



An Introduction to Statistics



Introduction to Statistics

I. What are Statistics?

- Procedures for organizing, summarizing, and interpreting information
- Standardized techniques used by scientists
- Vocabulary & symbols for communicating about data
- A tool box
 - How do you know which tool to use?
 - (1) What do you want to know?
 - (2) What type of data do you have?
 - Two main branches:
 - *Descriptive statistics*
 - *Inferential statistics*

Two Branches of Statistical Methods

- Descriptive statistics
 - Techniques for describing data in abbreviated, symbolic fashion
- Inferential statistics
 - Drawing inferences based on data. Using statistics to draw conclusions about the population from which the sample was taken.

Descriptive vs Inferential

A. Descriptive Statistics:

Tools for summarizing, organizing, simplifying data

Tables & Graphs

Measures of Central Tendency

Measures of Variability

Examples:

Average rainfall in Richmond last year

Number of car thefts in IV last quarter

Your college G.P.A.

Percentage of seniors in our class

B. Inferential Statistics:

Data from *sample* used to draw inferences about *population*

Generalizing beyond actual observations

Generalize from a sample to a population

Populations and Samples

- A parameter is a characteristic of a population
 - e.g., the *average* height of all Americans.
- A statistics is a characteristic of a sample
 - e.g., the *average* height of a sample of Americans.
- Inferential statistics infer population parameters from sample statistics
 - e.g., we use the average height of the sample to **estimate** the average height of the population

Symbols and Terminology:

Parameters = Describe POPULATIONS

Greek letters → μ σ^2 σ ρ

Statistics = Describe SAMPLES

English letters → \bar{X} s^2 s r

Sample will not be identical to the population

So, generalizations will have some error

Sampling Error = discrepancy between sample *statistic* and corresponding popl'n *parameter*

Statistics are Greek to me!

Statistical notation:

X = “score” or “raw score”

N = number of scores in population

n = number of scores in sample

Quiz scores for 5 Students:

X
4
10
6
2
8

X = Quiz score for each student

Statistics are Greek to me!

X = Quiz score for each student

Y = Number of hours studying

Summation notation:

“Sigma” = Σ

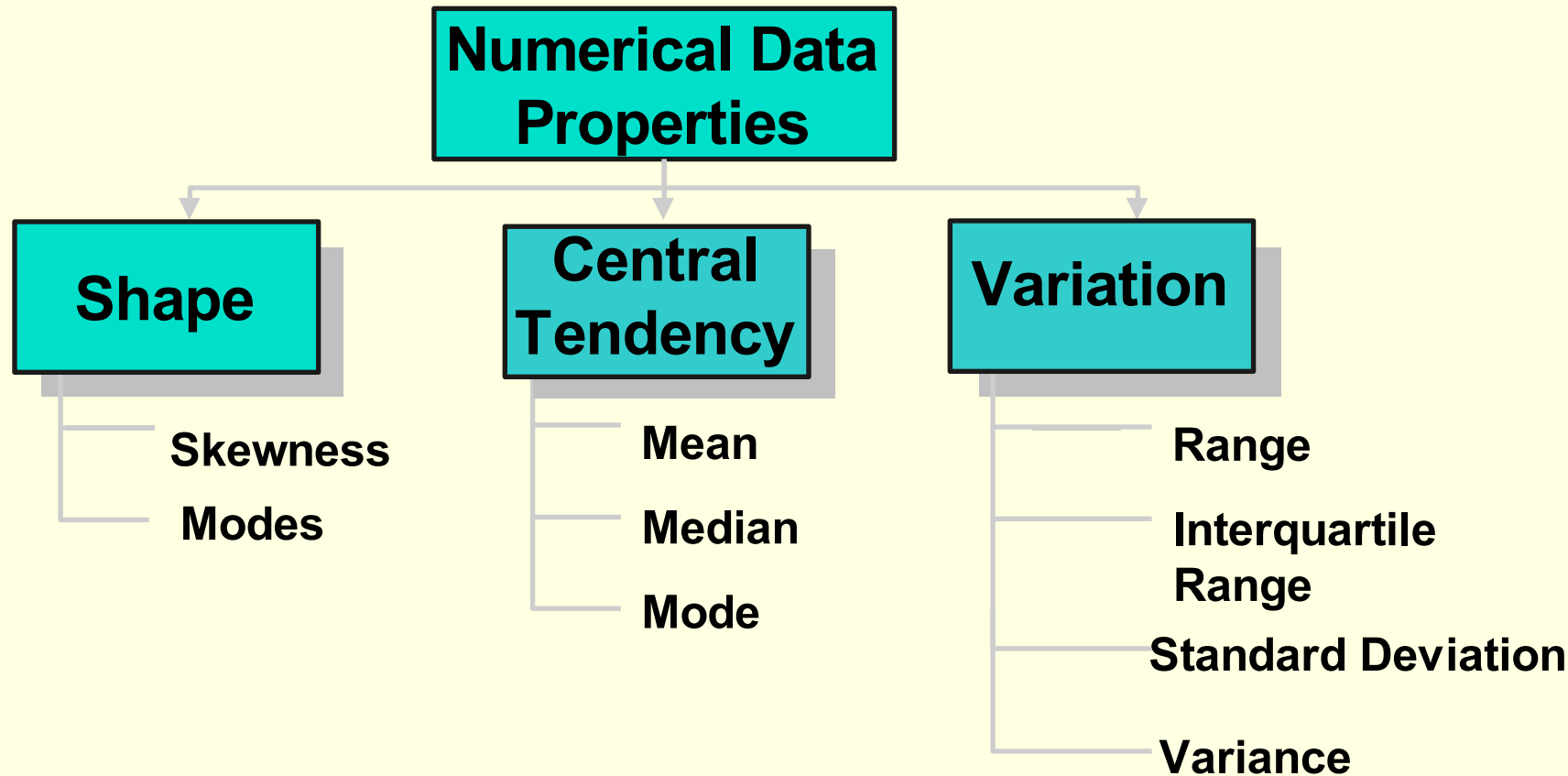
“The Sum of”

ΣX = add up all the X scores

ΣXY = multiply $X \times Y$ then add

X	Y
4	2
10	5
6	2
2	1
8	3

Descriptive Statistics



Ordering the Data: Frequency Tables

- Three types of frequency distributions (FDs):
 - (A) Simple FDs
 - (B) Relative FDs
 - (C) Cumulative FDs

- Why Frequency Tables?
 - Gives some order to a set of data
 - Can examine data for outliers
 - Is an introduction to **distributions**

A. Simple Frequency Distributions

QUIZ SCORES ($N = 30$)

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

Simple Frequency Distribution of Quiz Scores (X)

X	f
10	
9	
8	
7	4
6	5
5	5
4	4
3	3
2	2
1	1

$$\Sigma f = N = 30$$

Relative Frequency Distribution

Quiz Scores

X	<i>f</i>	<i>p</i>	%
10	1		
9	2		
8	3		
7	4	.13	13%
6	5	.17	17%
5	5	.17	17%
4	4	.13	13%
3	3	.10	10%
2	2	.07	7%
1	1	.03	3%

$\Sigma f = N = 30$ $\Sigma =$ $\Sigma =$

Cumulative Frequency Distribution

Quiz Score	<i>f</i>	<i>p</i>	%		<i>cf</i>	<i>c%</i>
10	1	.03	3%	30	100%	
9	2	.07	7%	29	97	
8	3	.10	10%	27	90	
7	4	.13	13%	24	80	
6	5	.17	17%	20	67	
5	5	.17	17%	15	50	
4	4	.13	13%	10	33	
3	3	.10	10%			
2	2	.07	7%			
1	1	.03	3%			

$\Sigma = 30$ $\Sigma = 1.0$ $\Sigma = 100\%$

Grouped Frequency Tables

Assign f s to *intervals*

Example: Weight for 194 people

Smallest = 93 lbs

Largest = 265 lbs

X (Weight)	f
255 - 269	1
240 - 254	4
225 - 239	2
210 - 224	6
195 - 209	3
180 - 194	10
165 - 179	24
150 - 164	31
135 - 149	27
120 - 134	55
105 - 119	24
90 - 104	7

$$\Sigma f = N = 194$$

Graphs of Frequency Distributions

- A picture is worth a thousand words!
- Graphs for numerical data:
 - Stem & leaf displays
 - Histograms
 - Frequency polygons
- Graphs for categorical data
 - Bar graphs

Making a Stem-and-Leaf Plot

- Cross between a table and a graph
- Like a grouped frequency distribution on its side
- Easy to construct
- Identifies each individual score
- Each data point is broken down into a “**stem**” and a “**leaf**.” Select one or more leading digits for the stem values. The trailing digit(s) becomes the leaves
- First, “stems” are aligned in a column.
- Record the leaf for every observation beside the corresponding stem value

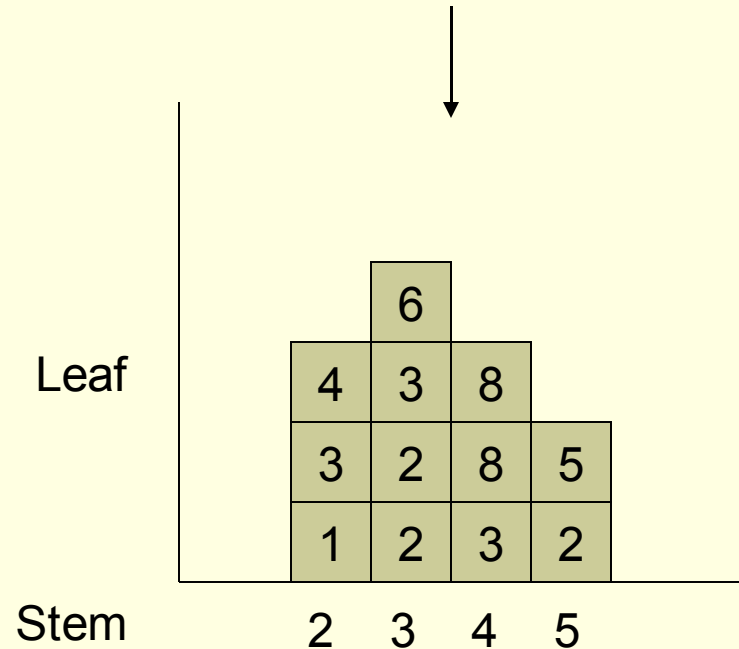
Stem and Leaf Display

DATA			STEM AND LEAF DISPLAY	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	6279
76	52	97	6	283
85	42	46	7	1643846
32	57	59	8	3521
56	73	74	9	37
74	81	76		

Stem and Leaf / Histogram

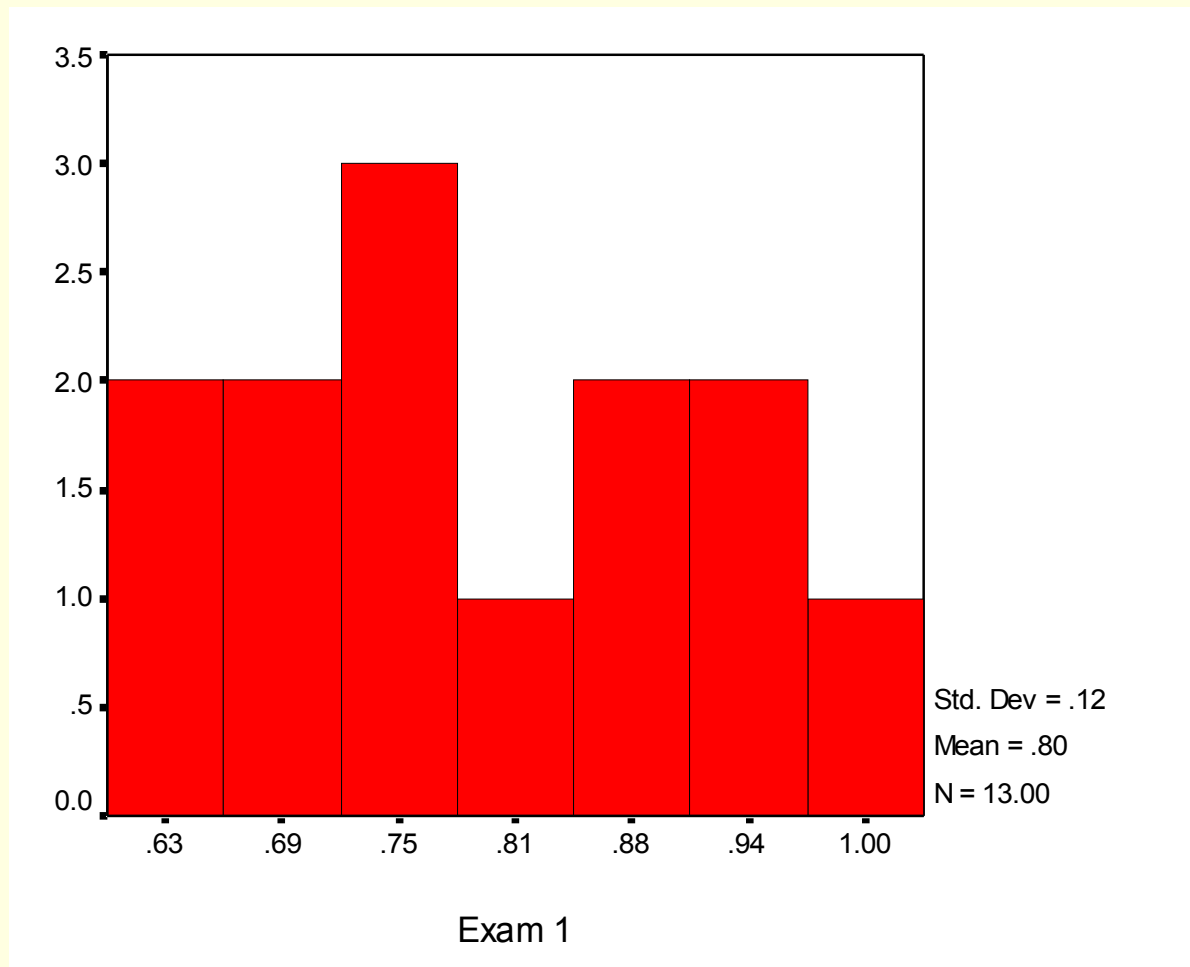
Stem	Leaf
2	1 3 4
3	2 2 3 6
4	3 8 8
5	2 5

By rotating the stem-leaf, we can see the shape of the distribution of scores.

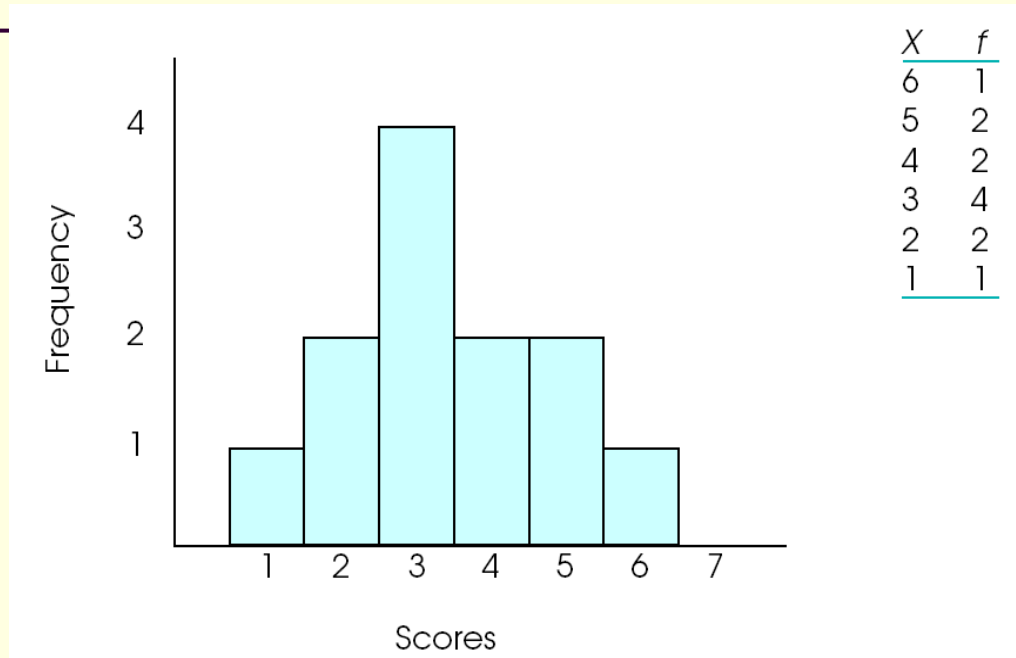


Histograms

■ Histograms



Histograms

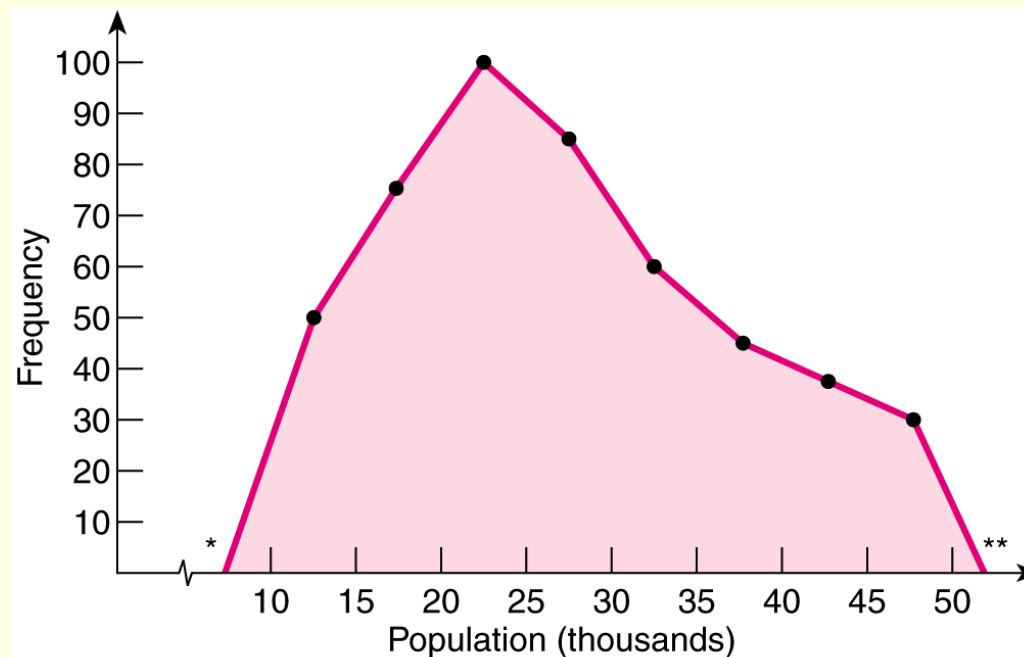


- f on y axis (could also plot p or %)
- X values (or midpoints of class intervals) on x axis
- Plot each f with a bar, equal size, touching
- ***No gaps between bars***

Frequency Polygons

■ Frequency Polygons

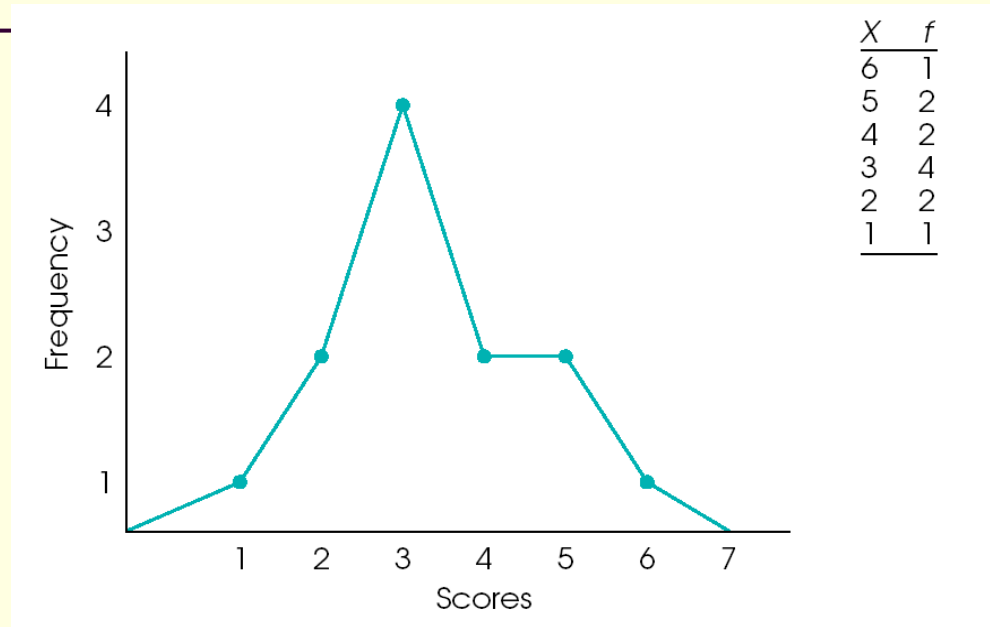
- Depicts information from a frequency table or a grouped frequency table as a **line graph**



* 4 cities had populations of less than 10,000.

** 5 cities had populations of 50,000 or greater.

Frequency Polygon



A smoothed out histogram

Make a point representing f of each value

Connect dots

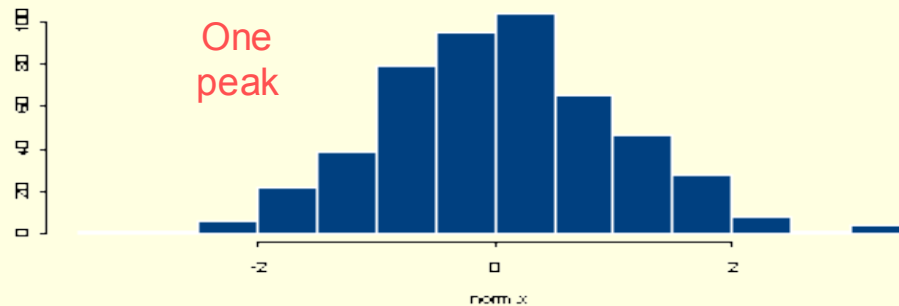
Anchor line on x axis

Useful for comparing distributions in two samples (in this case, plot p rather than f)

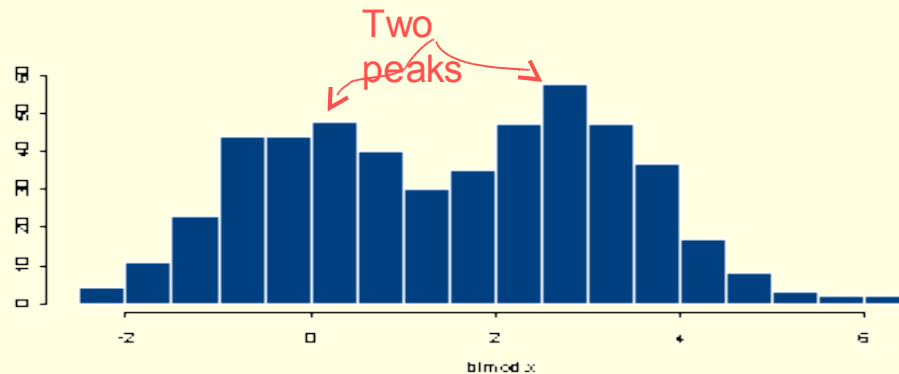
Shapes of Frequency Distributions

- Frequency tables, histograms & polygons describe how the frequencies are distributed
- Distributions are a fundamental concept in statistics

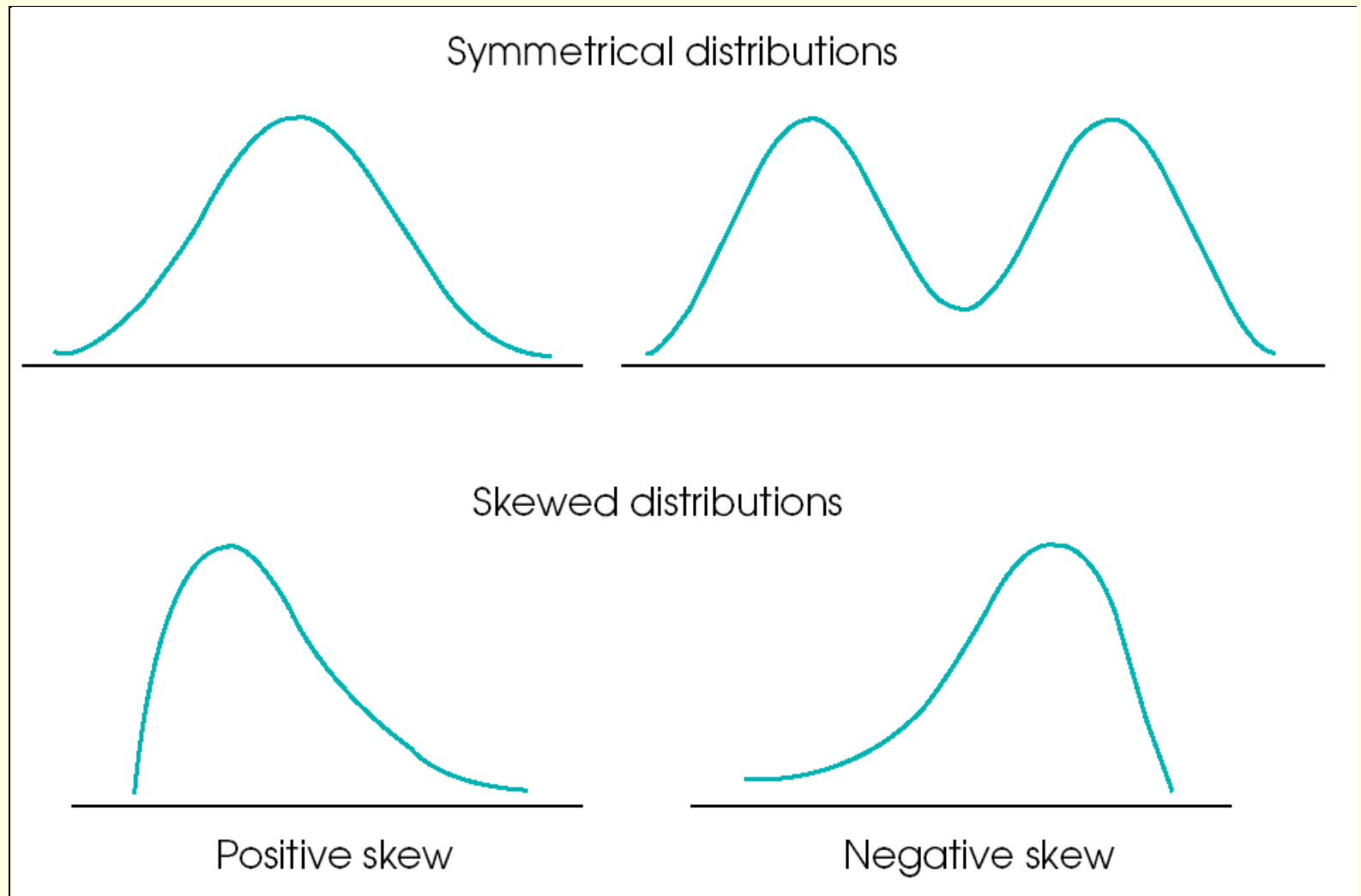
Unimodal



Bimodal



Typical Shapes of Frequency Distributions



Normal and Bimodal Distributions

(1) Normal Shaped Distribution

- Bell-shaped
- One peak in the middle (unimodal)
- Symmetrical on each side
- Reflect many naturally occurring variables

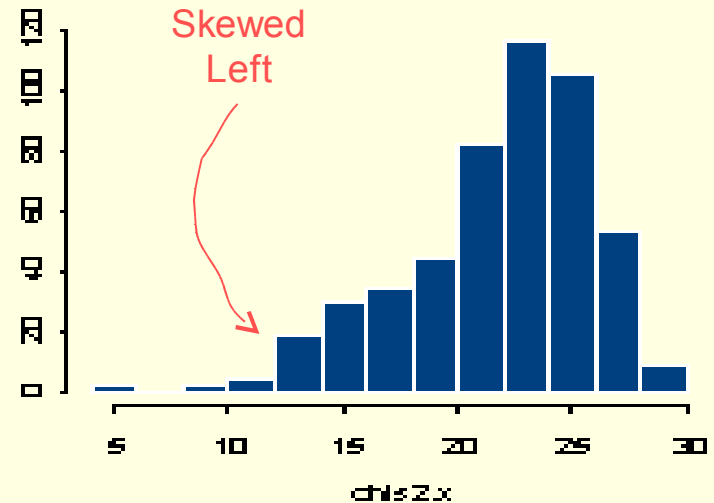
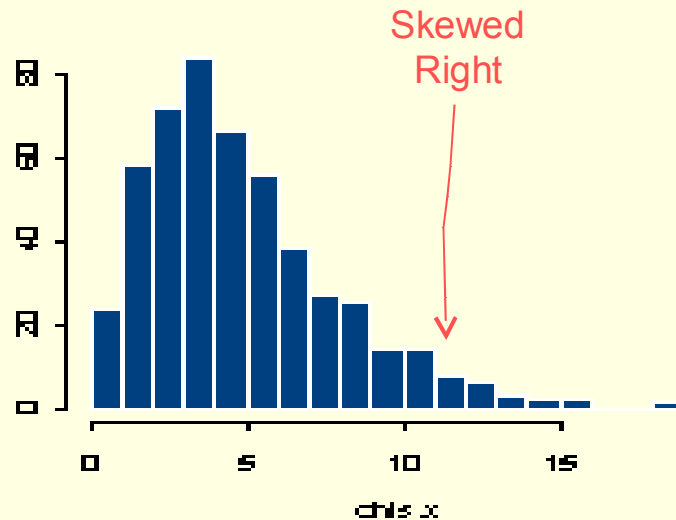
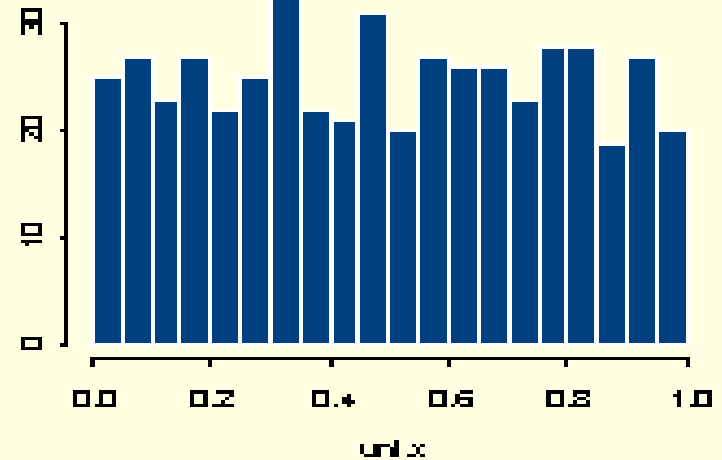
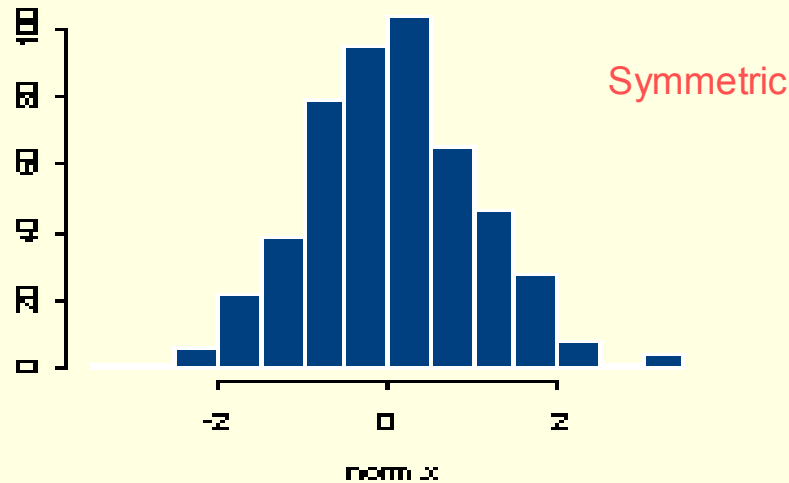
(2) Bimodal Distribution

- Two clear peaks
- Symmetrical on each side
- Often indicates two distinct subgroups in sample

Symmetrical vs. Skewed Frequency Distributions

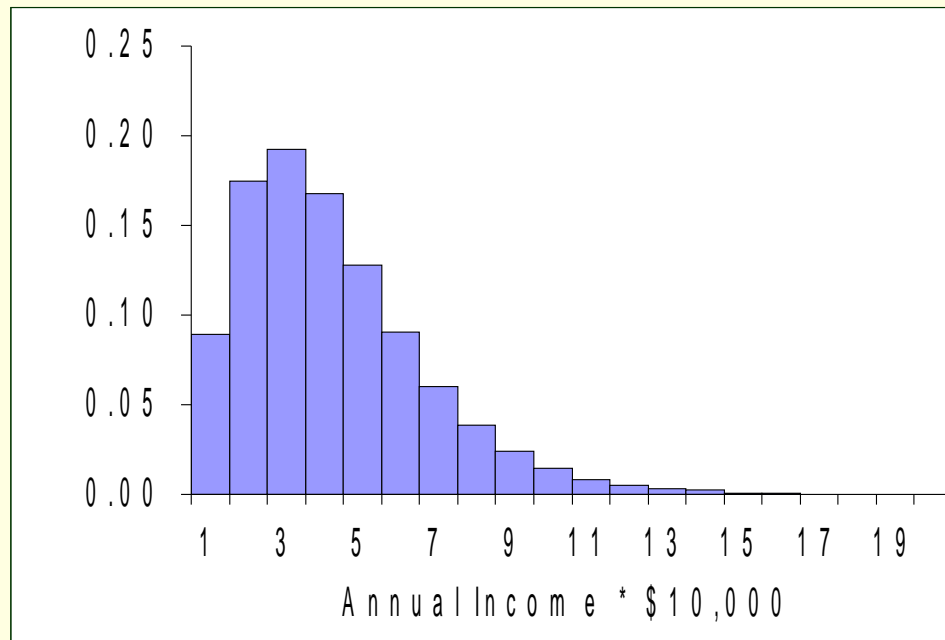
- Symmetrical distribution
 - Approximately equal numbers of observations above and below the middle
- Skewed distribution
 - One side is more spread out than the other, like a tail
 - Direction of the skew
 - Positive or negative (right or left)
 - Side with the fewer scores
 - Side that looks like a tail

Symmetrical vs. Skewed



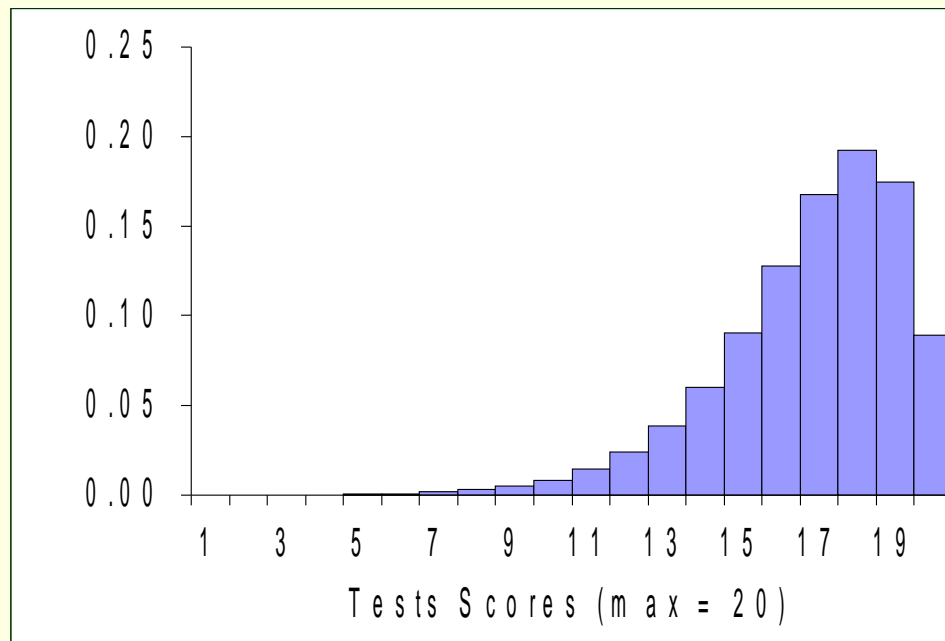
Skewed Frequency Distributions

- Positively skewed
 - AKA Skewed right
 - Tail trails to the right
 - ***** *The skew describes the skinny end* *****



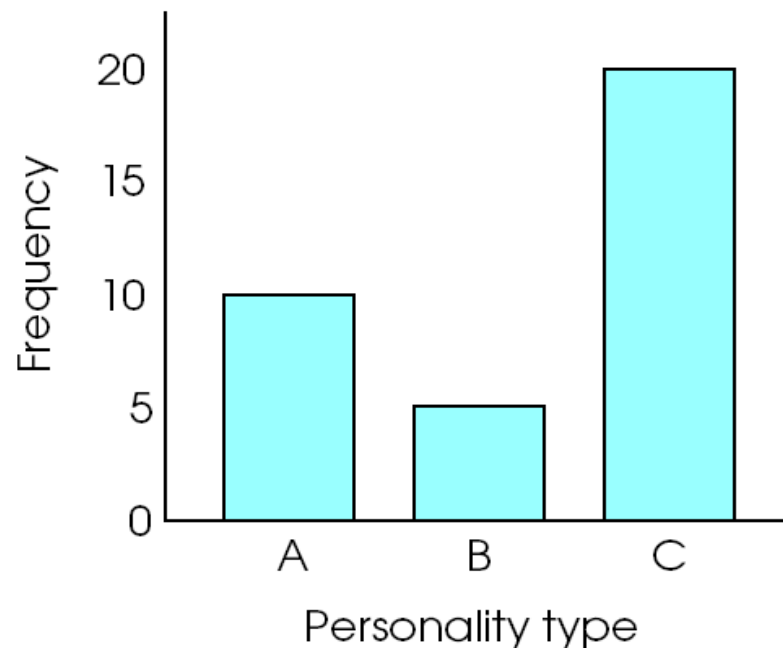
Skewed Frequency Distributions

- Negatively skewed
 - Skewed left
 - Tail trails to the left



Bar Graphs

- For categorical data
- Like a histogram, but with gaps between bars
- Useful for showing two samples side-by-side



Central Tendency

- Give information concerning the average or typical score of a number of scores
 - mean
 - median
 - mode

Central Tendency: The Mean

- The Mean is a measure of *central tendency*
 - What most people mean by “average”
 - Sum of a set of numbers divided by the number of numbers in the set

$$\frac{\cancel{1} \cancel{2} \cancel{3} \cancel{4} \cancel{5} \cancel{6} \cancel{7} \cancel{8} \cancel{9} \cancel{10}}{10} = \frac{55}{10} = 5.5$$

Central Tendency: The Mean

Arithmetic average:

Sample

$$\bar{X} = \frac{\sum x}{n}$$

Population

$$\mu = \frac{\sum x}{N}$$

~~1~~ ~~2~~ ~~3~~ ~~4~~ ~~5~~ ~~6~~ ~~7~~ ~~8~~ ~~9~~

$$\sum X / n = 5.5$$

Example

Student	(X) Quiz Score
Bill	5
John	4
Mary	6
Alice	5

$$\Sigma X =$$

$$n =$$

$$\bar{X} =$$

Central Tendency: The Mean

- Important conceptual point:
- The mean is *the balance point* of the data in the sense that if we took each individual score (X) and subtracted the mean from them, some are positive and some are negative. If we add all of those up we will get zero.



$$\sum (X - \bar{X}) = 0$$

Central Tendency: The Median

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves
- If *odd n*, middle value of sequence
 - if $X = [1, 2, 4, 6, 9, 10, 12, 14, 17]$
 - then 9 is the median
- If *even n*, average of 2 middle values
 - if $X = [1, 2, 4, 6, 9, 10, 11, 12, 14, 17]$
 - then 9.5 is the median; i.e., $(9+10)/2$
- Median is not affected by extreme values

Median vs. Mean

- Midpoint vs. balance point
- Md based on middle location/# of scores
- based on deviations/distance/balance
- Change a score, Md may not change
- Change a score, \bar{x} will always change

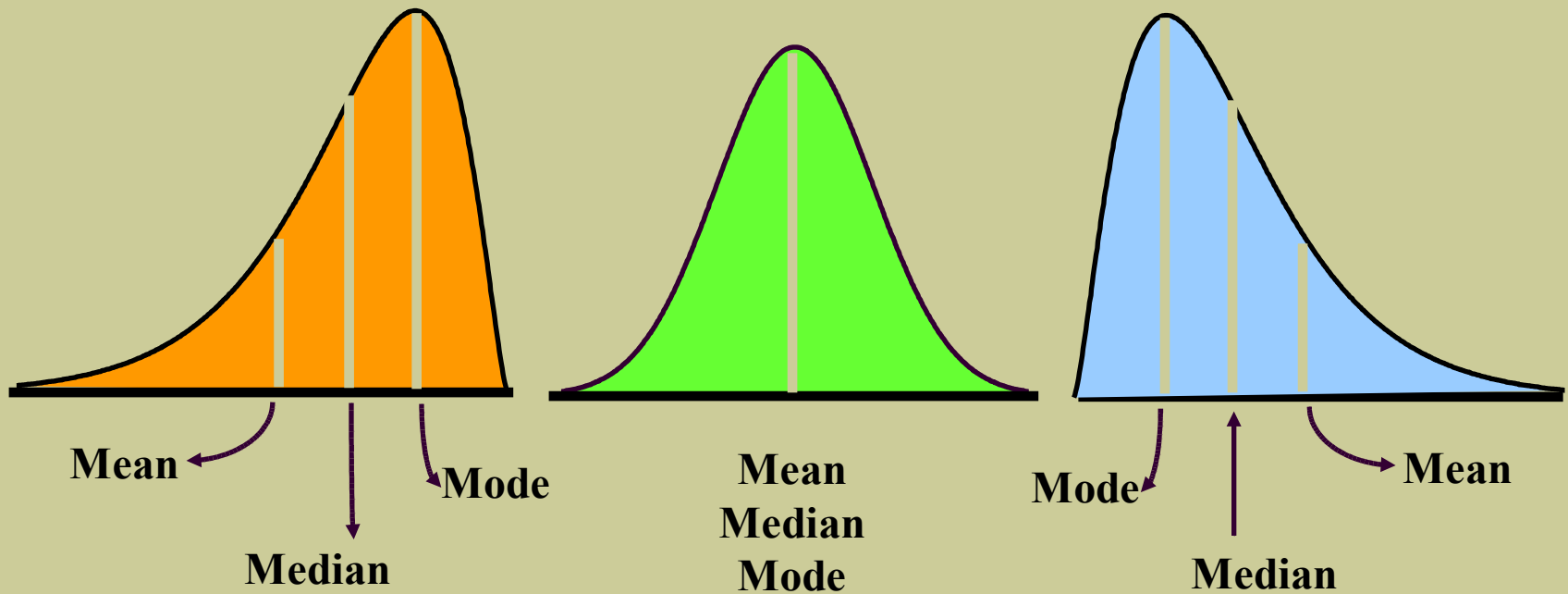
Central Tendency: The Mode

- The mode is the most frequently occurring number in a distribution
 - if $X = [1, 2, 4, 7, 7, 7, 8, 10, 12, 14, 17]$
 - then 7 is the mode
- Easy to see in a simple frequency distribution
- Possible to have no modes or more than one mode
 - *bimodal* and *multimodal*
- Don't have to be exactly equal frequency
 - *major mode, minor mode*
- Mode is not affected by extreme values

When to Use What

- Mean is a great measure. But, there are times when its usage is inappropriate or impossible.
 - Nominal data: Mode
 - The distribution is bimodal: Mode
 - You have ordinal data: Median or mode
 - Are a few extreme scores: Median

Mean, Median, Mode



**Negatively
Skewed**

**Symmetric
(Not Skewed)**

**Positively
Skewed**

Measures of Central Tendency

Overview

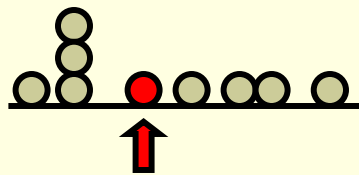
Central Tendency

Mean

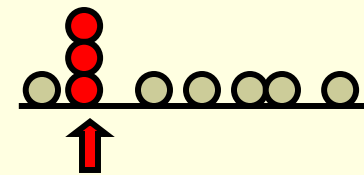
Median

Mode

$$\bar{X} = \frac{\sum X}{N}$$



Midpoint of ranked values



Most frequently observed value

Class Activity

- Complete the questionnaires
- As a group, analyze the classes data from the three questions you are assigned
 - compute the appropriate measures of central tendency for each of the questions
 - Create a frequency distribution graph for the data from each question

Variability

- Variability
 - How tightly clustered or how widely *dispersed* the values are in a data set.
- Example
 - Data set 1: [0,25,50,75,100]
 - Data set 2: [48,49,50,51,52]
 - Both have a mean of 50, but data set 1 clearly has greater *Variability* than data set 2.

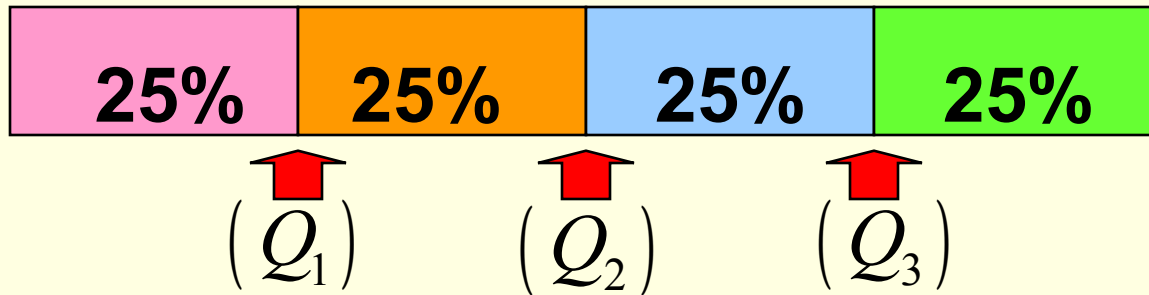
Variability: The Range

- The *Range* is one measure of variability
 - The range is the difference between the maximum and minimum values in a set
- Example
 - Data set 1: [1,25,50,75,100]; R: $100 - 1 + 1 = 100$
 - Data set 2: [48,49,50,51,52]; R: $52 - 48 + 1 = 5$
 - *The range ignores how data are distributed and only takes the extreme scores into account*

$$\text{■ } \text{RANGE} = (X_{\text{largest}} - X_{\text{smallest}}) + 1$$

Quartiles

- Split Ordered Data into 4 Quarters



- Q_1 = first quartile

- Q_2 = second quartile = Median

- Q_3 = third quartile

Variability: Interquartile Range

- Difference between third & first quartiles
 - Interquartile Range = $Q_3 - Q_1$
- Spread in middle 50%
- Not affected by extreme values

Standard Deviation and Variance

- How much do scores deviate from the mean?

- **deviation** = $X - \mu$

X	X- μ
1	
0	
6	
1	

$$\mu = 2 \quad \Sigma (X - \mu) =$$

- Why not just add these all up and take the mean?

Standard Deviation and Variance

■ Solve the problem by squaring the deviations!

X	X- μ	(X- μ) ²
1	-1	1
0	-2	4
6	+4	16
1	-1	1

$$\mu = 2$$

Variance = $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$

Standard Deviation and Variance

- Higher value means greater variability around μ
- Critical for inferential statistics!
- But, not as useful as a purely descriptive statistic
 - *hard to interpret “squared” scores!*
- Solution → un-square the variance!

Standard Deviation =
$$\sigma = \sqrt{\frac{\sum (X - u)^2}{N}}$$

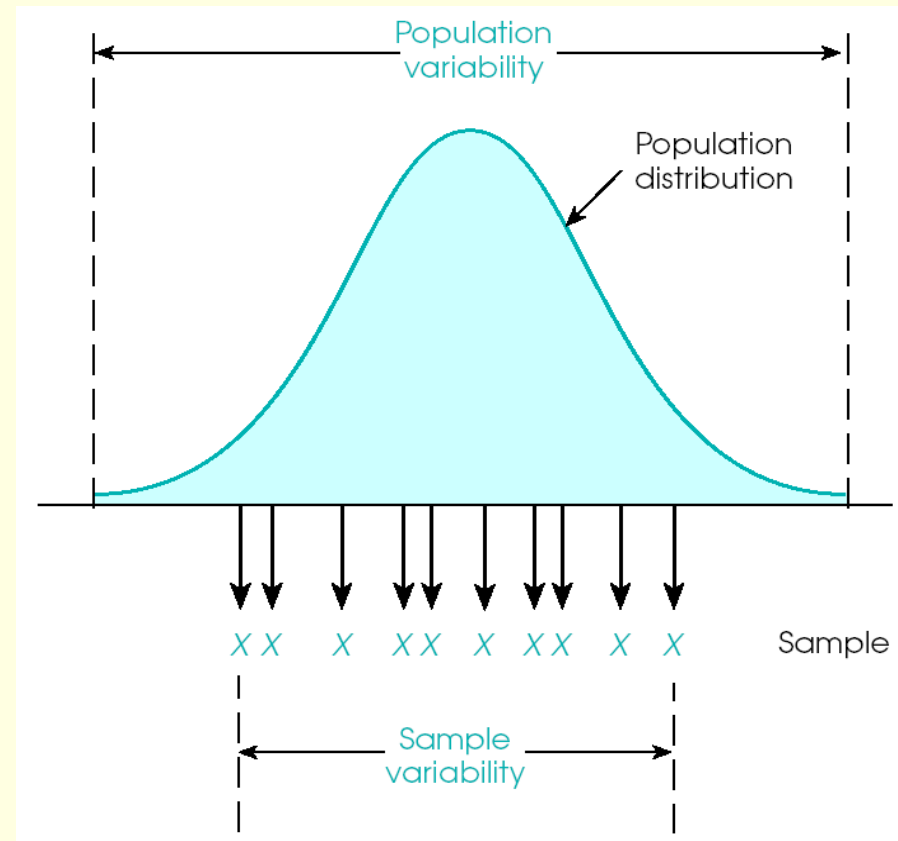
Variability: Standard Deviation

- The Standard Deviation tells us approximately how far the scores vary from the mean on average
- estimate of average deviation/distance from μ
- small value means scores clustered close to μ
- large value means scores spread farther from μ
- Overall, most common and important measure
- extremely useful as a descriptive statistic
- extremely useful in inferential statistics

The typical deviation in a given distribution

Sample variance and standard deviation

- Sample will tend to have less variability than popl'n
- if we use the population formula, our sample statistic will be ***biased***
- will tend to ***underestimate*** popl'n variance



Sample variance and standard deviation

- Correct for problem by adjusting formula

$$s^2 = \frac{\sum (X - M)^2}{n - 1}$$

- *Different symbol:* s^2 vs. σ^2
- *Different denominator:* $n-1$ vs. N
- $n-1 =$ “**degrees of freedom**”
- Everything else is the same
- Interpretation is the same

Definitional Formula:

Variance: $s^2 = \frac{\sum (X - \bar{X})^2}{n - 1} = \frac{SS}{n - 1} = \frac{SS}{df}$

- *deviation*
- *squared-deviation*
- **'Sum of Squares' = SS**
- *degrees of freedom*

Standard Deviation: $s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{SS}{n - 1}} = \sqrt{\frac{SS}{df}}$

Variability: Standard Deviation

- let $X = [3, 4, 5, 6, 7]$

- $\bar{X} = 5$

- $(X - \bar{X}) = [-2, -1, 0, 1, 2]$

↑ subtract \bar{x} from each number in X

- $(X - \bar{X})^2 = [4, 1, 0, 1, 4]$

↑ squared deviations from the mean

- $\Sigma (X - \bar{X})^2 = 10$

↑ sum of squared deviations from the mean (SS)

- $\Sigma (X - \bar{X})^2 / n - 1 = 10 / 5 = 2.5$

↑ average squared deviation from the mean

- $\sqrt{\Sigma (X - \bar{X})^2 / n - 1} = \sqrt{2.5} = 1.58$

↑ square root of averaged squared deviation

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$