

# Lecture One

## Section 1.1 – Statistical Thinking

### Definitions

**Data** are collections of observations (such as measurements, genders, survey responses)

**Statistics** is the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

A **population** is the complete collection of all individuals (scores, people, measurements, and so on) to be studied. The collection is complete in the sense that it includes all of the individuals to be studied.

A **census** is the collection of data from every member of the population.

A **sample** is a subcollection of members selected from a population.

### Key Concept

This section introduces basic principles of statistical thinking used throughout this book. Whether conducting statistical analysis of data that we have collected, or analyzing a statistical analysis done by someone else, we should not rely on blind acceptance of mathematical calculation. We should consider these factors:

- ✓ Context of the data
- ✓ Source of the data
- ✓ Sampling method
- ✓ Conclusions
- ✓ Practical implications

### Context

There is no description of what the values represent, where they came from, and why they were collected.

### Example

Table 1 – Data Used for Analysis

<b><i>x</i></b>	56	67	57	60	64
<b><i>y</i></b>	53	66	58	61	68

The data in the table consists of pair ( $x$ ,  $y$ ) of values that has a **before** weight and an **after** weight for one particular student. An understanding of the context will directly affect the statistical procedure used. The key issue is whether the changes in weight appear to support or contradict that college students gain 15 lb. during their freshman year.

## ***Source of data***

Consider the source of the data, and consider whether that source is likely to be objective or there is some incentive to be biased.

### ***Example***

The researchers have no incentive to distort or spin results to support some self-serving position. This may have nothing to gain or lose by distorting results. They were not paid by a company that could profit from favorable results. We can be confident that these researchers are unbiased and they did not distort results.

## ***Sampling Method***

If we are collecting sample data for a study, the sampling method that we choose can greatly influence the validity of our conclusions.

### ***Example***

The weights in the table are from a larger sample of weights in data set. All 217 students who participated in the assessment were invited for a follow-up, and 67 of those students responded and were measured again. This sample is a voluntary respond sample. The researchers wrote that “the sample obtained was not random and may have introduced self-selection bias”. They elaborated on the potential for bias by specifically listing particular potential sources of bias, such as the response of “only those students who felt comfortable enough with there to be measured both times.”

## ***Conclusions***

- ✓ Make statements that are clear to those without an understanding of statistics and its terminology.
- ✓ Avoid making statements not justified by the statistical analysis.

### ***Example***

The table lists before and after weight of five subjects taken from Data Set. Those weights were analyzed with conclusions included in “Changes in Body Weight and Fat Mass of Men and Women in the First Year of College: A study of the ‘Freshman 15,’ “ by Hoffman, Policastro, Quick and Lee, *Journal of American College*, Vol. 55, No 1. The investigators concluded that the freshman year of college is a time during which weight gain occurs. But the investigators went on to state that in the small nonrandom group studied, the weight gain was less than 15 pounds, and this amount was not universal. Therefore, they concluded that the weight gain is a myth.

## ***Practical Implications***

In addition to clearly stating conclusions of the statistical, we should also identify any practical implications of the results.

### ***Example***

In their analysis of the data collected, the researchers point out some practical implications of their results. They wrote that “it is perhaps most important for students to recognize that seemingly minor and perhaps even harmless changes in eating or exercise behavior may result in large changes in weight and body fat mass over an extended period of time.” Beginning freshman college students should recognize that there could be serious health consequences resulting from radically different diet and exercise routines.

- ✓ There exist some *statistical significance* yet there may be NO *practical significance*. Common sense might suggest that the finding does not make enough of a difference to justify its use or to be practical.

### ***Example***

In a test of the Atkins weight loss program, 40 subjects using that program has a mean weight loss of 2.1 lb. after one year (based on data), *Journal of American Medical Association*, Vol. 2935, No 1. Using methods of statistical analysis, we can conclude that the mean weight loss of 2.1 is statistically significant/ That is, based in statistical criteria, the diet appears to be effective. However, using common sense, it does not seem worthwhile to pursue a weight loss program resulting in such relatively insignificant results. Someone starting a weight loss program would likely want to lose considerably more than 2.1 lb. Although the mean weight loss of 2.1 lb. is statistically significant, it does not have practical significance. The statistical analysis suggests that the weight loss program is effective, but practical considerations suggest that the program is basically ineffective.

## ***Statistical Significance***

Consider the likelihood of getting the results by chance.

If results could easily occur by chance, then they are *not statistically significant*.

If the likelihood of getting the results is so small, then the results are *statistically significant*.

### ***Example***

The Genetics and IVF Institute in Fairfax, Virginia developed a technique called MicroSort, which supposedly increases the chances of a couple having a baby girl. In a preliminary test, researchers located 14 couples who wanted baby girls. After using the MicroSort technique, 13 of them had girls and one couple had a boy. After obtaining these results, we have two possible conclusions:

1. The MicroSort technique is not effective and the result of 13 girls in 14 births occurred chance.
2. The MicroSort technique is effective, and couples who use the technique are more likely to have baby girls, as claimed by the Genetics and IVF Institute.

When choosing between the two possible explanations for the results, statisticians consider the *likelihood* of getting the results by chance. They are able to determine that if the MicroSort technique has no effect, then there is about 1 chance in 1000 of getting results like those obtained here. Because that likelihood is so small, statisticians conclude that the results are statistically significant, so it appears that the MicroSort technique is effective.

### ***Example***

Suppose the couples had 8 baby girls in 14 births. We can see that 8 baby girls are more than the 7 girls that we would expect with an ineffective treatment. However, statisticians can determine that if the MicroSort technique has no effect, then there are roughly two chances in five of getting 8 girls in 14 births. Unlike the one chance in 1000 from the preceding example, two chances in five indicate that the results could easily occur by chance. This would indicate that the result of 8 girls in 14 births is not *statistically significant*. With 8 girls in 14 births, we would not conclude that the technique is effective, because it is so easy (two chances in five) to get the results with an ineffective treatment or no treatment.

## Exercises      Section 1.1 – Statistical Thinking

Use common sense to determine whether the given event is (a) *impossible*; (b) *possible, but very unlikely*; (c) *possible and likely*.

1. Giants best the Denver Broncos in the Super Bowl by a score of 120 to 98.
2. While driving to his home in Connecticut, David was ticketed for driving 205 *mi/h* on a highway with a speed limit of 55 *mi/h*.
3. Thanksgiving Day will fall on a Monday next year.
4. When each of 25 statistics students turns on his or her TI-84 Plus calculator, all 25 calculators operate successfully.

*Nicotine Amounts from Menthol and King-  
Size Cigarettes*

<i>x</i>	1.1	0.8	1.0	0.9	0.8
<i>y</i>	1.1	1.7	1.7	1.1	1.1

The *x*-values are nicotine amounts (in *mg*) in different 100 *mm* filtered, non-light menthol cigarettes; the *y*-values are nicotine amounts (in *mg*) in different king-size non-filtered, non-menthol, and non-light cigarettes.

5. Each *x* value associated with the corresponding *y* value in some meaningful way? If the *x* and *y* values are not matched, does it make sense to use the difference between each *x* value and the *y* value that is the same column?
6. The Federal Trade Commission obtained the measured amounts of nicotine in the table. Is the source of the data likely to be unbiased?
7. Note that the table lists measured nicotine amounts from two different types of cigarette. Given these data, what issue can be addressed by conducting a statistical analysis of the values?
8. One of Gregor Mendel's famous hybridization experiments with peas yielding 580 off spring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the off spring peas should have yellow pods. Do the results of the experiment differ from Mendel's claimed rate of 25% by an amount that is statistically significant?
9. In a Gallup poll of 1038 randomly selected adults, 85% said that secondhand smoke is somewhat harmful or very harmful, but a representative of the tobacco industry claims that only 50% of adults believe that secondhand smoke is somewhat harmful or very harmful. Is there statistically significant evidence against the representative's claim? Why or why not?

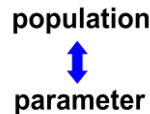
## Section 1.2 – Types of Data

### Key Concept

The subject of statistics is largely about using sample data to make inferences (or generalizations) about an entire population. It is essential to know and understand the definitions that follow.

### Definition

**Parameter** is a numerical measurement describing some characteristic of a population



**Statistic** is a numerical measurement describing some characteristic of a sample



### Example

There are exactly 100 Senators in the 109<sup>th</sup> Congress of the US, and 55% of them are Republicans. The figure of 55% is a **parameter** because it is based on the entire population of all 100 Senators.

### Example

In 1936, *Literary Digest* polled 2.3 million adults in the US, and 57% said that they would vote for Alf Landon for the presidency. That figure of 57% is a **statistic** because it is based on a sample, not the entire population of all adults in the US.

### Definition

**Quantitative** (or numerical) data consists of *numbers* representing counts or measurements.

**Example:** The weights of supermodels

**Example:** The ages (in years) of survey respondents

**Categorical** (or qualitative or attribute) data consists of names or labels (representing categories)

**Example:** The genders (male/female) of professional athletes

**Example:** Shirt numbers on professional athletes uniforms - substitutes for names.

### Working with Quantitative Data

Quantitative data can further be described by distinguishing between **discrete** and **continuous** types.

**Discrete data** result when the number of possible values is either a finite number or a ‘countable’ number (i.e. the number of possible values is 0, 1, 2, 3, . . .)

### **Example**

The number of eggs that hens lay are **discrete** data because they represent counts.

**Continuous (numerical) data** result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

### **Example**

The amounts of milk from cows are continuous data because they are measurements that can assume any value over a continuous span. During a year, a cow might yield an amount of milk that can be any value between 0 and 7000 liters. It would be possible to get 2.343115 gallons per day

## **Levels of Measurement**

Another way to classify data is to use levels of measurement. Four of these levels are discussed in the following slides.

**Nominal level of measurement** characterized by data that consist of names, labels, or categories only, and the data cannot be arranged in an ordering scheme (such as low to high)

### **Example**

- ✓ Survey responses *yes, no, undecided*
- ✓ The political party affiliations of survey respondents (Democrat, Republican, Independent, other)

**Ordinal level of measurement** involves data that can be arranged in some order, but differences between data values either cannot be determined or are meaningless

### **Example**

A college professor assigns **grades** of *A, B, C, D, or F*. These grades can be arranged in order, but we can’t determine differences between the grades. For example, we know that *A* is higher than *B* (so there is an ordering), but we cannot subtract *B* from *A* (so the difference cannot be found).

### **Example**

*U.S. News and World Report* ranks colleges. Those ranks (first, second, third, and so on) determine an ordering. However, the differences between ranks are meaningless. For example, a difference of “second

minus first” might suggest  $2 - 1 = 1$ , but this difference of 1 is meaningless because it is not an exact quantity that can be compared to other such differences. The difference between Harvard and Brown cannot be quantitatively compared to the difference between Yale and Johns Hopkins.

**Interval level of measurement** like the ordinal level, with the additional property that the difference between any two data values is meaningful, however, there is no natural zero starting point (where none of the quantity is present)

### **Example**

The years 2000, 1776, and 1492. Time did not begin in the year 0, so the year 0 is arbitrary instead of being a natural zero starting point representing “no time”.

### **Example**

Body temperatures of 98.2°F and 98.6°F are examples of data at this interval level of measurement. Those values are ordered, and we can determine their difference of 0.4°F. However, there is no natural starting point. The value of 0°F might seem like a starting point, but it is arbitrary and does not represent the total absence of heat.

**Ratio level of measurement** the interval level with the additional property that there is also a natural zero starting point (where zero indicates that none of the quantity is present); for values at this level, differences and ratios are meaningful

### **Example**

Prices of college textbooks (\$0 represents no cost, a \$100 book costs twice as much as a \$50 book)

### **Example**

Distances (in km) traveled by cars (0 km represents no distance traveled, and 400 km is twice as far as 200 km.

<b>Levels of Measurement</b>		
<b>Ratio</b>	There is a natural zero starting point and ratios are meaningful	Distances
<b>Interval</b>	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless	Body temperatures
<b>Ordinal</b>	Categories are ordered, but differences can't be found or are meaningless	Ranks of colleges
<b>Nominal</b>	Categories only. Data cannot be arranged in an ordering scheme	Eye colors

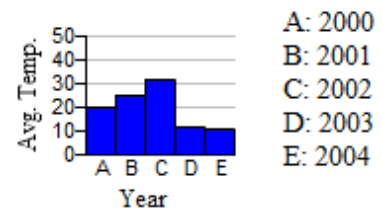


## **Exercises**      **Section 1.2 – Types of Data**

1. In a large sample of households, the median annual income per household for high school graduates is \$19,856 (based on data from the U.S. Census Bureau). Determine if it is a statistic or a parameter.
2. Among the Senators in the current Congress, 44% are Democrats. Is it a statistic or a parameter?
3. A study of all 2223 passengers aboard the Titanic found that 706 survived when it sank. Is it a statistic or a parameter?
4. If the areas of the 50 states are added and the sum is divided by 50, the result is 196,533 square kilometers. Is it a statistic or a parameter?
5. The average (mean) atomic weight of all elements in the periodic table is 134,355 unified mass units. Is it a statistic or a parameter?
6. Determine whether the given value is parametric or a statistic
  - a) One of greatest baseball hitters of all time has a career batting average of 0.366
  - b) A sample of employees is selected and it is found that 50% own a vehicle
  - c) A survey of 42 out of hundreds in a dining hall showed that 17 enjoyed their meal
7. Suppose a survey of 568 women in the U.S. found that more than 61% are the primary investor in their household.
  - a) Describe the survey represents the descriptive branch of statistic
  - b) Make an inference based on the results of the survey
8. In the recent study, volunteers who had 8 hours of sleep were three times more likely to answer questions correctly on a math test than were sleep-deprived participants.
  - a) Identify the sample used in the study
  - b) What is the sample's population
  - c) Which part of the study represents the descriptive branch of statistics
  - d) Make an inference based on the results of the study
9. Determine whether the data set is a population or a sample. Explain your reasoning  
The salary if each baseball player in a league
10. In a poll, 1,004 adults in a country were asked whether they favor or oppose the use of “federal tax dollars to fund medical research using stem cells obtained from human embryos.” Among the responders, 48% said that they were in favor. Describe the statistical study
  - a) What is the population?
  - b) Identify the sample
11. A study shows that the obesity rate among boys ages 2 to 19 has increased over the past several years
  - a) Make an inference based on the results of this study?
  - b) What is wrong with this type of reasoning

12. In the Literary Digest poll, Landon received 16,679,583 votes. Is it from a discrete or continuous data set?
13. The amount of nicotine in a Marlboro cigarette is 1.2 mg. Is it from a discrete or continuous data set?
14. The volume of cola in a can of regular coke is 12.3 oz. Is it from a discrete or continuous data set?
15. When a woman is randomly selected and measured for blood pressure, the systolic blood pressure is found to be 61 mm Hg.
16. Types of movies (drama, comedy, adventure, documentary, etc.). Determine which of the four levels of measurement (nominal, ordinal, interval, ratio) is most appropriate.
17. Critic ratings of movies on a scale from 0 star to 4 stars. Determine which of the four levels of measurement (nominal, ordinal, interval, ratio) is most appropriate
18. Ranks of cars evaluated by Consumer's Union. Determine which of the four levels of measurement (nominal, ordinal, interval, ratio) is most appropriate
19. The newspaper USA Today published a health survey, and some readers completed the survey and returned it. Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population.
20. A Gallup poll of 1012 randomly surveyed adults found that 9% of them said cloning of humans should be allowed. Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population.
21. Some people responded to this request: "Dial 1-900-PRO-LIFE to participate in a telephone poll on abortion. (\$1.95 per minute. Average call: 2 minutes. You must be 18 years old.)" Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population
22. In the Born Loser cartoon strip by Art Sansom, Brutus expresses joy over an increase in temperature from  $1^{\circ}$  to  $2^{\circ}$ . When asked what is so good about  $2^{\circ}$ , he answers that "it's twice as warm as this morning." explain why Brutus is wrong yet again.
23. A group of students develops a scale for rating the quality of cafeteria food, with 0 representing "neutral: not good and not bad." Bad meals are given negative numbers and good meals are given positive numbers, with the magnitude of the number corresponding to the severity of badness or goodness. The first three meals are rated as 2, 4, and  $-5$ . What is the level of measurement for such rating? Explain your choice.
24. Suppose that a study based on a sample from a targeted population shows that people who own a fax machine have more money than people who do not
  - a) Make an inference based on the results of this study?
  - b) What might this inference incorrectly imply?

25. Determine whether the statement is true or false, rewrite it as a true statement
- Data at the ordinal level are quantitative only
  - More types of calculations can be performed with data at the nominal level than with data at the interval level
26. Determine whether the variable is qualitative or quantitative
- Favorite film
  - Population of country of origin
  - Gallons of water in a swimming pool
  - Model of car driven
  - Distance in miles to nearest school
  - Time in hours that a light bulb lasts
  - Number of students at a high school
27. The region of a country with the highest per capita income for the past six years is shown below  
Northeast      Southern      Southwest      Southeast      Northern      Western
- Determine whether the data are qualitative or quantitative and identify the data set's level of measurement
  - What is the data set's level of measurement?
28. The region of a country with the six highest level of food production last year are shown below  
1. Eastern    2. Southwest    3. Western    4. Southeast    5. Northwest    6. Southern
- Determine whether the data are qualitative or quantitative and identify the data set's level of measurement
  - What is the data set's level of measurement?
29. The region of a country with the six highest level of food production last year are shown below  
22.8    26.4    24.1    22.2    21.6    21.1    25.8    21.5    24.6
- Determine whether the data are qualitative or quantitative and identify the data set's level of measurement
  - What is the data set's level of measurement?
30. The graph shows the average temperature in an arctic city, in degree Fahrenheit, for certain years. Identify the level of measurement of the data listed on the horizontal axis in the graph
31. Identify the level of measurement of the data:
- Temperature
  - Age
  - Family history of illness
  - Pain level (scale of 0 to 10)



## Section 1.3 – Critical Thinking

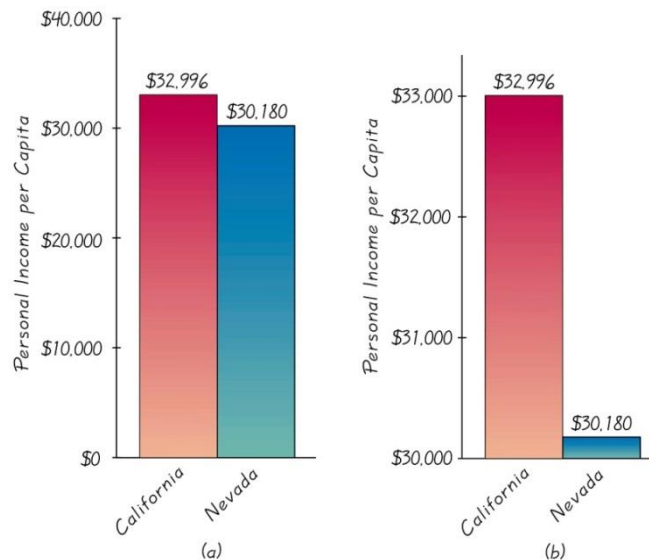
### Misuses of Statistics

1. Evil intent on the part of dishonest people.
2. Unintentional errors on the part of people who don't know any better.

We should learn to distinguish between statistical conclusions that are likely to be valid and those that are seriously flawed.

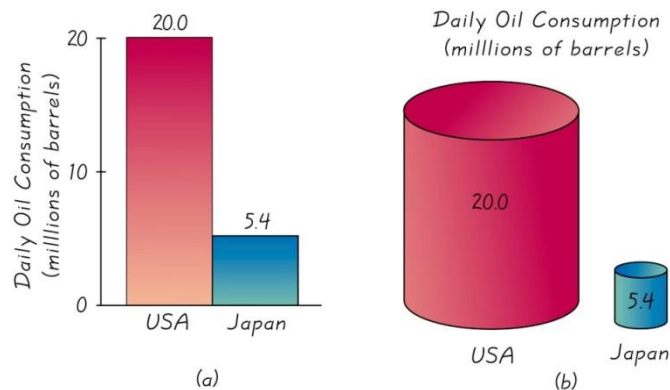
### Graphs/Misuse of Graphs

Statistical data are often presented in visual form- that is, in graphs. Data represented graphically must be interpreted carefully.



To correctly interpret a graph, you must analyze the numerical information given in the graph, so as not to be misled by the graph's shape. READ labels and units on the axes!

### Pictographs



Part (b) is designed to exaggerate the difference by increasing each dimension in proportion to the actual amounts of oil consumption.

## ***Bad Samples***

Some samples are bad in the sense that the method used to collect the data dooms the sample, so that it is likely to be somehow biased.

### ***Voluntary response sample (or self-selected sample)***

Is one in which the respondents themselves decide whether to be included. In this case, valid conclusions can be made only about the specific group of people who agree to participate and not about the population.

### ***Example***

Newsweek magazine ran a survey about Napster Web site, which had been providing free access to downloading copies of music CDs. Readers were asked this question: “will you still use Napster if you have to pay a fee?” Readers could register their responses on the Web site. Among the 1873 responses received, 19% said yes, it is still cheaper than buying CDs. Another 5% said yes, they felt more comfortable using it with a charge. When Newsweek or anyone else runs a poll on the Internet, individuals decide themselves whether to participate, so they constitute a voluntary response sample. But people with strong opinions are more likely to participate, so it is very possible that the responses are not representative of the whole population.

## **Correlation and Causality**

Concluding that one variable *causes* the other variable when in fact the variables are linked. Two variables may seem linked, smoking and pulse rate, this relationship is called correlation. Cannot conclude the one causes the other.

*Correlation does not imply causality.*

## **Small Samples**

Conclusions should not be based on samples that are far too small.

### ***Example***

The Children’s Defense Fund published *Children out of school in America*, in which it was reported that among secondary school students suspended in one region, 67% were suspended at least three times. But that figure is based on a sample of only three students! Media reports failed to mention that this sample size was so small.

## **Percentages**

Misleading or unclear percentages are sometimes used. For example, if you take 100% of a quantity, you take it all. If you have improved 100%, then are you perfect?! 110% of an effort does not make sense.

### ***Example***

To find a percentage of an amount, drop the % symbol and divide the percentage value by 100, then multiply.

$$6\% \text{ of } 1200 \text{ responses} = \frac{6}{100} \times 1200 = \underline{72}$$

### ***Example***

To *convert from fraction to a percentage*, divide the denominator into the numerator to get an equivalent decimal number, then multiply by 100 and affix the % symbol.

$$\frac{3}{4} = 0.75 \rightarrow 0.75 \times 100 = \underline{75\%}$$

### ***Example***

To *convert from decimal to a percentage*, multiply by 100%.

$$0.250 \rightarrow 0.250 \times 100\% = \underline{25\%}$$

### ***Example***

To *convert from percentage to a decimal number*, delete the % symbol and divide by 100.

$$85\% = \frac{85}{100} = \underline{0.85}$$

## **Loaded Questions**

If survey questions are not worded carefully, the results of a study can be misleading. Survey questions can be “loaded” or intentionally worded to elicit a desired response.

Too little money is being spent on “welfare” versus too little money is being spent on “assistance to the poor.” Results: 19% versus 63%

### ***Example***

See the following actual “yes” response rates for the different wordings of a question:

97% yes: “Should the President have the line item veto to eliminate waste?”

57% yes: “Should the President have the line item veto, or not?”

## Order of Questions

Questions are unintentionally loaded by such factors as the order of the items being considered.

### *Example*

These questions are from a poll conducted in Germany:

Would you say that traffic contributes more or less to air pollution than industry?

*Results:* 45% blamed traffic; 27% blamed industry.

Would you say that industry contributes more or less to air pollution than traffic?

*Results:* industry – 57%; traffic – 24%

## Nonresponse

A *nonresponse* occurs when someone either refuses to respond to a survey question or is unavailable. People who refuse to talk to pollsters have a view of the world around them that is markedly different than those who will let poll-takers into their homes.

## Missing Data

Results can sometimes be dramatically affected by missing result. Sometimes sample data values are missing because of random factors such subjects may drop out for reasons unrelated to the study, but some data are missing because of special factors, such as the tendency of people with low incomes are less likely to report their incomes. US Census suffers from missing people (tend to be homeless or low income).

## Self-Interest Study

Some parties with interest to promote will sponsor studies. For example, Kiwi Brands, a maker of shoe polish, commissioned a study that resulted printed in some newspapers.

Be wary of a survey in which the sponsor can enjoy monetary gain from the results. When assessing validity of a study, always consider whether the sponsor might influence the results.

## Precise Numbers

Because as a figure is precise, many people incorrectly assume that it is also *accurate*. A precise number can be an estimate, and it should be referred to that way.

## Deliberate Distortion

Some studies or surveys are distorted on purpose. The distortion can occur within the context of the data, the source of the data, the sampling method, or the conclusions.

## **Exercises**      **Section 1.3 – Critical Thinking**

1. Typical surveys involve about 500 people to 2000 people. When author Sheri Hire wrote *Woman and Love: A Cultural Revolution in Progress*, she based conclusions on a relatively large sample of 4500 replies that she received after mailing 100,000 questionnaires to various women's groups. Are her conclusions likely to be valid in the sense that they can be applied to the general population of all women? Why or why not?
2. Based on a study showing that college graduates tend to live longer than those who do not graduate from college, a researcher concludes that studying causes people to live longer. Use critical thinking to develop an alternative or correct conclusion.
3. In the judicial case *U.S. v. City of Chicago*, a minority group failed the Fire Captain Examination at a much higher rate than the majority group. Conclusion: The exam is biased and causes members of the minority group to fail at a much higher rate. Use critical thinking to develop an alternative or correct conclusion.
4. When Harris Interactive surveyed 1013 adults, 91 % of them said that they washed their hands after using a public restroom. But when 6336 adults were observed, it was found that 82% actually did wash their hands. How can we explain the discrepancy? Which percentage is more likely to accurately indicate the true rate at which people wash their hands in a public restroom?
5. The Internet service provider AOL ran a survey of its users and asked if they preferred a real Christmas tree or a false one. AOL received 7073 responses, and 4650 of them preferred a real tree. Given that 4650 is 66% of the 7073 responses; can we conclude that about 66% of people who observe Christmas prefer a real tree? Why or why not?
6. After the last national census was conducted, the *Poughkeepsie Journal* ran this front-page headline: "281,421,906 in America." What is wrong with this headline?
7. The *Statistical Abstract of the United States* includes the average per capita income for each of the 50 states. When those 50 values are added, then divided by 50, the result is \$29,672.52. Is \$29,672.52 the average per capita income for all individuals in the U.S.? Why or why not?
8. The author surveyed students with this request: "Enter your height in inches," Identify two major problems with this request.
9. Convert the fraction  $\frac{5}{8}$  to an equivalent percentage.
10. Convert the fraction  $\frac{227}{773}$  to an equivalent percentage. Express the answer to the near tenth of a percent.
11. Convert 23.4% to an equivalent decimal.



12. Convert 83% to an equivalent decimal.
13. What is 37% of 500?
14. What is 5% of 5020?
15. Convert 0.127 to an equivalent percentage.
16. Convert 0.045 to an equivalent percentage.
17. In a Gallup poll, 49% of 734 surveyed Internet users said that they shop on the Internet frequently or occasionally. What is the actual number of Internet users who said that they shop on the internet frequently or occasionally.
18. In a Gallup poll of 976 adults, 68 said that they have a drink every day. What is the percentage of respondents who said that they have a drink every day?
19. Among 976 adults surveyed, 32% said that they never drink. What is the actual number of surveyed adults who said that they never drink?
20. A New York Times editorial criticized a chart caption that described a dental rinse as one that “reduces plaque on teeth by over 300%.” What is wrong with that statement?
21. In an ad for the Club, a device used to discourage car thefts, it was stated that “The club reduces your odds of car theft by 400%.” What is wrong with this statement?

## Section 1.4 – Collecting Sample Data

### Basics of Collecting Data

Statistical methods are driven by the data that we collect. We typically obtain data from two distinct sources: *observational studies* and *experiment*.

**Observational study** observing and measuring specific characteristics without attempting to modify the subjects being studied

#### Example

The Literary Digest poll in which respondents were asked who they would vote for in the presidential election is an observational study. The subjects were asked for their choices, but they were not given any type of treatment.

**Experiment** apply some treatment and then observe its effects on the subjects; (subjects in experiments are called experimental units)

#### Example

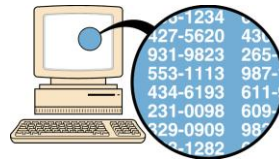
In the largest public health experiment ever conducted, 200,745 children were given a treatment consisting of the Salk vaccine, while 201,229 other children were given a placebo. The Salk vaccine injections constitute a treatment that modified the subjects, so this is an example of an experiment.

**Simple Random Sample** of  $n$  subjects selected in such a way that every possible sample of the same size  $n$  has the same chance of being chosen

**Random Sample** members from the population are selected in such a way that each individual member in the population has an equal chance of being selected

**Probability Sample** selecting members from a population in such a way that each member of the population has a known (but not necessarily the same) chance of being selected

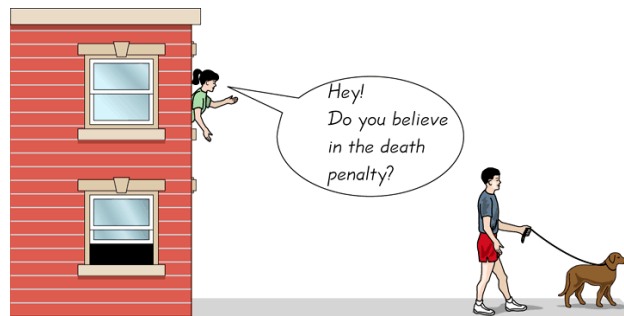
**Random Sampling** selection so that each individual member has an equal chance of being selected



**Systematic Sampling** Select some starting point and then select every  $k$ th element in the population



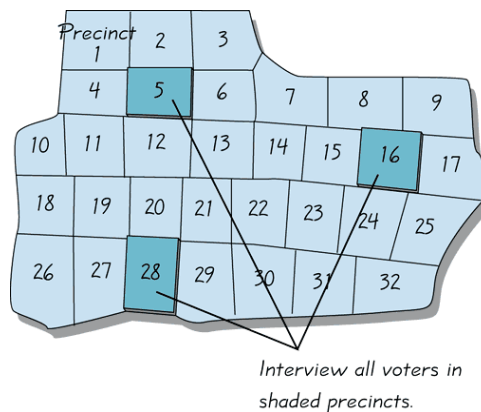
**Convenience Sampling** use results that are easy to get



**Stratified Sampling** subdivide the population into at least two different subgroups that share the same characteristics, then draw a sample from each subgroup (or stratum)



**Cluster Sampling** divide the population area into sections (or clusters); randomly select some of those clusters; choose all members from selected clusters



## Multistage Sampling

Collect data by using some combination of the basic sampling methods

In a multistage sample design, pollsters select a sample in different stages, and each stage might use different methods of sampling.

### *Example*

The U.S. government's unemployment statistics are based on surveyed households. It is impractical to personally visit each member of a simple random sample, because individual households would be spread all over the country. Instead, the U.S. Census Bureau and the Bureau of Labor Statistics combine to conduct a survey called the Current Population Survey. This survey obtains data describing such factors as unemployment rates, college enrollments, and weekly earning amounts. The survey incorporates a multistage sample design, roughly following these steps:

1. The surveys partition the entire United States into 2007 different regions called *primary sampling units* (PSU). The primary sampling units are metropolitan areas, large counties, or groups of smaller counties.
2. The surveyors select a sample of primary sampling units in each of the 50 states. For the Current Population Survey, 792 of the primary sampling units are used. (All of the 432 primary sampling units with the largest populations are used, and 360 primary sampling units are randomly selected from the other 1575.)
3. The surveyors partition each of the 792 selected primary sampling units into blocks, and they then use stratified sampling to select a sample of blocks.
4. In each selected block, surveyors identify clusters of households that are close to each other. They randomly select clusters, and they interview all households in the selected clusters.

This multistage sample design includes random, stratified, and cluster sampling at different stages. The end result is a complicated sampling design, but it is much more practical and less expensive than using a simpler design, such as using a simple random sample.

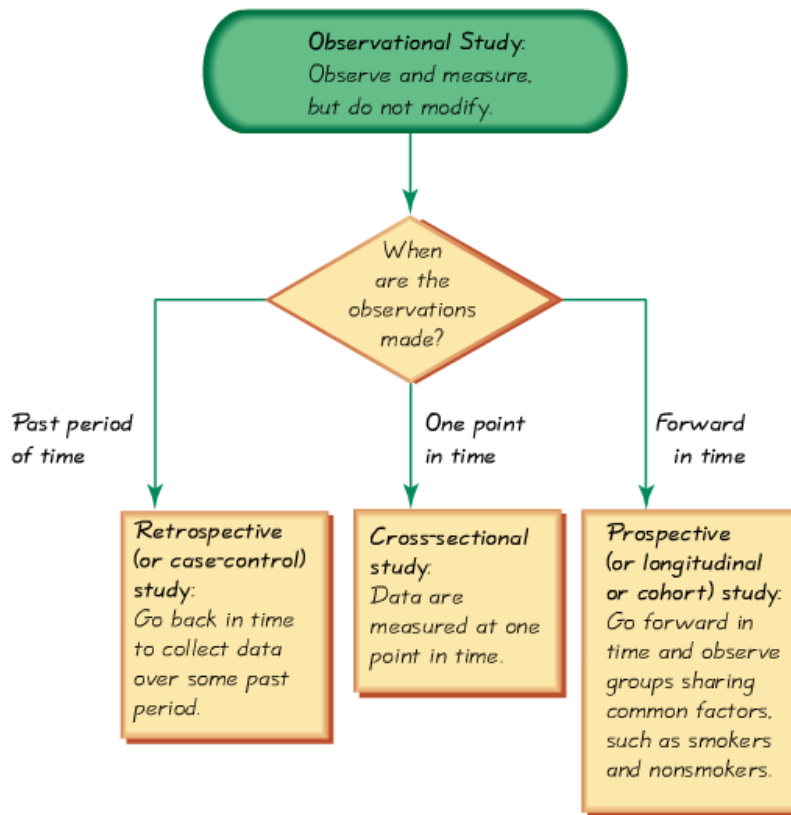
### *Definition*

In a ***cross-sectional study***, data are observed, measured, and collected at one point in time.

In a ***retrospective*** (or ***case-control***) **study**, data are collected from the past by going back in time (through examination of records, interviews, and so on).

In a ***prospective*** (or ***longitudinal*** or ***cohort***) **study**, data are collected in the future from groups sharing common factors (called cohorts).

## Design of Experiments



### Example

In 1954, a large-scale experiment was designed to test the effectiveness of the Salk vaccine in preventing polio, which had killed or paralyzed thousands of children. In that experiment, 200,745 children were given a treatment consisting of Salk vaccine injections, while a second group of 201,229 children were injected with a placebo that contained no drug. The children being injected did not know whether they were getting the Salk vaccine or the placebo. Children were assigned to the treatment or placebo group a process of random selection, equivalent to flipping a coin. Among the children given the Salk vaccine, 33 later developed paralytic polio, but among the children given a placebo, 115 later develop paralytic polio.

**Randomization** is used when subjects are assigned to different groups through a process of random selection. The logic is to use chance as a way to create two groups that are similar.

**Replication** is the repetition of an experiment on more than one subject. Samples should be large enough so that the erratic behavior that is characteristic of very small samples will not disguise the true effects of different treatments. It is used effectively when there are enough subjects to recognize the differences from different treatments.

Use a sample size that is large enough to let us see the true nature of any effects, and obtain the sample using an appropriate method, such as one based on *randomness*.

**Blinding** is a technique in which the subject doesn't know whether he or she is receiving a treatment or a placebo. Blinding allows us to determine whether the treatment effect is significantly different from a placebo effect, which occurs when an untreated subject reports improvement in symptoms.

**Double-Blind** : Blinding occurs at two levels:

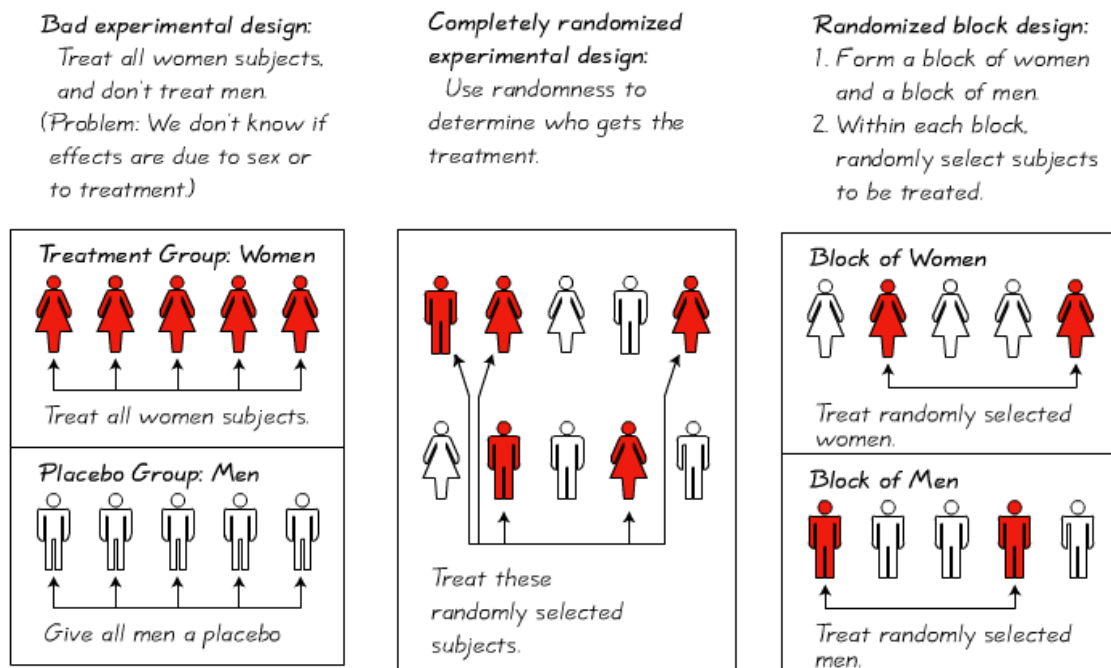
1. The subject doesn't know whether he or she is receiving the treatment or a placebo
2. The experimenter does not know whether he or she is administering the treatment or placebo

**Confounding** occurs in an experiment when the experimenter is not able to distinguish between the effects of different factors. Try to plan the experiment so that confounding does not occur.

**Completely Randomized Experimental Design** assign subjects to different treatment groups through a process of random selection

**Randomized Block Design** a block is a group of subjects that are similar, but blocks differ in ways that might affect the outcome of the experiment.

**Rigorously Controlled Design** carefully assign subjects to different treatment groups, so that those given each treatment are similar in ways that are important to the experiment



**Matched Pairs Design** compare exactly two treatment groups using subjects matched in pairs that are somehow related or have similar characteristics

## ***Summary***

Three very important considerations in the design of experiments are the following:

1. Use *randomization* to assign subjects to different groups
2. Use replication by repeating the experiment on enough subjects so that effects of treatment or other factors can be clearly seen.
3. *Control the effects of variables* by using such techniques as blinding and a completely randomized experimental design

## ***Errors***

No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.

***Sampling error*** the difference between a sample result and the true population result; such an error results from chance sample fluctuations

***Nonsampling error*** sample data incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)

## ***Exercises***     **Section 1.4 – Collecting Sample Data**

1. A Gallup poll surveyed 1018 adults by telephone, and 22% of them reported that they smoked cigarettes within the past year. Determine whether the description corresponds to an observation study or an experiment.
2. In a morally and criminally wrong study, 399 black men with syphilis were not given a treatment that could have cured them. The intent was to learn about the effects of syphilis on black men. The subjects were initially treated with small amounts of bismuth, neoarsphenamine, and mercury, but those treatments were replaced with aspirin. Determine whether the description corresponds to an observation study or an experiment.
3. A student of the author collected measurements of arm lengths from her family members. Identify what type is used: random, systematic, convenience, stratified, or cluster.
4. On the day of the last presidential election, ABC News organized an exit poll in which specific polling stations were randomly selected and all voters were surveyed as they left the premises. Identify what type is used: random, systematic, convenience, stratified, or cluster.
5. The author was an observer at a town of Poughkeepsic Police sobriety checkpoint at which every fifth driver was stopped and interviewed. (He witnessed the arrest of a former student.) Identify what type is used: random, systematic, convenience, stratified, or cluster.
6. You observed professional wine taster working at the Consumer's Union testing facility in NY. Assume that a taste test involves three different wines randomly selected from each of five different wineries. Identify what type is used: random, systematic, convenience, stratified, or cluster.
7. The U.S. Department of Corrections collects data about returning prisoners by randomly selecting five federal prisons and surveying all of the prisoners in each of the prisons. Identify what type is used: random, systematic, convenience, stratified, or cluster.
8. You instructor surveyed all of his students to obtain sample consisting of the number of credit cards students possess. Identify what type is used: random, systematic, convenience, stratified, or cluster.
9. In a study of college programs, 820 students are randomly selected from those majoring in communications, 1463 students are randomly selected from those majoring in business, and 760 students are randomly selected from those majoring in history. Identify what type is used: random, systematic, convenience, stratified, or cluster.
10. Pharmacists typically fill prescriptions by scooping a sample of pills from a larger batch that is in stock. A pharmacist thoroughly mixes a large batch of Lipitor pills, then selects 30 of them. Does this sampling plan result in a random sample? Simple random sample? Explain.
11. A quality control engineer selects every 10,000<sup>th</sup> M&M plain candy that is produced. Does this sampling plan result in a random sample? Simple random sample? Explain.



12. NBC News polled reactions to the last presidential election by surveying adults who were approached by a reporter at a location in N.Y. City. Does this sampling plan result in a random sample? Simple random sample? Explain.
13. A classroom consists of 36 students seated in six different rows, with six students in each row. The instructor rolls a die to determine a row, then rolls the die again to select a particular student in the row. This process is repeated until a sample of 6 students is obtained. Does this sampling plan result in a random sample? Simple random sample? Explain.
14. A computer company employs 100 software engineers and 100 hardware engineers. The personnel manager randomly selects 20 of the software engineers and 20 of the hardware engineers and questions them about career opportunities within the company. Does the sampling plan result in a random sample? Simple random sample? Explain.
15. A polling company obtains an alphabetical list of names of voters in a precinct. They select every 20<sup>th</sup> person from the list until a sample of 100 is obtained. They then call these 100 people. Does the sampling plan result in a random sample? Simple random sample? Explain.
16. What is an inherent zero? Describe three examples of data sets that have inherent zeros and three that do not.
17. What is the different between a random sample and a simple random sample?
18. Determine whether the statement is true or false. If false, rewrite it as a true statement
  - a) In a randomized block design, subjects with similar characteristics are divided into blocks, and then, within each block, randomly assigned to treatment groups.
  - b) Using a systematic sample guarantees that members of each group within a population will be sampled.
  - c) The method for selected a stratified sample is to order a population in some way and then select members of the population at regular intervals.
19. Which method of data collection should be used to collect data for the following study
  - a) A study of the health of 148 kidney transplant patients at a hospital.
  - b) A study of the effect on the taste of a snack food made with a sugar substitute
  - c) A study of how fast a virus would spread in a herd of cattle.
20. A pharmaceutical company wants to test the effectiveness of a new allergy drug. The company identifies 250 females 30-35 years old who suffer from severe allergies. The subjects are randomly assigned into two groups. One group is given the new allergy drug and the other is given a placebo that looks exactly like the new allergy drug. After six months, the subjects' symptoms are studied and compared
  - a) Identify the experimental units and treatment used in this experiment.
  - b) Identify a potential problem with the experiment design being used and suggest a way to improve it.
  - c) How could this experiment be designed to be a double-blind?

21. What type of sampling is used: random, stratified, convenience, cluster, systematic, in the following?
- a) Toyota wants to administer a satisfaction survey to its current customers. Using their customer database, the company randomly selects 80 customers and asks them about their level of satisfaction with the company
  - b) To determine her power usage, Dan divides up his day into three parts: morning, afternoon, and evening. He then measures his power usage at 3 randomly selected times during each part of the day.
  - c) A newspaper asks its readers to call in their opinion regarding the number of books they have read this month.
  - d) Toshiba wants to administer a satisfaction survey to its current customers. Using their customer database, the company randomly selects 80 customers and asks them about their level of satisfaction with the company.
  - e) An education researcher randomly selects 48 middle schools and interviews all the teachers at each school.
  - f) A market researcher selects 500 drivers under 30 years of age and 500 drivers over 30 years of age.
  - g) To avoid working late, a quality control analyst simply inspects the first 100 items produced in a day.
22. Determine whether you would take a census or use a sampling to collect data for the study described:
- a) The average credit card debt of the 65 employees of a company
  - b) The most popular grocery store among the 40,000 employees of a company
23. Determine if the survey question is biased. If the question is biased, suggest a better wording
- a) Why drinking fruit juice good for you?
  - b) Why is eating ice cream bad for you?
24. A company has been rating television programs for more than 60 years. It uses several sampling procedures, but its main one is to track the viewing patterns of 20,000 households. These contain more than 45,000 people and are chosen to form a cross section of the overall population. The households represent various locations, ethnic groups, and income brackets. The data gathered from the sample of 20,000 households are used to draw inferences about the population of all households in the U.S.
- a) What strata are used in the sample?
  - b) Why is it important to have a stratified sample for these ratings?
  - c) Observation studies are sometimes referred to as natural experiments. Explain what this means
25. Some polling agencies ask people to call a telephone number and give their response to a question
- a) What is an advantage of this type of survey?
  - b) What is disadvantage of this type of survey?
  - c) Identify the sampling technique used.

26. A computer company employs 100 software engineers and hardware engineers. The personnel manager randomly selects 20 of the software engineers and 20 of the hardware and questions them about career opportunities within the company. Does this sampling plan result in a random sample? Simple random sample? Explain.

## Section 1.5 – Frequency Distributions and Histograms

### Definition

A **frequency distribution** (or **frequency table**) shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number of data values in each of the categories.

### Example

Consider pulse rate measurements (in beats per minute) obtained from a simple random sample of 40 males and another simple random sample of 40 females, with the results listed in the table below.

**Pulse Rates (*beats per minute*) of Females and Males**

<b>Females</b>																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
<b>Males</b>																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

The frequency distribution summarizing the pulse rate of females listed in table below.

**Pulse Rates of Females**

<b>Pulse Rate</b>	<b>Frequency</b>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

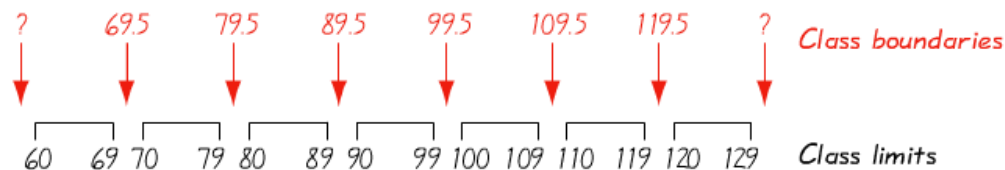
The frequency for a particular class is the number of original values that fall into that class. That is the frequency of 12, indicating that 12 of the original pulse rates are between 60 and 69 beats per minute.

### Definition

**Lower class limits** are the smallest numbers that can belong to the different classes (60, 70, 80, 90, 100, 110, and 120) (from previous table).

**Upper class limits** are the largest numbers that can belong to the different classes (table has upper class limits of 69, 79, 89, 99, 109, 119, 129.)

**Class boundaries** are the numbers used to separate the classes, but without the gaps created by the class limits. So the complete list of class boundaries is 59.5, 69.5, 79.5, 89.5, 99.5, 109.5, 119.5, 129.5.



**Class midpoints** are the values in the middle of the classes. From the table: 64.5, 74.5, 84.5, 94.5, 104.5, 114.5, and 124.5.) Each class midpoint is found by adding the lower class limits to the upper class limits and dividing the sum by 2.

Class width is the difference between two consecutive lower class limits or two consecutive lower class boundaries in a frequency distribution. (Table uses a class width of 10.)

## Procedure for Constructing a Frequency Distribution

The steps for constructing the frequency distributions are as follows:

1. Determine the number of classes. The number of classes should be between 5 and 20, and the number you select might be affected by the convenience of using round numbers.
2. Calculate the class width.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

Round this result to get a convenient number (usually round up). If necessary, change the number of classes so that they use convenient values.

3. Choose either the minimum data value or a convenient value below the minimum data value as the first lower class limit.
4. Using the first lower class limit and the class width, list the other lower class limits. (Add the class width to the first lower class limit to get the second lower class limit. Add the class width to the second lower class limit to get the third lower class limit, and so on.)
5. List the lower class limits in a vertical column and then enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

### Example

Using the pulse rate of females in previous table, follow the above procedure to construct the frequency distribution

#### Solution

**Step 1:** The number of desired classes is 7.

**Step 2:**  $\text{Class width} \approx \frac{124 - 60}{7} \approx 9.142857 \approx 10$

**Step 3:** The minimum data value is 60 and it is also a convenient number.

**Step 4:** Add the class width of 10 to 60 to get 70 (second lower class limit)

$70 + 10 = 80$ ,  $80 + 10 = 90$ ,  $90 + 10 = 100$ ,  $100 + 10 = 110$ , and  $110 + 10 = 120$ .

60 –
70 –
80 –
90 –
100 –
110 –
120 –

**Step 5:** List the lower class limits vertically as shown in the margin. From this list, we identify the corresponding upper class limits as 69, 79, 89, 99, 109, 119, and 129.

**Step 6:** Enter a tally mark for each data value in the appropriate class. Then add the tally marks to find the frequencies.

## Relative Frequency Distribution

A variation of the basic frequency distribution is a *relative frequency distribution*.

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequency}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequency}} \times 100\%$$

<i>Pulse Rate</i>	<i>Frequency</i>	<i>Relative Frequency</i>
60 – 69	12	$\frac{12}{40} \times 100\% = 30\%$
70 – 79	14	$\frac{14}{40} \times 100\% = 35\%$
80 – 89	11	$\frac{11}{40} \times 100\% = 27.5\%$
90 – 99	1	$\frac{1}{40} \times 100\% = 2.5\%$
100 – 109	1	$\frac{1}{40} \times 100\% = 2.5\%$
110 – 119	0	0
120 – 129	1	$\frac{1}{40} \times 100\% = 2.5\%$
	<b>40</b>	

## Cumulative Frequency Distribution

The *cumulative frequency* for a class is the sum of the frequencies for that class and all previous classes. The *cumulative frequency distribution* based on the frequency distribution.

<i>Pulse Rate</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
60 – 69	12	12
70 – 79	14	12 + 14 = 26
80 – 89	11	26 + 11 = 37
90 – 99	1	37 + 1 = 38
100 – 109	1	38 + 1 = 39
110 – 119	0	39 + 0 = 39
120 – 129	1	39 + 1 = 40

## Normal Distribution

- The *frequencies* start low, then increase to one or two high frequencies, then decrease to a low frequency.
- The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

### Example

IQ scores from 1000 adults were randomly selected. The results are summarized in the frequency distribution table

<i>IQ Score</i>	<i>Frequency</i>	<i>Normal Distribution</i>
50 – 69	24	← Frequencies start low
70 – 89	228	
90 – 109	490	← Increase to a maximum, ...
110 – 129	232	
130 – 149	26	← Decrease to become low again

The frequencies start low, then increase to a maximum frequency of 490, then decrease to low frequencies. Also, the frequencies are roughly symmetric about the maximum frequency of 490. It appears that the distribution is approximately a normal distribution.

### Example

The frequency distribution in the table summarizes the last digits of the pulse rates of females.

If the pulse rates are measured by counting the number of heartbeats in 1 minute, we expect that the last digits should occur with frequencies that are roughly the same. But note that the frequency distribution shows that the last digits are all even numbers; there are no odd numbers present. This suggests that the pulse rates were not counted for 1 minute. Upon further examination of the original pulse rates, we can see that every original value is multiply of four, suggesting that the number of heartbeats was counted for 15 seconds, then that count was multiplied by 4. It's fascinating and interesting that we are able to deduce something about the measurement procedure through an investigation of characteristics of the data.

<i>Last Digit</i>	<i>Frequency</i>
0	9
1	0
2	8
3	0
4	6
5	0
6	7
7	0
8	10
9	0

## Example

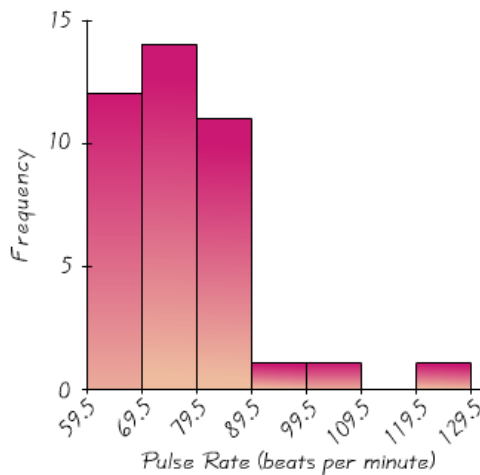
College undergraduate enrollments table shows the distribution of undergraduate college student enrollments among the four categories of colleges (based on data from the U.S. National Center for Education Statistic). The sum of the relative frequencies is 100.3%, which is slightly different from 100% because of rounding errors.

<i>College</i>	<i>Relative Frequency</i>
<i>Public 2 - Year</i>	36.8%
<i>Public 4 - Year</i>	40.0%
<i>Private 2 - Year</i>	1.6%
<i>Private 4 - Year</i>	21.9%

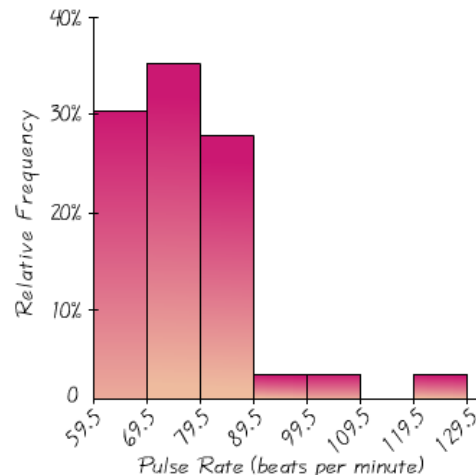
## Histograms

### Definition

A histogram is a graph consisting of bars of equal width drawn adjacent to each other (without gaps). The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values.



**Frequency Histogram**



**Relative Frequency Histogram**

**Horizontal Scale for Histogram:** Use class boundaries or class midpoints.

**Vertical Scale for Histogram:** Use the class frequencies.

### Relative Frequency Histogram

A relative frequency histogram has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies (as percentages or proportions) instead of actual frequencies.



## *Exercises*

## Section 1.5 – Frequency Distributions and Histograms

1. Identify the class width, class midpoints, and class boundaries for the given frequency distribution. Then construct the cumulative frequency distribution that corresponds to the frequency distribution.

<i>Tar (mg) in Nonfiltered Cigarettes</i>	<i>Frequency</i>
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

<i>Tar (mg) in Filtered Cigarettes</i>	<i>Frequency</i>
2 – 5	2
6 – 9	2
10 – 13	6
14 – 17	15

c)

<i>Weights (lb) of Discarded Metal</i>	<i>Frequency</i>
0.00 – 0.99	5
1.00 – 1.99	26
2.00 – 2.99	15
3.00 – 3.99	12
4.00 – 4.99	4

d)

<i>Weights (lb) of Discarded Plastic</i>	<i>Frequency</i>
0.00 – 0.99	14
1.00 – 1.99	20
2.00 – 2.99	1
3.00 – 3.99	4
4.00 – 4.99	2
5.00 – 5.99	1

2. In a study, researchers treated 570 people who smoke with either nicotine gum or a nicotine patch. Among those treated with nicotine gum, 191 continued to smoke and the other 59 stopped smoking. Among those treated with nicotine patch, 263 continued to smoke and the other 57 stopped smoking (based on data from the Center for Disease Control and Prevention). Construct the relative frequency distribution.
3. Heights of statistics students were obtained by the author as part of a study conducted for class. The last digits of those heights are listed below. Construct a frequency distribution with 10 classes. Based on the distribution, do the heights appear to be reported or actually measured? What do you know about the accuracy of the results?

00000000011233345555555555555555668889

4. Listed below are amounts of strontium-90 (in millibecquerels) in a simple random sample of baby teeth obtained from Pennsylvania residents born after 1979. Construct a frequency distribution with eight classes. Begin with a lower class limit of 110, and use a class width of 0. Cite a reason why such data are important.

155	142	149	130	151	163	151	142	156	133
138	161	128	144	172	137	151	166	147	163
145	116	136	158	114	165	169	145	150	150
150	158	151	145	152	140	170	129	188	156

5. Refer to the data below and use the 40 voltage measurements from the generator. Construct a frequency distribution with seven classes. Begin with a lower class limit of 123.9 volts, and use a class width of 0.20 volt. Using a very loose interpretation of the relevant criteria, does the result appear to have a normal distribution?

Day	Home	Generator	UPS	Day	Home	Generator	UPS
1	123.8	124.8	123.1	21	124.0	125.0	123.8
2	123.9	124.3	123.1	22	123.9	124.7	123.8
3	123.9	125.2	123.6	23	123.6	124.9	123.7
4	123.3	124.5	123.6	24	123.5	124.9	123.8
5	123.4	125.1	123.6	25	123.4	124.7	123.7
6	123.3	124.8	123.7	26	123.4	124.2	123.8
7	123.3	125.1	123.7	27	123.4	124.7	123.8
8	123.6	125.0	123.6	28	123.4	124.8	123.8
9	123.5	124.8	123.6	29	123.3	124.4	123.9
10	123.5	124.7	123.8	30	123.3	124.6	123.8
11	123.5	124.5	123.7	31	123.5	124.4	123.9
12	123.7	125.2	123.8	32	123.6	124.0	123.9
13	123.6	124.4	123.5	33	123.8	124.7	123.9
14	123.7	124.7	123.7	34	123.9	124.4	123.9
15	123.9	124.9	123.0	35	123.9	124.6	123.6
16	124.0	124.5	123.8	36	123.8	124.6	123.2
17	124.2	124.8	123.8	37	123.9	124.6	123.1
18	123.9	124.8	123.1	38	123.7	124.8	123.0
19	123.8	124.5	123.7	39	123.8	124.3	122.9
20	123.8	124.6	123.7	40	123.8	124.0	123.0

6. As part of the Garbage Project at the University of Arizona, the discarded garbage for 62 households was analyzed. Refers to the 62 weights from table below and construct a frequency distribution. Begin with a lower class of 1.00 lb., and use a class width of 4.00 lb. Do the weights of discarded paper appear to have a normal distribution?

2.41	11.08	9.45	5.88
7.57	12.43	12.32	8.26
9.55	6.05	20.12	12.45
8.82	13.61	7.72	10.58
8.72	6.98	6.16	5.87
6.96	14.33	7.98	8.78
6.83	13.31	9.64	11.03
11.42	3.27	8.08	12.29
16.08	6.67	10.99	20.58
6.38	17.65	13.11	12.56
13.05	12.73	3.26	9.92
11.36	9.83	1.65	3.45
15.09	16.39	10	9.09
2.8	6.33	8.96	3.69
6.44	9.19	9.46	2.61
5.86	9.41		

7. a) Refer to the data below for the FICO credit rating scores. Construct a frequency distribution beginning with a lower class limit of 400, and use a class width of 50. Does the result appear to have a normal distribution? Why or why not?

708	713	781	809	797	793	711	681	768	611	698	729	829
836	768	532	657	559	741	792	701	753	745	681	594	744
598	693	743	444	502	739	755	835	714	517	787	706	752
714	497	636	637	797	568	714	618	830	579	818	722	783
751	731	850	591	802	756	689	789	654	617	849	604	630
628	692	779	756	782	760	503	784	798	611	709	661	579
591	834	694	795	660	651	696	638	697	732	796	753	782
635	795	519	682	824	603	709	777	664				

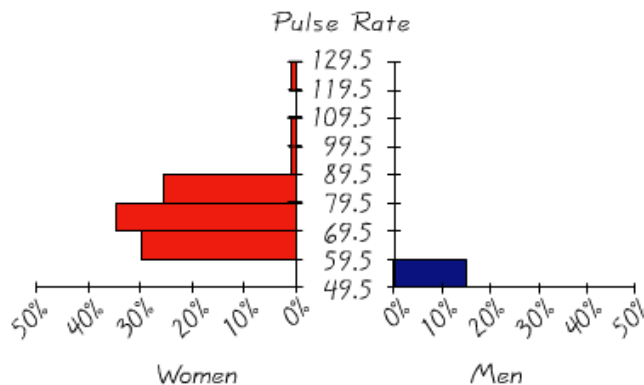
- b) Use the table to construct a histogram. Does the result appear to be normal distribution? Why or why not?

8. a) Refer to the data in the table below. Construct a frequency distribution. Begin with lower class limit of 6.0000 g, and use a class width of 0.0500 g.

6.2771	6.2371	6.1501	6.0002	6.1275	6.2151
6.2866	6.0760	6.1426	6.3415	6.1309	6.2412
6.1442	6.1073	6.1181	6.1352	6.2821	6.2647
6.2908	6.1661	6.2674	6.2718	6.1949	6.2465
6.3172	6.1487	6.0829	6.1423	6.1970	6.2441
6.3669	6.0775	6.1095	6.1787	6.2130	6.1947
6.1940	6.0257	6.1719	6.3278		

- b) Use the table to construct a histogram.

9. When using histograms to compare two data sets, it is sometimes difficult to make comparisons by looking back and forth between the two histograms. A back-to-back relative frequencies histogram uses a format that makes the comparison much easier. Instead of frequencies, we should use relative frequencies (percentages or proportions) so that the comparisons are not distorted by different sample sizes. Complete the back-to-back relative frequency histograms shown below by using the data below. Then use the result to compare the two data sets.



**Pulse Rates (*beats per minute*) of Females and Males**

<b><i>Females</i></b>																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
<b><i>Males</i></b>																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

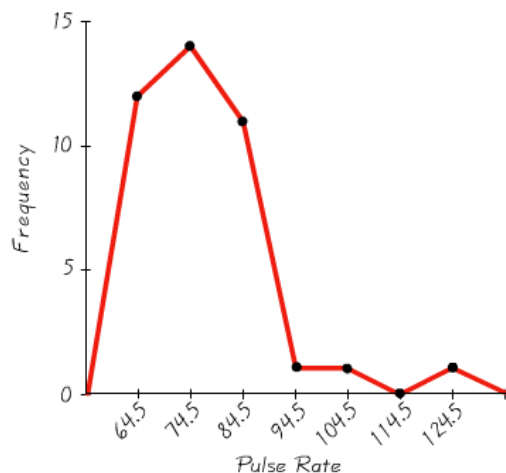
## Section 1.6 – Statistical Graphics & Bad Graphs

### Frequency Polygon

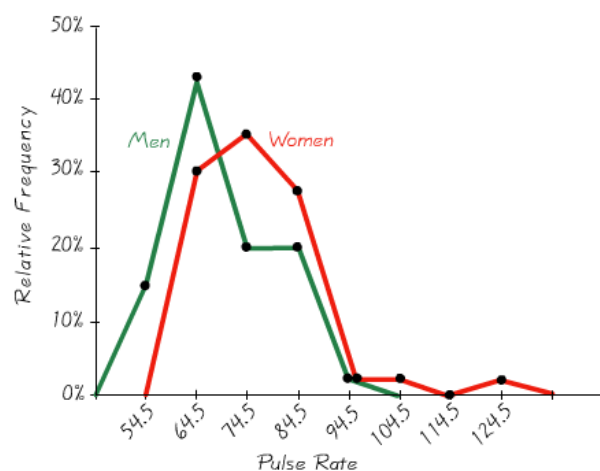
- A **frequency polygon** uses line segments connected to points located directly above class midpoint values.

### Example

For the frequency polygon corresponding to the pulse rates of women summarized in the frequency distribution. The heights of the points correspond to the class frequencies, and the line segments are extended to the right and left so that the graph begins and ends on the horizontal axis. Just as it is easy to construct a histogram from a frequency distribution table, it is also easy to construct a frequency polygon from a frequency distribution table.



**Frequency Polygon: Pulse Rates of Women**

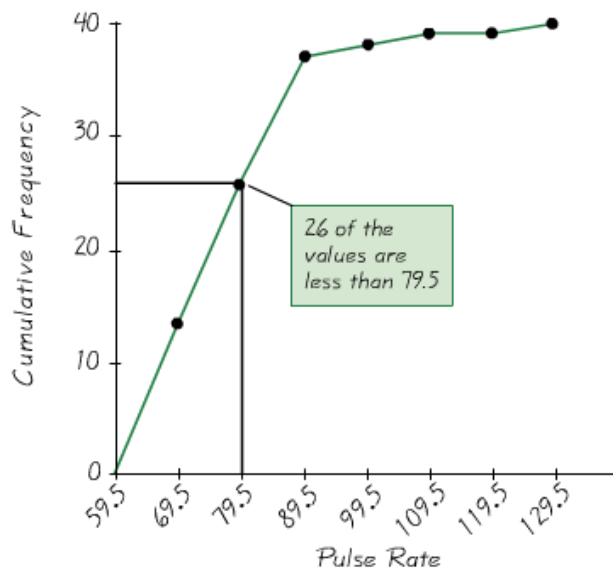


**Relative Frequency Polygons: Pulse Rates of Women and Men**

- A variation of the basic frequency polygon is the **relative frequency polygon**, which uses relative frequencies (proportions or percentages) for the vertical scale.

### Ogive (“oh-jive”)

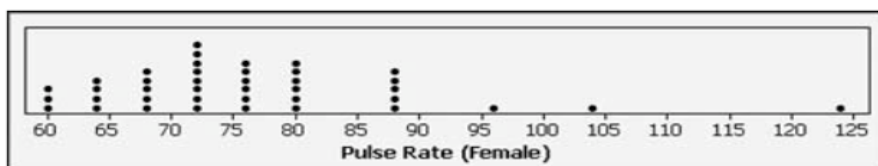
**Ogives** are useful for determining the number of values below some particular value. An **ogive** is a line graph that depicts *cumulative* frequencies. An ogive uses class boundaries along the horizontal scale, and cumulative frequencies along the vertical scale.



*Orgive*

## Dot Plots

A dotplot consists of a graph in which each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.



## Stemplots

A *stemplot* (or *stem-and-leaf plot*) represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit)

**Example** of pulse

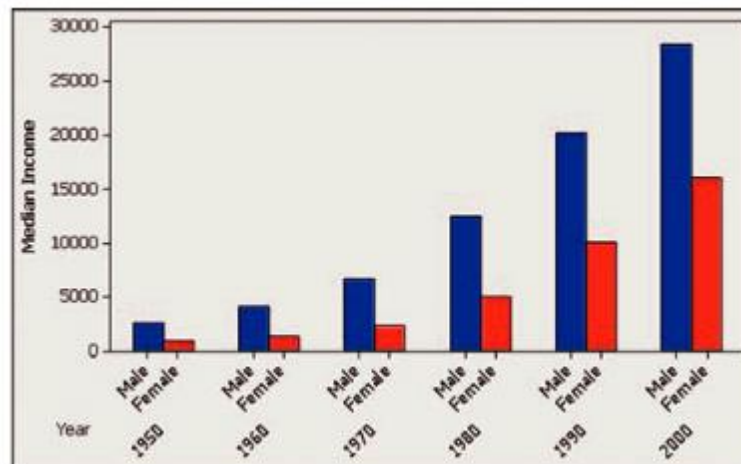
<i>Stem</i> (tens)	<i>Leaves</i> (units)	
6	000444488888	← Data values are 60, 60, 60, 64, 64, ..., 68
7	22222222666666	← Data values are 72, 72, ...
8	00000088888	← Data values are 80, 80, ....
9	6	← Data value is 96
10	4	← Data value is 104
11		
12	4	← Data value is 124

## Bar Graph

Uses bars of equal width to show frequencies of categories of qualitative data. Vertical scale represents frequencies or relative frequencies. Horizontal scale identifies the different categories of qualitative data. A **multiple bar graph** has two or more sets of bars, and is used to compare two or more data sets.

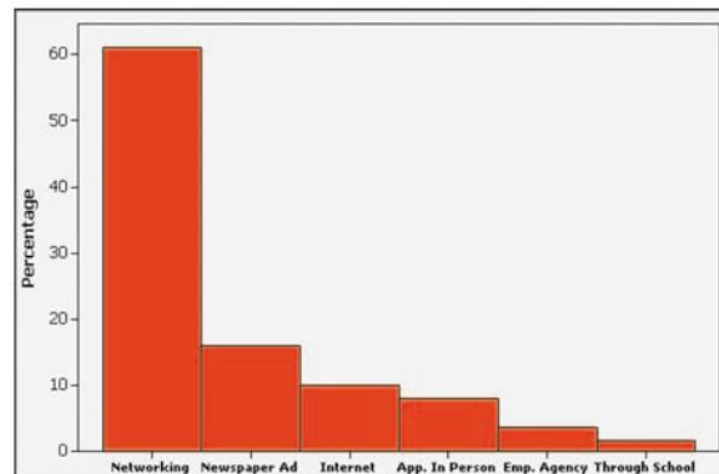
### *Example*

Median Income of Males and Females



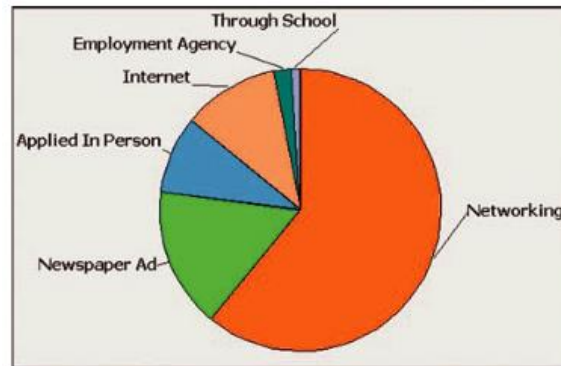
## Pareto Chart

A bar graph for qualitative data, with the bars arranged in descending order according to frequencies



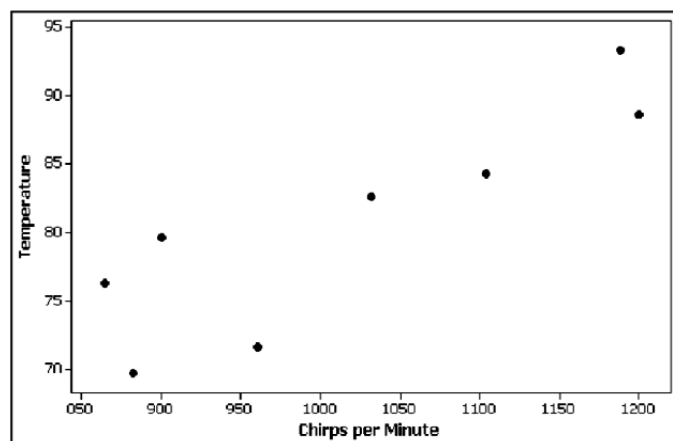
## Pie Chart

A graph depicting qualitative data as slices of a circle, size of slice is proportional to frequency count



## Scatter Plot (or Scatter Diagram)

A plot of paired  $(x,y)$  data with a horizontal  $x$ -axis and a vertical  $y$ -axis. Used to determine whether there is a relationship between the two variables

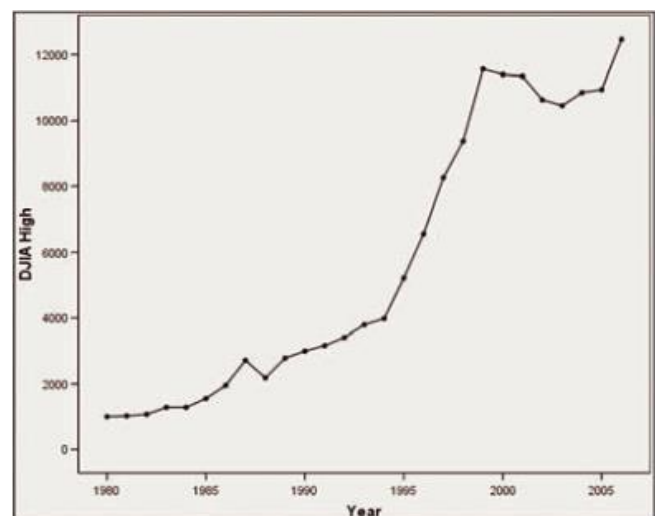


## Time-Series Graph

Data that have been collected at different points in time: *time-series data*.

### Example

The accompanying SPSS-generated time-series graph shows the yearly high values of the Dow Jones Industrial Average (DJIA) for the N.Y. Stock Exchange. This graph shows a steady increase between the years 1980 and 2007, but the DJIA high values have not been so consistent in more recent years.





## Bad Graphs

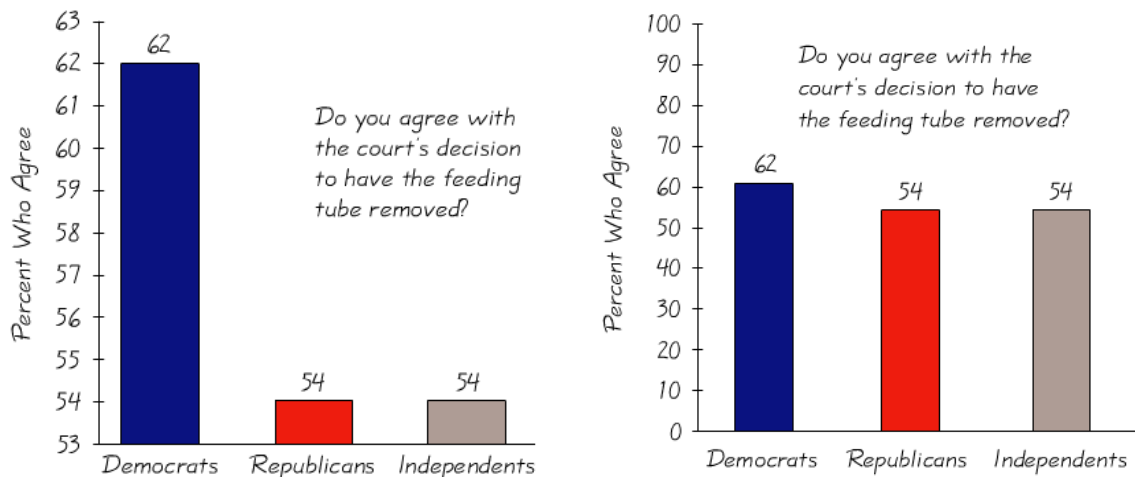
Some graphs are bad in the sense that they contain errors. Some are bad because they are technically correct, but misleading. It is important to develop the ability to recognize bad graphs and identify exactly how they are misleading.

### Nonzero Axis

Some graphs are misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.

### Example

The results of a CNN poll regarding the case of Schiavo is as shown in graph below



**Survey Results by Party**

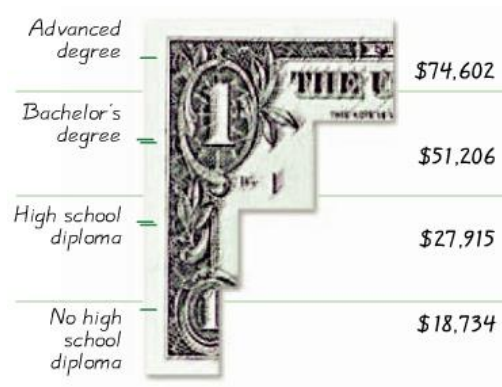
This graph (on the left) creates the incorrect impression that significantly more Democrats agreed with the court's decision than Republicans or Independents. Since graph depicts the data objectively, it creates the more correct impression that the differences are not very substantial. Many people complained that it was deceptive, so CNN posted a modified graph similar to figure on the right.

### Pictographs

Drawings of objects, called pictographs, are often misleading. Three-dimensional objects - money bags, stacks of coins, army tanks (for army expenditures), people (for population sizes), barrels (for oil production), and houses (for home construction) are commonly used to depict data. These drawings can create false impressions that distort the data.

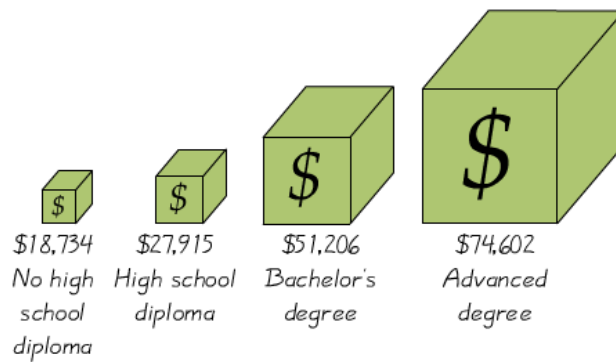
If you double each side of a square, the area does not merely double; it increases by a factor of four; if you double each side of a cube, the volume does not merely double; it increases by a factor of eight. Pictographs using areas and volumes can therefore be very misleading.

### Example



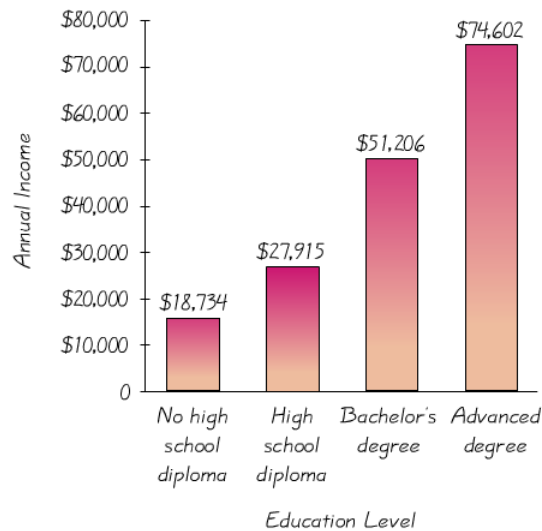
This picture is not misleading because the bars have same width, but it is somewhat too busy, and too difficult to understand.

### Example



It depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.

### Example



It depicts the data in a fair, objective, unencumbered by distracting features.

## Exercises Section 1.6 – Statistical Graphics & Bad Graphs

1. Given listed amounts of Strontium-90 (in millibecquerels) in a simple random sample of baby teeth.

155 142 149 130 151 163 151 142 156 133 138 161 128 144 172  
 137 151 166 147 163 145 116 136 158 114 165 169 145 150 150  
 150 158 151 145 152 140 170 129 188 156

- Construct a dot plot of the amounts of Strontium-90. What does the dot plot suggest about the distribution of those amounts?
  - Construct a stemplot of the amounts of Strontium-90. What does the stemplot suggest about the distribution of those amounts?
  - Construct a frequency polygon of the amounts of Strontium-90. For the horizontal axis, use the midpoints of the class intervals in the frequency distribution: 110-119, 120-129, 130-139, ..., 180-189.
  - Construct an ogive of the amounts of Strontium-90. For the horizontal axis, use the class boundaries corresponding to the class limits. How many of the amounts are below 150 millibecquerels?
2. Use the 62 weights of discarded plastic listed in Data set below

0.27 1.41 2.19 2.83 2.19 1.81 0.85 3.05 3.42 2.10 2.93 2.44 2.17 1.41 2.00  
 0.93 2.97 2.04 0.65 2.13 0.63 1.53 4.69 0.15 1.45 2.68 3.53 1.49 2.31 0.92  
 0.89 0.80 0.72 2.66 4.37 0.92 1.40 1.45 1.68 1.53 1.44 1.44 1.36 0.38 1.74  
 2.35 2.30 1.14 2.88 2.13 5.28 1.48 3.36 2.83 2.87 2.96 1.61 1.58 1.15 1.28  
 0.58 0.74

- Construct a dot plot of the weights of discarded plastic. What does the dot plot suggest about the distribution of the weights?
  - Construct a stemplot of the weights of discarded plastic. What does the stemplot suggest about the distribution of the weights?
  - Construct a frequency polygon of the weights of discarded plastic. For the horizontal axis, use the midpoints of the class intervals: 0.00-0.99, 1.00-1.99, 2.00-2.99, 3.00-3.99, 4.00-4.99, 5.00-5.99.
  - Construct an ogive of the weights of discarded plastic. For the horizontal axis, use these classes boundaries: -0.005, 0.995, 1.995, 2.995, 3.995, 4.995, 5.995. How many of the weights are below 4 lb.?
3. In 1965, Intel cofounder Gordon Moore proposed what has since become known as Moore's law: the number of transistors per square inch on integrated circuits with double approximately every 18 months. The table below lists the number of transistors per square inch (in thousands) for several different years. Construct a time-series graph of the data.

Year	1971	1974	1978	1982	1985	1989	1993	1997	1999	2000	2002	2003
Transistors	2.3	5	29	120	275	1180	3100	7500	24,000	42,000	220,000	410,000

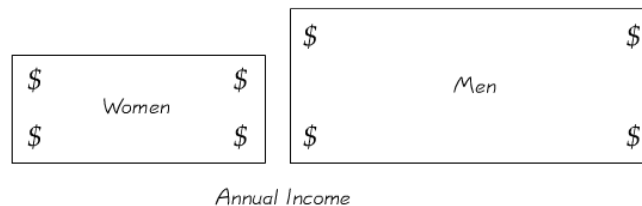
4. The following table shows the numbers of cell phone subscriptions (in thousands) in the U.S. for various years. Construct a time-series graph of the data. “Linear” growth would result in a graph that is approximately a straight line. Does the time-series graph appear to show linear growth?

Year	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003	2005
Number	340	1231	3509	7557	16,009	33,786	55,312	86,047	128,375	158,722	207,900

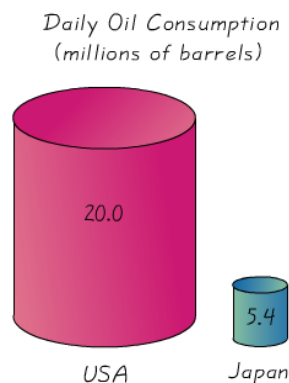
5. The following table lists the marriage and divorce rates per 1000 people in the U.S. for selected years since 1900 (based on data from the Department of Health and Human Services). Construct a multiple bar graph of the data. Why do these data consist of marriage and divorce rates rather than total numbers of marriages and divorces? Comment on any trends that you observe in these rates, and give explanations for these trends.

Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Marriage	9.3	10.3	12.0	9.2	12.1	11.1	8.5	10.6	10.6	9.8	8.3
Divorce	0.7	0.9	1.6	1.6	2.0	2.6	2.2	3.5	5.2	4.7	4.2

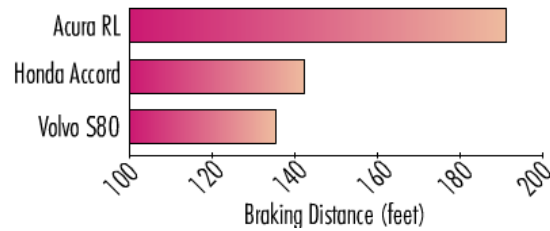
6. Assume that, as a newspaper reporter, you must graph data showing that increased smoking causes an increased risk of lung cancer. Given that people might be helped and lives might be saved by creating a graph that exaggerates the risk of lung cancer, is it ethical to construct such a graph?
7. The accompanying graph depicts average full-time incomes of women and men aged 18 and over. For a recent year, those incomes were \$37,197 for women and \$53,059 for men (based on data from the U.S. Census Bureau). Does the graph make a fair comparison of the data? Why or why not? If the graph distorts the data, construct a fair graph.



8. The accompanying graph uses cylinders to represent barrels of oil consumed by the U.S. and Japan. Does the graph distort the data or does it depict the data fairly? Why or why not? If the graph distorts the data, construct a graph that depicts the data fairly.



9. The accompanying graph shows the braking distances for different cars measured under the same conditions. Describe the ways in which this graph might be deceptive. How much greater is the braking distance of the Acura RL than the braking distance of the Volvo S80? Draw the graph in a way that depicts the data more fairly.



10. Use the data to create a stemplot

The midterm test scores for the seventh-period typing class are listed below

85 77 93 91 74 65 68 97 88 59 74 83 85 72 63 79

11. Use the data to create a stemplot

Twenty-four workers were surveyed about how long it takes them to travel to work each day. The data below are given in minutes

20 35 42 52 65 20 60 49 24 37 23 24  
22 20 41 25 28 27 50 47 58 30 32 48

12. Find the original data from the stemplot

<i>Stem</i>	<i>Leaves</i>
76	2 6 7
77	2 4 9
78	1 7

## Section 1.7 – Measures of Center

### Characteristics of center

Measures of center include mean and median, as tools for analyzing data. Not only determine the value of each measure of center, but also interpret those values.

### Definition

A *measure of center* is a value at the center or middle of a data set

### Mean

#### Definition

The *arithmetic mean*, or the *mean*, of a set of data is the measure of center found by adding the data values and dividing the total by the number of data values. (is also called the *average*)

$$\text{mean} = \frac{\sum x}{n} \quad \begin{array}{l} \leftarrow \text{sum of all data values} \\ \leftarrow \text{number of data values} \end{array}$$

$$\bar{x} = \frac{\sum x}{n} \quad \text{is the mean of a set of } \textit{sample values}.$$

$$\mu = \frac{\sum x}{N} \quad \text{is the mean of all values in a } \textit{population}.$$

### Notation

$\Sigma$  denotes the *sum* of a set of values.

$x$  is the *variable* usually used to represent the individual data values.

$n$  represents the *number of data values* in a *sample*. (*sample size*)

$N$  represents the *number of data values* in a *population*.

### Example

Find the mean of these first five word counts from men: 27,531; 15,684; 5,638; 27,997; and 25,433

#### Solution

$$\begin{aligned} \bar{x} &= \frac{\sum x}{n} = \frac{27,531 + 15,684 + 5,638 + 27,997 + 25,433}{5} \\ &= \frac{102,283}{5} \\ &= 20,456.6 \end{aligned}$$

The mean of the first five word counts is 20,456.6 words.

## Median

### Definition

The **median** of a data set is the measure of center that is the **middle value** when the original data values are arranged in order of increasing (or decreasing) magnitude. The median is often denoted by  $\tilde{x}$  (*x-tilde*)

### Finding the Median

First **sort** the values (arrange them in order), then follow one of these

1. If the number of data values is odd, the median is the number located in the exact middle of the list.
2. If the number of data values is even, the median is found by computing the mean of the two middle numbers.

In order of **even** number of values: 5.40; 1.10; 0.42; 0.73; 0.48; 1.10

$$\begin{array}{cccccc} 0.42 & 0.48 & 0.73 & 1.10 & 1.10 & 5.40 \\ & & \uparrow & \uparrow & & \\ & & \text{Median} = \frac{0.73 + 1.10}{2} = \underline{0.195} \end{array}$$

In order of **odd** number of values: 5.40; 1.10; 0.42; 0.73; 0.48; 1.10; 0.66

$$\begin{array}{ccccccc} 0.42 & 0.48 & 0.66 & 0.73 & 1.10 & 1.10 & 5.40 \\ & & & \uparrow & & & \\ & & & \text{Median} = \underline{0.73} \end{array}$$

### Example

Find the median for this sample of data values: 27,531; 15,684; 5,638; 27,997; and 25,433

#### Solution

First sort the data: 5,638    15,684    25,433    27,531    27,997

Median is 25,433

### Example

Find the median for this sample of data values: 27,531, 15,684, 5,638, 27,997, 25,433 and 8,077

#### Solution

First sort the data: 5,638    8,077    15,684    25,433    27,531    27,997

Median is  $= \frac{15,684 + 25,433}{2} = \underline{20,558.5}$

## Mode

### Definition

The **mode** of a data set is the value that occurs with the greatest frequency

A data set can have one, more than one, or no mode

<b>Bimodal</b>	two data values occur with the same greatest frequency
<b>Multimodal</b>	more than two data values occur with the same greatest frequency
<b>No Mode</b>	no data value is repeated

### Example

a) Find the mode of: 5.40 1.10 0.42 0.73 0.48 1.10  
Mode is 1.10

b) Find the mode of: 27 27 27 55 55 55 88 88 99  
Mode is 27 & 55 (**bimodal**)

c) Find the mode of: 1 2 3 4 5 6 7  
No Mode

## Midrange

### Definition

The midrange is the value midway between the maximum and minimum values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by 2:

$$\text{Midpoint} = \text{Midrange} = \frac{\text{minimum data value} + \text{maximum data value}}{2}$$

### Example

Find the midrange of these values: 27,531; 15,684; 5,638; 27,997; and 25,433

### Solution

$$\begin{aligned}\text{Midrange} &= \frac{\text{minimum data value} + \text{maximum data value}}{2} \\ &= \frac{5,638 + 27,997}{2} \\ &= \underline{16,817.5}\end{aligned}$$



## Critical Thinking

- Think about whether the results are reasonable.
- Think about the method used to collect the sample data.

### Example

For each of the following, identify a major reason why the mean and median are not meaningful statistics

- Zip codes: 1260, 77573, 77574, 90210, 77550
- Ranks of stress levels from different jobs: 2, 3, 1, 7, 9

### Solution

- The zip codes don't measure or count anything. The numbers are actually labels for geographic locations.
- The ranks reflect an ordering, but they don't measure or count anything. The rank of 1 might come from a job that has a stress level substantially greater than the stress level from the job with a rank of 2, so the different numbers don't correspond to the magnitude of the stress levels.

## Beyond the Basics of Measures of Center

### Mean from a Frequency Distribution

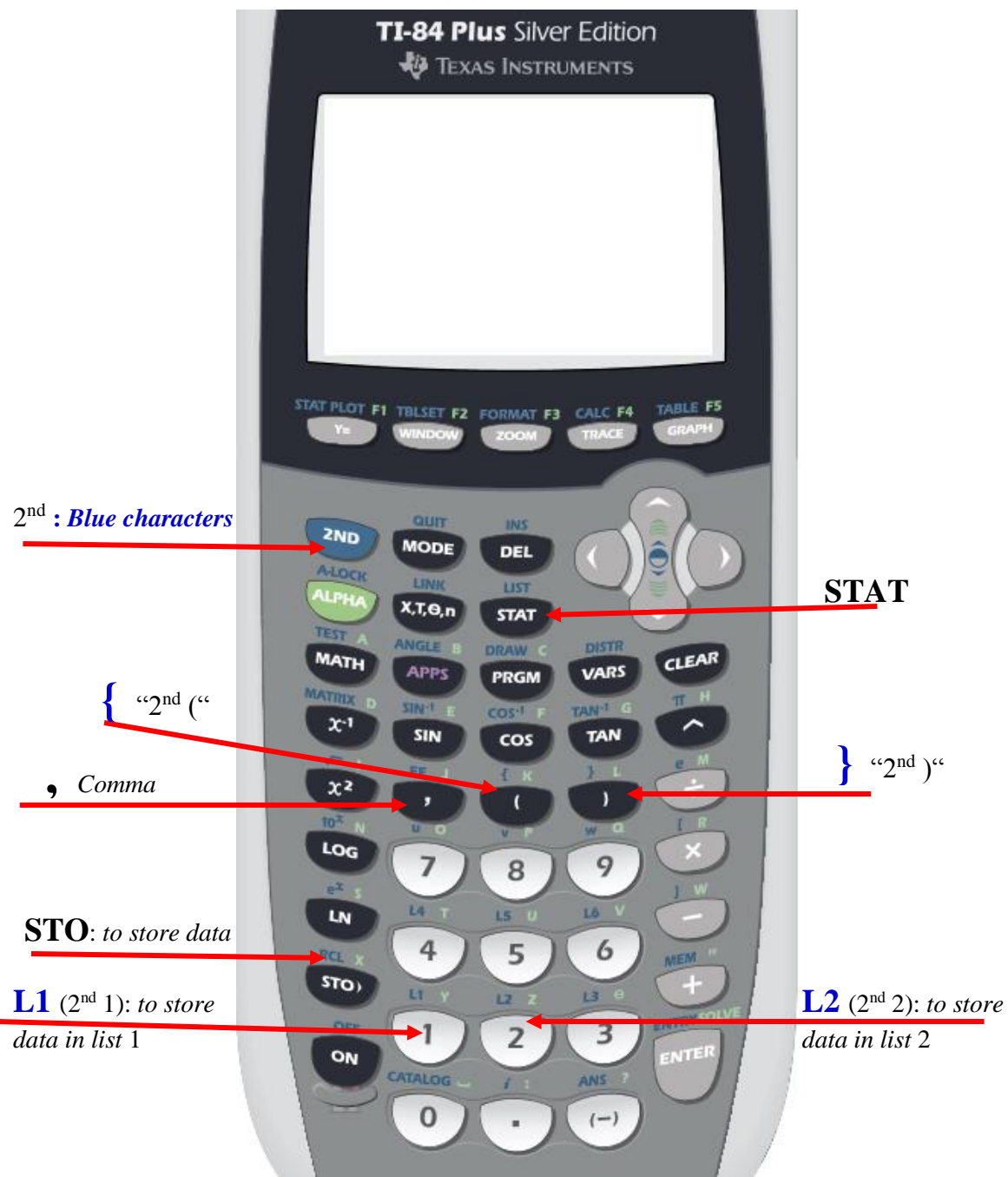
Assume that all sample values in each class are equal to the class midpoint.

*multiply each frequency and class midpoint, then add the products*

Mean from frequency distribution:  $\bar{x} = \frac{\sum (f \cdot x)}{\sum f}$  ← sum of frequencies

### Example

Word Counts	Frequency $f$	Class Midpoint $x$	$f \cdot x$
0 – 9,999	45	4,999.5	229,977.
10,000 – 19,999	90	14,999.5	1,349,995
20,000 – 29,999	40	24,999.5	999,980
30,000 – 39,999	7	34,999.5	244,996.5
40,000 – 49,999	3	44,999.5	134,996.5
Totals:	$\Sigma f = 186$		$\Sigma f \cdot x = 2,959,907$
			$\bar{x} = \frac{\sum (f \cdot x)}{\sum f} = \frac{2,959,907}{186} = \underline{15,913.5}$



```

0000 CALC TESTS
01Edit...
02:SortA(
03:SortD(
04:ClrList
05:SetUpEditor
  
```





L1	L2
4995.5	45.0
15000	90.0
25000	40.0
35000	7.0
45000	3.0
-----	-----
L2(1)=45	



```

EDIT 0000 TESTS
011-Var Stats
  
```

**OR**

To store values: click 2<sup>nd</sup>  – “(“  ‘{‘




Type number follow by comma, after you are done entering all the numbers.

Click 2<sup>nd</sup> then “)” to close ‘}’

Click STO (to store the numbers in a list).


Click 2<sup>nd</sup> then 1 for list 1 (L1)

2 for list 2 (L2)

Click on  → CALC → 1   1-Var Stats

```
1-Var Stats
x=15972.5
Σx=2954907.5
Σx²=6.1E10
Sx=8668.1
σx=8644.6
↓n=185.0
```

 **Mean Value**

## Weighted Mean

When data values are assigned different weights, we can compute a **weighted mean**.

$$\text{weighted mean: } \bar{x} = \frac{\sum (w \cdot x)}{\sum w}$$

### Example

In her first semester of college, a student of the author took five courses. Her final grades along with the number of credits for each course were: *A* (3 credits); *A* (4 credits); *B* (3 credits); *C* (3 credits) and *F* (1 credit). The grading system assigns quality points to letter grades as follows: *A* = 4; *B* = 3; *C* = 2; *D* = 1; *F* = 0. Compute her grade point average.

### Solution

Weights = number of credits:  $w = 3, 4, 3, 3, 1$ .

Replace A, B, C, D, and F with the corresponding quality points:  $x = 4, 4, 3, 2, 0$ .

$$\begin{aligned}\bar{x} &= \frac{\sum (w \cdot x)}{\sum w} \\ &= \frac{(3 \times 4) + (4 \times 4) + (3 \times 3) + (3 \times 2) + (1 \times 0)}{3 + 4 + 3 + 3 + 1} \\ &= \frac{43}{14} \\ &= \underline{3.07}\end{aligned}$$

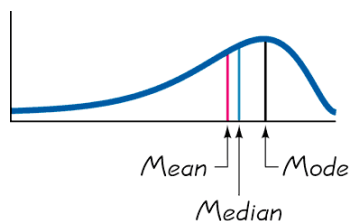
## Skewed and Symmetric

### Definition

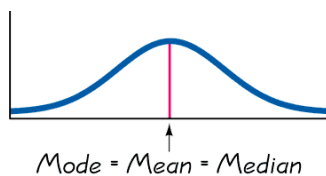
A distribution of data is **skewed** if it is not symmetric and extends more to one side than the other.

A distribution of data is **symmetric** if the left half of its histogram is roughly a mirror image of its right half.

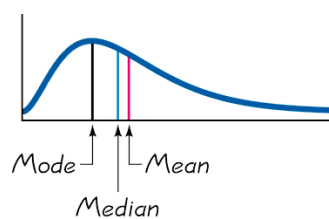
The mean and median cannot always be used to identify the shape of the distribution.



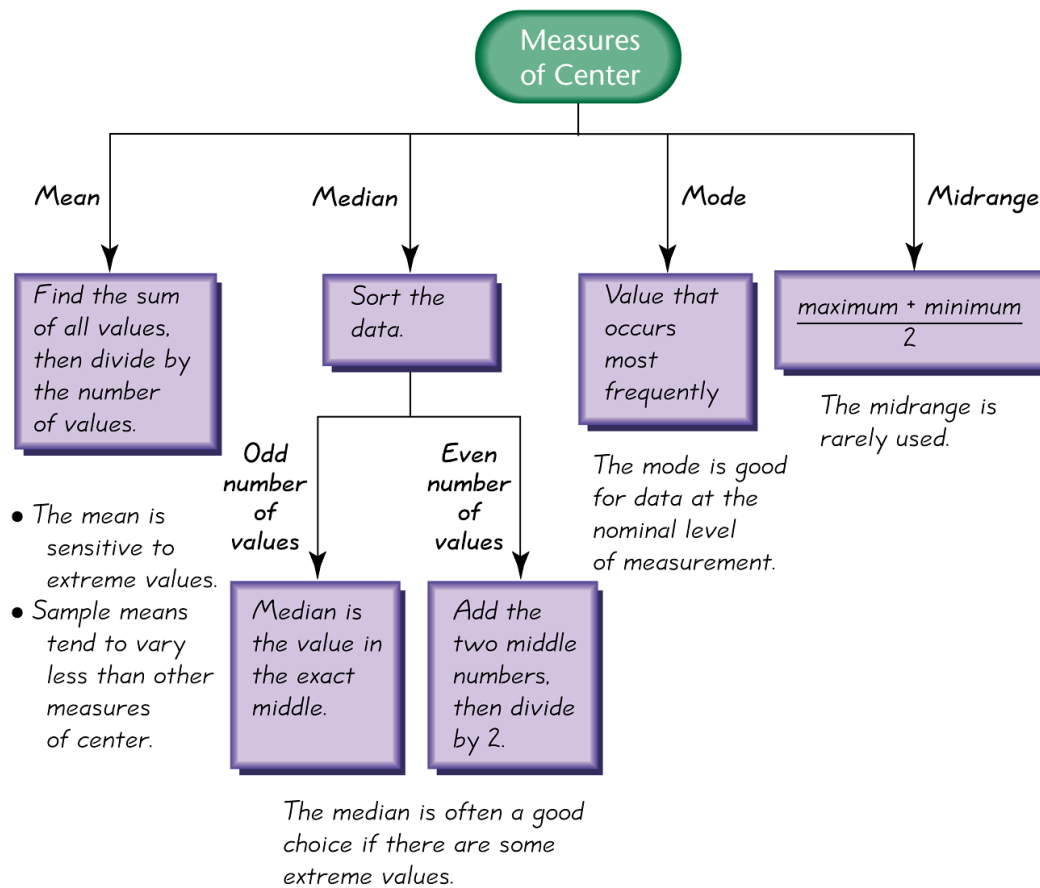
*Skewed to the Left (Negatively)*



*Symmetric*



*Skewed to the Right (Positively)*



## **Exercises**      **Section 1.7 – Measures of Center**

1. In what sense are the mean, median, mode and midrange measures of “center”?
2. A headline in USA Today stated that “Average family income drops 2.3%.” What is the role of the term average in statistics? Should another term be used in place of average?
3. In an editorial, the Poughkeepsie Journal printed this statement: “The median price – the price exactly in between the highest and lowest -- ...” Does that statement correctly describes the median? Why or why not?
4. A simple random sample of pages from Merriam-Webster’s Collegiate Dictionary, 11th edition, was obtained. Listed below are the numbers of words defined on those pages. Given that this dictionary has 1459 pages with defined words, estimate the total number of defined words in the dictionary.  
51   63   36   43   34   62   73   39   53   79  
Find the   a) mean   b) median   c) mode   d) midrange  
e) Is that estimate likely to be an accurate estimate of the number of words in the English language?
5. The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those tests, with the measurements given in hic (standard head injury condition units).  
774   249   1210   546   431   612  
Find the   a) mean   b) median   c) mode   d) midrange  
e) According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?
6. The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below  
\$7448   \$4911   \$9051   \$6374   \$4277  
Find the   a) mean   b) median   c) mode   d) midrange  
e) Do the different measures of center differ very much?
7. Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA’s Space Transport System (space shuttle).  
73   95   235   192   165   262   191   376   259   235   381   331   221   244   0  
Find the   a) mean   b) median   c) mode   d) midrange  
e) How might that duration time be explained?

8. Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.

448 242 231 246 246 293 280 227 213 262 239 213 258 255 257

Find the a) mean b) median c) mode d) midrange

e) Is there on time that is very different from the others?

9. Listed below are numbers of satellites in orbit from different countries.

158 17 15 17 7 3 5 1 8 3 4 2 4 1 2 3 1 1 1 1 1 1 1 1

Find the a) mean b) median c) mode d) midrange

e) Does on country have an exceptional number of satellites?

f) Can you guess which country has the most satellites?

10. Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

a) Find the mean and median for each then compare the two sets of results.

b) Does it make much difference if the tickets are purchased 30 days in advance or 1 day in advance?

11. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (from recent winners)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the mean and median for each then compare the two sets of results.

12. Find the mean of the data summarized in the given frequency distribution.

a)

<i>Tar (mg) in Nonfiltered Cigarettes</i>	<i>Frequency</i>
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

<i>Pulse Rates of Females</i>	<i>Frequency</i>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

13. A student of the author earned grades of B, C, B, A, and D. Those courses has these corresponding numbers credit hours: 3, 3, 4, 4, and 1. The grading system assigns quality points to letter grades as follows: A = 4; B = 3; C = 2; D = 1; F = 0. Compute the grade point average (GPA) and round the result with two decimal places. If the Dean's list requires a GPA 3.00 or greater, did this student make the Dean's list?
14. A student of the author earned grades of 92, 83, 77, 84, and 82 on her five regular tests. She earned grades of 88 on the final exam and 95 on her class projects. Her combined homework grade was 77. The five regular tests count for 60% of the final grade, the final exam counts for 10%, the project counts for 15%, and homework counts for 15%. What is her weighted mean grade? What letter grade did she earn? (A, B, C, D, or F)
15. You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework). What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?
16. During a quality assurance check, the actual coffee contents (in ounces) of six jars of instant coffee were recorded as 6.03, 5.59, 6.40, 6.00, 5.99, and 6.02.
- Find the mean and the median of the coffee content.
  - The third value was incorrectly measured and is actually 6.04. Find the mean and median of the coffee content again.
  - Which measure of central tendency, the mean or the median, was affected more by the data entry error?
17. The table below shows the U.S. exports (in billions of dollars) to 19 countries for a recent year.

<b><i>U.S. Exports</i></b> (in billions of dollars)		
Canada: 261.1	Mexico: 151.2	Germany: 54.5
Taiwan: 24.9	Netherlands: 39.7	China: 69.7
Australia: 22.2	Malaysia: 12.9	Switzerland: 22.0
Saudi Arabia: 12.5	United Kingdom: 53.6	Japan: 65.1
South Korea: 34.7	Singapore: 27.9	France: 28.8
Brazil: 32.3	Belgium: 28.9	Italy: 15.5
Thailand: 9.1		

- Find the mean and the median.
- Find the mean and median without the U.S. exports to Canada. Which measure of central tendency, the mean or the median, was affected more by the elimination of the Canadian exports?
- The U.S. Exports to India were \$17.7 billion. Find the mean and median with the Indian exports added to the original data set. Which measure of central tendency was affected more by adding the Indian exports?



## Section 1.8 – Measures of Variation

### Basic Concepts of Variation

#### Range

##### *Definition*

The **range** of a set of data values is the difference between the maximum data value and the minimum data value.

$$\text{Range} = (\text{maximum data value}) - (\text{minimum data value})$$

*It is very sensitive to extreme values; therefore not as useful as other measures of variation.*

##### *Example*

India has 1 satellite used for military and intelligence purposes, Japan has 3, and Russia has 14. Find the range of the sample values of 1, 3, and 14.

##### Solution

$$\begin{aligned}\text{Range} &= (\text{maximum data value}) - (\text{minimum data value}) \\ &= 14 - 1 \\ &= 13.0\end{aligned}$$

##### *Definition*

The **standard deviation** of a set of sample values, denoted by  $s$ , is a measure of variation of values about the mean. It is a type of average deviation of values from the mean that is calculated.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad \text{Sample standard deviation}$$

$$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n - 1)}} \quad \text{Sample standard deviation}$$

### Standard Deviation - Important Properties

- The standard deviation is a measure of variation of all values from the mean.
- The value of the standard deviation  $s$  is usually positive.
- The value of the standard deviation  $s$  can increase dramatically with the inclusion of one or more outliers (data values far away from all others).
- The units of the standard deviation  $s$  are the same as the units of the original data values.

### Example

Find the standard deviation of the numbers: 7, 9, 18, 22, 27, 29, 32, 40.

**Solution**

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

$$= \sqrt{\frac{5132 - 8(23)^2}{8-1}}$$

$$\approx 11.34$$



## Standard Deviation of a Population

The standard deviation  $\sigma$  (lowercase sigma) of a population is given by the formula

Population standard deviation  $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$

## Variance of a Sample and a Population

### Definition

The variance of a set of values is a measure of variation equal to the square of the standard deviation

Sample variance:  $s^2$  square of the standard deviation  $s$ .

Population variance:  $\sigma^2$  square of the population standard deviation  $\sigma$ .

## Notation

$s =$  *sample* standard deviation

$$s^2 = \text{sample variance}$$

$\sigma$  = *population* standard deviation

$$\sigma^2 = \text{population variance}$$

## Unbiased Estimator

The sample variance  $s^2$  is an ***unbiased estimator*** of the population variance  $\sigma^2$ , which means values of  $s^2$  tend to target the value of  $\sigma^2$  instead of systematically tending to overestimate or underestimate  $\sigma^2$ .

## Using and Understanding Standard Deviation

**Range Rule of Thumb** is based on the principle that for many data sets, the vast majority (such as 95%) of sample values lie within two standard deviations of the mean

### *Interpreting a Known Value of the Standard Deviation*

Informally define **usual** values in a data set to be those that are typical and not too extreme. Find rough estimates of the minimum and maximum “usual” sample values as follows:

$$\text{Minimum “usual” value} = (\text{mean}) - 2 \times (\text{standard deviation})$$

$$\text{Maximum “usual” value} = (\text{mean}) + 2 \times (\text{standard deviation})$$

### **Estimating a Value of the Standard Deviation $s$**

To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{\text{range}}{4}$$

Where range = (maximum value) – (minimum value)

### ***Example***

The Wechsler Adult intelligence Scale involves an IQ test designed so that the mean score is 100 and the standard deviation is 15. Use the range rule thumb to find the minimum and maximum “usual” IQ scores. Then determine whether an IQ score of 135 would be considered “unusual”

### **Solution**

$$\text{Mean} = 100$$

$$\text{Standard deviation} = 15$$

$$\begin{aligned}\text{Minimum “usual” value} &= (\text{mean}) - 2 \times (\text{standard deviation}) \\ &= 100 - 2(15) \\ &= 70\end{aligned}$$

$$\begin{aligned}\text{Maximum “usual” value} &= (\text{mean}) + 2 \times (\text{standard deviation}) \\ &= 100 + 2(15) \\ &= 130\end{aligned}$$

### Example

Use the range of thumb to estimate the standard deviation of the sample of 100 FICO credit rating scores listed in the table below.

708	713	781	809	797	793	711	681	768	611	698	836	768
532	657	559	741	792	701	753	745	681	598	693	743	444
502	739	755	835	714	517	787	714	497	636	637	797	568
714	618	830	579	818	654	617	849	798	751	731	850	591
802	756	689	789	628	692	779	756	782	760	503	784	591
834	694	795	660	651	696	638	635	795	519	682	824	603
709	777	829	744	752	783	630	753	661	604	729	722	706
594	664	782	579	796	611	709	697	732				

### Solution

Those scores have a minimum of 444 and a maximum of 850.

$$\begin{aligned}s &\approx \frac{\text{range}}{4} \\&= \frac{850 - 444}{4} \\&= 101.5\end{aligned}$$

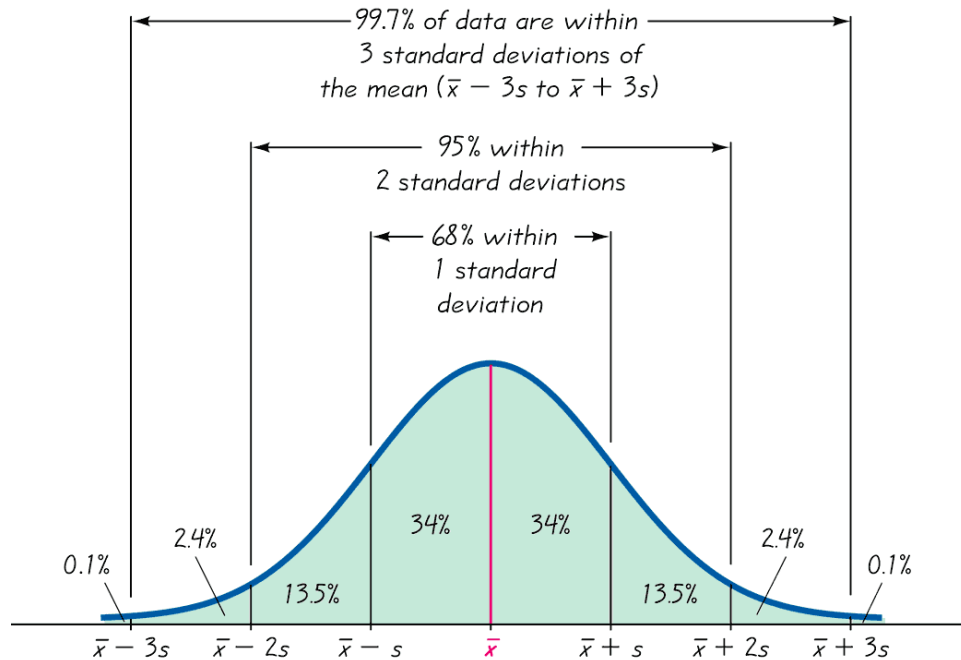
### Properties of the Standard Deviation

- ✓ The standard deviation measures the **variation** among data values
- ✓ Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- ✓ Has the same units of measurement as the original data
- ✓ For many data sets, a value is *unusual* if it differs from the mean by more than two standard deviations
- ✓ When comparing variation in two different data sets, compare the standard deviation only if they use the same scale and units, and they have means that are approximately the same.

## Empirical (or 68-95-99.7) Rule

Another concept that is helpful in interpreting the value of a standard deviation is the *empirical rule*. For data sets having a distribution that is approximately bell shaped, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.



### Example

Empirical Rule IQ scores have a bell-shaped distribution with mean of 100 and a standard deviation of 15. What percentages of IQ scores are between 70 and 130?

### Solution

$$130 = 100 + 15 + 15$$

$$70 = 100 - 15 - 15$$

70 and 130 are each exactly 2 standard deviation away from the mean 100.

$$2 \text{ standard deviation} = 2s = 2(15) = 30$$

Therefore, 2 standard deviation from the mean is

$$100 - 30 = 70$$

$$100 + 30 = 130$$

The empirical rule tells us that about 95% of all values are within 2 standard deviation of the mean, so about 95% of all IQ scores are between 70 and 130.

## Chebyshev's Theorem

The proportion (or fraction) of any set of data lying within  $K$  standard deviations of the mean is always at least  $1 - \frac{1}{K^2}$ , where  $K$  is any positive number greater than 1.

- For  $K = 2$ , at least  $3/4$  (or 75%) of all values lie within 2 standard deviations of the mean.
- For  $K = 3$ , at least  $8/9$  (or 89%) of all values lie within 3 standard deviations of the mean.

### *Example*

Chebyshev's Theorem IQ scores have a mean of 100 and a standard deviation of 15. What can we conclude from Chebyshev's Theorem?

### Solution

We can conclude that:

At least  $\frac{3}{4}$  (or 75%) of IQ scores are within 2 standard deviation of the mean (between 70 and 130).

At least  $\frac{8}{9}$  (or 89%) of IQ scores are within 3 standard deviation of the mean (between 55 and 145).

## Standard Deviation Defined

For a particular data value of  $x$ , the amount of deviation is  $x - \bar{x}$ . Those deviations could be a negative numbers, and the sum could be zero. To get statistic that measures variation (instead of always zero), we need to avoid canceling out of negative and positive numbers. We can get the ***mean absolute deviation*** (or **MAD**), which is the mean distance of the data from the mean:

$$\text{mean absolute deviation} = \frac{\sum |x - \bar{x}|}{n}$$

## Definition

The **coefficient of variation** (or **CV**) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

$$\text{Sample} \\ CV = \frac{s}{\bar{x}} \cdot 100\%$$

$$\text{Population} \\ CV = \frac{\sigma}{\mu} \cdot 100\%$$

## Example

Compare the variation in heights of men to the variation in weights of men, using these sample results obtained from data below

Men heights

70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6	63.0	68.3	73.1	67.6
68.0	71.0	61.3	76.2	66.3	69.7	65.4	70.0	62.9	68.5	68.3	69.4	69.2	68.0
71.9	66.1	72.4	73.0	68.0	68.7	70.3	63.7	71.1	65.6	68.3	66.3		

Men weights

169.1	144.2	179.3	175.8	152.6	166.8	135.0	201.5	175.2	139.0	156.3	186.6	191.1
151.3	209.4	237.1	176.7	220.6	166.1	137.4	164.2	162.4	151.8	144.1	204.6	193.8
172.9	161.9	174.8	169.8	213.3	198.0	173.3	214.5	137.1	119.5	189.1	164.7	170.1
151.0												

## Solution

The heights yield:  $\bar{x} = 68.34$  in. and  $s = 3.02$  in.

The weights yield:  $\bar{x} = 172.55$  lb. and  $s = 26.33$  lb.

```
1-Var Stats
x̄=68.335
Σx=2733.400
Σx²=187142.480
sx=3.020
σx=2.982
n=40.000
```

**Heights**

$$\text{Heights} \quad CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{3.02}{68.34} \cdot 100\% = \underline{4.42\%}$$

$$\text{Weights} \quad CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{26.33}{172.55} \cdot 100\% = \underline{15.26\%}$$

We can see that heights (with  $CV = 4.42\%$ ) have considerably less variation than weights (with  $CV = 15.26\%$ ). This makes intuitive sense, because the weights among men vary much more than heights.

It is very rare to see two adult men with one of them being twice as tall as the other, but it is much more common to see two men with one of them weighing twice as much as the other.

## **Exercises**      **Section 1.8 – Measures of Variation**

1. In statistics, how do variation and variance differ?
2. Collegiate Dictionary has 1459 pages of defined words. Listed below are the numbers of defined words per page for a simple random sample of those pages. If we use this sample as a basis for estimating the total number of defined words in the dictionary.  
51   63   36   43   34   62   73   39   53   79
  - a) Find the range, variance, and standard deviation.
  - b) How does the variation of these numbers affect our confidence on the accuracy of the estimate?
3. The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those tests, with the measurements given in hic (standard head injury condition units).  
774   249   1210   546   431   612
  - a) Find the range, variance, and standard deviation
  - b) According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?
4. The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below  
\$7448   \$4911   \$9051   \$6374   \$4277
  - a) Find the range, variance, and standard deviation
  - b) Do the different measures of center differ very much?
5. Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA's Space Transport System (space shuttle).  
73   95   235   192   165   262   191   376   259   235   381   331   221   244   0
  - a) Find the range, variance, and standard deviation
  - b) How might that duration time be explained?
6. Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.  
448   242   231   246   246   293   280   227   213   262   239   213   258   255   257
  - a) Find the range, variance, and standard deviation
  - b) Is there on time that is very different from the others?
7. Listed below are numbers of satellites in orbit from different countries.  
158   17   15   17   7   3   5   1   8   3   4   2   4   1   2   3   1   1   1   1   1   1   1
  - a) Find the range, variance, and standard deviation
  - b) Does on country have an exceptional number of satellites?



8. Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

Find the coefficient of variation for each of the two sets of data, then compare the variation.

9. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (from recent winners)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the coefficient of variation for each of the two sets of data, then compare the variation.

10. Find the Standard Deviation from the frequency distribution and find the standard deviation of sample summarized in a frequency distribution table by using the formula

$$s = \sqrt{\frac{n \left[ \sum (f \cdot x^2) \right] - \left[ \sum (f \cdot x) \right]^2}{n(n-1)}}, \text{ where } x \text{ represents the class midpoint, } f \text{ represents the class frequency, and } n \text{ represents the total number of sample values.}$$

a)

<b><i>Tar (mg) in Nonfiltered Cigarettes</i></b>	<b><i>Frequency</i></b>
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

<b><i>Pulse Rates of Females</i></b>	<b><i>Frequency</i></b>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

11. Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using the empirical rule, what is the approximate percentage of women between
- 154 cm and 168 cm?
  - 147 cm and 175 cm?
12. The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts, and the voltages have a bell-shaped distribution. Using the empirical rule, what is the approximate percentage of voltage amounts between
- 124.4 volts and 125.6 volts?
  - 124.1 volts and 125.9 volts?

13. The mean value of land and buildings per acre from a sample of farms is \$1,500, with a standard deviation of \$200. Using the empirical rule, estimate the percent of farms whose land and building values per acre are between \$1,300 and \$1,700. (Assume the data set has a bell-shaped distribution.)
14. The mean value of land and buildings per acre from a sample of farms is \$2,400, with a standard deviation of \$450. Using the empirical rule, between what two values do about 95% of the data lie? (Assume the data set has a bell-shaped distribution.)
15. Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using Chebyshev's Theorem, what do we know about the percentage of women with heights that are within 2 standard deviations of the mean? What are the minimum and maximum heights that are within 2 standard deviations of the mean?
16. The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts. Using Chebyshev's Theorem, what do we know about the percentage of voltage amounts that are within 3 standard deviations of the mean? What are the minimum and maximum voltage amounts that are within 3 standard deviations of the mean?
17. The mean time in a women's 400-meter dash is 57.07 seconds, with a standard deviation of 1.05 seconds. Apply Chebyshev's Theorem to the data using  $k = 2$ . Interpret the results.
18. The number of gallons of water consumed per day by a small village are listed. Make a frequency distribution (using five classes) for the data set. Then approximate the population mean and the population standard deviation of the data set.  

167	180	192	173	145	151	174	175	178	160
195	224	244	146	162	146	177	163	149	188
19. To get the best deal on a microwave oven, Jeremy called six appliance stores and asked the cost of a specific model. The prices he was quoted are listed below:  

\$325	\$384	\$156	\$210	\$219	\$284
-------	-------	-------	-------	-------	-------

Find the variance for the given data.
20. Compare the variation in heights to the variation in weights of thirteen-year old girls. The heights (in inches) and weights (in pounds) of nine randomly selected thirteen-year old girls as listed below  

Heights (inches):	59.3	61.2	62.6	64.7	60.1	58.3	64.6	63.7	66.1
Weights (pounds):	87	96	91	119	96	90	123	98	139

Find the coefficient of variation for each of the two sets of data, then compare the variation
21. The amount of Jen's monthly phone bill is normally distributed with a mean of \$56 and a standard deviation of \$9. What percentage of her phone bills are between \$29 and \$83? Use the empirical rule to solve.

## Section 1.9 – Measures of Relative Standing and Boxplots

This section introduces measures of relative standing, which numbers are showing the location of data values relative to the other values within a data set. They can be used to compare values from different data sets, or to compare values within the same data set. The most important concept is the  $z$  score. We will also discuss percentiles and quartiles, as well as a new statistical graph called the boxplot.

### Definition

A  $z$  score (or standardized value) is the number of standard deviations that a given value  $x$  is above or below the mean. The  $z$  score is calculated by using one of the following:

<i>Sample</i>	<i>Population</i>
$z = \frac{x - \bar{x}}{s}$	$z = \frac{x - \mu}{\sigma}$

### Example

Compare those two data values by finding  $z$  score.

Men heights

70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6	63.0	68.3	73.1	67.6
68.0	71.0	61.3	76.2	66.3	69.7	65.4	70.0	62.9	68.5	68.3	69.4	69.2	68.0
71.9	66.1	72.4	73.0	68.0	68.7	70.3	63.7	71.1	65.6	68.3	66.3		

Men weights

169.1	144.2	179.3	175.8	152.6	166.8	135.0	201.5	175.2	139.0	156.3	186.6	191.1
151.3	209.4	237.1	176.7	220.6	166.1	137.4	164.2	162.4	151.8	144.1	204.6	193.8
172.9	161.9	174.8	169.8	213.3	198.0	173.3	214.5	137.1	119.5	189.1	164.7	170.1
151.0												

### Solution

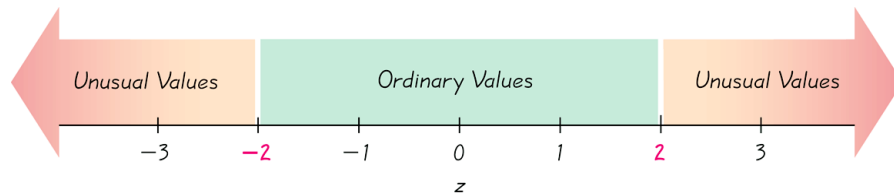
The heights:  $\bar{x} = 68.34$  in. and  $s = 3.02$  in.

$$z = \frac{x - \bar{x}}{s} = \frac{76.2 - 68.34}{3.02} = 2.60$$

The weights:  $\bar{x} = 172.55$  lb. and  $s = 26.33$  lb.

$$z = \frac{x - \bar{x}}{s} = \frac{237.1 - 172.55}{26.33} = 2.45$$

## Z Scores, Unusual Values, and Outliers



Whenever a value is less than the mean, its corresponding  $z$  score is negative

**Ordinary values:**  $-2 \leq z \text{ score} \leq 2$

**Unusual Values:**  $z \text{ score} < -2$  or  $z \text{ score} > 2$

## Percentiles

### Definition

Percentiles are measures of location. There are 99 percentiles denoted  $P_1, P_2, \dots, P_{99}$ , which divide a set of data into 100 groups with about 1% of the values in each group.

### Finding the Percentile of a Data Value

The process of finding the percentile that corresponds to a particular data value  $x$  is given by the following:

$$\text{percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

**(Round to the nearest whole number)**

### Example

The table below lists the 35 sorted budget amounts (in millions of dollars) from the simple random sample of movies. Find the percentile for the value of \$29 million

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

### Solution

From the table, there are 6 budget amounts less than 29, so

$$\text{percentile of } 29 = \frac{6}{35} \cdot 100 \approx 17$$

- ✓ The budget amount of \$29 million is the 17<sup>th</sup> percentile. This can be interpreted loosely as: The budget amount of \$29 million separates the lowest 17% of the budget amounts from the highest 83%.

## Notation

$$L = \frac{k}{100} \cdot n$$

- $n$  total number of values in the data set  
 $k$  percentile being used  
 $L$  locator that gives the position of a value  
 $P_k$   $k$ th percentile

## Example

The table below lists the 35 sorted budget amounts (in millions of dollars) from the simple random sample of movies. Find the value of the 90<sup>th</sup> percentile,  $P_{90}$

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

## Solution

$k = 90$  and  $n = 35$  because there are 35 data values.

$$\begin{aligned} L &= \frac{k}{100} \cdot n \\ &= \frac{90}{100} \cdot 35 \\ &\approx 32 \end{aligned}$$

The 32nd value is 150 that is,  $P_{90} = \$150$  million.

So, about 90% of the movies have budgets below \$150 million and about 10% of the movies have budgets above \$150 million.

## Example

The list of setting speed limits are recorded (in mi/h.) and listed below

68	68	72	73	65	74	73	72	68	65	65	73	66	71
68	74	66	71	65	73	59	75	70	56	66	75	68	75
62	72	60	73	61	75	58	74	60	73	58	74		

That section has a posted speed limit of 65 mi/h. Traffic engineers often establish speed limits by using the “85<sup>th</sup> percentile rule” whereby the speed limit is set so that 85% of drivers are at or below the speed limit.

- Find the 85<sup>th</sup> percentile of the listed speeds.
- Given that speed limits are usually rounded to a multiple of 5, what speed limit is suggested by these data? Explain your choice.
- Does the existing speed limit conform to the 85th percentile rule?

### Solution

Sorting the data

56	58	58	59	60	60	61	62	65	65	65	65	66	66	66	68	68	68	68	68
70	71	71	72	72	72	73	73	73	73	73	73	74	74	74	74	75	75	75	75

a)  $n = 40$  because there are 40 sample. To find the 85th percentile, then  $k = 85$ .

$$L = \frac{k}{100} \cdot n = \frac{85}{100} \cdot 40 = 34$$

That indicated that the 85<sup>th</sup> percentile is 34<sup>th</sup> speeds, the 34<sup>th</sup> speed is 74 mi/h.

b) A speed of 75 mi/h is the multiple of 5 closest to the 85<sup>th</sup> percentile, but it is probably safer to round down, so that a speed of 70 mi/h is the closest multiple of 5 below the 85<sup>th</sup> percentile.

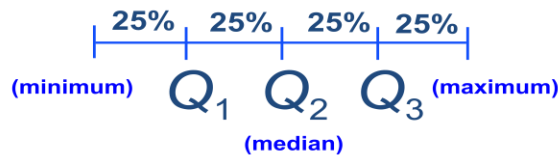
c) The existing speed limit of 65 mi/h is below the speed limit determined by the 85<sup>th</sup> percentile rule, so the existing speed limit does not conform to the 85<sup>th</sup> percentile rule.

### Quartiles

#### Definition

Quartile are measures of location, denoted  $Q_1$ ,  $Q_2$ , and  $Q_3$ , which divide a set of data into four groups with about 25% of the values in each group.

- $Q_1$  (First Quartile) separates the bottom 25% of sorted values from the top 75%.
- $Q_2$  (Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.
- $Q_3$  (Third Quartile) separates the bottom 75% of sorted values from the top 25%.



### Example

Find the value of the first quartile  $Q_1$ .

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

### Solution

Finding  $Q_1$  is the same as finding  $P_{25}$

$$L = \frac{k}{100} \cdot n = \frac{25}{100} \cdot 35 = 8.75 \approx 9.$$

Therefore, the first quartile is given by  $Q_1 = \$35$  million.

- ✓ Interquartile range (or **IRQ**) =  $Q_3 - Q_1$
- ✓ Semi-interquartile range =  $\frac{Q_3 - Q_1}{2}$
- ✓ Midquartile =  $\frac{Q_3 + Q_1}{2}$
- ✓ 10–90 percentile range =  $P_{90} - P_{10}$

### Definition

For a set of data, the **5-number summary** consists of the minimum value; the first quartile  $Q_1$ ; the median (or second quartile  $Q_2$ ); the third quartile,  $Q_3$ ; and the maximum value.

A **boxplot** (or **box-and-whisker-diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile,  $Q_1$ ; the median; and the third quartile,  $Q_3$ .

### Example

Use the movie budget amount to find the 5-number summary.

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

### Solution

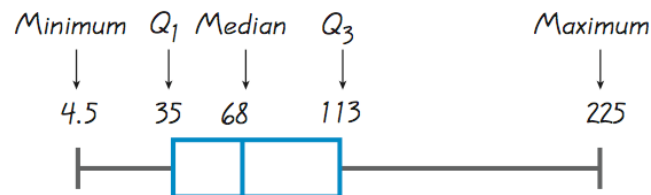
From the table:

The minimum is \$4.5 million and the maximum is \$225 million.

$$L = \frac{25}{100} \cdot 35 \approx 9 \Rightarrow P_{25} = 35 \rightarrow Q_1 = \$35 \text{ million}$$

$$L = \frac{50}{100} \cdot 35 \approx 18 \Rightarrow P_{50} = 68 \rightarrow Q_2 = \$68 \text{ million.}$$

$$L = \frac{75}{100} \cdot 35 \approx 27 \Rightarrow P_{75} = 113 \rightarrow Q_3 = \$113 \text{ million.}$$



### Procedure for Constructing a Boxplot

1. Find the 5-number summary consisting of the minimum value,  $Q_1$ , the median,  $Q_3$ , and the maximum value.
2. Construct a scale with values that include the minimum and maximum data values.
3. Construct a box (rectangle) extending from  $Q_1$  to  $Q_3$ , and draw a line in the box at the median value.
4. Draw lines extending outward from the box to the minimum and maximum data values.

### Outliers and Modified Boxplots

#### Definition

An **outlier** is a value that lies very far away from the vast majority of the other values in a data set.

For purposes of constructing *modified boxplots*, we can consider outliers to be data values meeting specific criteria.

In modified boxplots, a data value is an outlier if it is . . .  
above  $Q_3$  by an amount greater than  $1.5 \times \text{IQR}$

or

below  $Q_1$  by an amount greater than  $1.5 \times \text{IQR}$

#### Example

The pulse rates of females listed below

76	72	88	60	72	68	80	64	68	68	80	76	68	72	96
72	68	72	64	80	64	80	76	76	76	80	104	88	60	76
72	72	88	80	60	72	88	88	124	64					

Use the data to construct a modified boxplot.

#### Solution

60	60	60	64	64	64	64	68	68	68	68	68	72	72	72	72
72	72	72	72	76	76	76	76	76	76	80	80	80	80	80	80
88	88	88	88	88	96	104	124								

From the table: The minimum is 60 and the maximum is 124.

$$L = \frac{25}{100} \cdot 40 \approx 10 \Rightarrow P_{25} = 68 \rightarrow Q_1 = 68$$

$$L = \frac{75}{100} \cdot 40 \approx 30 \Rightarrow P_{75} = 80 \rightarrow Q_3 = 80$$

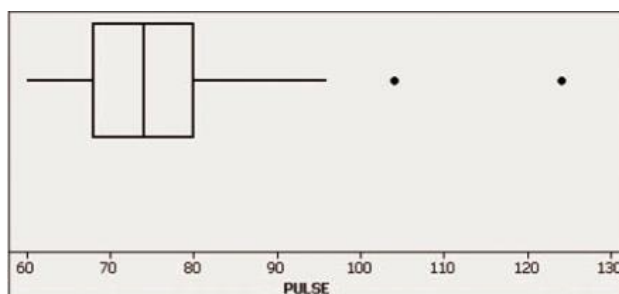
Interquartile range (or **IRQ**) =  $Q_3 - Q_1 = 80 - 68 = 12$

For pulse rates above the third quartile of 80 by an amount that is greater than



$1.5 \times IQR = 1.5 \times 12 = 18$  so high outliers are greater than 98.

The pulse rates of 104 and 124 satisfy this condition, so those two values are outliers.



Store the list values in L1.

Press [2<sup>nd</sup>] (Y=) [STAT PLOT] [1]

Press [ENTER] to turn on the stat plot.

Scroll down to Type, then right 4 times first pictures in 2<sup>nd</sup> row.

Be sure that Xlist is L1

Allow Freq: 1

Press [GRAPH], Press [ZOOM] and select 9, and press [ENTER]

## Exercises    **Section 1.9 – Measures of Relative Standing and Boxplots**

1. When Reese Witherspoon won an Oscar as Best Actress for the movie *Walk the Line*, her age was converted to a  $z$ -score of  $-0.61$  when included among the ages of all other Oscar-winning Best Actress at the time of this writing. Was her age above the mean or below the mean? How many standard deviations away from the mean is her age?
2. Hoffman was 38 years of age when he won a Best Actor Oscar for his role in *Capote*. The Oscar-winning Best Actors have a mean age of 43.8 years and a standard deviation of 8.9 years.
  - a) What is the difference between Hoffman's age and the mean age?
  - b) How many standard deviations is that (the difference found in part (a))?
  - c) Convert Hoffman's age to a  $z$ -score.
  - d) If we consider "usual" ages to be those that convert to  $z$ -scores between  $-2$  and  $2$ , is Hoffman's age usual or unusual?
3. Eruptions of the Old Faithful geyser have duration times with a mean of 245.0 sec and a standard deviation of 36.4 sec. One eruption had a duration time of 110 sec.
  - a) What is the difference between a duration time of 110 sec and the mean?
  - b) How many standard deviations is that (the difference found in part (a))?
  - c) Convert duration time of 110 sec to a  $z$ -score.
  - d) If we consider "usual" ages to be those that convert to  $z$ -scores between  $-2$  and  $2$ , is a duration time of 110 sec usual or unusual?
4. Human body temperatures have a mean of  $98.20^{\circ}\text{F}$  and a standard deviation of  $0.62^{\circ}\text{F}$ . Convert each given temperature to a  $z$ -score and determine whether it is usual and unusual.
  - a)  $101.00^{\circ}\text{F}$
  - b)  $96.90^{\circ}\text{F}$
  - c)  $96.98^{\circ}\text{F}$
5. Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1840 on the SAT test or a score of 26.0 on the ACT test? Why?
6. Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1190 on the SAT test or a score of 16.0 on the ACT test? Why?
7. Use the given sorted values, which are the numbers of points scored in the Super Bowl for a recent period of 24 years. Find the percentile corresponding to the given number of points  
36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75
  - a) 47
  - b) 65
  - c) 54
  - d) 41

8. For the given data, find the indicated percentile or quartile

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

- a)  $P_{20}$                       c)  $P_{50}$                       e)  $P_{25}$                       g)  $Q_1$   
b)  $P_{80}$                       d)  $P_{75}$                       f)  $P_{95}$                       h)  $Q_3$

9. The number of hours of television watched per day by a sample of 28 people

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

- a) Find the data set's first, second, and third quartiles.  
b) Draw a box-and-whisker plot that represents the data set.  
c) About 75% of the people watched no more than how many hours of television per day?  
d) What percent of the people watched more than 4 hours of television per day?  
e) If you randomly selected one person from the sample, what is the likelihood that the person watched less than 2 hours of television per day? Write your answer as a percent.

10. The hourly earnings (in dollars) of a sample of 25 railroad equipment manufacturers

15.6 18.75 14.6 15.8 14.35 13.9 17.5 17.55 13. 14.2 19.05 15.35 15.2  
19.45 15.95 16.5 16.3 15.25 15.05 19.1 15.2 16.22 17.75 18.4 15.25

- a) Find the data set's first, second, and third quartiles.  
b) Draw a box-and-whisker plot that represents the data set.  
c) About 75% of the manufacturers made less than \$15.80 per hour?  
d) What percent of the manufacturers made more than \$15.80 per hour?  
e) If you randomly selected one manufacturer from the sample, what is the likelihood that the manufacturer made less than \$15.80 per hour? Write your answer as a percent.

11. A certain brand of automobile tire has a mean life span of 35,000 miles, with a standard deviation of 2250 miles. (Assume the life spans of the tires have a bell-shaped distribution)

- a) The life spans of three randomly selected tires are 34,000 miles, 37,000 miles, and 30,000 miles. Find the  $z$ -score that corresponds to each life span. According to the  $z$ -scores, would the life spans of any of these tires be considered unusual?  
b) The life spans of three randomly selected tires are 30,500 miles, 37,250 miles, and 35,000 miles. Using the Empirical Rule, find the percentile that corresponds to each life span.

12. The life spans of species of fruit fly have a bell shaped distribution, with mean of 33 days and a standard deviation of 4 days.

- a) The life spans of three randomly selected fruit flies are 34 days, 30 days, and 42 days. Find the  $z$ -score that corresponds to each life span and determine if any of these life spans are unusual.  
b) The life spans of three randomly selected fruit flies are 29 days, 41 days, and 25 days. Using the Empirical Rule, find the percentile that corresponds to each life span.

13. Find the  $Q_1$  and  $Q_3$  for the given data: 49 52 52 52 74 67 55 55

14. Find the  $Q_1$  and  $Q_3$  for the given weights (in pounds) of 30 newborn babies listed below:

5.5 5.7 5.8 6.0 6.1 6.1 6.3 6.4 6.5 6.6  
6.7 6.7 6.7 6.9 7.0 7.0 7.0 7.1 7.2 7.2  
7.4 7.5 7.7 7.7 7.8 8.0 8.1 8.1 8.3 8.7

15. Find the percentile for the data value:

113 125 117 111 119 121 111 109 116 113 117 127 109 113 115 110  
Data value: 119

16. The test scores of 40 students are listed below:

30 35 43 44 47 48 54 55 56 57 59 62 63 65 66 68 69 69 71 72  
72 73 74 76 77 77 78 79 80 81 81 82 83 85 89 92 93 94 97 98

Find  $P_{56}$