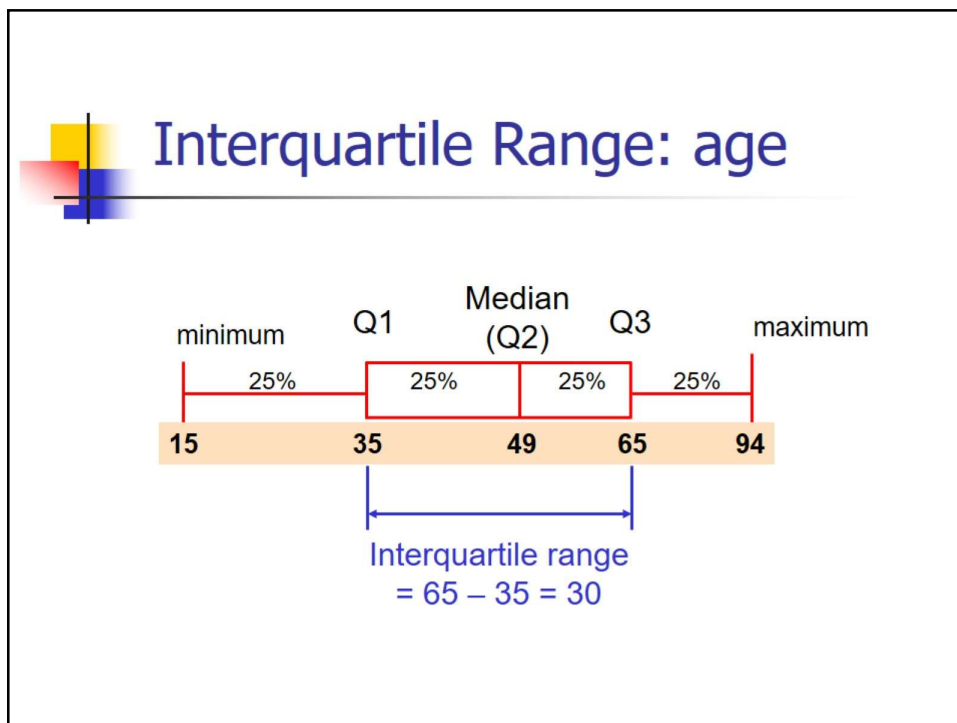
38




Interquartile Range

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

39



40




Variance

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

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

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

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

43

Calculation Example: Sample Standard Deviation

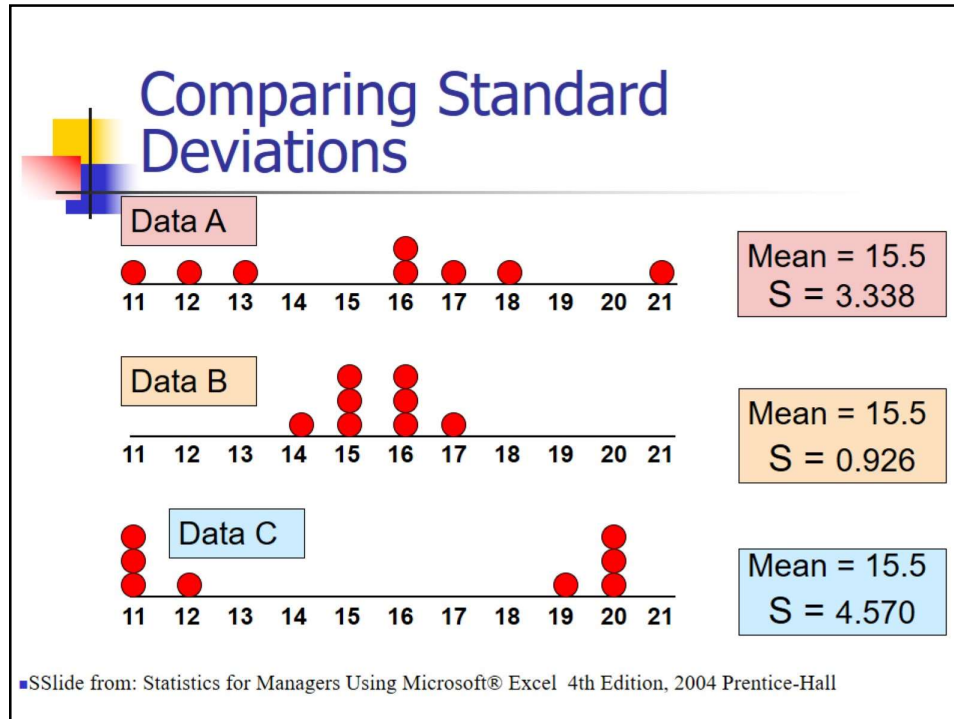
Age data (n=8) : 17 19 21 22 23 23 23 38

n = 8 Mean = \bar{X} = 23.25

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

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

44



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

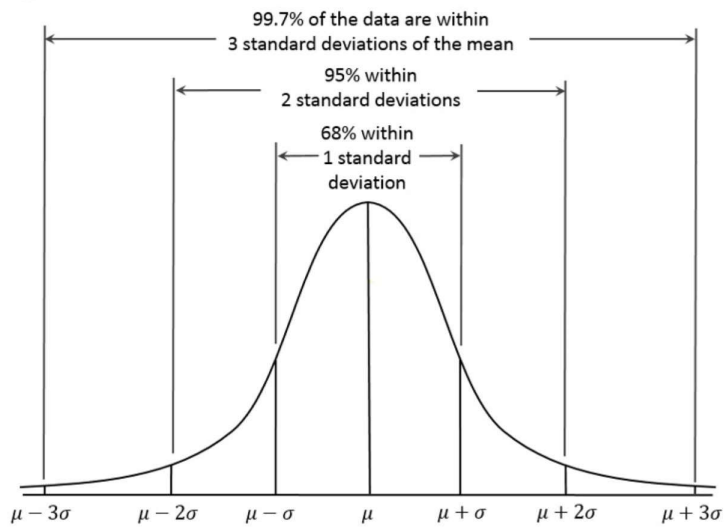
46

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


No matter what μ and σ are, the area between $\mu - \sigma$ and $\mu + \sigma$ is about 68%; the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 95%; and the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ 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

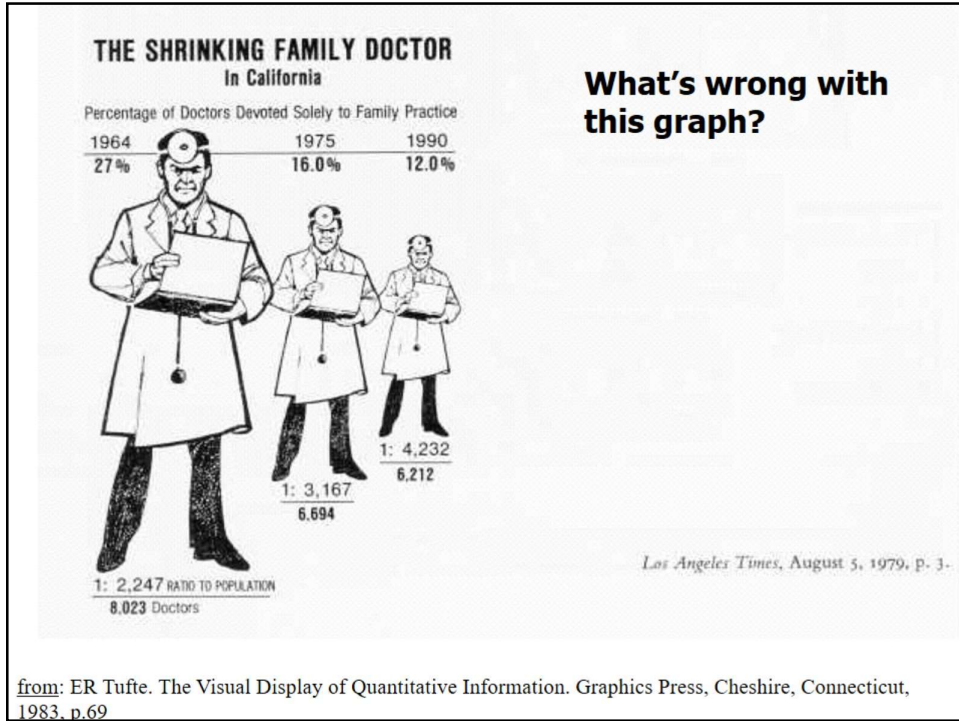
- S^2 = Sample variance
- S = Sample standard dev
- σ^2 = Population (true or theoretical) variance
- σ = Population standard dev.
- \bar{X} = Sample mean
- μ = Population mean
- IQR = interquartile range (middle 50%)

49

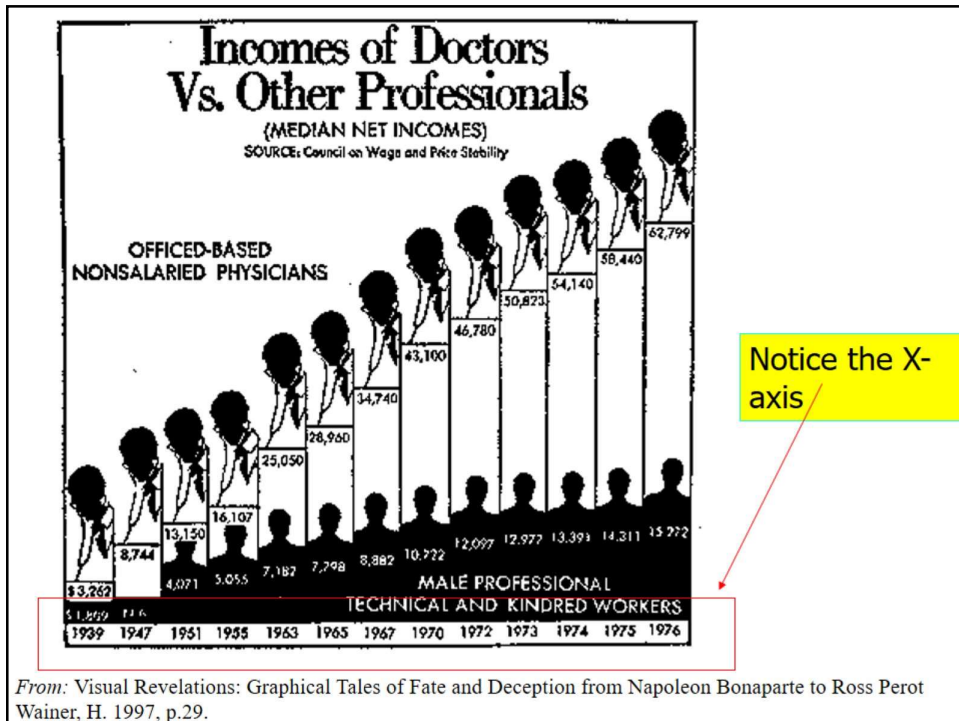


Examples of bad graphics

50

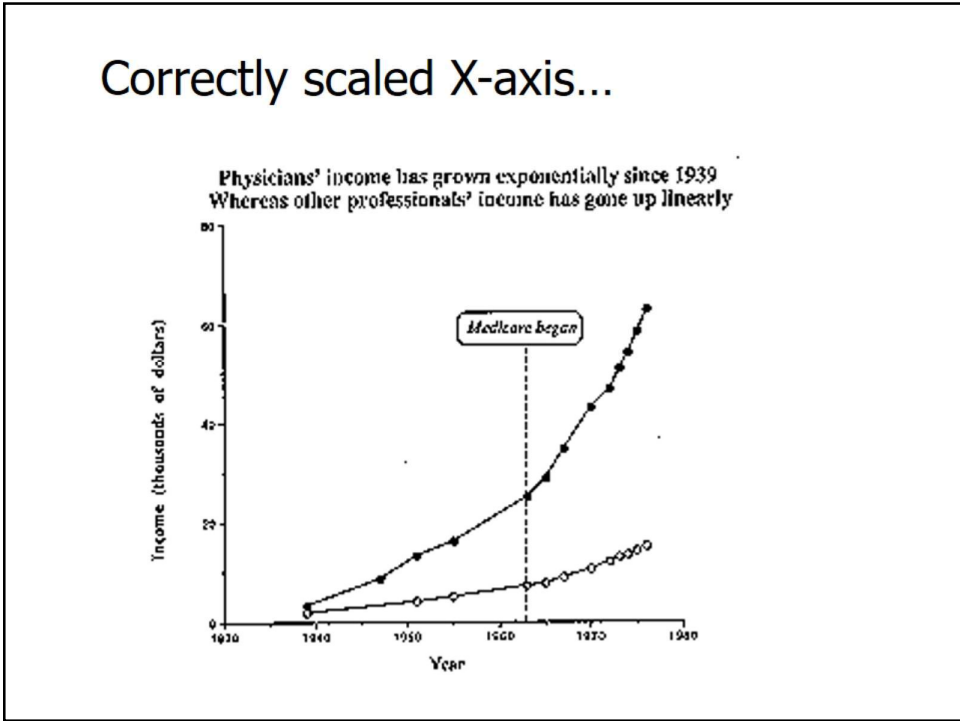


51

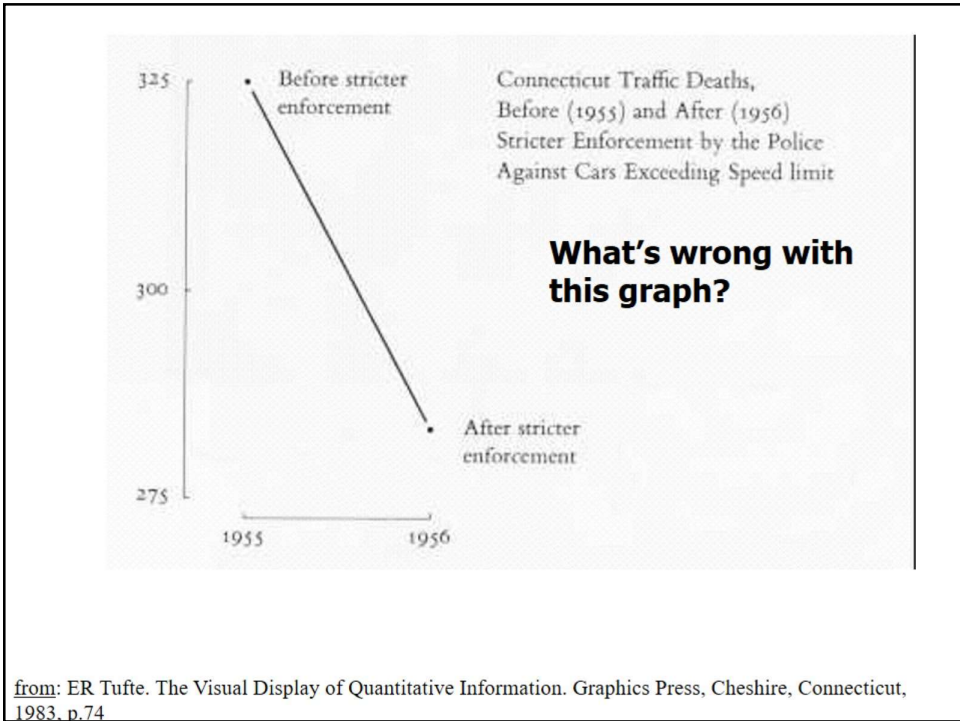


52

Correctly scaled X-axis...

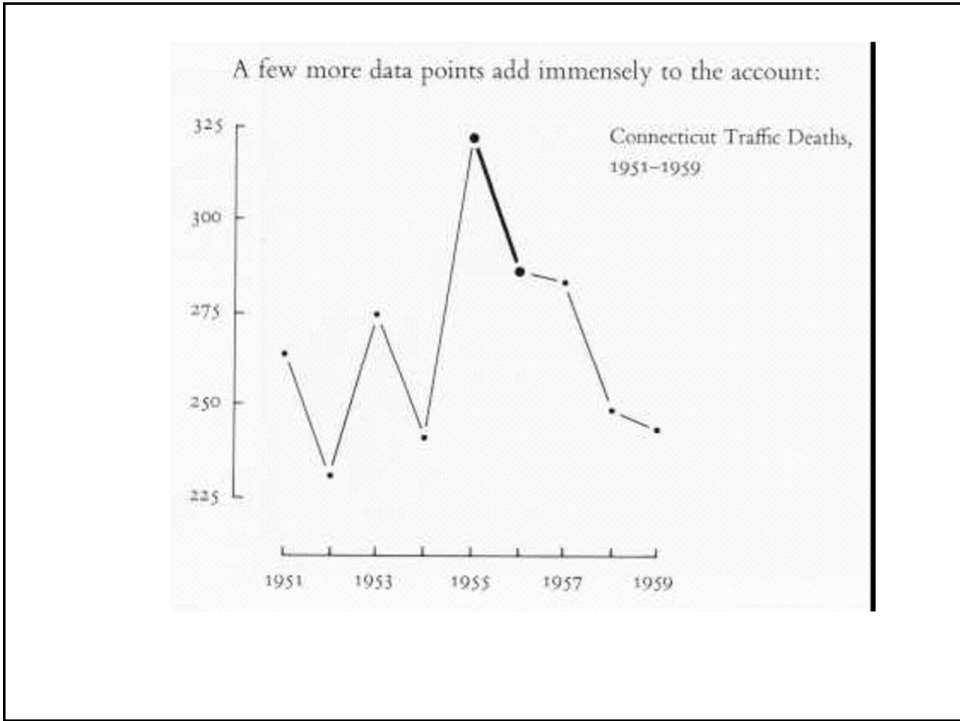


53

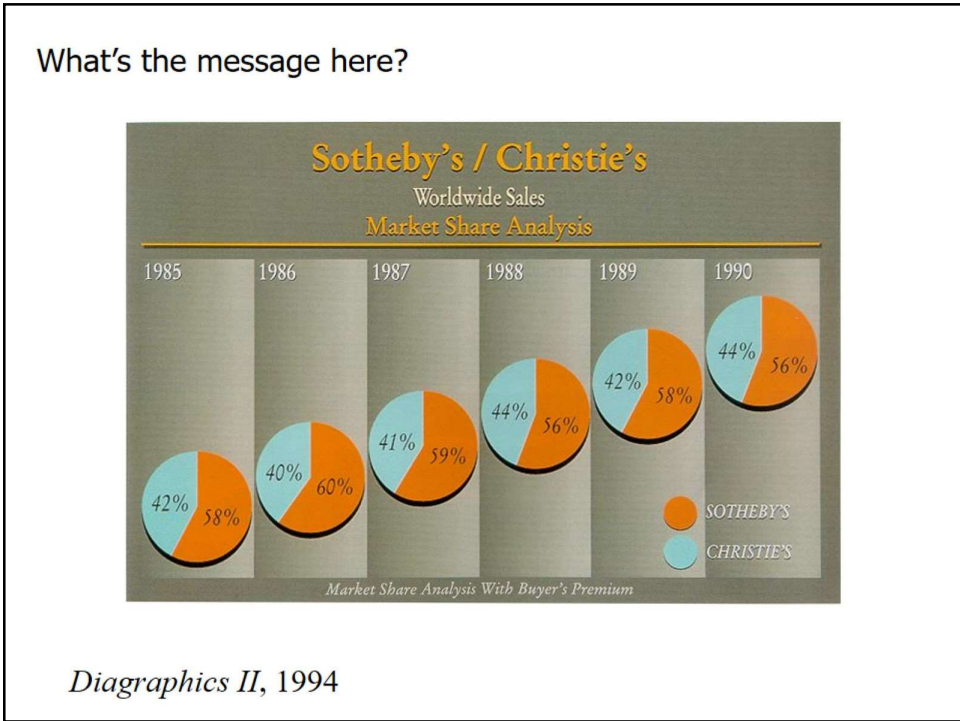


from: ER Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983, p.74

54

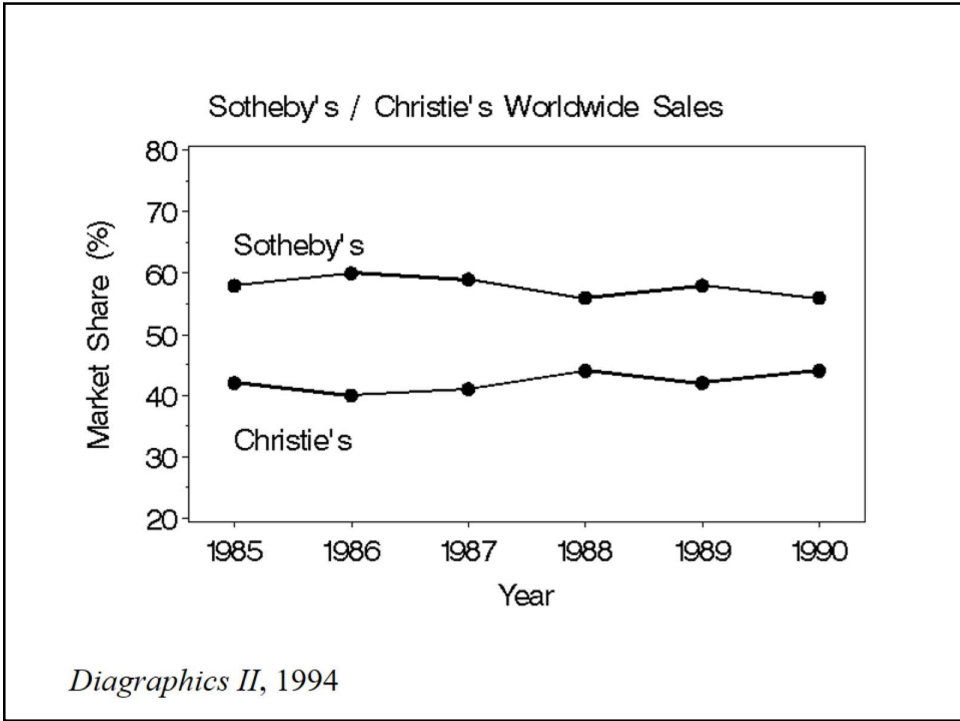


55



Diagraphics II, 1994

56



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.

Bankruptcies in Western New York*

1990/2,869	
1989/2,489	
1988/2,000	
1987/1,860	
1986/1,699	
1985/1,461	

Source: U.S. Bankruptcy Court

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

*Monroe, Wayne, Livingston, Ontario, Steuben, Chemung, Schuyler, Yates and Seneca counties.

David Cowles Democrat and Chronicle

**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.

Bankruptcies in Western New York*

Year	Bankruptcies
1989	2,489
1988	2,000
1987	1,860
1986	1,699

*Monroe, Wayne, Livingston, Ontario, Steuben, Chemung, Schuyler, Yates and Seneca counties.

Source: U.S. Bankruptcy Court David Cowies Democrat and Chronicle

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