Variables and Descriptive Statistics

Donna Kritz-Silverstein, Ph.D.
Professor
Department of Family & Preventive Medicine
University of California, San Diego

(858) 534-1818
dsilverstein@ucsd.edu

Last Class…

- **Scientific method**
  - Hypothesis testing
- **Sampling Strategies**
- **Types of study designs**

Purpose . . .

- Variables
  - Variable types
  - Types of data
  - Scales of measurement

Purpose…

- Descriptive statistics
- Categorical variables
  - rates, %
- Continuous variables
  - Measures of Central Tendency
  - Measures of dispersion
- Distributions, normal, skewed
- Data display for descriptive statistics

Introduction

- Overview of statistical techniques
- Includes most major types of statistical analyses needed to analyze your data
- Focus
  - Practical considerations
  - Applied data analysis

Variables

- **Variable** =

  Any characteristic that can vary

Examples: Height, weight, age, behaviors, attitudes, presence of specific disease, clinical measurements, physical measures

Variables

- **Independent Variable (IV) =**
  - Variable that is changing or manipulated
  - Presumed cause
- **Dependent Variable (DV) =**
  - Response
  - Outcome

Variables

- In experiments,
  - Independent Variable =
    Variable being manipulated by experimenter
  - Dependent Variable =
    Is observed or measured for variation as a presumed result of the variation in the IV

Variables

- In observational studies,
  - IV = variable that "logically" has some effect on the DV
  - Example =

    Research on smoking & lung cancer

Which is IV? DV?

- The hypothesis for a study:
  - There will be a significant difference in anatomy grades of DS who participated in an intensive study summer program as an undergrad compared to DS who did not participate in an intensive study program

Which is IV? DV?

- **Study Description:**
  - **Dean of faculty at a dental school is concerned about the turnover in faculty—many of them leave their positions before completing 2 years of their 5 year contract. The Dean wants to identify factors that predict commitment to teaching and decides to assess attitudes toward students and amount of money owed on loans**

Variables

- **Any variable can serve as the IV in one study, and the DV or outcome in another**

Examples:

Does use of fluoride prevent tooth decay?

IV=fluoride DV=caries

Does parents education level predict use of fluoride in children?

IV=education **DV=fluoride**

Variables

- **Confounding**
  - –A distortion in an observed relationship between an exposure & outcome brought about by a third variable

Variables

- **Confounders**
  - Associated w/ both independent & dependent variables (eg., age in study of diabetes & AD)
  - Variables that can affect or bias observed results ("Lurking variables")

Variables

- Synergism
  - The interaction of 2 causal variables so that the combined effect is greater than the sum of their effects
  - Example =effect of both smoking and drinking on cognitive function is greater in combination

Variables

- **Effect modification (interaction)**
  - The direction or strength of an association between 2 variables differs according to a third variable
  - Example =coffee & cognitive function—sex modifies the association (women positive association, men no association)
  - Hypertension & sex—modified by age (under 45y, men more likely to have HTN; after 45y, women more likely to have HTN)

Types of Data

- **Discrete data**
  - Categorical data
  - Has limited set of values
  - May be qualitative
  - Examples: eye color, blood type, gender, presence/absence of diseases, yes/no data

Types of Data

- **Continuous data**
  - Has values that range along a continuum
  - Quantitative
  - Examples: age, body mass index, blood pressure, # teeth
  - Can always take continuous data & convert to categories

Scales of Measurement

- **Nominal scales**
  - Named categories
  - No particular order (1 isn't any more than another)
  - Examples: eye color, hair color, gender

Scales of Measurement

- **Ordinal scales**
  - Ordered categories
  - Distance between categories is unequal
  - Examples: 1st place, 2nd place, 3rd place;
    rate heath compared to others –
    better, the same, worse;
    mild, mod, severe perio disease

Scales of Measurement

- **Interval (continuous) scales**
  - Equal distance between data points
  - No true zero
  - Examples: Fahrenheit temperature – (distance 10° & 20° =distance 20° & 30°)

Scales of Measurement

- **Ratio scales**
  - Equal intervals between data points
  - Has true zero
  - Best type of scale
  - Examples: blood pressure, # teeth

Scales of Measurement

- **Order of scales**
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Each successive scale has all characteristics of the previous one

Data Analysis

- **Statistics** = describes & presents collected data in a meaningful way

- 2 types of statistics
  - **Descriptive statistics** = describes the sample, summarizes who is in sample
  - **Inferential Statistics** = infer things about population based on sample

Descriptive statistics

- For Categorical variables
  - Nominal scale
    - Rate (% of total sample with that characteristic)
    - Example: total sample=150
      - 15 have diabetes=10%
    - 80 dental students

      64 exercise 3x/week = 80%

Descriptive statistics

- For continuous variables
- Measures of central tendency
  - Mean
  - Median
  - Mode

Descriptive statistics

- **Measures of central tendency**
  - **Mean** = average = $\frac{\Sigma x}{N}$

Where x=scores; N=total sample size

Scores:
55
95
95
78
96
81
63
82

Mean= $\frac{645}{8}$ = 80.625

Descriptive statistics

- Mean - properties
  - Very sensitive to small variations in scores
  - Outliers (extreme values) can cause large changes in the mean; won't give accurate picture of the population (eg., exam scores)
  - More powerful statistics use means

Descriptive statistics

- **Measures of central tendency**
  - **Median** = middle score, 50th percentile

    -Put into numerical order, middle score; if 2 middle scores, median= average of the two

Scores:	55	→	55
	95		63
	95		78
	78		81
	96		82
	81		95
	63		95
	82		96

$$\text{Mean} = \frac{645}{8} = 80.625$$

Median=81.5

Descriptive statistics

- Median
  - Advantages
    - Not as sensitive to outliers
    - Use for describing a variable where there are many outliers (eg., income)
  - Disadvantages
    - Statistics not as powerful

Descriptive statistics

- **Measures of central tendency**
  - **Mode** = Most frequently occurring score

Scores:		→	
	55		55
	95		63
	95		78
	78		81
	96		82
	81		95
	63		95
	82		96

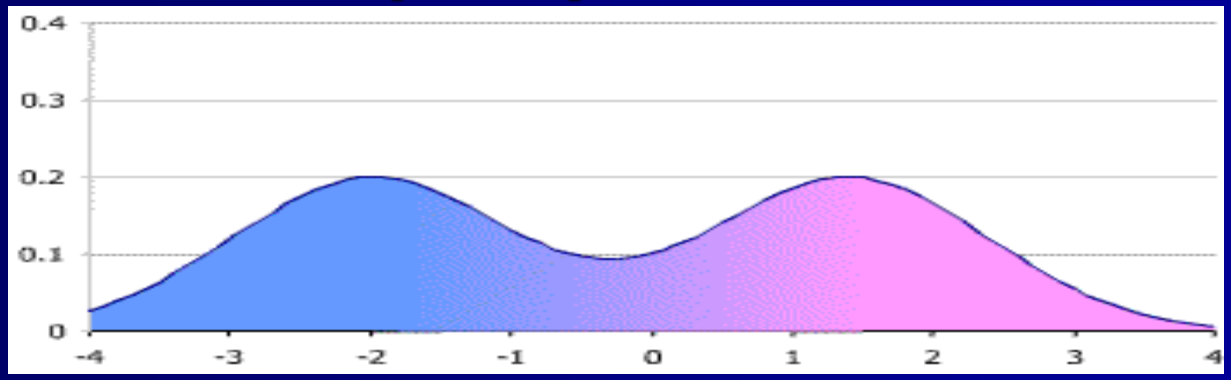
$$\text{Mean} = \frac{645}{8} = 80.625$$

Median = 81.5

Mode = 95

Descriptive statistics

- **Mode- properties**
  - **Distributions can have ≥1 mode**
  - **Bimodal distribution-** distribution with 2 different peaks
    2 distinct values that measurements center around
    example: heights of men & women
  - **Distributions can have no mode—all measures=frequency**

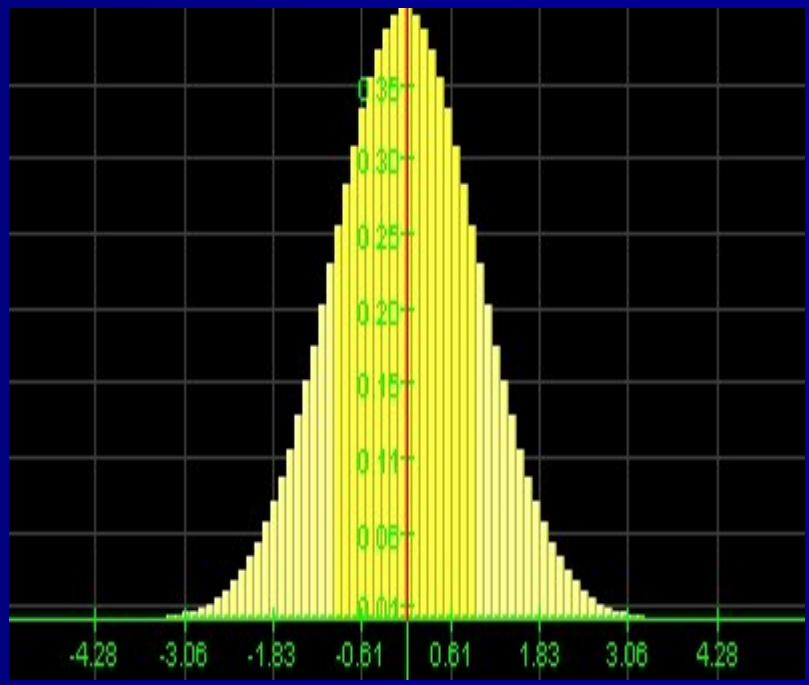


Descriptive statistics

- Measures of dispersion
  - Another way to describe the sample
  - Shows how far scores are scattered around the mean
    - Distributions
    - Range
    - Variance
    - Standard deviation

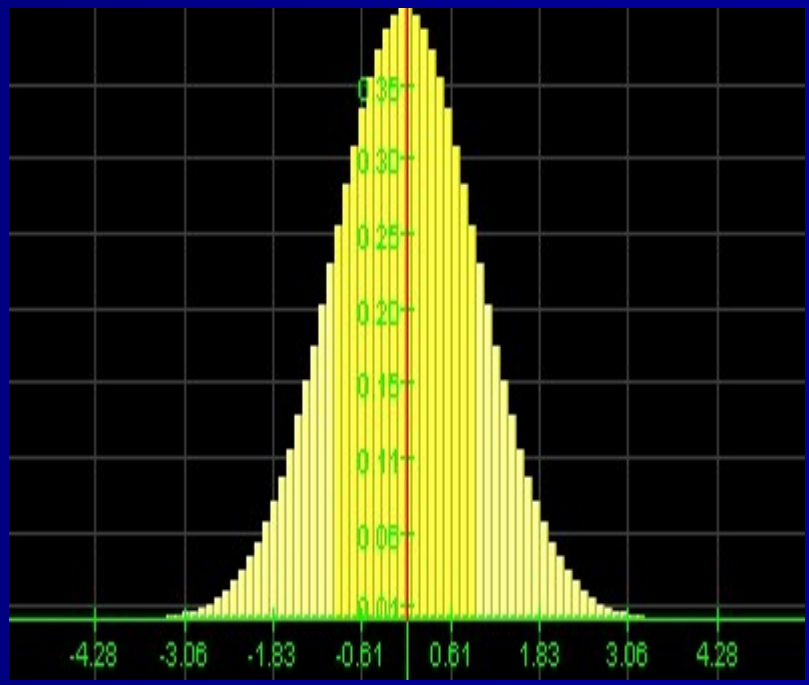
Distributions

- **Normal distribution**
  - *Bell shaped*
  - **Most data points fall in middle, w/ few very small & few very large values**
  - **Mean, Median & Mode all occur at the same score**



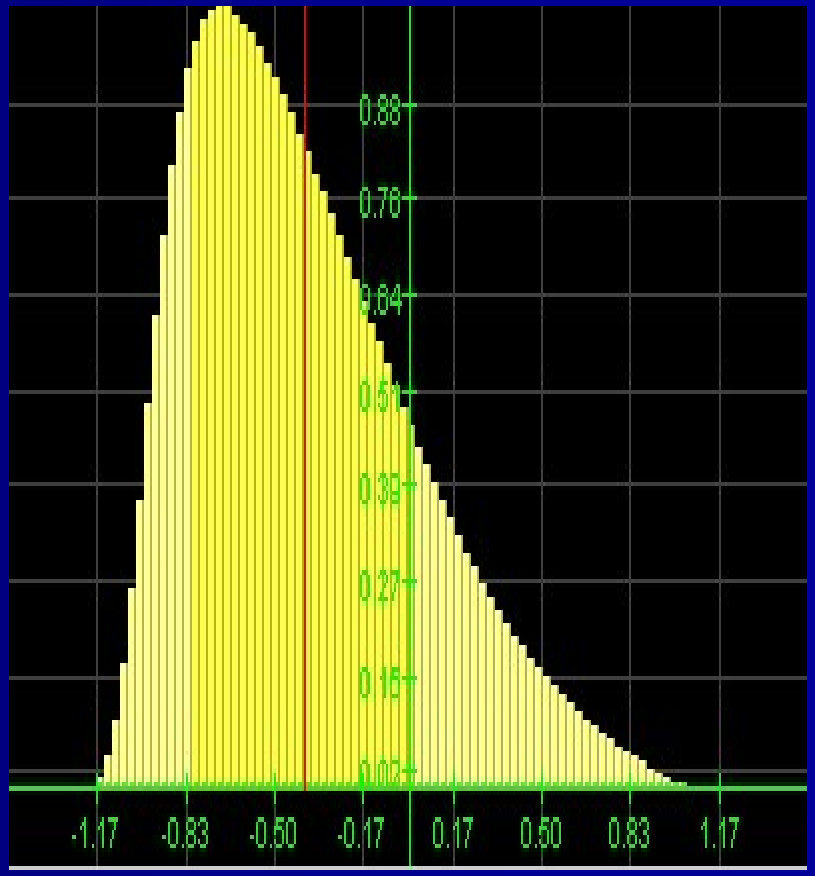
Distributions

- **Normal distribution**
  - **Mean, Median & Mode all occur at the same score**
  - **Symmetrical – each half = mirror image exactly half the scores occur above and half below mean**



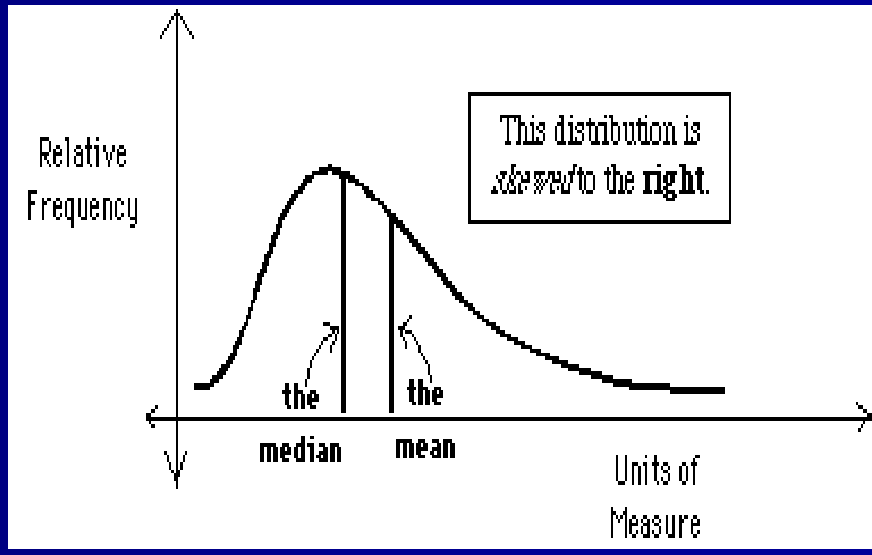
Distributions

- **Skewed to Right**
  - **looks like bell curve w/ *longer tail on right* and mound pushed to left**
  - **Most data points fall to left of middle & more very small than very large values**



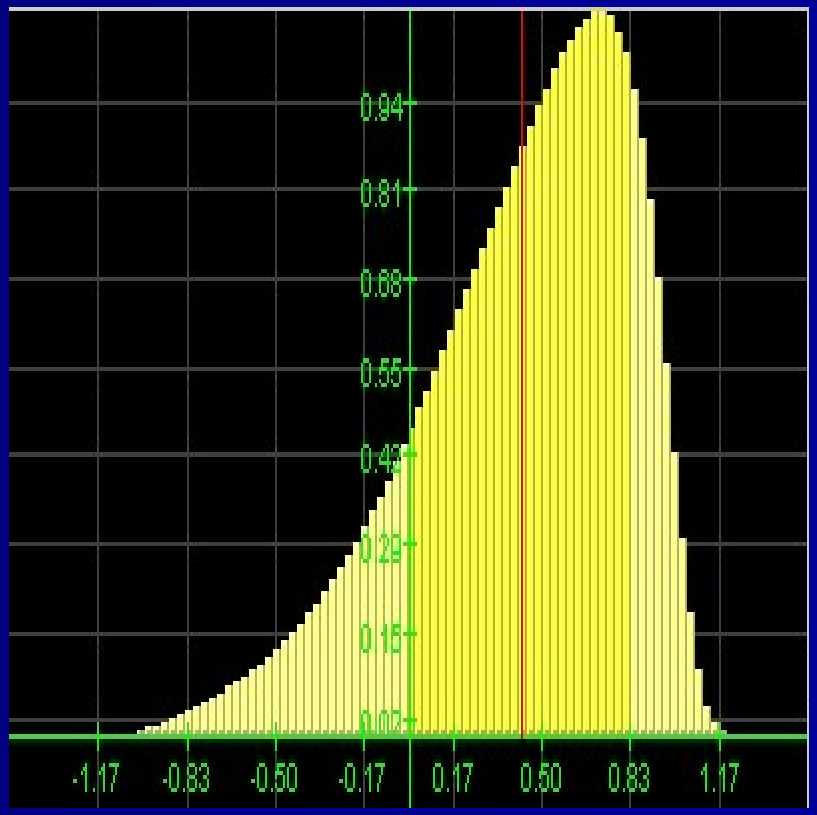
Distributions

- **Skewed to Right**
  - **Mean > median**
  - ***Positively* skewed**
  - **large extremes pull mean → the tail**
    (extremes high values)
  - **Median remains closer to center of the distribution**
  - **Ex: income, CRP**



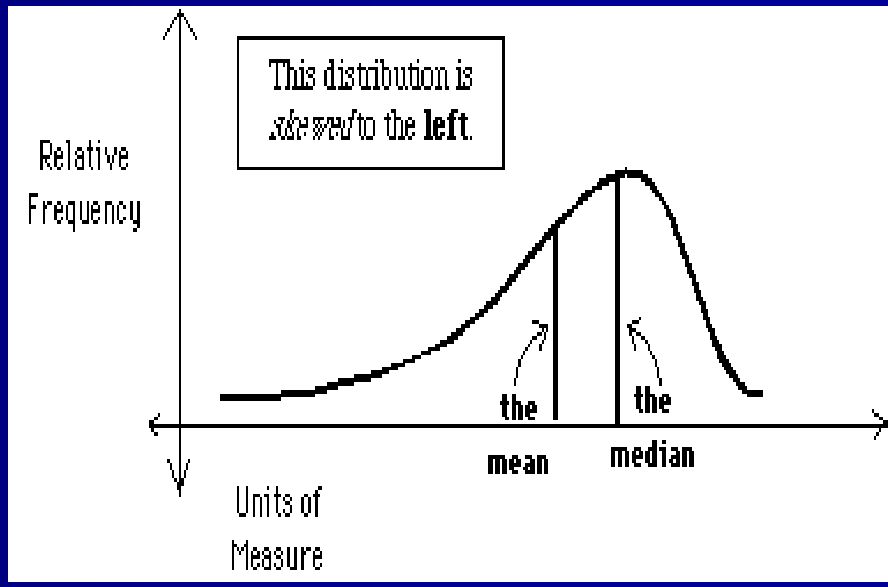
Distributions

- **Skewed to Left**
  - looks like a bell curve w/ a *longer tail on left* & mound pushed to right
  - Most data points fall to right of middle, & there are more very large than very small values



Distributions

- **Skewed to Left**
  - Mean < median
  - *Negatively* skewed
  - large extremes pull mean → the tail
    (extremes are low values)
  - Median remains closer to center of the distribution
  - Ex: Hormone assays



Distributions

- What if you have a skewed distribution?
  - Most statistics assume normality
    - **Fairly robust to violation of assumptions**
    - **But may not get accurate results if very skewed**
  - Data transformations-logs
    - **Pulls in extremes**
    - **Problem-logged values not clinically useful**
    - **Do statistics on logged values & p based on logs, but report unlogged means**
    - **Compare results of stats w/unlogged values**

Descriptive statistics

- Measures of dispersion
  - Describes the sample
  - Shows scatter of scores around mean
    - Distributions
    - Range
    - Variance
    - Standard deviation

Range

- **Range**
  - lowest to highest score/value
  - Use for continuous variables
  - Normally distributed, presenting mean
  - Example: age ranged from 18-60 months
    years in practice ranged from 1-25

Range

- **Interquartile range (IQR)**
  - Use w/ continuous data
  - Skewed data & presenting median
  - Divide sample into quartiles
  - IQR = 75^{th} – 25^{th} quartile
  - Tells where most values are located

Descriptive statistics

- Measures of dispersion
  - Describes the sample
  - Shows scatter of scores around mean
    - Distributions
    - Range
    - Variance
    - Standard deviation

Variance

- Shows dispersion (spread) of data points around mean
- The further away the data points are from the mean, the greater the variance

Variance

- Might think the variance = average difference of each score from the mean, summed together & ÷ by total # data points or $\frac{\Sigma (x - \text{mean})}{N}$

but,

- If normal distribution, then # data pts above mean = # data pts below mean
- averaging the difference of each score from the mean=0

Variance

- Average squared deviation from the mean
- Computational formula:

$$\text{Variance} = \frac{\Sigma (x - \text{mean})^2}{N-1}$$

Where Σ **= sum of; x = each score**
N=sample size or # values

***Note, formula above is for sample variance; to get population variance, use N**

Variance

- Example: **Community research project of teenaged mothers & their knowledge of early childhood caries**
- **12 teen mothers in study group**
- **Give survey to assess their knowledge & score it**

Variance

Mother	Score(%)		$(x-48)^2$	
1	45		9	
2	45	Mean= $\frac{580}{12}$=48.3%	9	
3	45		9	variance=
4	30	Median=45%	324	$\frac{2518}{12-1}$ = 228.9
5	35	mode=45%	169	
6	25		529	
7	40	Range=25 – 70	64	
8	50		4	
9	60	variance=$\frac{\Sigma(x\text{-mean})^2}{N-1}$	144	
10	65		289	
11	70		484	
12	70		484	
	Σ= 580		Σ= 2518	

Descriptive statistics

- Measures of dispersion
  - Describes the sample
  - Shows scatter of scores around mean
    - Distributions
    - Range
    - Variance
    - Standard deviation

Standard Deviation

- Average deviation from the mean, ignoring the sign of the difference
- The further away data points are from the mean, the greater the SD

Standard Deviation

- Computed as sq root of variance =

$$SD = \sqrt{\frac{\sum (x - \text{mean})^2}{N-1}}$$

- For population, use N; for sample, use N-1
- w/ large sample, difference bet N or N-1 is negligible

Standard Deviation

Mother	Score(%)	$(x-48)^2$
1	45	9
2	45	9
3	45	9
4	30	324
5	35	169
6	25	529
7	40	64
8	50	4
9	60	144
10	65	289
11	70	484
12	70	484
	Σ = 580	Σ = 2518

Mean= $\frac{580}{12}$=48.3%

Median=45%

mode=45%

Range=25 – 70

variance=$\frac{\Sigma(x\text{-mean})^2}{N-1}$

SD=sqrt variance

variance= $\frac{2518}{12-1}$ = 228.9

SD=sqrt 228.9 = 15.1

Standard Deviation

- SD useful to compare sets of data w/ the same mean but a different range
- Example: two data sets

  Set A=15, 15, 15, 14, 16

  Mean A = 15

  SD=sqrt 2/4=0.7

  Set B=2, 7, 14, 22, 30

  Mean B=15

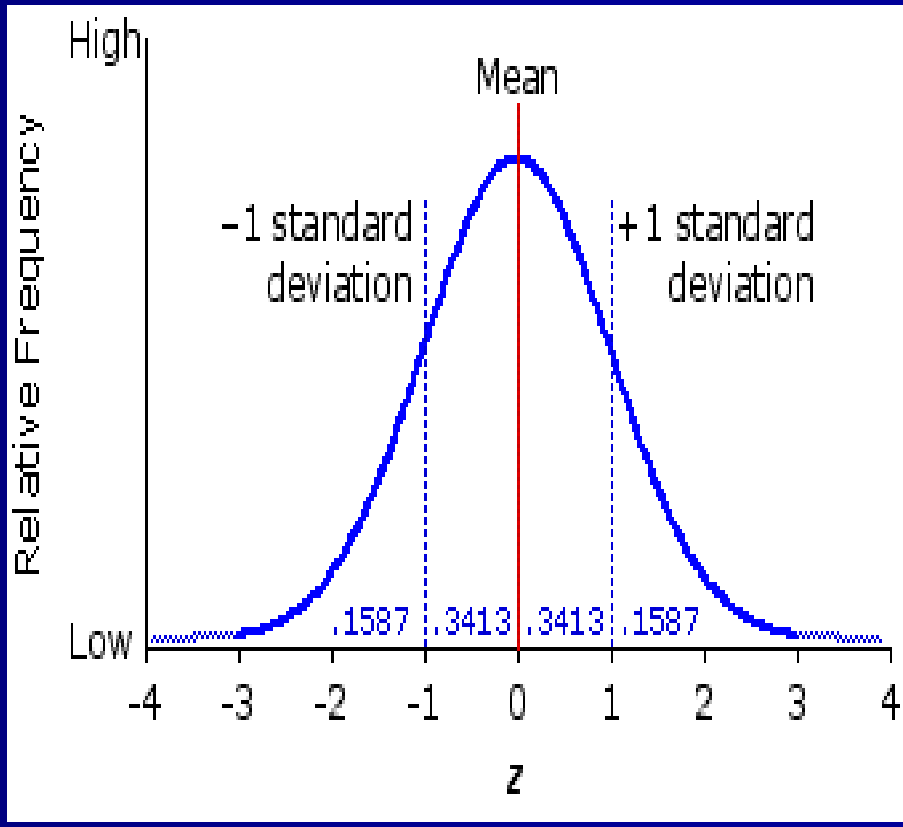
  SD=sqrt 508/4=11.3

  Set B-more spread out
- Low SD= values are not spread

  High SD= values very spread out

Standard Deviation

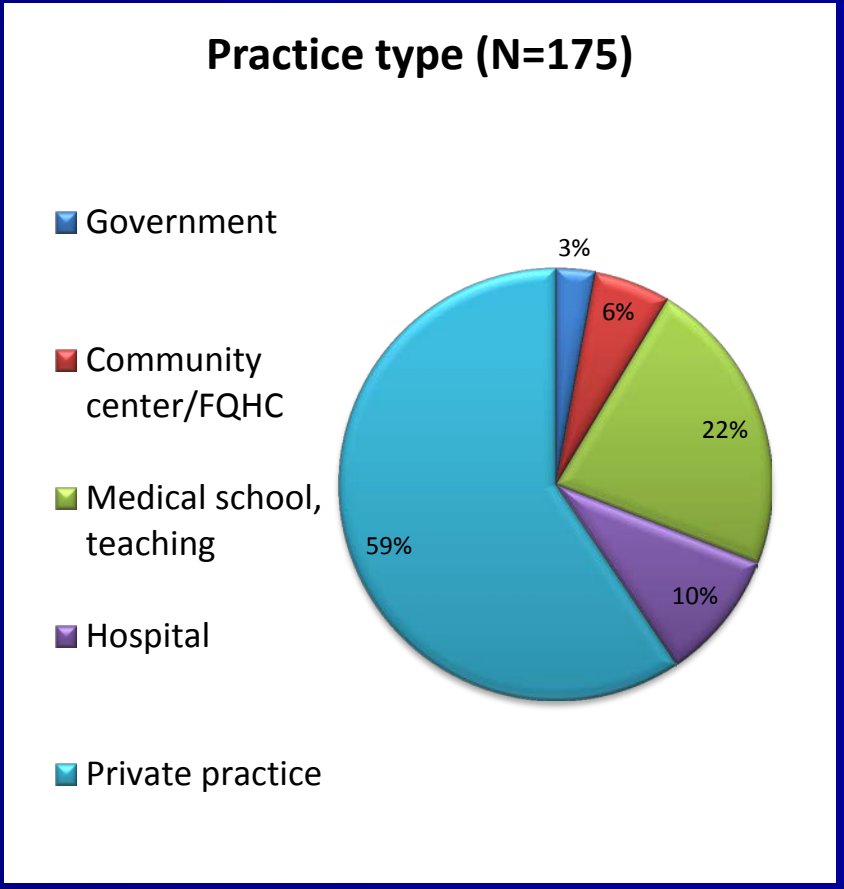
- Normal Distribution
  - **68% within ±1 SD of the mean**
  - **95% within ±2 SD of the mean**
  - **99% within ±3 SD of the mean**
- Skewed Distribution
  - **Eliminate scores >3 SD above or below mean**



Data Display for Categorical data

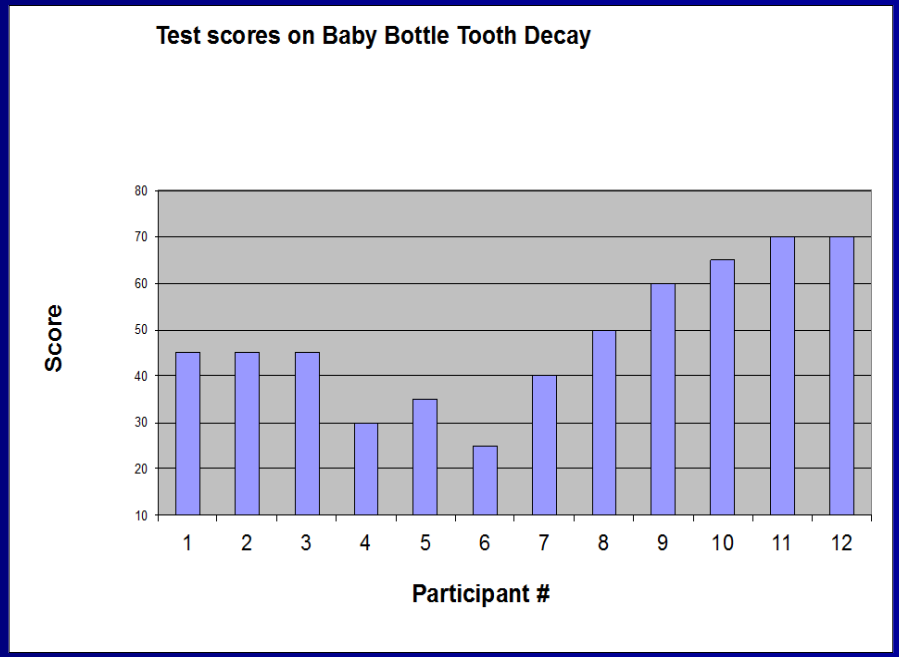
- **Pie chart--**

shows rates for different categories of a nominal variable



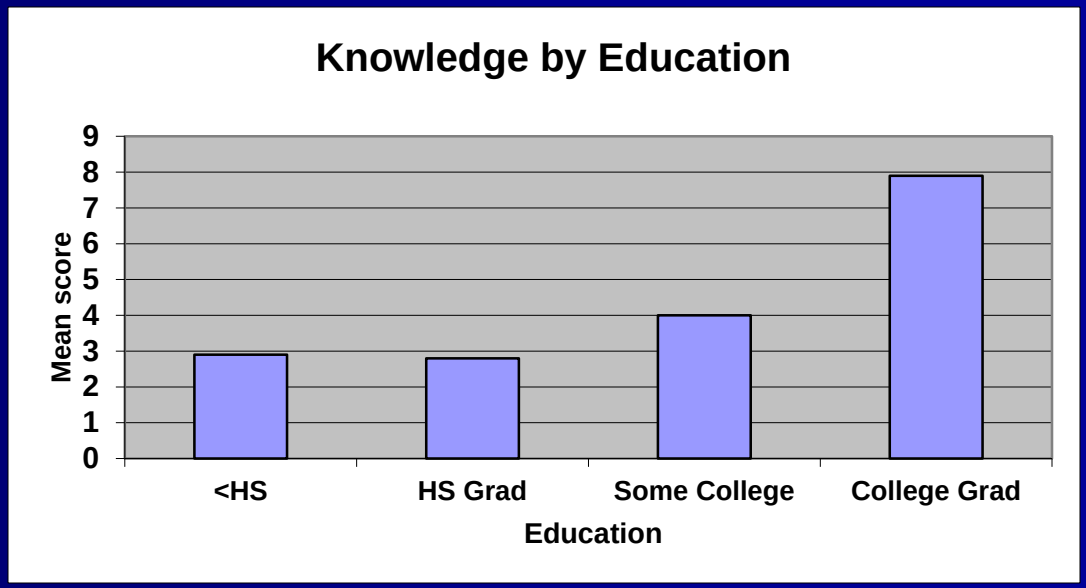
Data Display for Descriptive Statistics

- **Bar graphs**
  - Used to display nominal or ordinal data that are discrete in nature
  - Display can be horizontal



Data Display for Descriptive Statistics

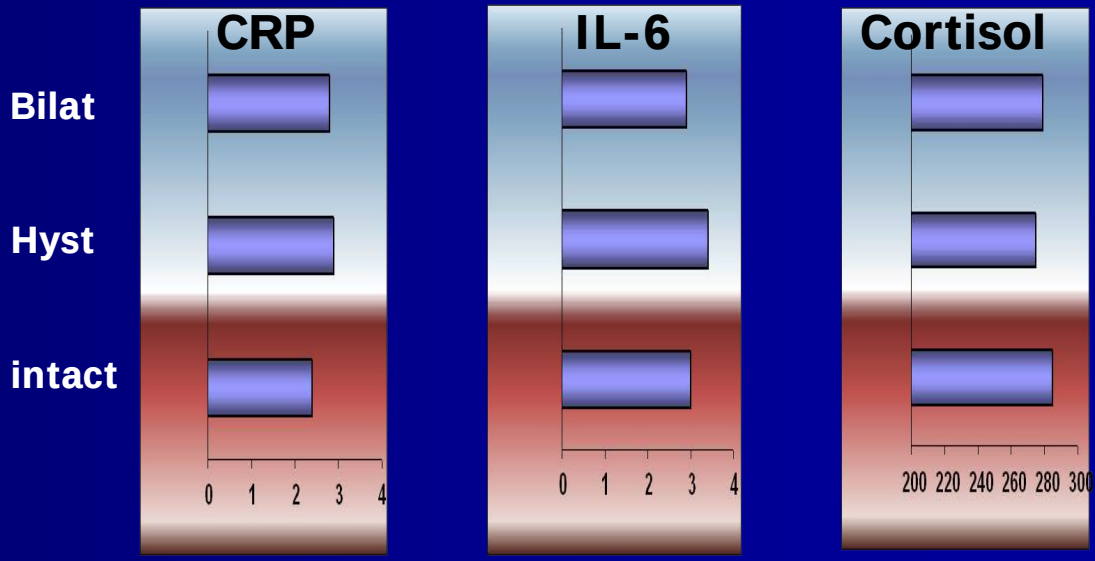
- **Bar graphs**
  - Can show comparisons of means of different groups



Comparisons of oral health knowledge scores between groups
Based on educational level

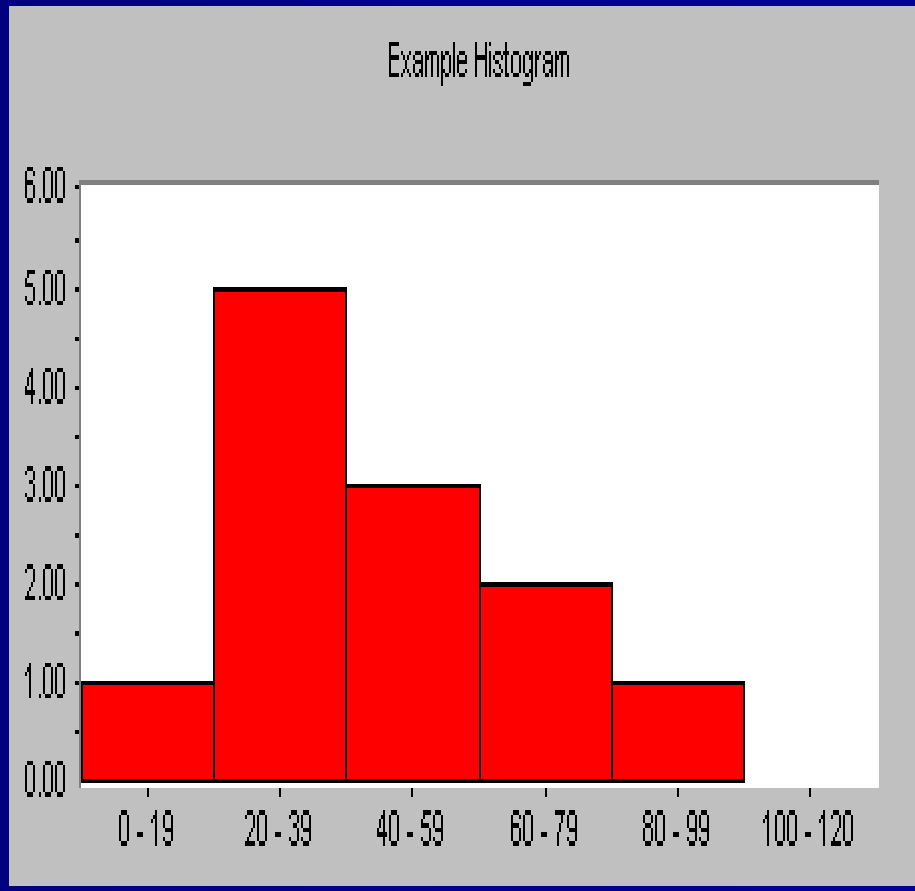
Data Display for Descriptive Statistics

- Bar graphs
  - Data display can be vertical



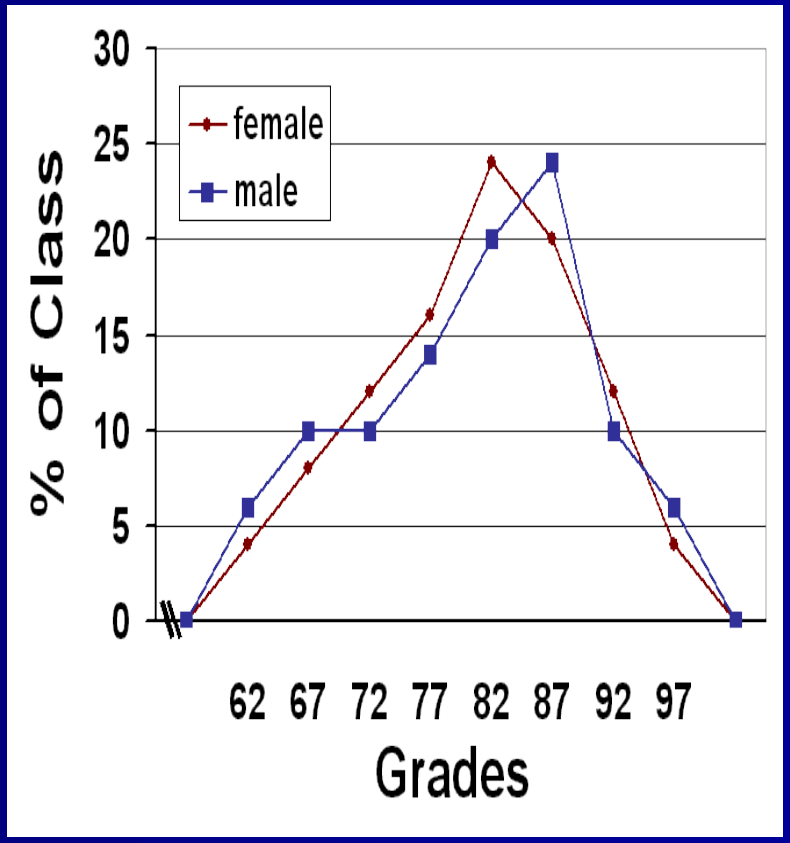
Data Display for Descriptive Statistics

- **Histogram**
  - Used to display interval or ratio scaled variables that are continuous
  - Bars have = width and touch each other indicating data are on a continuum

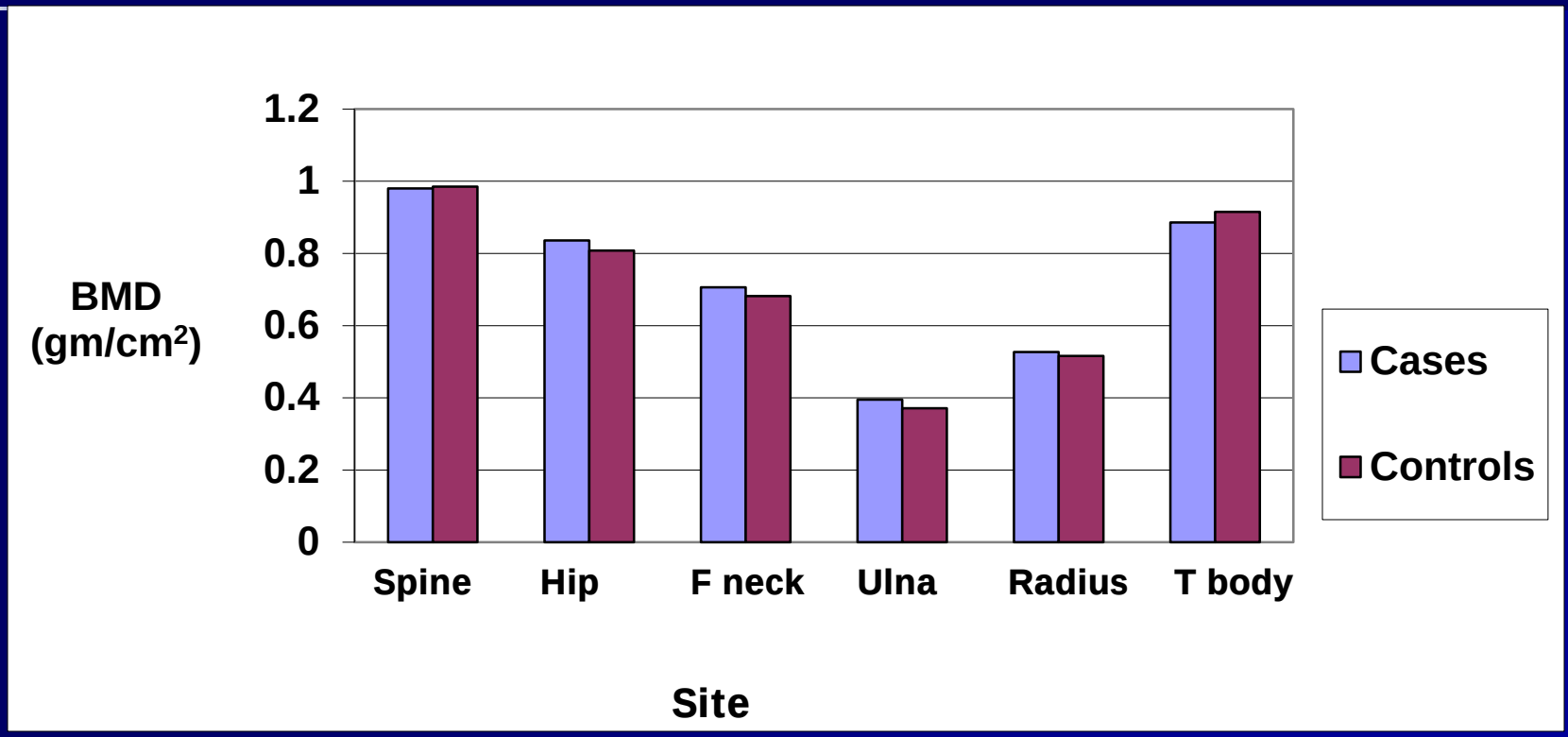


Data Display for Descriptive Statistics

- Frequency polygon
  - Used to display interval or ratio scaled variables that are continuous in nature
  - Shorthand way to present a histogram; use instead of histogram
  - Dots are put in the center of the top of each bar and connected



Data Display for Descriptive Statistics



- Histograms (bar graphs) show comparisons between groups (cases w/Br CA vs. controls) on means of multiple continuous variables (BMD at various sites)

Questions????

Thank You!