



College of Computer at Al-Lith

Lecture notes on:

Elements of statistics and probabilities



Dr Moeiz Miraoui
mfmiraoui@uqu.edu.sa

Course Title: Elements of statistics and probabilities

Course ID: 30042301-3

Prerequisites : Introduction to math II

Level: Undergraduate (computer Science) / Year 2 (1st Semester)

Semester: Fall 2019

Course Schedule:

Sunday : 4:00 PM-4:50 PM + 5:00 PM-5:50 PM

Tuesday: 5:00 PM-5:50 PM

Office Hours:

Sunday :10:00 AM-10:50 AM + 11:00 AM-11:50 AM

Instructor : Dr Moeiz Miraoui

E-mail : mfmiraoui@uqu.edu.sa

Grading plan

Midterm 1 : 25% (October 21st, 2019)

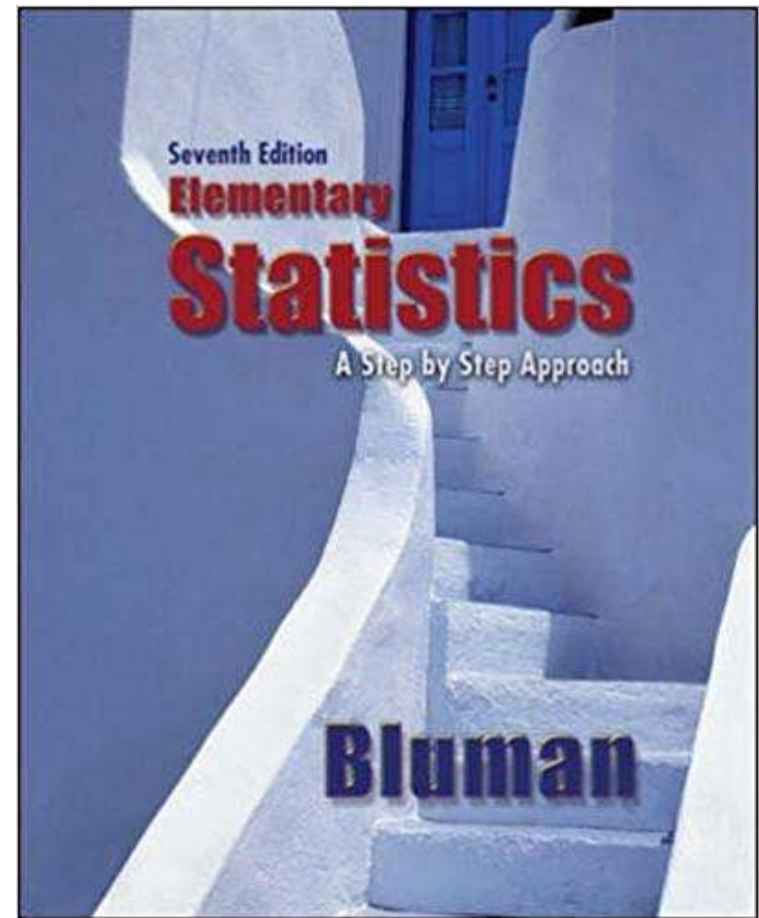
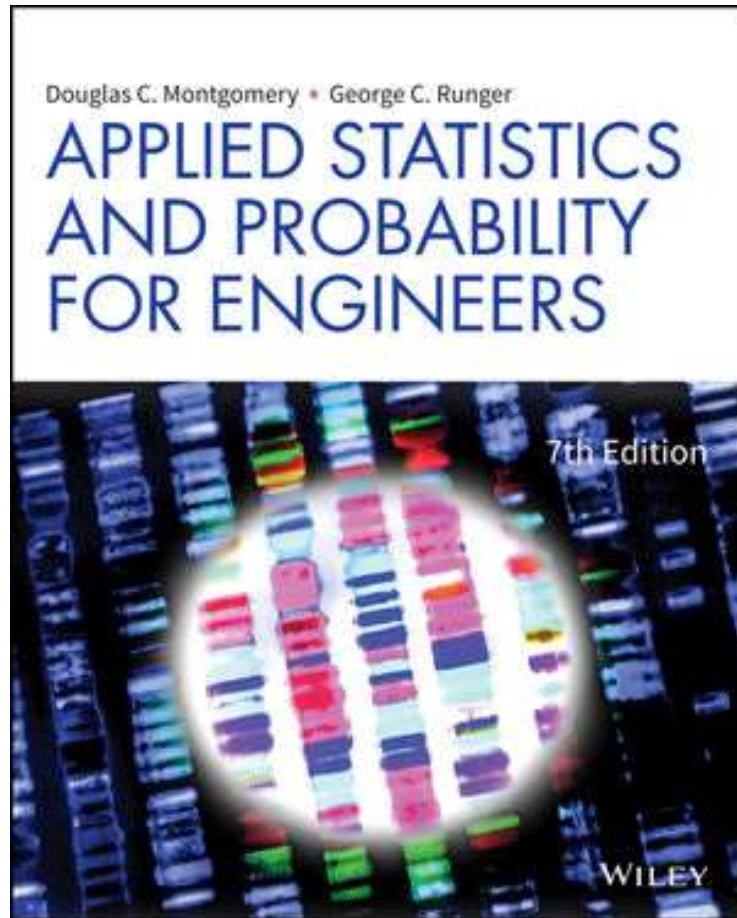
Midterm 2 : 25% (November 19th, 2019)

Final exam: 50% (December 2019 ... TBA)

Textbook/References

- ❑ Douglas C. Montgomery and George C. Runger, “Applied Statistics & Probability for Engineers”, Sixth Edition, John Wiley & Sons, 2014.
- ❑ Elementary Statistics: A Step by Step Approach, Allan Bluman, McGraw-Hill Science/Engineering/Math; 7 edition (October 27, 2008)

Reference Books



CONTENTS

- ❑ Definition and general view of statistics
- ❑ Organization and presentation of statistical data
- ❑ Data description (measure of central tendency and variation)
- ❑ Probability basics and counting rules
- ❑ Probability distributions
- ❑ Conditional probability - Independence of events and Bayes theorem
- ❑ Confidence Intervals and Sample Size
- ❑ Correlation and Regression

Chapter 1

Definition and general view of statistics

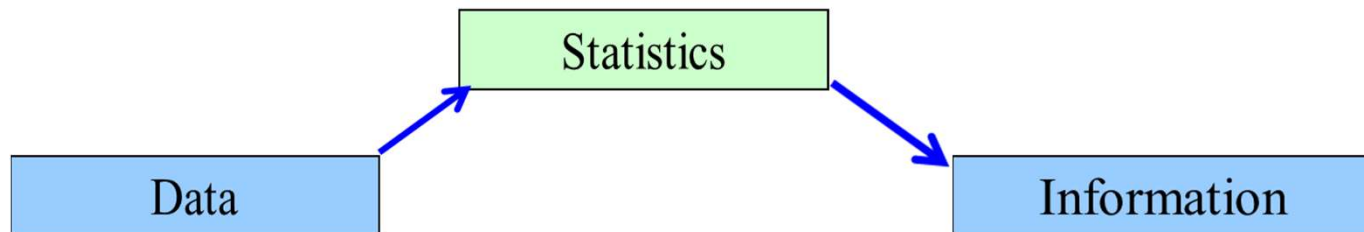
❑ Introduction

- ❑ You may be familiar with probability and statistics through radio, television, newspapers, magazines and the internet.
 - In Massachusetts, 36% of adults aged 25 and older have at least a bachelor's degree
 - Toddlers need an average of 13 hours of sleep per day.
 - The average in-state college tuition and fees for 4-year public college is \$5836
- ❑ statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments
- ❑ **Statistics** is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

□ Why study statistics

1. Data are everywhere
2. Statistical techniques are used to make many decisions that affect our lives
3. No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions effectively

Statistics is a way to get information from data



❑ Basic terms

- ❑ **Population:** A collection, or set, of individuals or objects or events whose properties are to be analyzed. Two kinds of populations: finite or infinite.
- ❑ **Sample:** A subset of the population.
- ❑ **Variable:** A characteristic about each individual element of a population or sample.
- ❑ **Data (singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.
- ❑ **Data (plural):** The set of values collected for the variable from each of the elements belonging to the sample.
- ❑ **Experiment:** A planned activity whose results yield a set of data.
- ❑ **Parameter:** A numerical value summarizing all the data of an entire population.
- ❑ **Statistic:** A numerical value summarizing the sample data.

□ Basic terms

Example: A college dean is interested in learning about the average age of faculty. Identify the basic terms in this situation.

The *population* is the age of all faculty members at the college.

A *sample* is any subset of that population. For example, we might select 10 faculty members and determine their age.

The *variable* is the “age” of each faculty member.

One *data* would be the age of a specific faculty member.

The *data* would be the set of values in the sample.

The *experiment* would be the method used to select the ages forming the sample and determining the actual age of each faculty member in the sample.

The *parameter* of interest is the “average” age of all faculty at the college.

The *statistic* is the “average” age for all faculty in the sample.

❏ Basic terms

Exercise: A study conducted at Manatee Community College revealed that students who attended class 95 to 100% of the time usually received an A in the class. Students who attended class 80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class.

Based on this information, attendance and grades are related. The more you attend class, the more likely you will receive a higher grade.

1. What are the variables under study?
2. What are the data in the study?
3. Are descriptive, inferential, or both types of statistics used?
4. What is the population under study?
5. Was a sample collected? If so, from where?
6. From the information given, comment on the relationship between the variables.

❑ Kinds of variables

❑ **Qualitative, or Attribute, or Categorical, Variable:** A variable that categorizes or describes an element of a population.

Note: Arithmetic operations, such as addition and averaging, are *not* meaningful for data resulting from a qualitative variable.

❑ **Quantitative, or Numerical, Variable:** A variable that quantifies an element of a population.

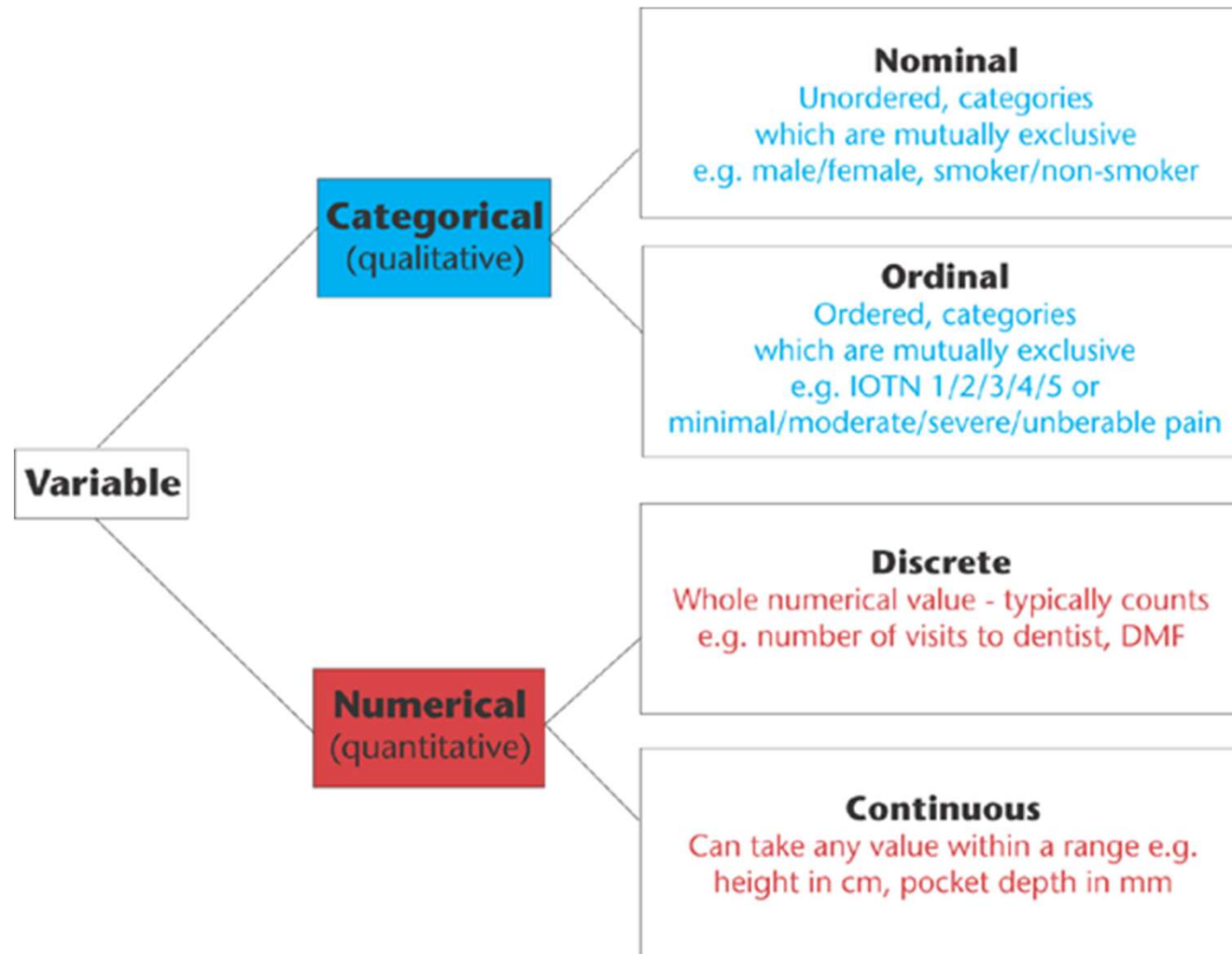
Note: Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.

□ Kinds of variables

Example: Identify each of the following examples as qualitative or quantitative variables.

1. The residence hall for each student in a statistics class. (qualitative)
2. The amount of gasoline pumped by the next 10 customers at the local petrol station. (quantitative)
3. The color of the baseball cap worn by each of 20 students. (qualitative)
5. The length of time to complete a mathematics homework assignment. (quantitative)
6. The state of a car when making the routine technical inspection. (qualitative)

□ Kinds of variables



Exercise:

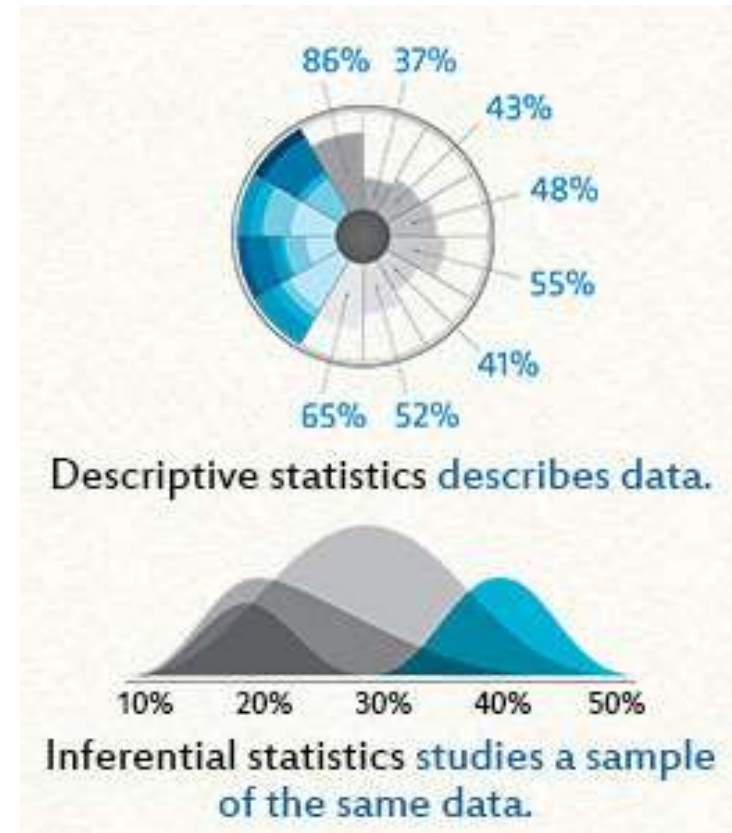
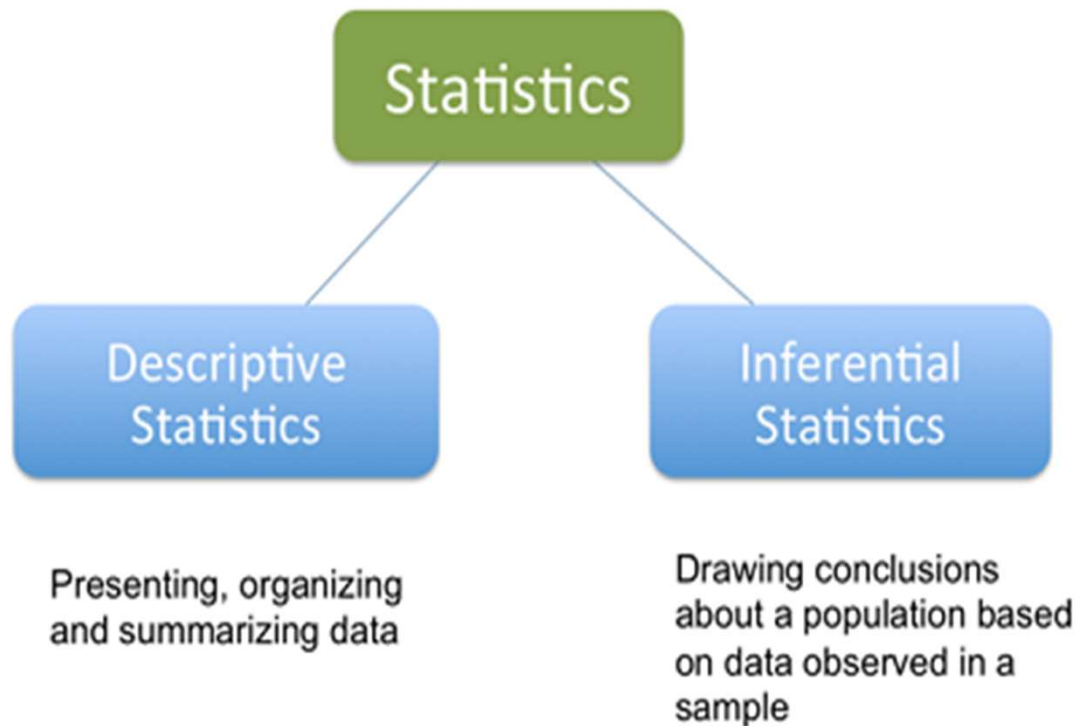
Transportation Safety

The chart shows the number of job-related injuries for each of the transportation industries for 1998.

Industry	Number of injuries
Railroad	4520
Intercity bus	5100
Subway	6850
Trucking	7144
Airline	9950

- 1. What are the variables under study?
- 2. Categorize each variable as quantitative or qualitative.
- 3. Categorize each quantitative variable as discrete or continuous.
- 4. Identify the level of measurement for each variable.
- 5. The railroad is shown as the safest transportation industry. Does that mean railroads have fewer accidents than the other industries? Explain.
- 6. What factors other than safety influence a person's choice of transportation?

□ Areas of statistics



❑ Data Collection and Sampling Techniques

- First problem a statistician faces: how to obtain the data.
- It is important to obtain *good*, or *representative*, data.
- Inferences are made based on statistics obtained from the data.
- Inferences can only be as good as the data.

❑ Data Collection and Sampling Techniques

- First problem a statistician faces: how to obtain the data.
- It is important to obtain *good*, or *representative*, data.
- Inferences are made based on statistics obtained from the data.
- Inferences can only be as good as the data.
- The data may be collected for the whole population or for a sample only
(**mostly used: less time & less costly**)

❑ Data Collection and Sampling Techniques

- A sampling method is called **biased** if it systematically favors some outcomes over others.
- You NEVER want to use a biased sampling method!
- The sampling method is very important because it can affect the validity of the data
- **Example:** Telephone sampling is common in marketing surveys. It is a biased sampling method because it will miss people who do not have a phone.

❑ Data Collection and Sampling Techniques

Process of data collection:

1. Define the objectives of the survey or experiment.

Example: Estimate the average life of an electronic component.

2. Define the variable and population of interest.

Example: Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate descriptive or inferential data-analysis techniques.

□ Data Collection and Sampling Techniques

There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

- Direct Observation
- Experiments, and
- Surveys.

The most common methods is through the use of surveys.

□ Data Collection and Sampling Techniques

There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

- Direct Observation
- Experiments, and
- Surveys.

The most common methods is through the use of surveys.

- Three of the most common methods are the telephone survey, the mailed questionnaire, and the personal interview.

❑ Data Collection and Sampling Techniques

❑ **Telephone surveys** have an advantage over personal interview surveys in that they are less costly.

people may be more candid in their opinions since there is no face-to-face contact.

A major drawback to the telephone survey is that some people in the population will not have phones or will not answer when the calls are made

❑ **Mailed questionnaire surveys** can be used to cover a wider geographic area and are less expensive to conduct.

Respondents can remain anonymous if they desire.

Disadvantages include a low number of responses and some people may have difficulty reading or understanding the questions.

❑ Data Collection and Sampling Techniques

❑ **Personal interview surveys** have the advantage of obtaining in-depth responses to questions from the person being interviewed.

One disadvantage is that interviewers must be trained in asking questions which makes it more costly

The interviewer may be biased in his or her selection of respondents.

- Researchers use samples to collect data and information about a particular variable from a large population.
- Using samples saves time and money
- four basic methods of sampling: **random, systematic, stratified, convenience,** and **cluster sampling**

❑ Data Collection and Sampling Techniques

- ❑ **Simple Random Sampling:** Every member of the population is equally likely to be selected)
- ❑ **Systematic Sampling:** Simple Random Sampling in an ordered systematic way, e.g. every 100th name in the yellow pages
- ❑ **Stratified Sampling:** Population divided into different groups from which we sample randomly
- ❑ **Cluster Sampling:** Population is divided into (geographical) clusters - some clusters are chosen at random - within cluster units are chosen with Simple Random Sampling
- ❑ **Convenience Sample:** Sample selected by taking the members of the population who are EASIEST to reach

❑ Data Collection and Sampling Techniques

❑ **Example:** You are doing a study to determine the opinion of students at your school regarding quality of education. Identify the sampling technique you are using if you select the samples listed.

❑ You select a class at random and question each student in the class.

Cluster sample

❑ You divide the student population with respect to majors and randomly select and question some students in each major.

Stratified sample

❑ You assign each student a number and generate random numbers. You then question each student whose number is randomly selected.

Simple random sample

❑ You assign each student a number and, after choosing a starting number, question every 25th student.

Systematic sample

❑ Data Collection and Sampling Techniques

❑ **Exercise:** Identify the sampling technique used

❑ 32 sophomores, 35 juniors, and 49 seniors are randomly selected from 230 sophomores, 280 juniors, 577 seniors at a certain high school.

❑ To ensure customer satisfaction, every 35th phone call received by customer service will be monitored.

❑ 3. A journalist goes to a campground to ask people how they felt about air pollution.

❑ 4. Calling randomly generated telephone numbers, a study asked 855 US adults which medical conditions could be prevented by their diet.

❑ 5. A pregnancy study in Chicago, randomly selected 25 communities from the metropolitan area, then interviewed all pregnant women in these communities.

❑ Data Collection and Sampling Techniques

Questionnaire design

- ❑ Keep the questionnaire as short as possible.
- ❑ Ask short, simple, and clearly worded questions.
- ❑ Start with demographic questions to help respondents get started comfortably.
- ❑ Use dichotomous (yes|no) and multiple choice questions.
- ❑ Use open-ended questions cautiously.
- ❑ Avoid using leading-questions.
- ❑ Pretest a questionnaire on a small number of people.
- ❑ Think about the way you intend to use the collected data when preparing the questionnaire.

❑ Data Collection and Sampling Techniques

❑ There are several different ways to classify statistical studies.

types of studies: *observational studies* and *experimental studies*.

- In an **observational study**, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
- In an **experimental study**, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Thank you

End of Chapter 1

Chapter 2

Organization and presentation of statistical data

❑ Date organization and presentation

The following table presents the ways in which a person's identity can be stolen

Lost or stolen wallet, checkbook, or credit card	38%
Friends, acquaintances	15
Corrupt business employees	15
Computer viruses and hackers	9
Stolen mail or fraudulent change of address	8
Online purchases or transactions	4
Other methods	11

Looking at the numbers presented in a table does not have the same impact as presenting numbers in a well-drawn chart or graph

The data should be organized in some **meaningful way**.

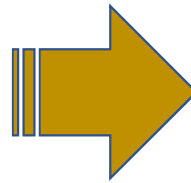
The most convenient method of **organizing** data is to construct a *frequency distribution*

The most useful method of **presenting** the data is by constructing *statistical charts and graphs*

❑ Data organization

- ❑ Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world.

49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	56	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74



Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62		10
63–69		10
70–76		5
77–83		10
84–90		5
		<hr/> Total 50

- ❑ A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.
- ❑ Two types of frequency distributions that are most often used are the *categorical frequency distribution* and the *grouped frequency distribution*

❑ Data organization

❑ The **categorical frequency distribution** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data

❑ **Example:** Twenty-five students were given a blood test to determine their blood type.

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

❑ Data organization

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

Step 1 Make a table as shown.

A Class	B Tally	C Frequency	D Percent
A			
B			
O			
AB			

- **Step 2** Tally the data and place the results in column B.
- **Step 3** Count the tallies and place the results in column C.
- **Step 4** Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100\%$$

❑ Data organization and presentation

Step 5 Find the totals for columns C (frequency) and D (percent). The completed table is shown.

A Class	B Tally	C Frequency	D Percent
A		5	20
B	//	7	28
O	//	9	36
AB		4	16
		Total 25	100

For the sample, more people have type O blood than any other type.

❑ Data organization and presentation

❑ **Grouped Frequency Distributions** is used When the range of the data is large, the data must be grouped into classes that are more than one unit in width

❑ **Example:** The following data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 U.S. states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

❑ Data organization and presentation

Step 1 Determine the classes.

Find the highest value and lowest value: $H = 134$ and $L = 100$.

Find the range: $R = \text{highest value} - \text{lowest value} = H - L$, so $R = 134 - 100 = 34$

Select the number of classes desired (usually between 5 and 20). In this case, 7

Find the class width by dividing the range by the number of classes.

$$\text{Width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9 \approx 5$$

Select a starting point for the lowest class limit. In this case, 100 is used.

Add the width to the lowest score, Keep adding until there are 7 classes

Find the class boundaries by subtracting 0.5 from each lower-class limit and adding 0.5 to each upper class limit

❑ Data organization and presentation

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies.

The completed frequency distribution is

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	/// ///	8
110–114	109.5–114.5	/// /// /// ///	18
115–119	114.5–119.5	/// /// ///	13
120–124	119.5–124.5	/// //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			<u>50</u>
			$n = \Sigma f = 50$

□ Date organization and presentation

Ages of Presidents at Inauguration

The data represent the ages of our Presidents at the time they were first inaugurated.

57	61	57	57	58	57	61	54	68
51	49	64	50	48	65	52	56	46
54	49	50	47	55	55	54	42	51
56	55	54	51	60	62	43	55	56
61	52	69	64	46	54			

1. Were the data obtained from a population or a sample? Explain your answer.
2. What was the age of the oldest President?
3. What was the age of the youngest President?
4. Construct a frequency distribution for the data. (Use your own judgment as to the number of classes and class size.)
5. Are there any peaks in the distribution?
6. Identify any possible outliers.
7. Write a brief summary of the nature of the data as shown in the frequency distribution.

❑ Date organization and presentation

Solution:

□ Data organization and presentation

□ **Exercise:** Listed are the weights of the NBA's top 50 players. Construct a grouped frequency distribution and a cumulative frequency distribution with 8 classes. Analyze the results in terms of peaks, extreme values, etc.

240	210	220	260	250	195	230	270	325	225
165	295	205	230	250	210	220	210	230	202
250	265	230	210	240	245	225	180	175	215
215	235	245	250	215	210	195	240	240	225
260	210	190	260	230	190	210	230	185	260

□ Date organization and presentation

Solution:

❑ Date presentation

- ❑ After you have organized the data into a frequency distribution, you can present them in graphical form
- ❑ It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions
- ❑ The three most commonly used graphs in research are
 1. The histogram.
 2. The frequency polygon.
 3. The cumulative frequency graph, or ogive (pronounced o-jive).

❑ Data presentation

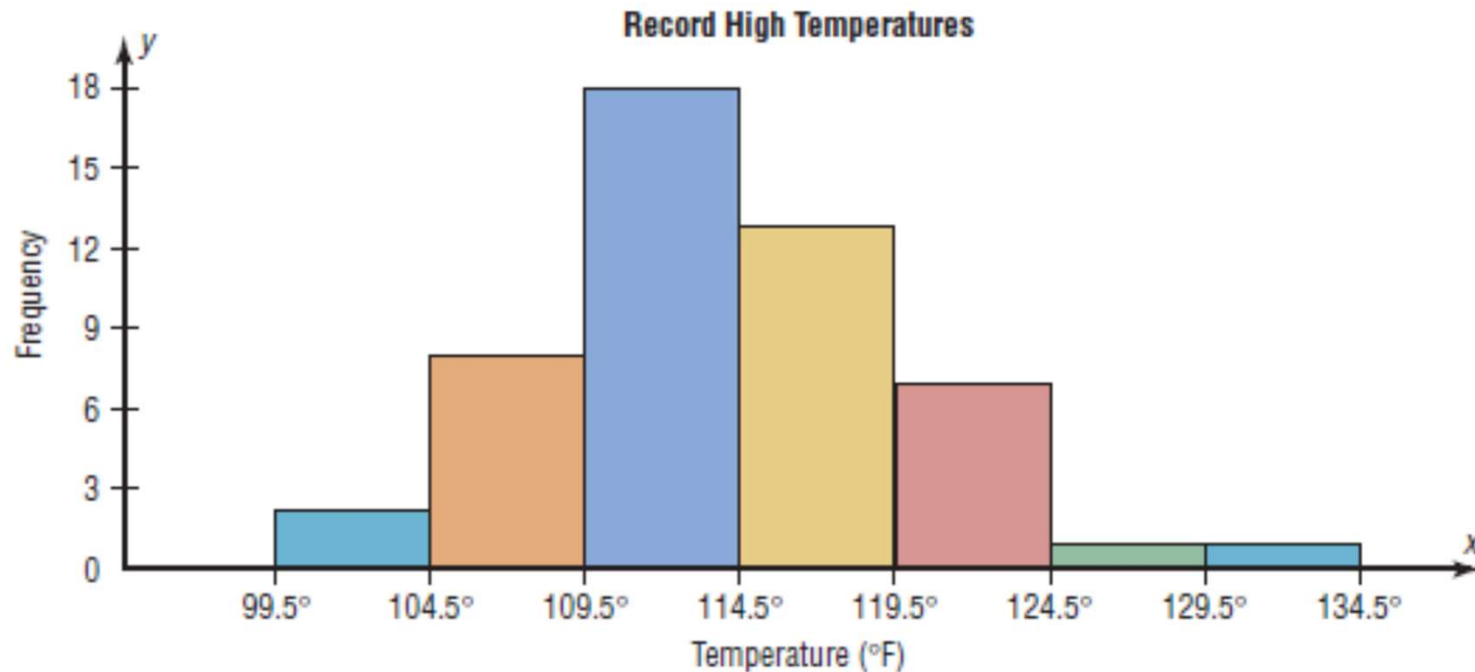
❑ The **histogram** is a graph that displays the data by using contiguous vertical bars(unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

Example: Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

❑ Data organization and presentation

- Represent the frequency on the y axis and the class boundaries on the x axis.
- Using the frequencies as the heights, draw vertical bars for each class



❑ Data organization and presentation

❑ The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

Step 1 Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

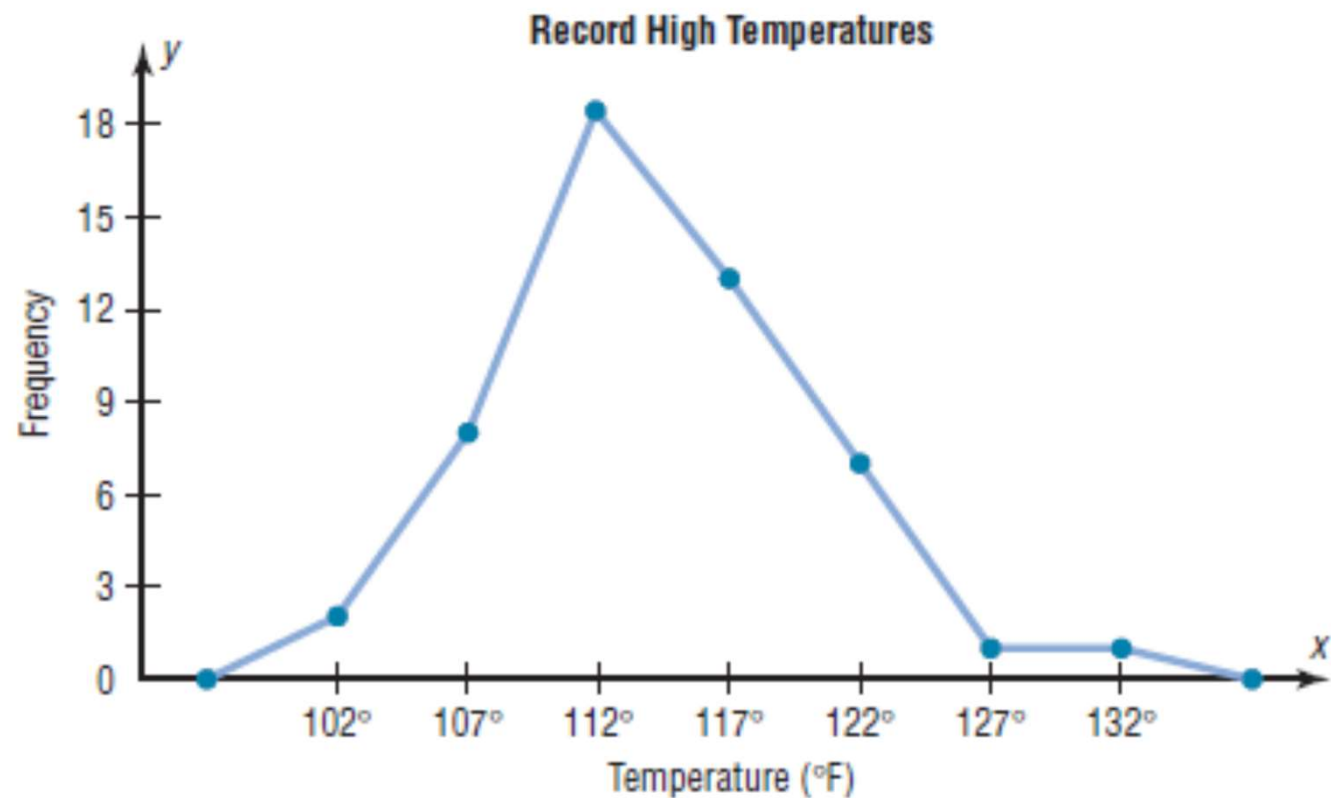
$$\frac{99.5 + 104.5}{2} = 102$$

$$\frac{104.5 + 109.5}{2} = 107$$

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

□ Date organization and presentation

Step 2 Using the midpoints for the x values and the frequencies as the y values, plot the points.



❑ Data presentation

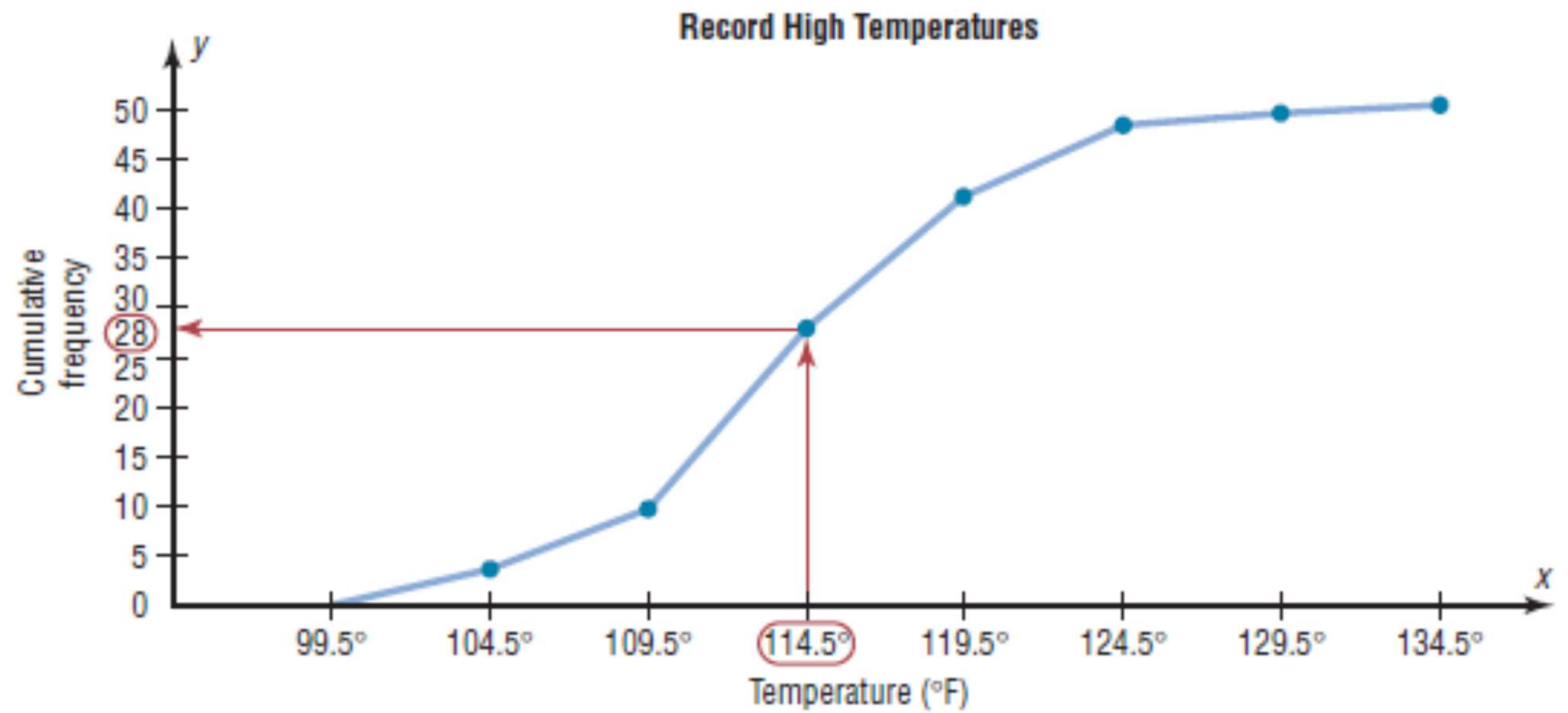
❑ The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

❑ A **cumulative frequency distribution** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary).

❑ The values are found by adding the frequencies of the classes less than or equal to the upper-class boundary of a specific class.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

□ Data presentation



❑ Data organization and presentation

❑ **Relative Frequency Graphs** are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class.

❑ To convert a frequency into a proportion or relative frequency:


- divide the frequency for each class by the total of the frequencies.
- The sum of the relative frequencies will always be 1.

Example: Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution (shown here) of the miles that 20 randomly selected runners ran during a given week.

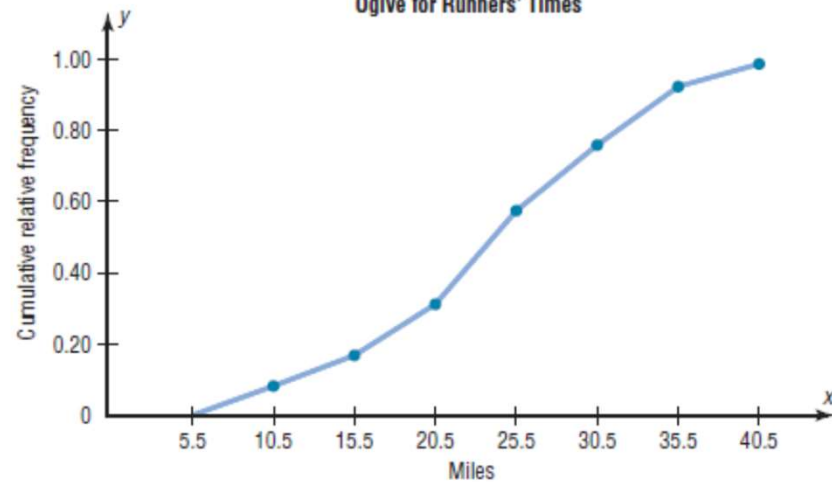
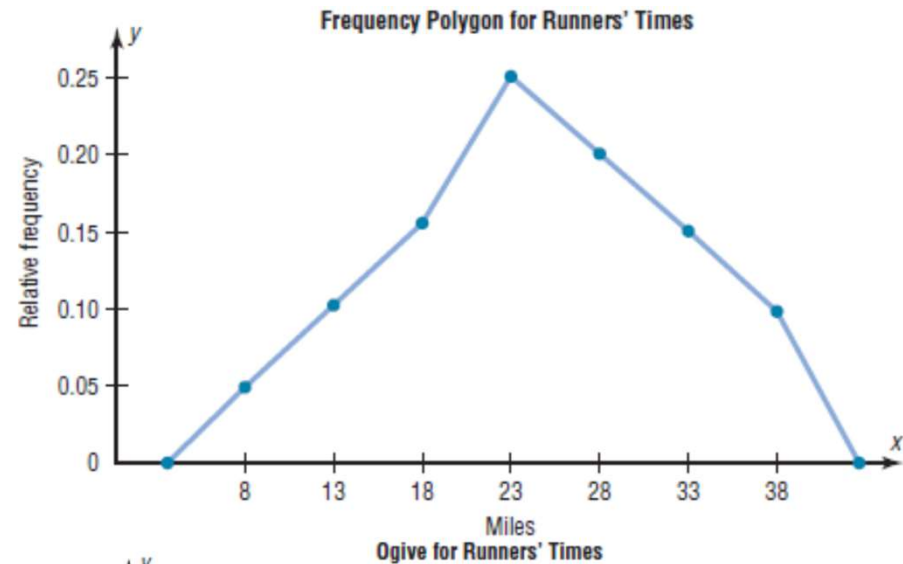
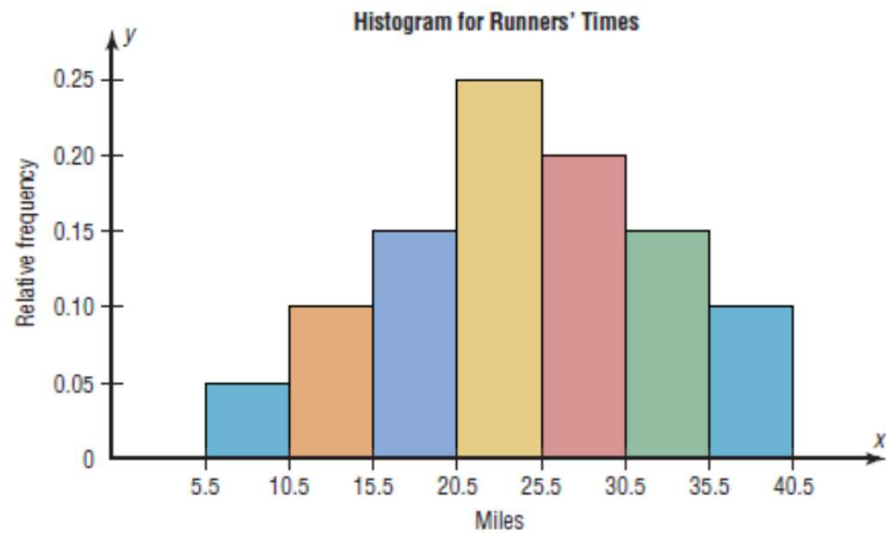
Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	<hr/> 20

□ Data organization and presentation

For class 5.5–10.5, the relative frequency is $\frac{1}{20} = 0.05$; for class 10.5–15.5, the relative frequency is $\frac{2}{20} = 0.10$; for class 15.5–20.5, the relative frequency is $\frac{3}{20} = 0.15$; and so on.

Class boundaries	Midpoints	Relative frequency		Cumulative frequency	Cumulative relative frequency	
5.5–10.5	8	0.05		Less than 5.5	0	0.00
10.5–15.5	13	0.10		Less than 10.5	1	0.05
15.5–20.5	18	0.15		Less than 15.5	3	0.15
20.5–25.5	23	0.25		Less than 20.5	6	0.30
25.5–30.5	28	0.20		Less than 25.5	11	0.55
30.5–35.5	33	0.15		Less than 30.5	15	0.75
35.5–40.5	38	0.10		Less than 35.5	18	0.90
		<u>1.00</u>		Less than 40.5	20	1.00

□ Date organization and presentation



□ Data presentation

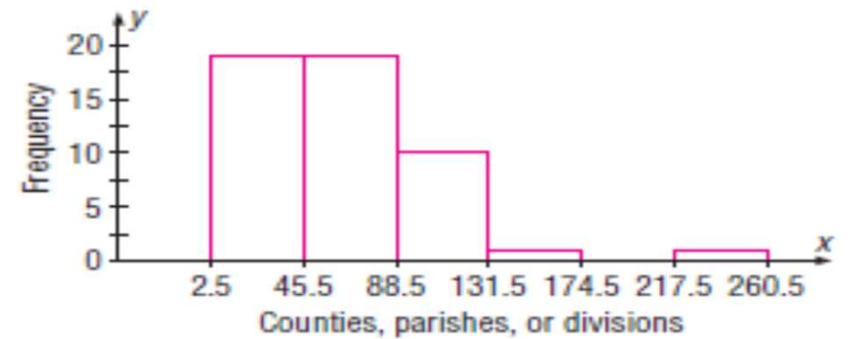
□ **Example:** number of counties for each of the 50 U.S. states is given below. Use the data to construct a grouped frequency distribution with 6 classes, a histogram, a frequency polygon, and an ogive. Analyze the distribution.

67	27	15	75	58	64	8	67	159	5
102	44	92	99	105	120	64	16	23	14
83	87	82	114	56	93	16	10	21	33
62	100	53	88	77	36	67	5	46	66
95	254	29	14	95	39	55	72	23	3

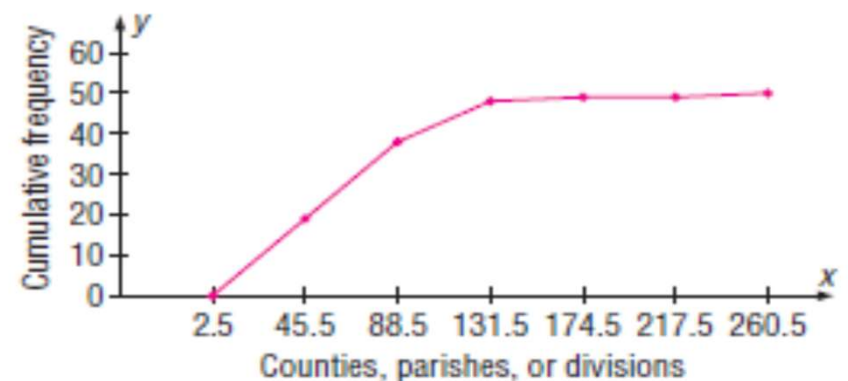
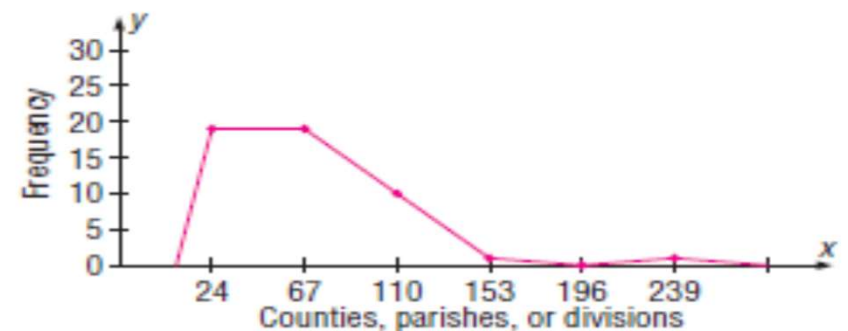
□ Data presentation

Limits	Boundaries	<i>f</i>
3–45	2.5–45.5	19
46–88	45.5–88.5	19
89–131	88.5–131.5	10
132–174	131.5–174.5	1
175–217	174.5–217.5	0
218–260	217.5–260.5	1
		<hr/> 50

	cf
Less than 2.5	0
Less than 45.5	19
Less than 88.5	38
Less than 131.5	48
Less than 174.5	49
Less than 217.5	49
Less than 260.5	50



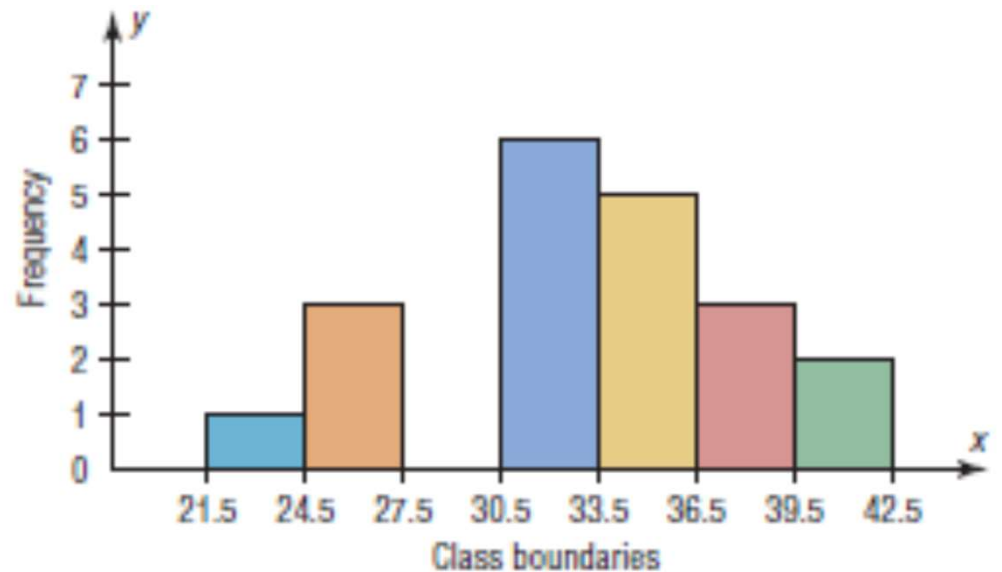
The distribution is positively skewed.



□ Data presentation

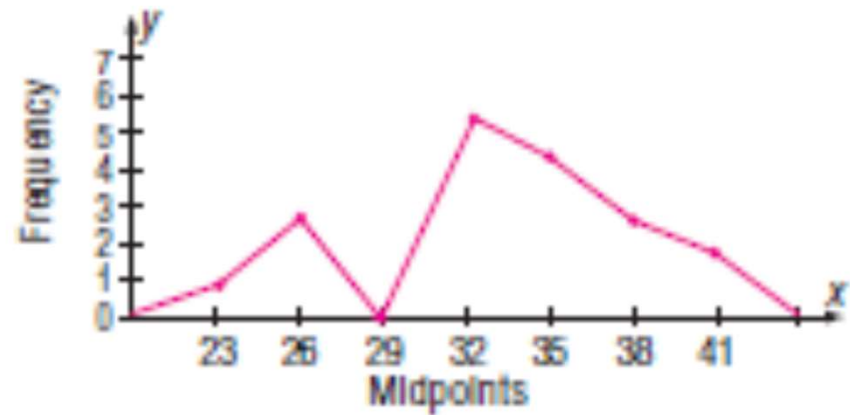
□ **Exercise:** Using the histogram shown here, do the following.

- Construct a frequency distribution; include class limits, class frequencies, midpoints, and cumulative frequencies.
- Construct a frequency polygon.
- Construct an ogive.

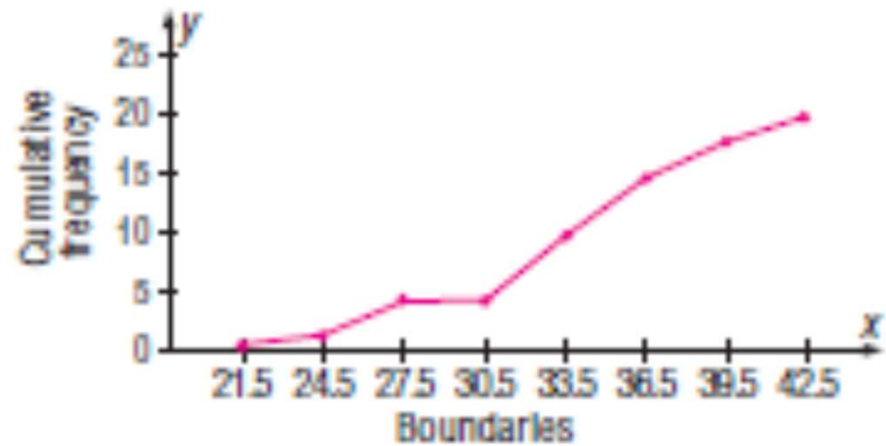


□ Data presentation

<i>a.</i> Limits	Boundaries	Midpoints	<i>f</i>
22–24	21.5–24.5	23	1
25–27	24.5–27.5	26	3
28–30	27.5–30.5	29	0
31–33	30.5–33.5	32	6
34–36	33.5–36.5	35	5
37–39	36.5–39.5	38	3
40–42	39.5–42.5	41	2



	<i>cf</i>
Less than 21.5	0
Less than 24.5	1
Less than 27.5	4
Less than 30.5	4
Less than 33.5	10
Less than 36.5	15
Less than 39.5	18
Less than 42.5	20

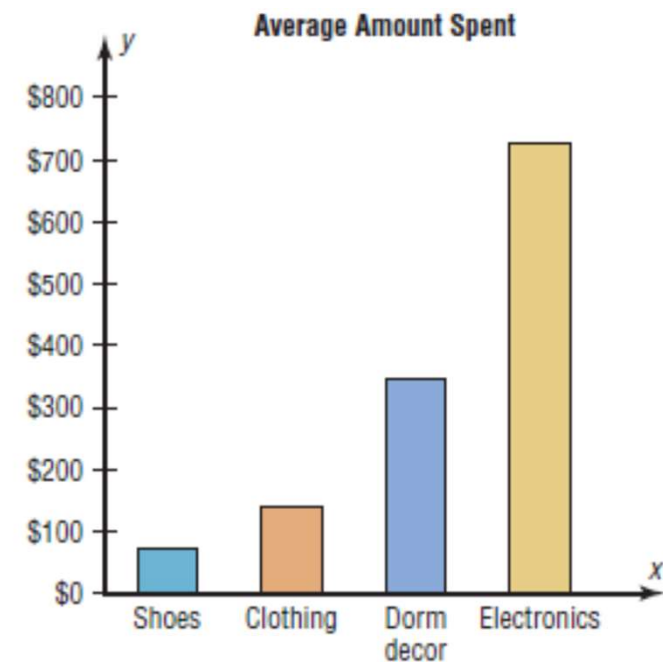
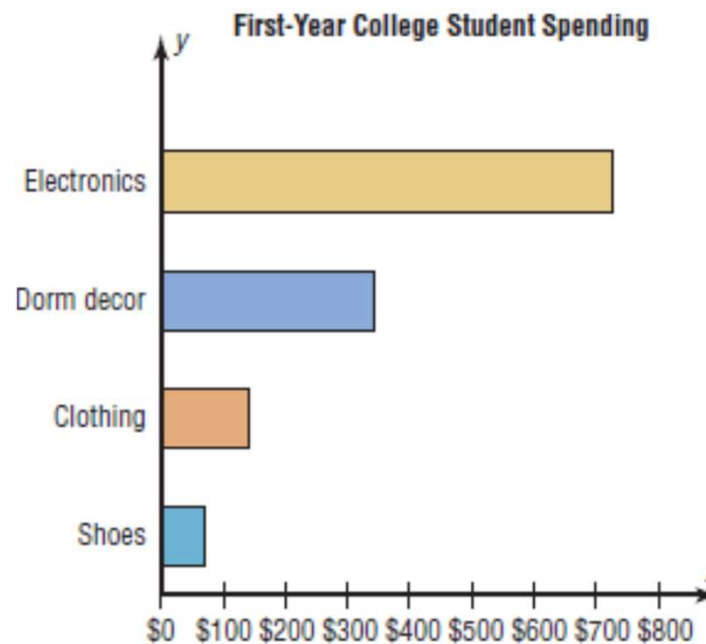


❑ Data presentation

❑ A **bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

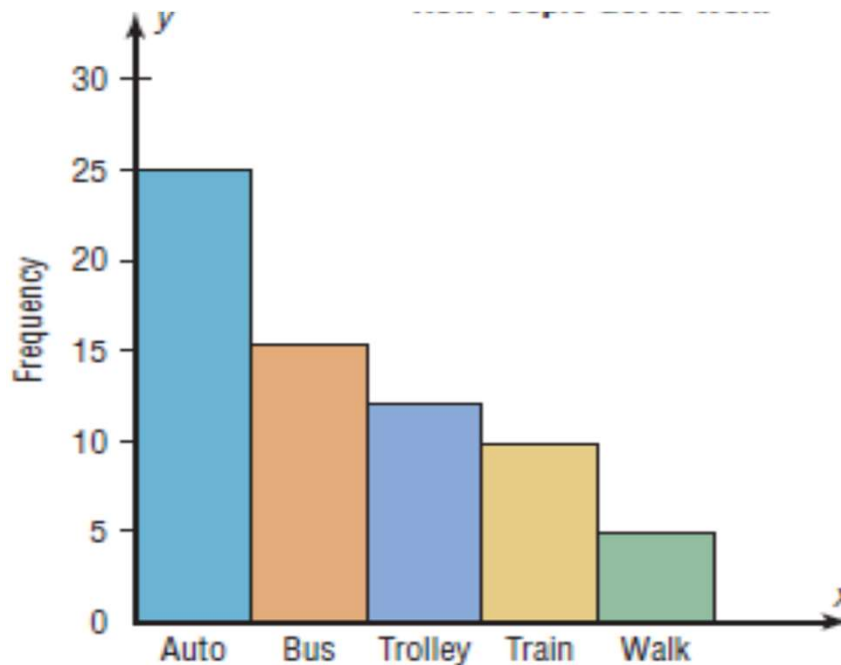
The table shows the average money spent by first-year college students. Draw a horizontal and vertical bar graph for the data.

Electronics	\$728
Dorm decor	344
Clothing	141
Shoes	72



❑ Data presentation

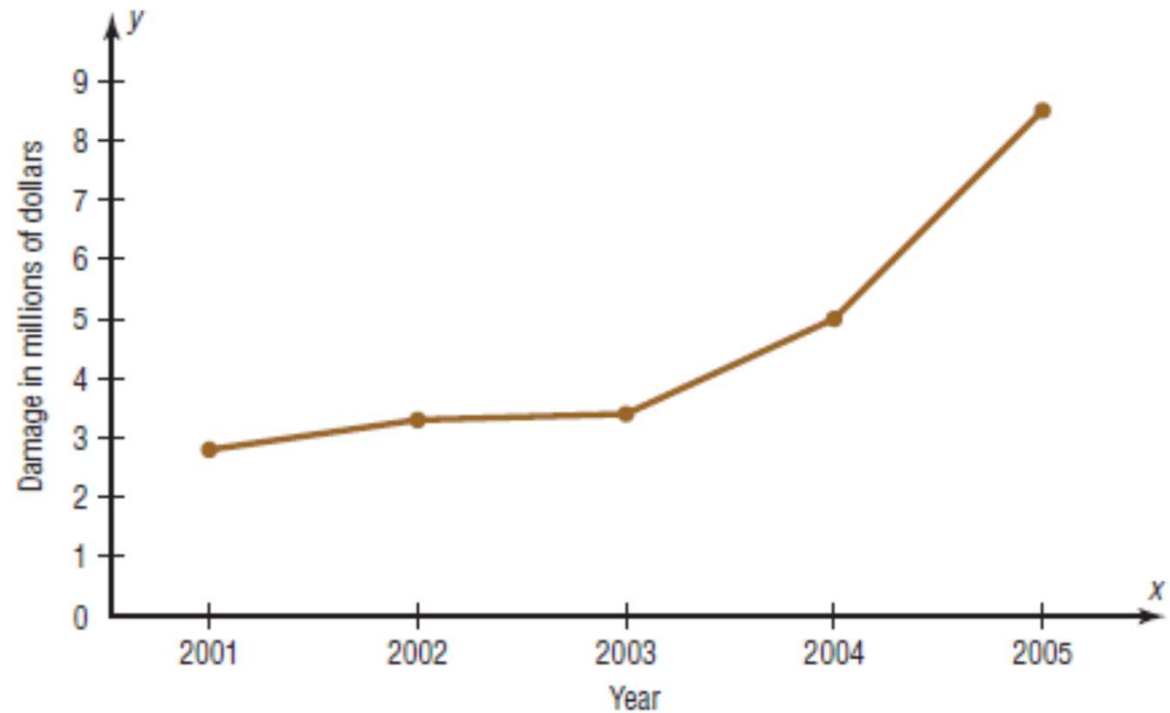
❑ A **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.



❑ Date presentation

❑ A **time series graph** represents data that occur over a specific period of time.

Year	Damage (in millions)
2001	\$2.8
2002	3.3
2003	3.4
2004	5.0
2005	8.5



❑ Data presentation

❑ A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Example: This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

Snack	Pounds (frequency)
Potato chips	11.2 million
Tortilla chips	8.2 million
Pretzels	4.3 million
Popcorn	3.8 million
Snack nuts	2.5 million
Total $n = 30.0$ million	

□ Date presentation

Step 1 Since there are 360 in a circle, the frequency for each class must be converted into a proportional part of the circle. This conversion is done by using the formula:

$\text{Degrees} = \frac{f}{n} \cdot 360^\circ$ where f = frequency for each class and n = sum of the frequencies. Hence, the following conversions are obtained. The degrees should sum to 360°

Potato chips	$\frac{11.2}{30} \cdot 360^\circ = 134^\circ$
Tortilla chips	$\frac{8.2}{30} \cdot 360^\circ = 98^\circ$
Pretzels	$\frac{4.3}{30} \cdot 360^\circ = 52^\circ$
Popcorn	$\frac{3.8}{30} \cdot 360^\circ = 46^\circ$
Snack nuts	$\frac{2.5}{30} \cdot 360^\circ = 30^\circ$
Total	<u>360°</u>

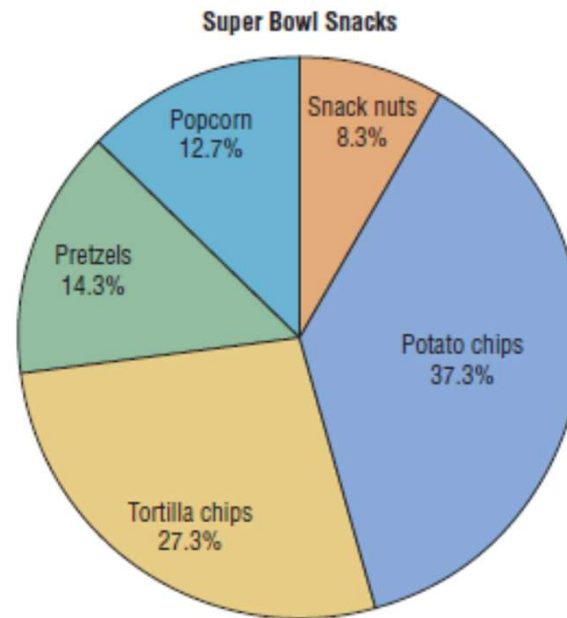
□ Data presentation

Step 2 Each frequency must also be converted to a percentage by using the formula:

$$\% = \frac{f}{n} \cdot 100\%$$

Hence, the following percentages are obtained. The percentages should sum to 100%.

Potato chips	$\frac{11.2}{30} \cdot 100\% = 37.3\%$
Tortilla chips	$\frac{8.2}{30} \cdot 100\% = 27.3\%$
Pretzels	$\frac{4.3}{30} \cdot 100\% = 14.3\%$
Popcorn	$\frac{3.8}{30} \cdot 100\% = 12.7\%$
Snack nuts	$\frac{2.5}{30} \cdot 100\% = 8.3\%$
Total	<u>99.9%</u>



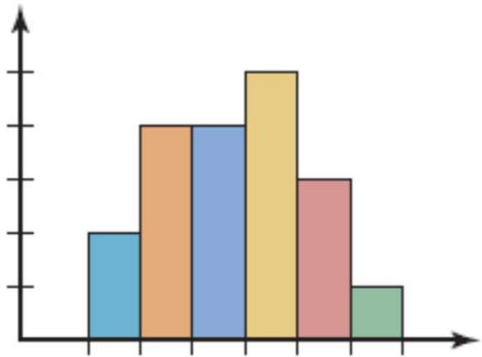
□ Data presentation

Exercise: Construct a pie graph showing the blood types of a group of students

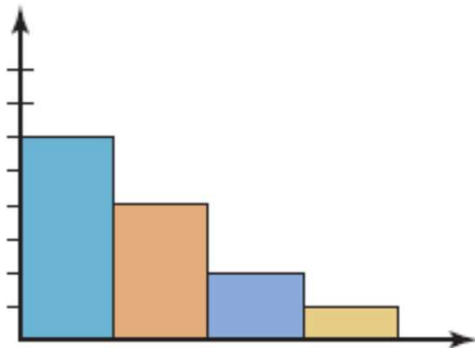
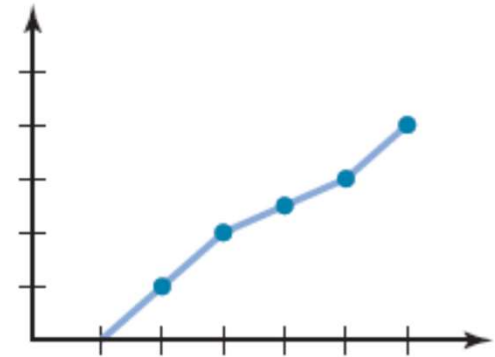
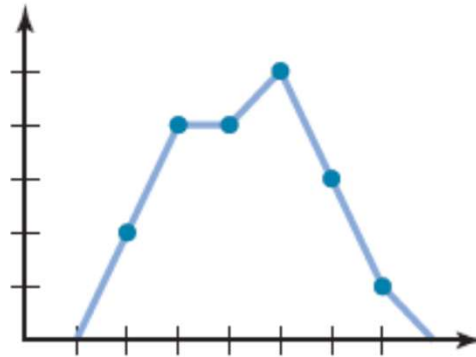
Class	Frequency	Percent
A	5	20
B	7	28
O	9	36
AB	4	16
	<u>25</u>	<u>100</u>

❑ Date presentation

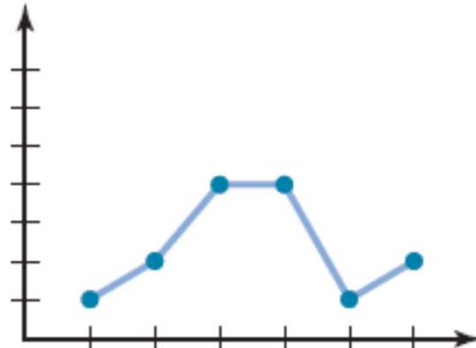
Summary



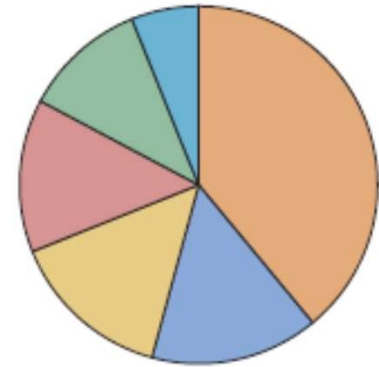
(a) Histogram; frequency polygon; ogive
Used when the data are contained in a grouped frequency distribution.



(b) Pareto chart
Used to show frequencies for nominal or qualitative variables.



(c) Time series graph
Used to show a pattern or trend that occurs over a period of time.



(d) Pie graph
Used to show the relationship between the parts and the whole.
(Most often uses percentages.)

❑ Data presentation

❑ A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

Example: the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

□ Data presentation

Step 1 Arrange the data in order: 02, 13, 14, 20, 23, 25, 31, 32, 32, 32, 32, 33, 36, 43, 44, 44, 45, 51, 52, 57

Step 2 Separate the data according to the first digit, as shown.

02 13, 14 20, 23, 25 31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45 51, 52, 57

Step 3 A display can be made by using the leading digit as the *stem* and the trailing digit as the *leaf*.

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

□ Date presentation

Exercise: An insurance company researcher conducted a survey on the number of car thefts in a large city for a period of 30 days last summer. The raw data are shown. Construct a stem and leaf plot by using classes 50–54, 55–59, 60–64, 65–69, 70–74, and 75–79.

Thank you

End of Chapter 2

Chapter 3

**Data description (measure of central tendency
and variation)**

□ Date Description

- Chapter 2 showed how you can gain useful information from raw data by organizing them into a frequency distribution and then presenting the data by using various graphs.
- This chapter shows the statistical methods that can be used to **summarize data**
- statisticians use samples taken from populations; however, when populations are small, it is not necessary to use samples since the entire population can be used to gain information
- A **statistic** is a characteristic or measure obtained by using the data values from a sample.
- A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

□ Data Description

The Mean

The **mean** is the sum of the values, divided by the total number of values. The symbol \bar{X} represents the sample mean.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

where n represents the total number of values in the sample.

For a population, the Greek letter μ (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

where N represents the total number of values in the population.

□ Data Description

Example: The data represent the number of days off per year for a sample of individuals

selected from nine different countries. Find the mean.

20, 26, 40, 36, 23, 42, 35, 24, 30

Solution

$$\bar{X} = \frac{\Sigma X}{n} = \frac{20 + 26 + 40 + 36 + 23 + 42 + 35 + 24 + 30}{9} = \frac{276}{9} = 30.7 \text{ days}$$

Hence, the mean of the number of days off is 30.7 days.

□ Date Description

Example: The table below shows the miles that 20 randomly selected runners ran during a given week.

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	<u>20</u>

Calculate the mean using the frequency distribution

□ Data Description

Solution

The procedure for finding the mean for grouped data is given here.

Step 1 Make a table as shown.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1		
10.5–15.5	2		
15.5–20.5	3		
20.5–25.5	5		
25.5–30.5	4		
30.5–35.5	3		
35.5–40.5	2		
	<u>20</u>		
	$n = 20$		

Step 2 Find the midpoints of each class and enter them in column C.

$$X_m = \frac{5.5 + 10.5}{2} = 8 \quad \frac{10.5 + 15.5}{2} = 13 \quad \text{etc.}$$

□ Date Description

Step 3 For each class, multiply the frequency by the midpoint, as shown, and place the product in column D.

$$1 \cdot 8 = 8 \quad 2 \cdot 13 = 26 \quad \text{etc.}$$

The completed table is shown here.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

Step 4 Find the sum of column D.

Step 5 Divide the sum by n to get the mean.

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

□ Date Description

The Median

The **median** is the midpoint of the data array. The symbol for the median is MD.

Example: The number of rooms in the seven hotels in downtown Pittsburgh is 713, 300, 618, 595, 311, 401, and 292. Find the median.

Solution

Step 1 Arrange the data in order.

292, 300, 311, 401, 595, 618, 713

Step 2 Select the middle value.

292, 300, 311, 401, 595, 618, 713



Median

Hence, the median is 401 rooms.

□ Date Description

Example: The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

Solution

656, 684, 702, 764, 856, 1132, 1133, 1303

↑

Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$MD = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

□ Date Description

The Mode

The value that occurs most often in a data set is called the **mode**.

Example: Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

Solution

It is helpful to arrange the data in order although it is not necessary.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5

Since \$10 million occurred 3 times—a frequency larger than any other number—the mode is \$10 million.

❑ Date Description

If each value occurs only once, there is no mode

Example: The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

Solution

Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

□ Date Description

The Midrange

The **midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$MR = \frac{\text{lowest value} + \text{highest value}}{2}$$

Example: In the last two winter seasons, the city of Brownsville, Minnesota, reported these

numbers of water-line breaks per month. Find the midrange.

2, 3, 6, 8, 4, 1

$$MR = \frac{1 + 8}{2} = \frac{9}{2} = 4.5$$

□ Date Description

The Weighted Mean

Find the **weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum wX}{\sum w}$$

where w_1, w_2, \dots, w_n are the weights and X_1, X_2, \dots, X_n are the values.

A student received the following grades, find the student's grade point average

Course	Credits (w)	Grade (X)
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

□ Data Description

Variance and Standard Deviation

The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

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

where

X = individual value

μ = population mean

N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ .

The corresponding formula for the population standard deviation is

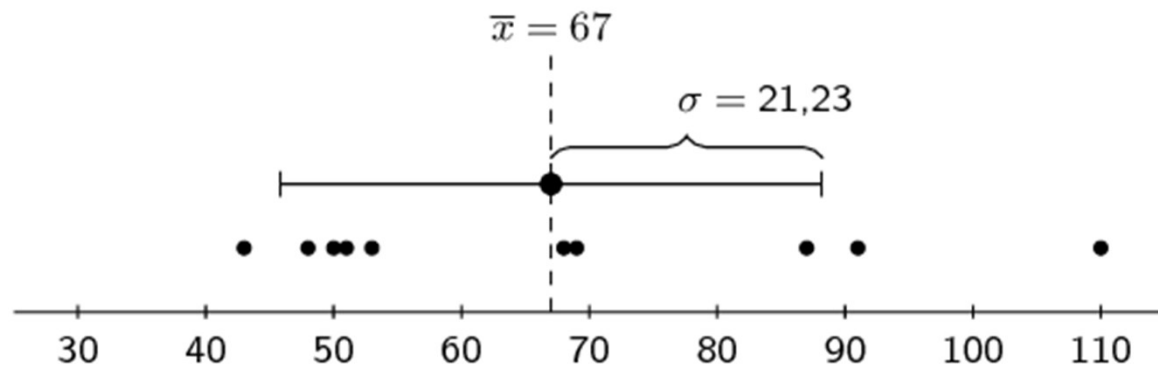
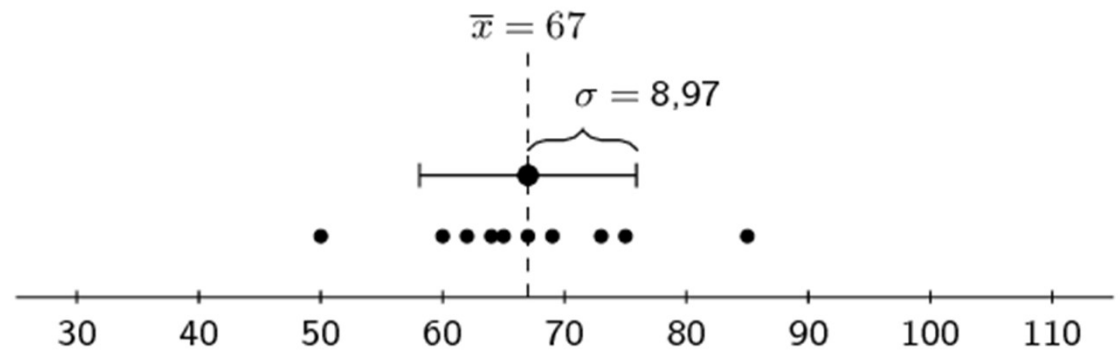
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

□ Data Description

Variance measures how far individuals in a group are spread out.

Conversely, **Standard Deviation** measures how much observations of a data set differs from its mean.

If the variance or standard deviation is large, the data are more dispersed.



□ Data Description

Variance and Standard Deviation for Grouped Data

Example: Find the variance and the standard deviation for the frequency distribution of the data in the data represent the number of miles that 20 runners ran during one week.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

□ Date Description

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is n . The sum of column D is $\sum f \cdot X_m$. The sum of column E is $\sum f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n - 1)}$$

Step 6 Take the square root to get the standard deviation.

□ Data Description

Coefficient of Variation

is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

The **coefficient of variation**, denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$\text{CVar} = \frac{s}{\bar{X}} \cdot 100\%$$

For populations,

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

□ Date Description

Example: The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5.

The mean of the commissions is \$5225, and the standard deviation is \$773.

Compare the variations of the two.

The coefficients of variation are

$$CVar = \frac{s}{X} = \frac{5}{87} \cdot 100\% = 5.7\% \quad \text{sales}$$

$$CVar = \frac{773}{5225} \cdot 100\% = 14.8\% \quad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

□ Data Description

Measures of Position

Used to locate the relative position of a data value in the data set

A **z score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is z . The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \bar{X}}{s}$$

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

The z score represents the number of standard deviations that a data value falls above or below the mean.

□ Date Description

Example: A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5.

Compare her relative positions on the two tests.

First, find the z scores.

For calculus the z score is $z = \frac{X - \bar{X}}{s} = \frac{65 - 50}{10} = 1.5$

For history the z score is $z = \frac{30 - 25}{5} = 1.0$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

□ Date Description

Revision

Twelve batteries were tested to see how many hours they would last. The frequency distribution is shown below.

Hours	Frequency
1–3	1
4–6	4
7–9	5
10–12	1
13–15	1

Find each of these.

a. Mean

c. Variance

z-score of

b. Modal class

d. Standard deviation

z-score of

Thank you

End of Chapter 3