Introductory Concepts:

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information.

Descriptive Stat: Involves collecting, presenting and characterizing data.

Inferential Stat: Involves using sample data to make generalizations about population (involves estimation and hypothesis testing).

Fundamental elements of statistics:

1. Experimental unit: object upon which we collect data
2. Population: all items of interest
3. Variable: characteristic of an individual experimental unit
4. Sample: subset of the units of a population

Example: Problem According to Variety (Aug. 10, 2010), the average age of viewers of television programs broadcast on CBS, NBC, and ABC is 51 years. Suppose a rival network (e.g., FOX) executive hypothesizes that the average age of FOX viewers is less than 51. To test her hypothesis, she samples 200 FOX viewers and determines the age of each.

a. Describe the population.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference.

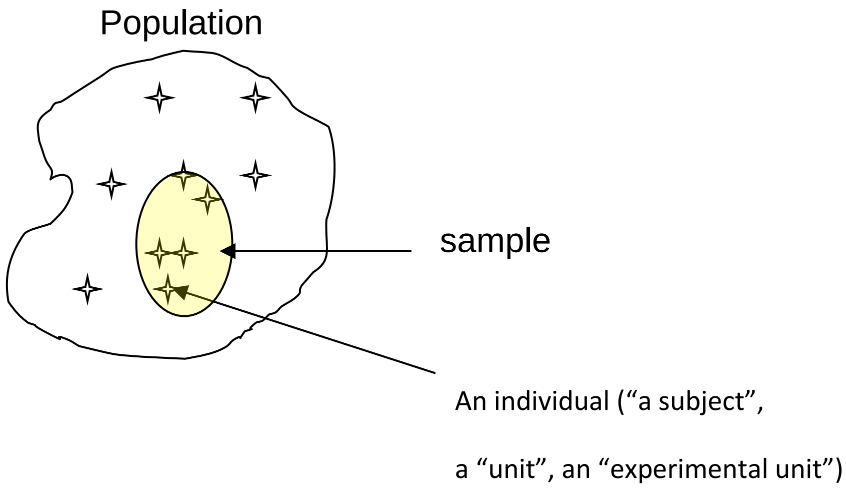
Data

Two Types of Data

- Qualitative - Categorical (Nominal):
- Quantitative - Measurable or Countable:

Examples…

Illustration: population vs sample


Population
sample
An individual (“a subject”,
a “unit”, an “experimental unit”)

Qualitative vs. Quantitative data:

Quantitative (numerical) data are measurements that can be placed on number line (age, height, time until next storm, unemployment rate, GPA, number of siblings etc.)

Qualitative (categorical) data cannot be measured using numbers. If numbers are present they only serve as labels (Student ID etc.) Qualitative data can be grouped into categories (political affiliation, ranking the movies, classifying the products as “good”, “fair”, “bad” etc.)

Examples:
Qualitative (categorical): Color, gender, name, PIN, phone number, etc.
Quantitative (numerical): Temperatures, salaries, exam scores (points) etc.

We use samples to make inferences about population.

Representative sample: sample has the characteristics of the population.
An n-elements Single Random Sample (a sample where every n-element subset of population has the same chance to be selected) is an example of a representative sample.

If a sample is not representative it is called biased and is useless.

Sampling methods:

Simple Random Sample (best): every possible sample size n has the same chance to be selected from the population
We use random sample generator to collect truly random samples.

(Class exercise: select a digit…)

Other sampling methods:

- Systematic
- Stratifying
- Cluster

Incorrect methods:

- Convenience sampling
- Voluntary sampling

Statistical biases:

- **Sampling, or selection bias** (a subset of the experimental units in the population is excluded so that these units have no chance of being selected for the sample.)
- **Measurement error** (inaccuracies in the values of the data recorded. In surveys, the error may be due to ambiguous or leading questions and the interviewer's effect on the respondent.)
- **Nonresponse** (the researchers conducting a survey or study are unable to obtain data on all experimental units selected for the sample.)

A **process** is a series of actions or operations that transforms inputs to outputs. A process produces or generates output over time.

Parameter: a numerical descriptive measure of a **population**. Often unknown. (Remember: P and P)

Statistic: a numerical descriptive measure of a sample. It is calculated from the observations in the sample. (Remember: S and S)

Misleading Statistics: Examples

A popular television program reported on several misleading (and possibly unethical) surveys in a "Fact or Fiction?" segment. The basic results from four of these studies are presented below.

a. Eating oat bran is a cheap and easy way to reduce cholesterol count. (Fact: Diet must consist of nothing but oat bran to achieve a slightly lower cholesterol count. Source: people who eat oat bran reported the cholesterol level.

b. Domestic violence causes more birth defects than all medical issues combined. (Fact: No study - false report).

c. Only 29% of high school girls are happy with themselves. (Fact: Of 3,000 high school girls, 29% responded "I am happy with the way I am". Most answered "Sort of true" and "Sometimes true”.)

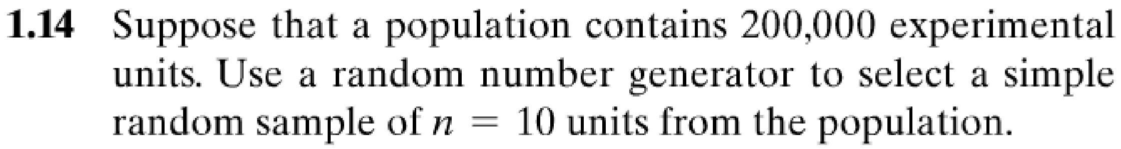
d. One in four children in a certain country under age 12 is hungry or at risk of hunger. (Fact: Based on responses to questions "Do you ever cut the size of meals" and "Do you ever eat less than you feel you should?)

e. 30% of employers would "definitely" or "probably" stop offering health coverage to employees if a government-sponsored act were passed. (Fact: Employers were asked leading questions that made it seem logical to them to stop offering insurance.)

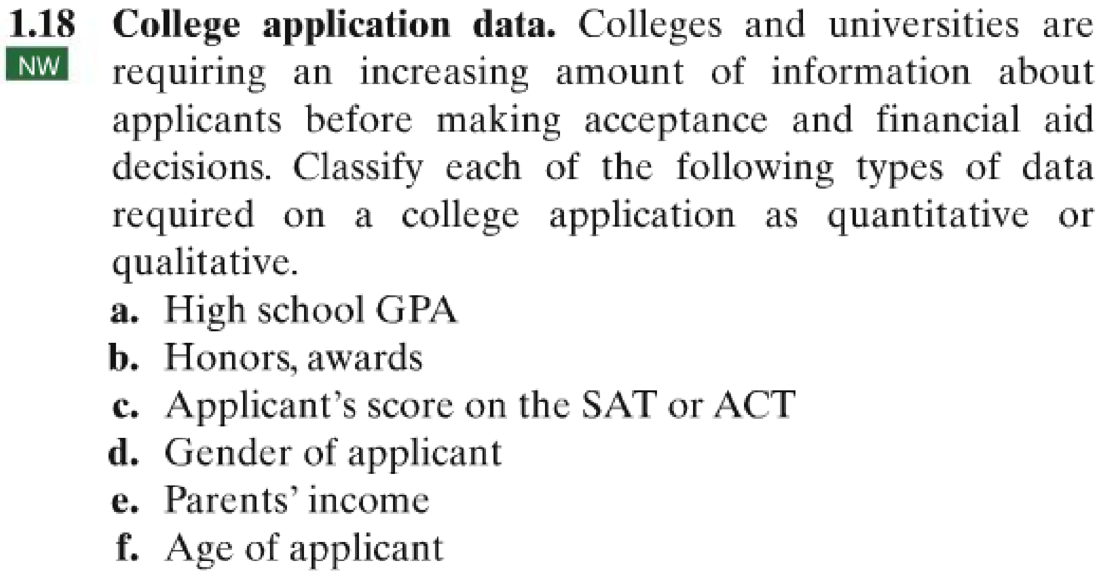
Obtaining data:

1. Data from a *published source*
2. Data from a *designed experiment*
3. Data from an *observational study*

Class exercises



1.14 Suppose that a population contains 200,000 experimental units. Use a random number generator to select a simple random sample of $n = 10$ units from the population.

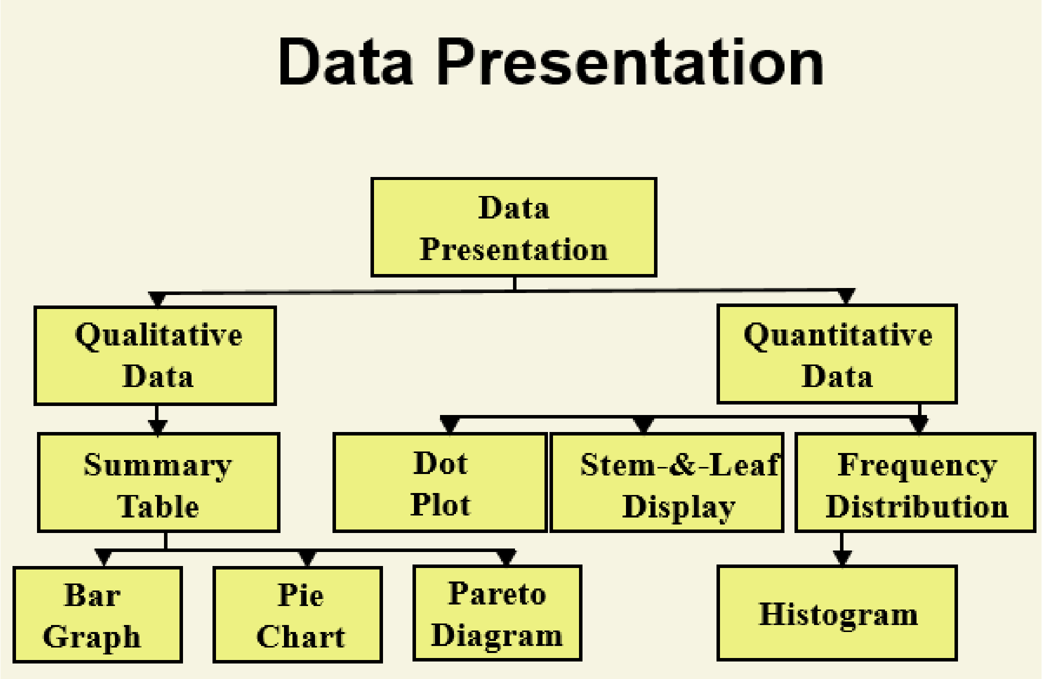


1.18 NW **College application data.** Colleges and universities are requiring an increasing amount of information about applicants before making acceptance and financial aid decisions. Classify each of the following types of data required on a college application as quantitative or qualitative.

a. High school GPA
b. Honors, awards
c. Applicant’s score on the SAT or ACT
d. Gender of applicant
e. Parents’ income
f. Age of applicant

Chapter 2 Describing Data

1. Describing Qualitative Data
2. Graphical Methods for Describing Quantitative Data
3. Numerical Measures of Central Tendency
4. Numerical Measures of Variability
5. Using the Mean and Standard Deviation or Median and IQR to Describe Data
6. Numerical Measures of Relative Standing
7. Methods for Detecting Outliers: Box Plots and *z*-scores
8. Graphing Bivariate Relationships
9. The Time Series Plot
10. Distorting the Truth with Descriptive Techniques



2.1 Describing Qualitative (Categorical) Data

Key concepts:

- **Class, frequency, relative frequency**
- **Bar graph**
- **Pie chart**
- **Pareto diagram**

- Bar Graphs - Heights of rectangles represent group frequencies. The bars have equal widths.
- Pie Charts - Categories represented as percentages of total and illustrated as the slices of a pie
- Pareto graphs - The bar graph is re-organized from the tallest bars to the shortest with possible exception "the others", which is always the last one.

A **class** is one of the categories into which qualitative data can be classified.

The **class frequency** is the number of observations in the data set falling into a particular class.

The **class relative frequency** is the class frequency divided by the total numbers of observations in the data set.

The **class percentage** is the class relative frequency multiplied by 100.

class percentage is the class relative frequency multiplied by 100.

Example: Construct a bar graph, pie chart, and Pareto diagram to describe the data. Compute relative frequencies.

Browser	**Mkt. Share (%)**
Firefox	14
Internet Explorer	81
Safari	4
Others	1

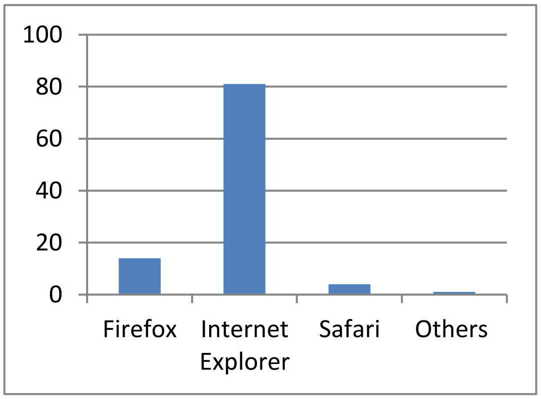
Classes: Firefox, IE, Safari, Others

Frequencies of Market Share: 14, 81, 4, 1

Relative frequencies:

……………………………

Bar Graph:



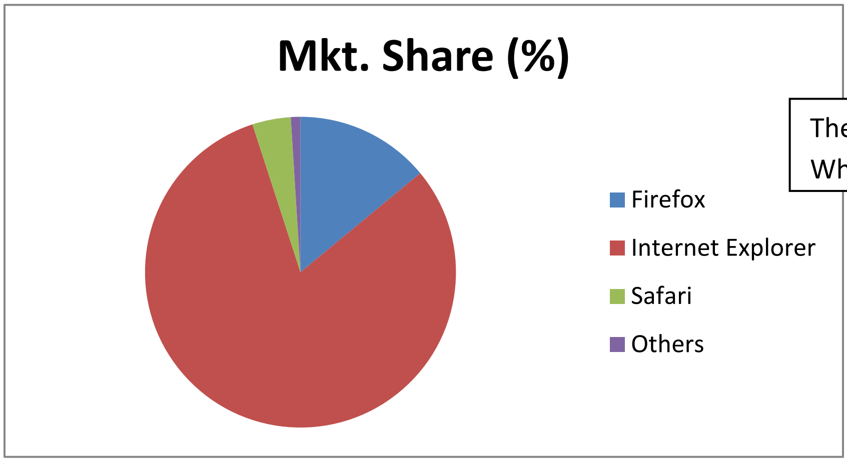
<u>Pie graph:</u>

Find the size of each slice

Example: how to find the size of the slice representing Firefox;

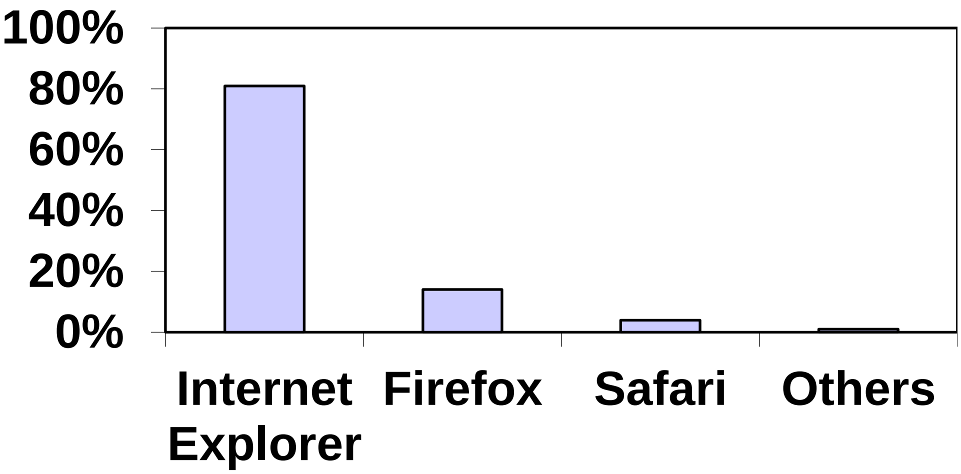
14/100*360^0=50.4^0 – the size of the central angle.

What is the size of the central angle for the slice representing Internet Explorer?



There is a problem with this graph…
What is it?

Pareto:



Classwork: **Section 2.1 p.49**

2.6 Who is to blame for rising health care costs? Rising health care costs are of major concern to Americans. A nationwide survey of 2,119 U.S. adults was conducted to elicit opinions on who is to blame for the rising costs (*The Harris Poll*, Oct. 28, 2008). The next table summarizes the responses to the question “When you think of the rising costs of health care, who do you think is most responsible?”

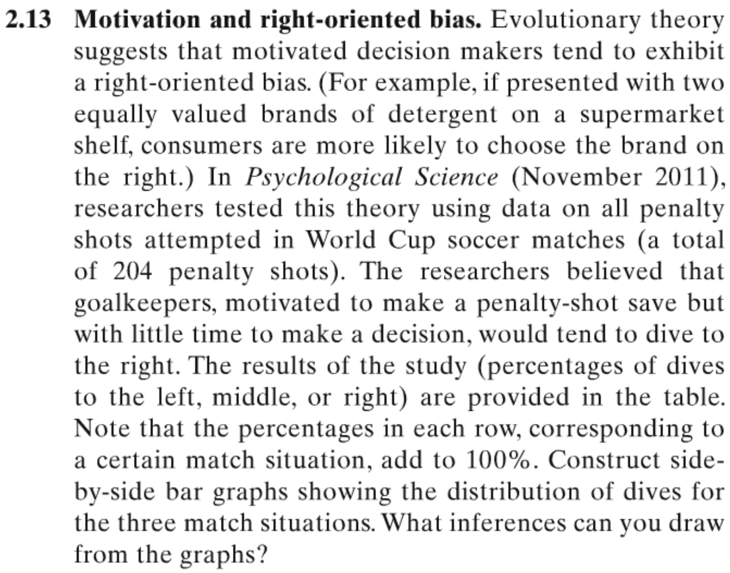
a. Compute the relative frequencies in each response category.
b. Construct a relative frequency bar graph for the data.
c. Convert the relative frequency bar graph into a Pareto diagram. Interpret the graph

Most Responsible for Rising Health Care Costs	Number Responding
Insurance companies	869
Pharmaceutical companies	339
Government	338
Hospitals	127
Physicians	85
Other	128
Not at all sure	233
Total	2,119

Solution: (load data to EXCEL)

Categorical data should be displayed with bar or pie graph.
Pie graphs are not helpful when having many categories, or when even small difference in frequencies is important to observe.

Exercise:



2.13 **Motivation and right-oriented bias.** Evolutionary theory suggests that motivated decision makers tend to exhibit a right-oriented bias. (For example, if presented with two equally valued brands of detergent on a supermarket shelf, consumers are more likely to choose the brand on the right.) In *Psychological Science* (November 2011), researchers tested this theory using data on all penalty shots attempted in World Cup soccer matches (a total of 204 penalty shots). The researchers believed that goalkeepers, motivated to make a penalty-shot save but with little time to make a decision, would tend to dive to the right. The results of the study (percentages of dives to the left, middle, or right) are provided in the table. Note that the percentages in each row, corresponding to a certain match situation, add to 100%. Construct side-by-side bar graphs showing the distribution of dives for the three match situations. What inferences can you draw from the graphs?

Match Situation	**Dive Left**	**Stay Middle**	**Dive Right**
Team behind	29%	0%	71%
Tied	48%	3%	49%
Team ahead	51%	1%	48%

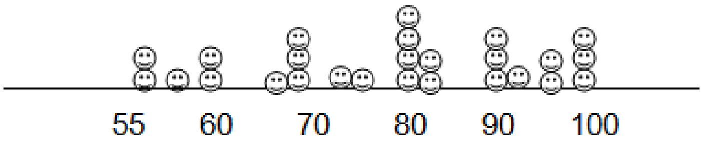
Chapter 2.2 Graphical Methods for Describing Quantitative Data

- **Dot Plots**
- **Stem-and-leaf**
- **Frequency Distribution Tables or Relative FDT**
- **Histograms**

Dot plots: Horizontal axis is a scale for the quantitative variable, e.g., percent. The numerical value of each measurement is located on the horizontal scale by a dot.

A dot plot is drawn by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.

Example: Test scores can be displayed as the columns of dots above a number line with data intervals as below:



Stem and leaf graph:

1. First, cut each data value into leading digits ("stems") and trailing digits ("leaves").

2. Use the stems to label the bins.

3. Use only one digit for each leaf—either round or truncate the data values to one decimal place after the stem.

For example, a data value of 147.3 would have 14 as the stem and 7 as the leaf.
Example: Data: 39, 50, 55, 57, 62, 63, 71,72, 72, 74, 75, 77, 82, 88, 89, 90, 92, 93,100

Stem: | Leaves:

3
4
5
6
7
8
9
10

Unlike a histogram, the stem-leaf graph preserves the original values of the data. But it cannot be easily made for a large set of data.

Making Frequency Distribution Table

How to – the steps:

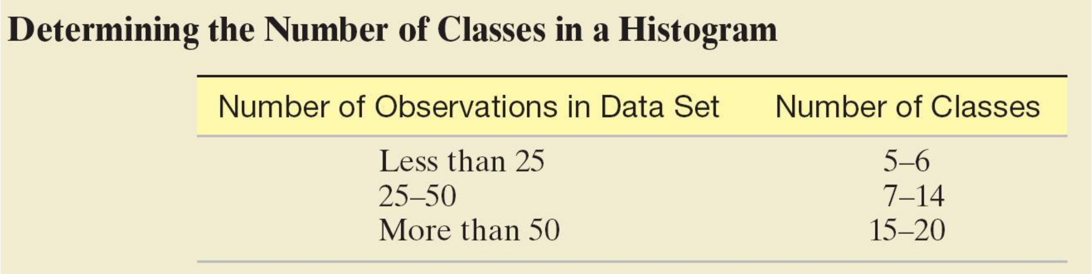
1. Determine range
2. Select number of classes *(usually* between 5 & 15 inclusive)
3. Compute class intervals (class width)
4. Determine class boundaries (limits)
5. Compute class midpoints
6. Count observations & assign to classes

Example:

90	85	89	90	83
89	90	89	85	89
87	87	84	81	82
83	86	86	90	82
81	82	83	84	89
85	86	85	81	89

The difference between maximum and minimum value (**<u>the range</u>**) is divided into the desired number of classes to find the **<u>class width</u>**. Try to give each class the same width.

There is no set rule for a number of classes. The text suggests the following (it is just a suggestion):



Determining the Number of Classes in a Histogram

Number of Observations in Data Set	Number of Classes
Less than 25	5–6
25–50	7–14
More than 50	15–20

We'll use 5 classes.

Maximum temperature: 90, Minimum temperature: 81

Range: 90-81=9 Number of classes: 5

Class width: 9/5=1.8. But we'll use more convenient number: Width= 2

The upper and lower class limits for given class are the largest and the smallest number in each class.

Class midpoint: the arithmetic average between the lower and upper class limit. Used in many programs instead of the class limits to mark a class.

Class limits	Class midpoint	Frequency
81-82	81.5	6
83-84	83.5	7
85-86	85.5	6
87-88	87.5	2
89-90	89.5	9
	Total:	30

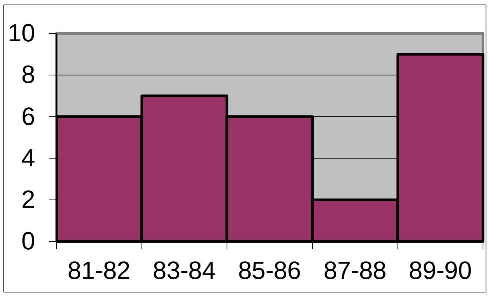
You can use Excel to make distribution tables (incorrectly called there "histograms")

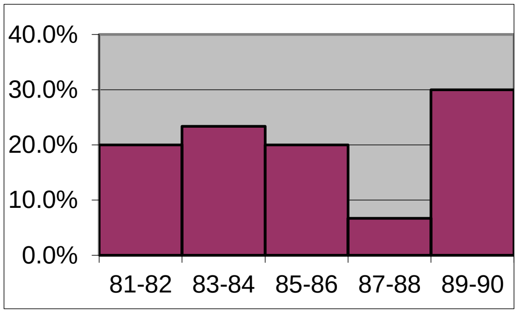
A histogram is a special kind of a bar graph in which the horizontal scale represents the classes of data values and the vertical scale represents the frequencies. **There are no gaps between the bars, and the widths of the bars are usually equal.**

Relative frequency distribution table:

Class	Class		Relative
limits	**midpoint**	**Frequency**	**Frequency**
81-82	81.5	6	20.0%
83-84	83.5	7	23.3%
85-86	85.5	6	20.0%
87-88	87.5	2	6.7%
89-90	89.5	9	30.0%

Here is the display of the data above, or the histogram. The histogram built on Relative Frequency table has exactly the same

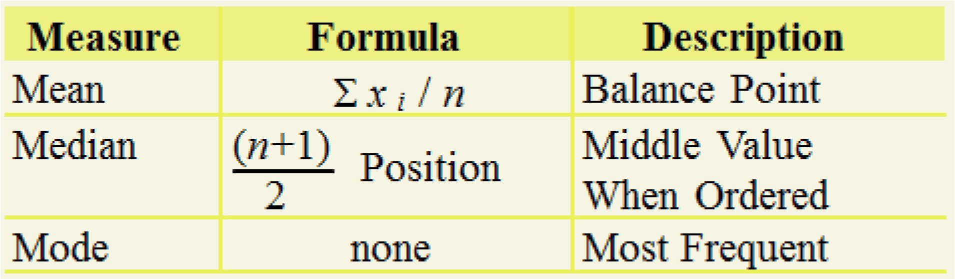




2.3 Numerical Measures of Central Tendency

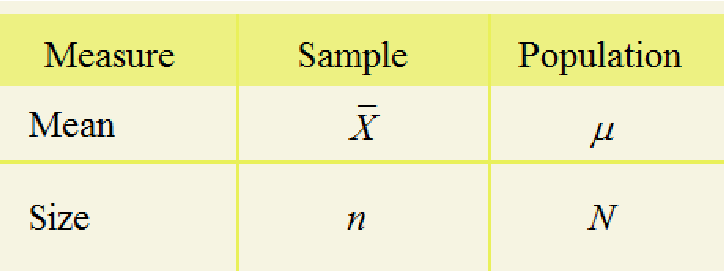
- **The central tendency** of the set of measurements–that is, the tendency of the data to cluster, or center, about certain numerical values
- **The variability** of the set of measurements–that is, the spread of the data.

Measures of Central Tendency



Measure	**Formula**	**Description**
Mean	$\Sigma x_i / n$	Balance Point
Median	$\frac{(n+1)}{2}$ Position	Middle Value When Ordered
Mode	none	Most Frequent

Notation:



Measure	Sample	Population
Mean	$\bar{X}$	μ
Size	n	N

Example: The population contains 6 employees of a small business The salaries are as follow: 8000, 10000, 11000, 12000, 25000, 90000.

a. Find the population mean annual salary for the list of all salaries below:

$$\mu = \frac{\sum_{1}^{6} x_i}{N} = \ldots$$

Select a sample of three employees, and then compute their mean salary:

$\bar{x}$ =…

Median: a **m**iddle value in ordered sequence

- If n is odd, take middle value of sequence
- If n is even, take the average of 2 middle values

Unlike the mean, the median is not affected by extreme values

Example:

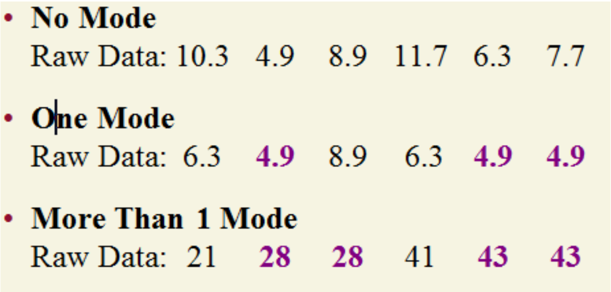
Find median annual salary for the list of all salaries below:

8000, 10000, 11000, 12000, 25000, 90000.

Mode:

1. Value that occurs most often
2. Not affected by extreme values
3. May be no mode or several modes
4. May be used for quantitative or qualitative data

Example:



- **No Mode**
  Raw Data: 10.3 4.9 8.9 11.7 6.3 7.7
- **One Mode**
  Raw Data: 6.3 **4.9** 8.9 6.3 **4.9** **4.9**
- **More Than 1 Mode**
  Raw Data: 21 **28** **28** 41 **43** **43**

Classwork: **Ch.2.4 #47**

- 2.47 Is honey a cough remedy? Refer to the *Archives of Pediatrics and Adolescent Medicine* (Dec. 2007) study of honey as a remedy for coughing, Exercise 2.30 (p. 61). Recall that the 105 ill children in the sample were randomly divided into three groups: those who received a dosage of an over-the-counter cough medicine (DM), those who received a dosage of honey (H), and those who received no dosage (control group). The coughing improvement scores (as determined by the children's parents) for the patients are reproduced in the accompanying table.

Honey Dosage:	12	11	15	11	10	13	10	4	15	16	9	14			
	10	6	10	8	11	12	12	8	12	9	11	15			
	10	15	9	13	8	12	10	8	9	5	12				
DM Dosage:	4	6	9	4	7	7	7	9	12	10	11	6	3	4	
	9	12	7	6	8	12	12	4	12	13	7	10			
	13	9	4	4	10	15	9								
No Dosage (Control):	5	8	6	1	0	8	12	8	7	7	1	6	7	7	12
	7	9	7	9	5	11	9	5	6	8					
	8	6	7	10	9	4	8	7	3	1	4	3			

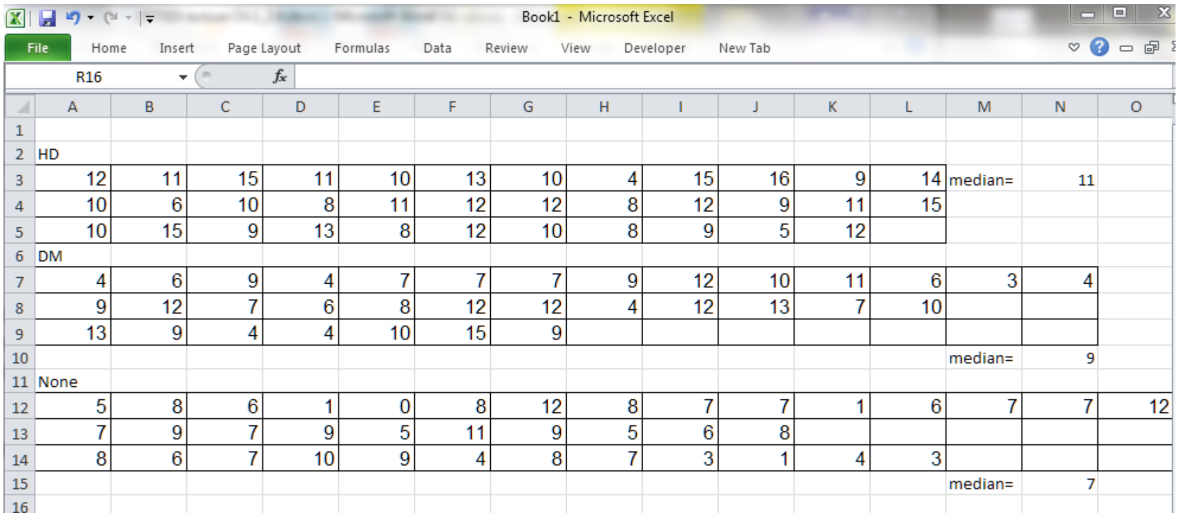
a. Find the median improvement score for the honey dosage group.

b. Find the median improvement score for the DM dosage group.

c. Find the median improvement score for the control group.

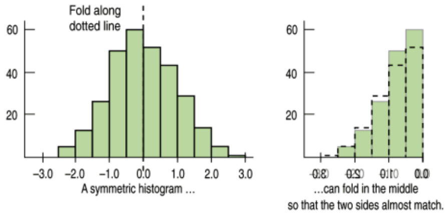
d. Based on the results, parts **a–c,** what conclusions can pediatric researchers draw? (We show how to support these conclusions with a measure of reliability in subsequent chapters.)

Solution (EXCEL)

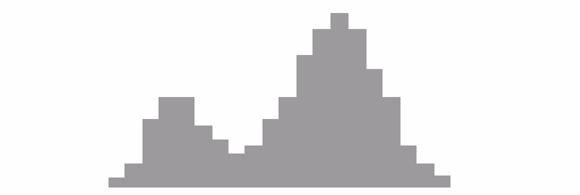


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2	HD														
3	12	11	15	11	10	13	10	4	15	16	9	14	median=	11	
4	10	6	10	8	11	12	12	8	12	9	11	15			
5	10	15	9	13	8	12	10	8	9	5	12				
6	DM														
7	4	6	9	4	7	7	7	9	12	10	11	6	3	4	
8	9	12	7	6	8	12	12	4	12	13	7	10			
9	13	9	4	4	10	15	9								
10													median=	9	
11	None														
12	5	8	6	1	0	8	12	8	7	7	1	6	7	7	12
13	7	9	7	9	5	11	9	5	6	8					
14	8	6	7	10	9	4	8	7	3	1	4	3			
15													median=	7	
16															

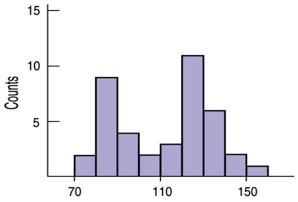
Shapes of the distribution of the data

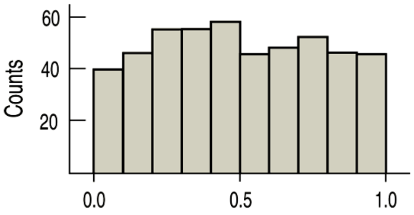


Symmetric:______________________________

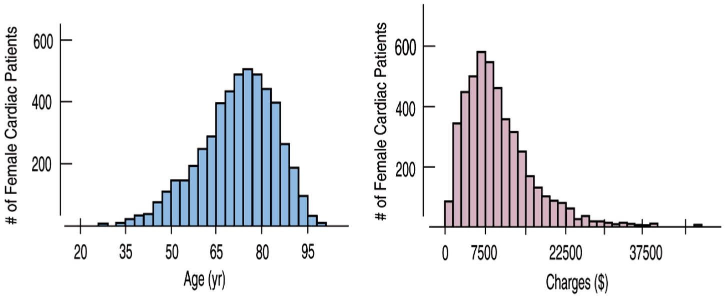


Bimodal (double-peaked) distribution





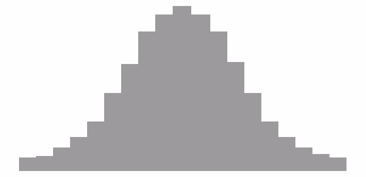
Uniform distribution



Skewed distribution

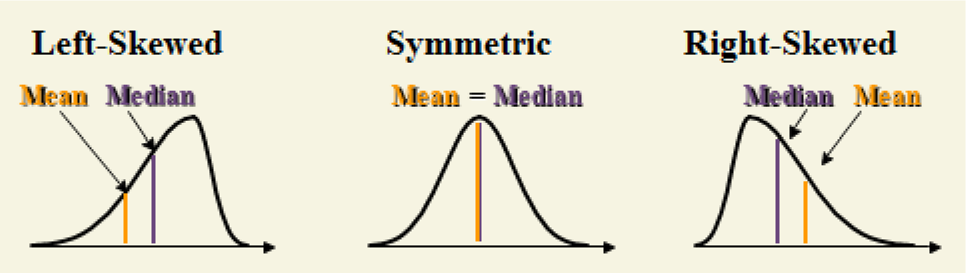
The (usually) thinner ends of a distribution are called the tails. If one tail stretches out farther than the other, the histogram is said to be skewed to the side of the longer tail. In the figure above, the histogram on the left is said to be skewed left, while the histogram on the right is said to be skewed right.

A special shape to remember:



Normal distribution

Mean vs. median location on the histograms of a skewed distribution



Example: Ch. 2.4 #49

Symmetric or skewed? Would you expect the data sets described below to possess relative frequency distributions that are symmetric, skewed to the right, or skewed to the left? Explain.

a. The salaries of all persons employed by a large university

b. The grades on an easy test

c. The grades on a difficult test

d. The amounts of time students in your class studied last week

e. The ages of automobiles on a used-car lot

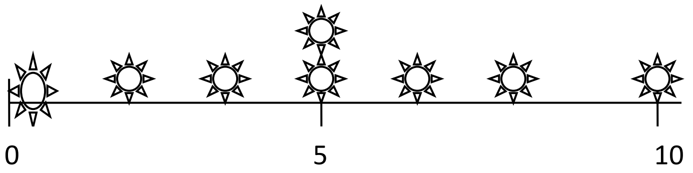
f. The amounts of time spent by students on a difficult examination (maximum time is 50 minutes)

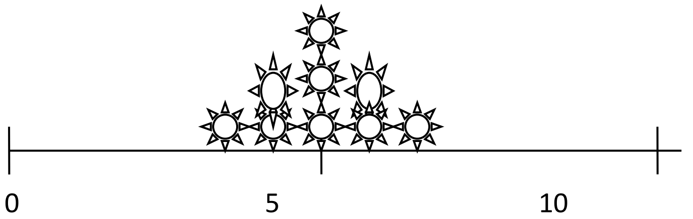
Chapter 2.4 Numerical Measures of Variability (Dispersion, Spread)

- **The range**
- **The variance**
- **The standard deviation**
- The **range, *R*,** of a variable is the difference between the largest data value and the smallest data values.
- Range = *R* = Largest Data Value – Smallest Data Value
- A more powerful measures of spread are **the variance and standard deviation**, which take into account how far *each* data value is from the mean.
- (Later on we will add one more powerful measure of spread: IQR)

Finding the sample variance and sample standard deviation

- Consider two sets of data:





A **deviation** is the distance that a data value is from the mean: $x - \bar{x}$.

Exercise: find the sum of all deviations for the following set of sample data:

1, 2, 3, 4, 5

Since averaging all deviations would give zero, we square each deviation and find an average of sorts for the deviations.

The **population variance** of a variable, denoted by **σ^2** is the sum of squared deviations from the population mean divided by the number of observations in the population, *N*.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}$$

The **sample variance**, notated by s^2, is found by summing the squared deviations and (almost) averaging them: $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$

The variance will play a role later in our study, but it is problematic as a measure of spread—it is measured in *squared* units!

x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
1			
2			

The **sample standard deviation, *s*,** is simply **the square root of the variance** and is measured in the same units as the original data.

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

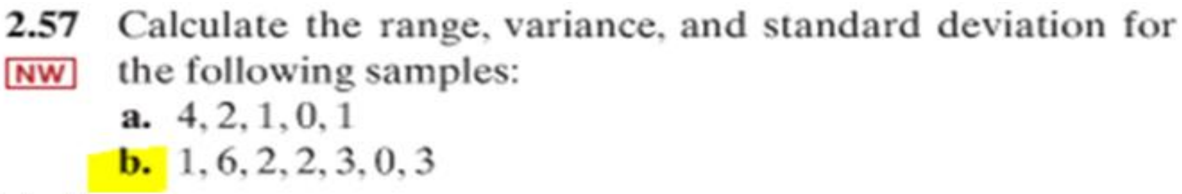
Population standard deviation:

$$\sigma=\sqrt{\frac{\sum(x-\mu)^2}{N}}$$

3			
4			
5			
		TOTAL	

- Standard deviation of a sample can be easily found on TI-83, but you must also learn how to use the formula.

Classwork:



2.57 NW Calculate the range, variance, and standard deviation for the following samples:

a. 4, 2, 1, 0, 1

b. 1, 6, 2, 2, 3, 0, 3

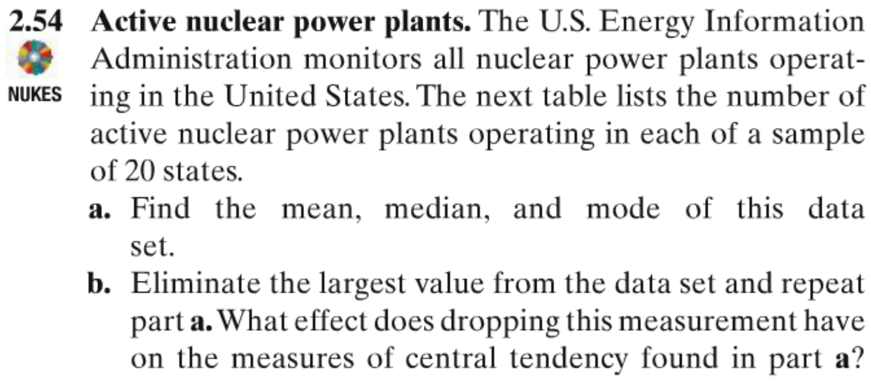
Notation:

Notation:

Population: use Greek letters: mean μ, standard deviation σ

Sample: use Latin letters: mean $\bar{x}$, standard deviation s

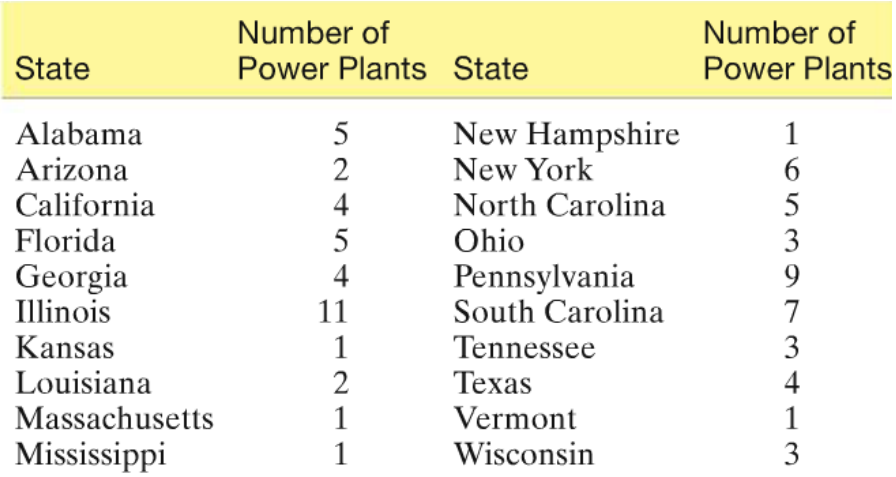
Exercise:



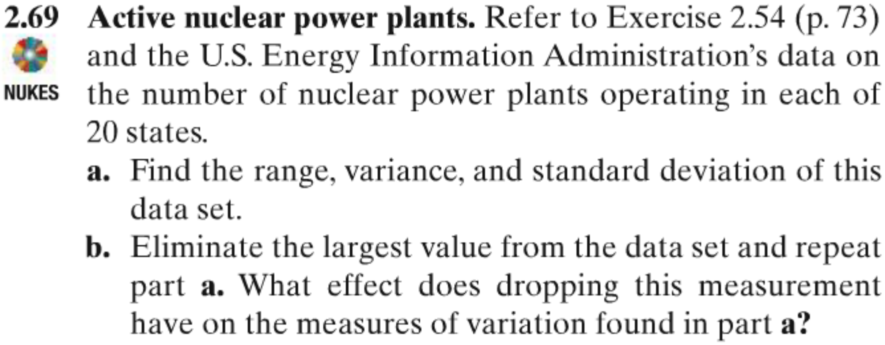
2.54 NUKES **Active nuclear power plants.** The U.S. Energy Information Administration monitors all nuclear power plants operating in the United States. The next table lists the number of active nuclear power plants operating in each of a sample of 20 states.

a. Find the mean, median, and mode of this data set.

b. Eliminate the largest value from the data set and repeat part **a.** What effect does dropping this measurement have on the measures of central tendency found in part **a**?



State	Number of Power Plants	State	Number of Power Plants
Alabama	5	New Hampshire	1
Arizona	2	New York	6
California	4	North Carolina	5
Florida	5	Ohio	3
Georgia	4	Pennsylvania	9
Illinois	11	South Carolina	7
Kansas	1	Tennessee	3
Louisiana	2	Texas	4
Massachusetts	1	Vermont	1
Mississippi	1	Wisconsin	3



2.69 **Active nuclear power plants.** Refer to Exercise 2.54 (p. 73) and the U.S. Energy Information Administration's data on the number of nuclear power plants operating in each of 20 states.

NUKES

a. Find the range, variance, and standard deviation of this data set.

b. Eliminate the largest value from the data set and repeat part **a.** What effect does dropping this measurement have on the measures of variation found in part **a?**