Math 10 MPS

Course Pack

Cheryl Jaeger Balm

1 Descriptive Statistics

1.1 Key terms

Statistics deals with the **collection**, **presentation** and **analysis** of data.

Example 1.1. We are interested in the average number of hours of sleep a De Anza student slept last night.

1. Collection...
2. Presentation
3. Analysis

In this example:

- Who did we ask? i.e. what people did we talk to?

  These people are called the **sample**.
- What people were we actually interested in?

  These people are called the **population**.
- What did we ask the people in our *sample*?

  The answer to this question is called the **variable**.
- What answers did people give?

  All of these different answers are called **data**.

Six Key Terms

– **Population:**

– **Sample:**

- A sample must be **representative**. A **representative sample** has the same characteristics as the population.

– **Parameter:**

– **Statistic:**

- What are the **parameter** and **statistic** in Example 1.1?

- Study tip: The **P**arameter goes with the **P**opulation and the **S**tatistic goes with the **S**ample.

– **Variable:**

– **Data:**

- There are 2 main types of variables.
  1. **Numerical**
  2. **Categorical**

You Try It

Example 1.2. You are interested in the proportion of registered voters in California who support a California greenhouse gas emissions law that requires the state to reduce its greenhouse gas emissions by at least 17% over the next decade. A survey of 834 registered voters in California is taken.

1. Population:
2. Sample:
3. Parameter:
4. Statistic:
5. Variable:
6. Data:

Example 1.3. You want to determine the average number of glasses of milk college students drink per day. In your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4.

1. Population:
2. Sample:
3. Parameter:
4. Statistic:
5. Variable:
6. Data:

1.2 Types of Data

There are 2 main types of data.

1. **Qualitative Data**
    - What is it?
    - Qualitative data corresponds to ____ variables.
    - Examples:

2. **Quantitative Data**
    - What is it?
    - Qualitative data corresponds to ____ variables.
    - Examples:
    - There are 2 main types of quantitative data

        (a) **Discrete quantitative data**

        (b) **Continuous quantitative data**

Example 1.4. Identify each of the following variables as having qualitative data, discrete quantitative data or continuous quantitative data.

1. Weight of your backpack
2. Number of books in your backpack
3. Brand of your backpack

You Try It

Example 1.5. Identify each of the following variables as having qualitative data, discrete quantitative data or continuous quantitative data. Are any of the variables *ambiguous*?

1. The number of shoes you own
2. Type of car you drive
3. Where you go on vacation
4. Distance from your house to the nearest grocery store
5. Number of classes you are taking this quarter
6. Political party preference
7. Weight of sumo wrestlers
8. Amount of money won playing poker

Summary Flow Chart

1.3 Levels of Measurement

How data is measured is called its level of measurement.

- **Nominal:** This is *qualitative data*. There is no implied order or ranking in the data.

  Examples

  - How a student commutes to school
  - A student's major

- **Ordinal:** This is data that can be *ranked*, but there is no set measurement associated with each of the different responses.

  Examples

  - 1st place, 2nd place, etc., such as in a race
  - Preference ratings such as 1 = strongly agree, 2 = agree, 3 = indifferent, 4 = disagree, 5 = strongly disagree

- **Interval:** This is *quantitative data* that is ordered and has *constant differences* for intervals, but there is no definite *natural* "0" value that can be used as the base point for comparisons. Differences can be measured, but ratios have no meaning.

  Example

  - Temperature in ^{o} Celsius or ^{o} Fahrenheit.

    These are arbitrary scales. Comparisons differ when using ^{o}C versus ^{o}F.

    80^{o}F is approximately 27^{o}C and 40^{o}F is approximately 4.5^{o}C. Is it twice as hot (80^{o}F is two times 40^{o}F) OR is it 6 times as hot (27^{o}C is about 6 times 4.5^{o}C)?

- **Ratio:** This is *quantitative data* that has constant differences AND has a *natural* "0" value that can be used as a base point for ratio comparisons

  Examples

  - *Distance* traveled from home to school.

    If Asaf travels 3 miles and Ben travels 6 miles, Ben travels twice as far as Asaf.

  - *Weights* of packages of flour in a supermarket.

    A 5 pound bag of flour weighs one half as much as a 10 pound bag of flour.

1.4 Sampling Methods

- Remember, a sample must be ______________________________
- If sampling is poorly done, it can lead to ______________________________

Exploration

Example 1.6. Which of the following seem like sampling methods that will give *representative samples*?

1. From the population of all U.S. residents, we sample 100 residents from each state.
2. From the population of California residents, we randomly choose 20 zip codes and sample every resident of those zip codes.
3. From the population of Cupertino residents, we sample people as they walk out of a certain Whole Foods store.

Five Sampling Methods (4 good and 1 bad)

1. **Simple Random Sample (SRS)**

   Step 1:

   Step 2:

   Example:

2. **Systematic Sample**

   Step 1:

   Step 2:

   Step 3:

   Example:

3. **Stratified Sample**

   Step 1:

   Step 2:

   Example:

4. **Cluster Sample**

   Step 1:

   Step 2:

   Step 3:

   Example:

5. **Convenience Sample**

   Step 1:

   Step 2:

   Example:

You Try It

Example 1.7. Determine the type of sampling used (SRS, stratified, systematic, cluster or convenience).

1. A soccer coach randomly selects 6 layers from a group of boys age 8-10, 7 players from a group of boys age 11-12, and 3 players from a group of boys age 13-14 to form a recreational soccer team.

2. From a list of high-tech companies, a pollster randomly selects 5 of them and then interviews all human resources personnel in those companies.

3. An engineering researcher interviews 50 randomly selected female engineers and 50 randomly selected male engineers.

4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

6. To determine how many pairs of jeans a college student owns, a student interviews classmates in his algebra class.

Calculator Instructions for TI-84

Creating a List

To store a list of data in L_1

1. Press the STAT key. Then press ENTER to EDIT your data lists.
2. If you have data in L_1, clear it as follows:
   (a) Use the arrow keys to highlight the top of the column L_1
   (b) Press CLEAR
   (c) Press ENTER
3. Enter your data into L_1 using the number keys, pressing ENTER after each entry.

Generating Random Numbers

To generate a single positive integer

1. Press MATH
2. Arrow over to PRB
3. Arrow down to randInt(, then press ENTER.
4. After the parentheses, enter: **lowest value, highest value**).
5. Your screen should read randInt(lowest value, highest value). Press ENTER.

To generate a list of n positive integers

- Follow the basic steps above to enter the command randInt(lowest value , highest value, n).

To store a list of n positive integers in L_1

1. Enter the command randInt(lowest value , highest value, n). Do not press ENTER yet.
2. Press STO. (This stands for "store".)
3. Press LIST by first pressing the blue button 2ND and then pressing STAT
4. Press ENTER while L_1 is highlighted.
5. Your screen should read randInt(lowest value, highest value, n) $\rightarrow L_1$. Press ENTER.
6. To view the list, press STAT then ENTER.

1.5 Survey Questions, Bias and Statistical Studies

It is very important to think critically about the validity and results of statistical studies rather than blindly believing the results of all studies. Bias occurs when a poll or survey produces results that do not reflect the true opinions or beliefs of the general population. This is often a result of the methods used to conduct the survey or the wording of the questions asked.

Common problems in statistics to beware of:

- A sample should be *representative* of the population. A sample that is not representative of the population is called **biased**. Self-selected samples and samples of small size might be biased.
- **Non-Response Bias:** In some surveys, members of the survey group may refuse to answer certain questions or may refuse to participate in the survey. This can particularly happen with phone surveys or mail-in surveys. This also can include when a group of people is excluded from participating in the poll, intentionally or not. For example, if a phone survey uses only randomly chosen land-line phone numbers, those who have only a cell phone would not have a chance to be included in the survey.
- Self-Funded or Self-Interest Studies
- Misleading use of data: Improperly displayed graphs, incomplete data, lack of context, or not enough information given to understand the data.
- Causality and Confounding
  - A relationship between two variables does not necessarily imply that one causes the other. They may both be affected by some other variable. **Correlation is not causation.**
  - **Confounding factors:** When the effects of multiple factors on a response can not be separated, it becomes difficult or impossible to draw valid conclusions about the effect of each factor.
- **Response Bias**
  - How questions are asked is very important in surveys. A survey question can be worded in such a way as to direct a person to answer in only one way. This can be intentional or unintentional. Questions should be asked in a way that is **fair** and not vague.
  - Often, when questions about controversial issues are asked, survey respondents may give answers contrary to their true beliefs in order to conform to a societal standard they believe is acceptable.
  - `https://www.qualtrics.com/blog/writing-survey-questions/`

Discussion

Example 1.8. Here are some questions from recent polls and surveys regarding same sex marriage. Discuss the issues of bias/fairness in each question. The decide whether the question is clear or ambiguous. If you determine the question is biased and/or ambiguous, identify the word(s) causing the bias and rewrite the questions to make them clear and fair.

1. Should states continue to discriminate against couples who want to marry and who are of the same gender?

2. Do you support marriage equality?

3. Should states be forced to legalize homosexual marriage over the wishes of a majority of the people?

4. Do you think marriages between same-sex couples should or should not be recognized by the law as valid, with the same rights as traditional marriages?

You Try It

Example 1.9. A large city is proposing a parcel tax to support education. Each property owner would be assessed a tax of $100 per property per year. The parcel tax will be voted on by voters in the next election. It will pass if 2/3 of the voters vote in favor of the tax.

1. Which survey below do you think would produce the most accurate prediction of the election results and why? Think about the **types of sampling** used and the **fairness of the questions** asked in your answer.

   I A group of parents and teachers supporting the parcel tax randomly select and call residents in the city. They identify themselves as members of the Parent Teachers Association for the school system and ask the person who answers the telephone call if they support the parcel tax.

   II A TV news station in the city conducts a Facebook survey. Viewers are asked whether they favor or oppose the tax and are given instructions to visit the TV station's Facebook page to respond with their opinion. The poll is publicized and responses are solicited by announcements on the TV station's evening news programs.

   III A professional polling organization conducts a survey by randomly calling selected residents in the city. If the resident is a registered voter, she is asked whether she favors the parcel tax, opposes the parcel tax, or has no opinion. These three choices are presented to the individual in random order, so that not all respondents hear the choices in the same order.

2. For each of the other two surveys, what problems do you think there might be with the information obtained and why?

   -
   -

Once we have picked a subject for study and identified our population, we need to collect data for the study. There are two primary ways that studies can be conducted: **observational studies** and **designed experiments**.

- **Observational study:**

- **Designed experiment:**

You Try It

Example 1.10. Read the following study design summaries then answer the questions below.

- **Study I:** Employees of a company are randomly divided into two groups. Group A gets classroom training from an instructor who is available to help and answer questions; Group B gets training via online software with an online discussion board available to get help and answers to questions.
- **Study II:** Researchers are studying whether retirement age affects the rate of memory problems in senior citizens. A survey of retired senior citizens showed that those who had retired earlier tended to have a higher incidence of memory problems after retirement than those who had retired at an older age.
- **Study III:** 300 randomly selected individuals are asked if they had been on a diet in the last 8 weeks and how much their weight has changed over the last 8 weeks. Weight change for dieters and non-dieters are compared.
- **Study IV:** 100 individuals are put on a low fat diet, 100 on a low carb diet and 100 eat their normal diet. Their weight change over an 8 week period is recorded.

1. For each of the above, determine whether it is an observational study or a designed experiment?

   I ______________________________

   II ______________________________

   III ______________________________

   IV ______________________________

2. What problem can you see in the design of Study II?

3. Which weight loss study (III or IV) do you think would give the best information about the effect of diet on weight loss? Why?

Observational studies do not allow a researcher to claim *causation*, only *association*. That is because of the possible presence of **lurking variables**. A lurking variable is a variable that was not considered in a study, but affects the results of the study.

Example 1.11. A widely reported study in May, 2012 looked at the association of coffee-drinking with mortality. The study followed over 400,000 men and women ages 50 71 years of age at baseline. Participants with cancer, heart disease, and stroke were excluded. Participants were asked their level of coffee consumption at the beginning of the study and then were monitored for mortality between 1995 and 2008. After adjusting for age, tobacco smoking, body weight and other factors, it was observed that those who drank 4 or 5 cups of coffee per day had lower mortality rates than those who drank less than 1 cup of coffee per day. *(Neal D. Freedman, Ph.D., Yikyung Park, Sc.D., Christian C. Abnet, Ph.D., Albert R. Hollenbeck, Ph.D., and Rashmi Sinha, Ph.D. N Engl J Med 2012; 366:1891-1904 May 17, 2012 DOI: 10.1056/NEJMoa1112010)*

Name 3 possible **lurking variables** in this study.

Vocabulary for Experiments:

Treatment	Placebo	Response
Control Group	Double Blind	

Example 1.12. A *designed experiment* is done to test a new drug that is supposed to relieve pain. The patients enrolled in the study are randomly divided into two groups.

- One group is given the new drug, called the ______________________________.
- The other group is called the ______________________________ group and is given the ______________________________ instead of the treatment.
- The ______________________________ of both groups is measured to determine if the drug is more effective at relieving pain than not receiving the drug.
- The study is ______________________________because neither the patient nor the doctor knows who is receiving the new drug and who is receiving the placebo.
- If a patient develops problems, the doctor works with the study administrator who knows who is receiving the drug and who is receiving the placebo. The doctor, the study administrator, and a statistician are part of a team of people who evaluate the effectiveness of the drug based on the results of the study.

2 Statistical Graphs

Graphs give us a way to ____

2.1 Histograms

A **histogram** is a special kind of bar graph displaying quantitative (numerical) data.

- Consecutive bars should be touching. There should not be a gap between *consecutive* bars.
- A gap should occur only if an interval does not have any data lying in it (so the bar height is zero).

Histograms are useful for *visualizing* large data sets. We will begin by learning to draw histograms *by hand*. We will then learn two different approaches to using a *graphing calculator* to draw histograms.

Example 2.1. Collect data and draw a *histogram* representing the number of classes each student in this class is currently taking. Be sure to use a *ruler*.

Note how the **axes** are labeled:

- Vertical axis =
- Horizontal axis =

Sometimes we want each bar of a histogram to represent a range of data instead of just one number. We call this range a **class**, and in this course we will only use class intervals of equal width.

Example 2.2. Life expectancy at birth, data from the U.S. Bureau of the Census 2005 International Data Base, includes 277 countries.

Life Expectancy: Interval Class Limits	**Interval Class Boundaries**	**Number of Countries**
30–39	29.5 to 39.5	6
40–49	39.5 to 49.5	25
50–59	49.5 to 59.9	19
60–69	59.5 to 69.5	38
70–79	69.5 to 79.5	120
80–89	79.5 to 89.5	19

Construct a histogram using this data.

Vocabulary:

- **Class Limits:** Lowest and highest possible data values in an interval
- **Class Boundaries:** Numbers used to separate the classes, but without gaps. *Boundaries use one more decimal place than the actual data values and class limits. This prevents data values from falling on a boundary, so no ambiguity exists about where to place a particular data value.*
- **Class Width:** Difference between two consecutive class boundaries

In Example 2.2, the first class *limits* are ____________________, the first class *boundaries* are ____________________, and the class *widths* are ____________________.

Calculator Instructions for TI-84

Creating a Histogram

1. **Enter your data**
   (a) If you are entering *each piece* of raw data (i.e. everything has *frequency* 1), enter your data into L_1.
   (b) If you are entering data *with frequencies*, enter the *data* into L_1 and the *frequencies* into L_2.
2. **Set up your plot**
   (a) Press STAT PLOT by first pressing the blue button 2ND and then pressing Y=
   (b) Press ENTER while Plot1 (or your desired plot) is highlighted
   (c) ENTER while On is highlighted
   (d) For Type, press ENTER while the histogram icon is highlighted
   (e) For Xlist, enter L_1 by pressing LIST (2ND then STAT) then pressing ENTER while L_1 is highlighted For Freq
      - If frequencies were entered in L_2, enter L_2 by pressing LIST (2ND then STAT) then pressing ENTER while L_2 is highlighted
      - If frequencies were not entered into L_2, enter the number 1.
3. **Set up your window** (*(DO NOT USE ZoomStat)*
   (a) First press Y= and make sure that no functions are entered and only Plot1 (or your desired plot) is highlighted.
   (b) Press WINDOW
      - Xmin= lower *boundary* of first interval
      - Xmax= upper *boundary* of last interval
      - Xscl= class width (scl stands for scale)
      - Ymin= 0
      - Ymax= greatest frequency (i.e. height of tallest bar
      - Yscl= tick mark spacing on y-axis
      - Xres= 1
   (c) Press GRAPH

Reading Your Histogram

- Once your histogram is graphed, press TRACE and look at the bottom of the screen.
  - min= lower boundary for the interval
  - max= upper boundary for the interval
  - n= bar height (i.e. frequency)
- Arrow left and right to see this information for other classes

Exploration

Example 2.3. Calories in Girl Scout Cookies

Cookie	**Cals**	**Cookie**	**Cals**
Savannah Smiles	28	Rah-Rah Raisins	60
Shortbread	30	Carmel deLites	65
Trefoils	32	Peanut Butter Patties	65
Cranberry Citrus Crisps	38	Samoas	70
Thin Mints	40	Tagalongs	70
Do-si-dos	53	Toffee-Tastic	70
Peanut Butter Sandwich	53	Lemonades	75
Trios	57	Thanks-a-Lot	75

1. Use the data above to complete the first two columns of the tabel below, using equal class widths.
2. Use your calculator to construct a histogram. Be sure to use a ruler when copying your histogram.
3. Use the TRACE button on your calculator to complete the third column in the table.

Cals per Cookie: Intervals (Limits)	**Interval Class Boundaries**	**Number of Cookie Types** (Freq.)
20–29		

Histogram:

You Try It

Example 2.4. Plants are being studied in a lab experiment. We are interested in the number of flowers on a plant. Our sample of 16 plants give the following data.

Number of flowers on the plant	1	2	3	4	5	7
Frequency:	4	5	3	2	1	1

1. What does the word "frequency" mean in the table above?

2. Use your calculator to construct a histogram of the data:

You Try It

Example 2.5. Community College Enrollment Fall 2014: Below is the data for the total number of students enrolled in the 27 community colleges comprising Bay and Interior Bay regions (Regions III and IV) of all CA community colleges, according to `http://datamart.cccco.edu/Students/Student_Term_Annual_Count.aspx`. Use this data to construct a histogram with class widths of 5,000 students, starting with the class 0–4,999.

Community College	**Enrollment**	**Community College**	**Enrollment**	**Community College**	**Enrollment**
Alameda	5,461	Los Medanos	8,689	Ohlone	11,065
Merritt	6,085	Mission	8,793	Chabot Hayward	13,177
Gavilan	6,298	San Jose City	8,906	Cabrillo	13,444
Berkeley City	6,312	San Mateo	8,922	Foothill	14,924
Cañada	6,315	Skyline	9,690	Diablo Valley	19,812
Marin	6,418	Hartnell	9,624	De Anza	22,715
Contra Costa	6,892	Evergreen Valley	8,953	San Francisco (non-credit)	23,159
Las Positas	8,364	West Valley	10,174	San Francisco	23,575
Monterey	8,464	Laney	10,747	Santa Rosa	26,288

Histogram:

Questions:

1. Which bar is tallest and what does this represent? Answer in a complete sentence.

2. What class does De Anza belong to?

2.2 Stem and Leaf Plots

Stem and leaf plot are sometimes used to organize small data sets. To make a stem and leaf plot:

1. Order the data from smallest to largest.

2. Divide each data point into the **leaf** (the last digit) and the **stem** everything that comes before the leaf).

3. Create the plot

Example 2.6. The following are the scores that students in a precalculus class got on their first exam. The data is already sorted from smallest to largest. Make a stem and leaf plot from the data.

23	42	49	49	53	55
55	61	63	67	68	68
69	69	72	73	74	78
80	83	88	88	88	90
94	94	94	94	96	100

An **outlier** is a piece of data that doesn't fit with the rest. We'll talk more specifically about finding outliers later, but based on your stem and leaf plot, do you think the data from Example 2.6 has any potential outliers?

Calculator Instructions for TI-84

Sorting data

To sort a list of data in L_1 from smallest to largest

1. Enter the data into L_1
2. Pre STAT
3. Arrow down to SortA(and press ENTER. (The "A" stands for ascending.)
4. Pres LIST (2ND then STAT) and the press ENTER while L_1 is highlighted.
5. Your screen should read SortA(L_1. Press ENTER.

You Try It

Example 2.7. The table below shows the number of baseball games won by each Major League Baseball Team in the 2015 regular season. Use it to create a stem and leaf plot for number of games won.

Team	**Wins**	**Team**	**Wins**
Arizona Diamondbacks	79	Atlanta Braves	67
Baltimore Orioles	81	Boston Red Sox	78
Chicago Cubs *(Go Cubs!)*	97	Chicago White Sox	76
Cincinnati Reds	64	Cleveland Indians	81
Colorado Rockies	68	Detroit Tigers	74
Houston Astros	86	Kansas City Royals	95
L.A. Angels of Anaheim	85	Los Angeles Dodgers	92
Miami Marlins	71	Milwaukee Brewers	68
Minnesota Twins	83	New York Mets	90
New York Yankees	87	Oakland Athletics	68
Philadelphia Phillies	63	Pittsburgh Pirates	98
San Diego Padres	74	San Francisco Giants	84
Seattle Mariners	76	St. Louis Cardinals	100
Tampa Bay Rays	80	Texas Rangers	88
Toronto Blue Jays	93	Washington Nationals	83

Do you see any potential outliers?

2.3 5-Number Summary and Boxplots

One way to summarize a data set is with the following **5-number summary**

- Minimum:
- Quartile 1:
- Median:
- Quartile 2:
- Maximum:

Example 2.8. Find the 5 number summary for the following data sets

1. Data: 1, 1, 3, 3, 5, 6, 7, 8, 9

2. Data: 1, 3, 5, 6, 7, 9

Calculator Instructions for TI-84

One-Variable Statistics

To find the 5-number summary for a set of data

1. Enter the data into L_1 (or the list of your choice). If you have a frequency list, enter it into L_2
2. Press STAT
3. Right-arrow over to CALC and then press ENTER while 1-Var Stats is highlighted.
   - List: Press LIST (2ND then STAT) and then press ENTER while L_1 is highlighted (or the list of your choice.
   - FreqList: If you have a frequency list, choose it by pressing LIST and then pressing ENTER while L_2 is highlighted. If you do not have a frequency list, leave this blank.
4. Arrow down to CALCULATE and press ENTER
5. Arrow down to the last 5 numbers, which are minX, Q_1, Med (median), Q_2 and maxX.

Exploration

Example 2.9. We are interested in the amount of money a student in this class spent on books for this quarter. We will first collect the data, and then use the 5-number summary to draw a **boxplot** of the data.

Data:

5-number summary:

- Minimum:
- Quartile 1:
- Median:
- Quartile 2:
- Maximum:

Boxplot *(use a ruler!)*:

The **interquartile range**, or **IQR**, is the span of the middle 50% of the data. Another way to think of this is that the IQR is the length of the box in the boxplot (without the whiskers). We can calculate the IQR with the formula

What is the IQR in Example 2.9?

Occasionally, a data set may contain values that are either unusually larger or smaller than most of the other data values. When this happens, we should question whether such data values were taken accurately. We call such values **outliers**. We can use the IQR to find outliers in a data set. A data value is deemed an outlier if it is more than 1.5× IQR above Q_3 or if it is more than 1.5× IQR below Q_1. Algebraically, we write:

Are there any outliers in our data in Example 2.9?

You Try It

Example 2.10. A class of 20 students had the following grades (out of 20 points) on a quiz.
2 5 8 10 12 12 12 14 14 14 15 15 17 17 17 18 20 20 20 20
Find the 5-number summary and use it to draw a boxplot for the quiz grade data.

Are there any outliers? If so, what are they?

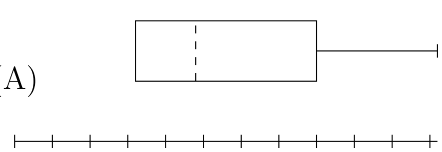
To summarize:

- The box shows where the middle 50% of the data values are located.
- The IQR is represented by the length of the box.
- The left WHISKER shows where the lowest 25% of the data values are located.
- The right WHISKER shows where the highest 25% of the data values are located.

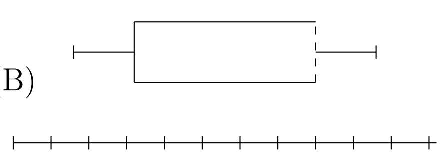
Exploration

Example 2.11. Explain what is strange about each boxplot and what it means.

(A)



(B)



3 Measures of Data

3.1 Measures of the Center of the Data: Mean, Median and Mode

Definitions:

- Median:

- Mode:

- Mean:

  - Symbols for mean:

The Law of Large Numbers: The larger a sample you take, the closer the sample mean $\overline{x}$ will be to the population mean μ.

Example 3.1. The data below shows the number of paperback books bought by shoppers at a bookstore.

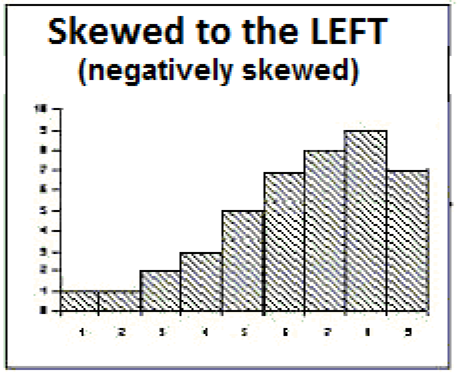
Number of books	Frequency (number of shoppers)
1	11
2	10
3	16
4	6
5	4
6	2
10	1

Find the mean, median and mode for the sample.

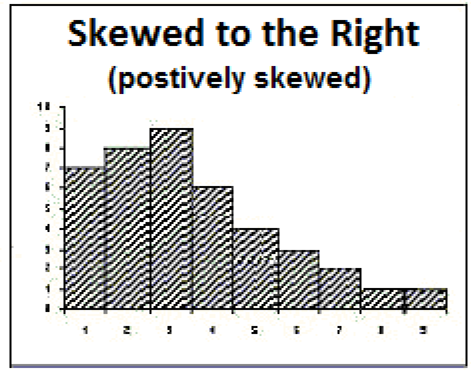
The 1-variable statistics (1-Var Stats) on your calculator will calculate the mean and median of a sample, but not the mode.

3.2 Shape of the Data: Skewness

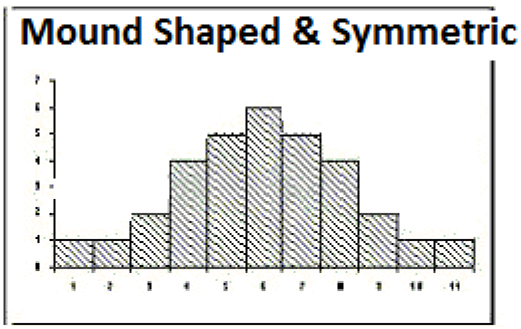
Shapes of Data Distributions



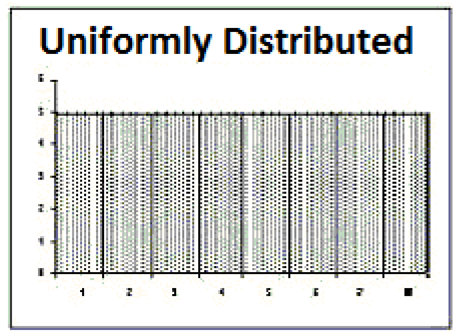
When data are skewed to the left generally the mean is less than the median

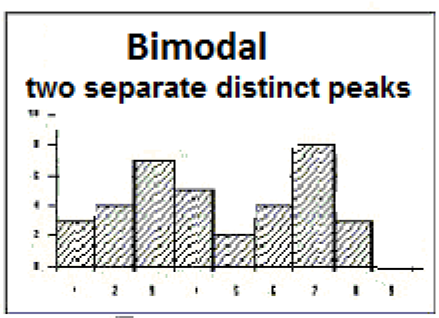


When data are skewed to the right generally the mean is greater than the median



For Symmetric data mean = median

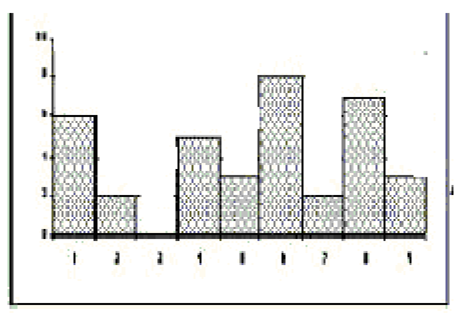




Distinct peaks appear as "hills" separated by a "valley"

Peaks do not need to be exactly the same height

IF data do not fit one of the descriptive terms for data, do not use a term that does not fit its shape.
Just describe what you see in the data if none of these descriptive terms apply



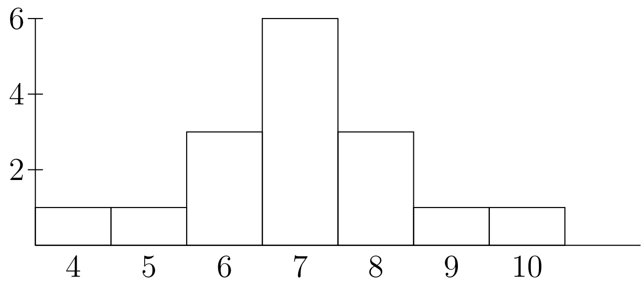
In general

- When data is **symmetrical**, mean = median = mode
- When data is **skewed to the left**, mean < median < mode
- When data is **skewed to the right**, mean > median > mode

If data are not skew, the *mean* (average) is usually the most appropriate measure of center of the data. If data are skew, the *median* is usually the most appropriate measure of center of the data.

Exploration

Example 3.2. For each of the histograms below, calculate the mean, median and mode of the data. Then determine if the data is symmetric or skewed in one direction, and which measure of the center would be most appropriate to use to describe the data.



Mean $\overline{x}$:

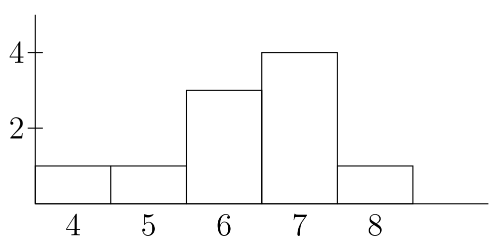
Median:

Mode:

Mean _____ Median _____ Mode

Skewed left, right or symmetric?

Best measure of the center?



Mean $\overline{x}$:

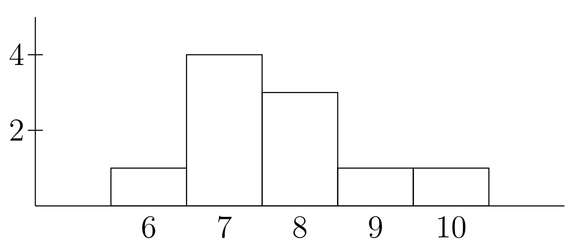
Median:

Mode:

Mean _____ Median _____ Mode

Skewed left, right or symmetric?

Best measure of the center?



Mean $\overline{x}$:

Median:

Mode:

Mean _____ Median _____ Mode

Skewed left, right or symmetric?

Best measure of the center?

3.3 Measures of the Spread of the Data

The **spread** of the data tells us how much *variation* there is in the data. The simplest way to measure this is the **range** of the data.

Range = Maximum Value – Minimum Value

Exploration

Example 3.3. A random sample of 6 students in a class were asked their age, and the answers given were 18, 23, 23, 24, 24, 32. What are the range, mean and median of this data?

Another class was sampled, and this time the ages given were 19, 19, 19, 19, 19, 33. What are the range, mean and median of this data?

The **standard deviation** measures variation (spread) in the data by finding the distances (deviations) between each data value and the mean (average). Standard deviation is the most common measure of spread in statistics. **Variance** is the square of the standard deviation and is also commonly used in statistics, but we will focus on *standard deviation* in this course.
Notation:

- Population standard deviation:

- Sample standard deviation:

Both population standard deviation σ and sample standard deviation s_X are part of your calculators 1-variable statistics (1-Var Stats), and are denoted σx and Sx, respectively.

Exploration

Example 3.4. Calculate the sample standard deviation s_X for each of the samples in Example 3.3. What do these numbers tell you about the *spread* of the data?

Discussion Your instructor will now show you how to calculate the sample standard deviation *by hand* for the first class in Example 3.3.

Standard deviation gives us a second way to calculate potential outliers. Any data value that is more than 3 standard deviations above or below the mean is a potential **outlier**.

You Try It

Example 3.5. A class of 20 students has a quiz every week. For the sixth week quiz, the grades are

2, 5, 8, 10, 12, 12, 12, 14, 14, 14, 15, 15, 17, 17, 17, 18, 20, 20, 20, 20

Find the mean $\overline{x}$ and the sample standard deviation s_X for the quiz scores. Are there any potential outliers?

3.4 Location of data: z-scores

A data point's **z-score** tells us how far away the data value is from the mean, measured in "units" of standard deviations. It is sometimes referred to as a *measures of relative standing*, and it describes the *location* of a data value as "how many standard deviations above or below the mean".
The z-score is calculated by:

Example 3.6. Anna is in the class whose quiz scores were given in Example 3.5. She scored 18 points on the quiz.

1. How many *points* above the average (mean) was Anna's score?

2. What is the z-score for Anna's grade on the quiz?

3. How many *standard deviations* above the average (mean) was Anna's score?

4. Did Anna perform better on the quiz when compared to the other students in her class? Use the z-score to explain and justify your answer.

Example 3.7. Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	GPA	School Mean GPA	School St. Dev.
John	2.5	2.0	1.0
Ali	77	75	10

Are high or low z-scores good or bad? It depends on the context of the problem. Read the problem carefully. Think about the context and the meaning of the numbers for that problem.

- **Positive z-scores correspond to numbers that are larger than the average.**
  - Higher than average is good for exam scores and salaries
  - Higher than average is bad for airline ticket costs or waiting time for a bus to arrive.
  - High z-scores are good for race speeds (fast) but bad for race times (slow).
- **Negative z-scores correspond to numbers that are smaller than the average.**
  - Lower than average is bad for exam scores and salaries.
  - Lower than average is good for airline ticket costs or waiting time for a bus to arrive.
  - Small z-scores are bad for race speeds (slow) but good for race times (fast).
- In some contexts, no value judgment applies; such as the number of children in a family.

You Try It

Example 3.8. The air at an industrial site is tested for a sample of 30 days to measure the level of two pollutants: Nitrogen Dioxide, NO_2, and Particulate Matter $PM_{2.5}$. (NO_2 and $PM_{2.5}$ are measured in different units, have different "safe" levels, and different effects on public health, so are not directly comparable.)
Suppose that for today's pollution readings:

- The level of NO_2 is 0.5 standard deviations below its average level: $z =$ _____
- The level of $PM_{2.5}$ is 0.8 standard deviations below its average level: $z =$ _____

Compare today's pollution levels for NO_2 and $PM_{2.5}$ to the average readings for the 30 day sample at this site. Which of today's pollutant levels would be considered better for this site? Explain by completing the sentence:

Today the level for pollutant __________ is better because...

(Note: Data underlying this example came from http://www.epa.gov/air/criteria.html

---

You Try It

Example 3.9. Here are wait times, in minutes, for a sample of 50 people waiting in line at the Department of Motor Vehicles (DMV).

Wait time at DMV (in minutes)	**Frequency** (number of people)
12	4
15	2
18	6
20	3
24	5
25	7
27	6
30	5
32	6
38	4
45	2

Find the mean and sample standard deviation. Then find the z-scores for the shortest and longest wait times shown. Write the interpretations in complete sentences in the context of the problem for each of these z-scores.

4 Frequency Tables and Percentiles

We have already been using the word "frequency" throughout these notes, and, in fact, we have already seen *frequency tables*. Now it is time to talk about these things in more depth.

- **Frequency** = count
- **Relative frequency** = proportion = $\dfrac{\text{freuqency}}{\text{number of observations}}$
- **Cumulative relative frequency (CRF)** = sum of relative frequencies for all data up to and including current interval

4.1 Relative Frequency Tables

Exploration

Example 4.1. We are interested in the number of days per week that a De Anza student in this class works. Collect the data from your classmates, complete the table, and answer the questions.

Data	Freq.	Rel. Freq.	CRF
0			
1			
2			
3			
4			
5			
6			
7			
Total			

1. What percent of students work 0 days per week?
2. What percent of students work 1 to 3 days per week?
3. What percent of students work less than 5 days per week?
4. What percent of students work at least 5 days per week?
5. What percent of students work at most 5 days per week?

You Try It

Example 4.2. Recall Example 2.4 where we were studying the number of flowers on a plant and had the following data.

Number of flowers	**Freq.**	**Rel. Freq.**	**CRF**
1	4		
2	5		
3	3		
4	2		
5	1		
7	1		

Complete the table above then answer the following questions.

1. What percent of plants had 3 flowers?

2. What percent of plants had at most 3 flowers?

3. What percent of plants had more than 3 flowers?

4. What percent of plants had at least 5 flowers?

You Try It

Example 4.3. Twenty students were asked how many hours on average they worked per day. The results were as follows:

5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3

What is the variable? $X =$ ________________________________

What are the data? ________________________________

Complete the frequency table.

Data	**Freq.**	**Rel. Freq.**	**CRF**

1. What percentage of students work 5 hours per day?

2. What percentage of students work between 3 and 6 hours per day, inclusive? In other words, what percentage of students are such that $3 \leq X \leq 6$?

3. What percentage of students work at most 4 hours per day?

4. What percentage of students work less than 6 hours per day ?

5. What per percentage of students work more than 5 hours per day?

4.2 Percentiles

The n^{th} **percentile** is a number that $n\%$ of the data is below. The median and quartiles that we used to make boxplots are examples of *percentiles.*

	Percentile
First Quartile Q_1	
Median	
Third Quartile Q_3	

Example 4.4. Let's use the 20 quiz scores from Example 3.5 to visualize a few percentiles.

The 50^{th} percentile:

2	5	8	10	12	12	12	14	14	14	15	15	17	17	17	18	20	20	20	20

The 40^{th} percentile:

2	5	8	10	12	12	12	14	14	14	15	15	17	17	17	18	20	20	20	20

The 20^{th} percentile:

2	5	8	10	12	12	12	14	14	14	15	15	17	17	17	18	20	20	20	20

The third quartile:

2	5	8	10	12	12	12	14	14	14	15	15	17	17	17	18	20	20	20	20

Write a sentence explaining the value of the third quartile.

How to calculate a percentile using a relative frequency table:

- If you see the percent as a decimal in the CRF column, average that data value (in the first column) and the next higher data value.
- If you **don't** see the percent as a decimal in the last column, take the data value (in the first column) for the next higher percent.

Example 4.5. 50 students were asked the number of hours of sleep they had gotten before. Complete the table and answer the questions.

Hrs of sleep	**Freq.**	**Rel. Freq.**	**CRF**
4	2		
5	5		
6	7		
7	12		
8	14		
9	7		
10	3		

1. What is the 28^{th} percentile?

2. What is the 75^{th} percentile? What is another name for this?

3. What is the median?

4. What is the first quartile?

5. What is the 80^{th} percentile?

You Try It

Example 4.6. Use the frequency table to answer the questions.

Data	**Freq.**	**Rel. Freq.**	**CRF**
5	1	0.1	0.1
6	2	0.2	0.3
7	2	0.2	0.5
8	3	0.3	0.8
9	1	0.1	0.9
10	1	0.1	1.0

1. What is the 50^{th} percentile?

2. What is the 25^{th} percentile (Q_1)?

3. What is the 70^{th} percentile?

4. What is the 51^{st} percentile?

The raw data is shown in the table below. Mark each of the percentiles that you calculated on the data.

5	6	6	7	7	8	8	8	9	10

5 Probability

5.1 Probability Basics

A **probability** is the long-range ______________________________________ of an outcome. In other words, a *probability* is the *likelihood* or *chance* that an outcome will happen. A probability is a number between _____ and _____, inclusive.

Exploration

Example 5.1. Suppose we flip a coin several times. We are interested in how many times we get heads.

Number of flips	**Number of heads**	**Rel. Freq.** $= \dfrac{\#\text{ of heads}}{\#\text{ of flips}}$

As we flip more and more times, the *relative frequency* **stabilizes** at about ________________________.

- **Theoretical probability:**

- **Experimental probability:**

Even if we calculate the *probability* of an event occurring, we can never predict anything with complete accuracy because **randomness** exists in the world.

Probability Vocabulary Terms

- **Experiment:** an activity conducted under controlled circumstances.
  - Examples:

- **Outcome:** a possible result of an experiment.
  - Examples:

- **Sample Space:** the set of all possible outcomes of an experiment.
  - Examples:

- **Event:** a specified set of outcomes of an experiment - usually indicated by capital letters: A, B, etc.
  - Examples: For rolling a die, A = {1, 3, 4} would be an event; B = {5} is an event.
- **Probability of an event:** the long-term relative frequency of the event; i.e., the percentage of times the event approaches if the experiment is repeated many, many times.
  - Examples: P(A) = 0.5; P(B) = 0.1667

In general,

$$\text{P(A)} = \frac{\text{Number of ways to get the desired outcome(s)}}{\text{Number of all possible outcomes}}$$

Exploration

Example 5.2. Rolling 1 Die: Complete the table. Leave all probabilities as unreduced fractions.

Sample space S = {1, 2, 3, 4, 5, 6}

Event	odd	even	2 or 4	≤ 4	≥ 5
Event	A = {1, 3, 5}	B =	C =	D =	E =
Probability	$P(A) = \frac{3}{6}$	P(B) =	P(C) =	P(D) =	P(E) =

Compound events use **AND** or **OR** to relate to relate two or more events. We can also you **NOT** to consider an event not occurring.

- **AND: A and B** means *both* event A and event B occur, i.e. the outcome satisfies both events A and B. This includes items in common to both (or in the *intersection* of) A and B
- **OR: A or B** means *either* event A occurs or event B occurs *or both* occur, i.e. the outcome satisfies event A or B or both. This includes the *union* of items from these events.
- **NOT: A'** means event A does *not* occur, and is called the **complement of A**.

Example 5.3. Use the events in Exercise 5.2 to complete the following table.

Event	**Probability**
A and D = { }	P(A and D) =
C and E = { }	P(C and E) =
A or D = { }	P(A or D) =
C or E = { }	P(C or E) =
C' = { }	P(C') =

Discussion

Example 5.4. Complete the table below using data from your classmates. The sample space is:

S = students in this class

The events are:

C = students who own a cat
D = students who own a dog

Event	**Probability**
S and C =	
C and D =	
S or C =	
S' =	
C' =	

Complement Rule:

$$P(A) + P(A') =$$

In other words,

$$P(A') =$$

Addition Rule:

$$P(A \text{ or } B) =$$

Two events are **mutually exclusive** if they cannot *both* happen. In other words, A and B are *mutually exclusive* if P(A and B) = _____.

Remember, probabilities are always between _____ and _____, in other words,

5.2 Conditional Probability

a **conditional probability** is the probability that event A occurs *if we know* that outcome B has occurred. In other words, we want to know the probability of A occurring *given the condition* tat B has occurred. We write **P(A | B)**, and we say *"the probability of A given B"*.

Discussion

Example 5.5. Let's use the same events from 5.4 to find some *conditional* probabilities. Recall, the sample space is:

S = students in this class

The events are:

C = students who own a cat
D = students who own a dog

- P(C | D) =

- P(D | C) =

Conditional Probability Rule:

$$P(A \mid B) =$$

Multiplication Rule:

$$P(A \text{ and } B) =$$

Exploration

Example 5.6. Rolling 1 Die: Answer the following based on the events in Example 5.2

1. P(odd $\mid \leq 4$) =

2. P(even $\mid \geq 5$) =

3. P($\geq 5 \mid$ even) =

It's important to note the following.

- For AND, OR events, the order of listing the events does not matter and can be switched.
  - P(A and B) = P(B and A)
  - P(A or B) = P(B or A)
- **For CONDITIONAL PROBABILITY the order is important.**
  - $\mathbf{P(A \mid B) \neq P(B \mid A)}$ in most situations.

Two events are **independent** if the fact that one has occurred does not affect the probability of the other. In other words, A and B are independent if $\mathbf{P(A \mid B) = P(A)}$. Since *independent* and *mutually exclusive* are often confused, let's write the definition of each below.

1. **Mutually exclusive:**

   Example:

2. **Independent:**

   Example:

Exploration

Example 5.7. Decide whether each of the following are **independent**.

1. Repeated tosses of a coin.

2. Selecting 2 cards consecutively from a deck of 52 cards, **without replacement**.

3. Selecting 2 cards from a deck of cards, **with replacement**.

4. The numbers that show on each of two dice when tossed

Decide whether each of the following are **mutually exclusive**.

5. Selecting a card from a deck of cards and getting a card that is black and a diamond.

6. Selecting a card from a deck of cards and getting a card that is black and a king.

7. Being a day student and being a night student at De Anza College.

8. Being a full-time student and being a part-time student at De Anza College.

Test for Independence: Two events, A and B, are *independent* if any one of these three statements is true:

1. P(A | B) = P(A)
2. P(B | A) = P(B)
3. P(A and B) = P(A) · P(B)

It can be shown algebraically (using the multiplication rule) that if any one of these statements is true, then all three of them are true. So, to test if two events are independent, *we only need to chose one of these statements and show it is true.*

Example 5.8. Suppose we roll a pair of dice. Once die is black and the other is orange. The events we are interested in are:

$$O_2 = \text{the orange die is a 2}$$
$$D = \text{doubles}$$

There are three ways we can show that these two events are *independent*.

1. $\mathrm{P}(O_2|D) = \mathrm{P}(O_2)$

2. $\mathrm{P}(D|O_2) = \mathrm{P}(D)$

3. $\mathrm{P}(O_2 \text{ and } D) = \mathrm{P}(O_2) \cdot \mathrm{P}(D)$

Which of these calculations was easiest? Which was hardest?

You Try It

Example 5.9. Given the sample space of current full-time De Anza students, consider the events

M = taking a math class
S = taking a science class

Now, suppose it is known that P(M) = 0.6, P(S) = 0.5 and P(M and S) = 0.3

1. Are M and S independent? Justify your answer with a probability calculation.

2. Are M and S mutually exclusive? Justify your answer with a probability calculation.

You Try It

Example 5.10. Carlos plays soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in the next game.

A = he makes a goal on the first shot
B = he makes a goal on the second shot

Carlos tends to shoot in streaks. It is known that if he makes a goal, then the probability he will make a goal on his next shot is 0.90. Calculate the following.

1. P(A)

2. P(B)

3. P(B | A)

4. Are B and A independent?

5. What is the probability he makes both goals? This means, P(A and B) =?

6. Are A and B mutually exclusive?

7. What is the probability he makes the first goal or the second goal? That is, P(A or B) =?

Summary of Probability Rules

- P(A or B) = P(A) + P(B) – P(A and B)
- P(A') = 1 – P(A)
- $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$
- $P(A \text{ and } B) = P(A \mid B) \cdot P(B)$
- Independent events:
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
  - $P(A \text{ and } B) = P(A) \cdot P(B)$
- Mutually exclusive events: P(A and B) = 0

Example 5.11. In the US, based on federal poverty level guidelines:

Events: C = person is a child L = person lives in poverty

15% of the population live in poverty (are poor). P(__________) = __________

24% of the population are children P(__________) = __________

22% of children in live in poverty. P(__________) = __________

Find the probability that a randomly US resident is a child AND lives in poverty.

You Try It

Example 5.12. At a medical clinic patients can call or use the online website appointment system to make appointments.

- 40% of patients request an urgent appointment
- 30% of patients use the website appointment system to make appointments
- 10% of all patients use the website appointment system and request an urgent appointment

Events:

U = appointment is urgent W = appointment is made using website

1. Find the probability that the appointment is urgent given that a patient uses the website.

2. Find the probability that a patient uses the website to make an appointment if the appointment is urgent.

3. Find the probability that the appointment is urgent OR that a patient uses the website.

5.3 Contingency Tables

A **contingency table** displays data for two variables. This table shows the number of individuals or items in each category. We can use the data in the table to find probabilities. **All probabilities EXCEPT conditional probabilities have the grand total in the denominator.**

Conditional Probabilities: The condition limits you to a particular row or column in the table. Condition says "IF" we look only at a particular row or column, find the probability **The denominator will be the total for the row or column** in the table that corresponds to the condition.

Example 5.13. Fill in the missing values in the contingency table for hair color and type, then calculate the probabilities.

	Brown	Yellow	Black	Red	Total
Wavy	20		15	3	43
Straight	80	15		12	
Total					215

1. P(Wavy)

2. P(Brown or Yellow)

3. P(Wavy and Brown)

4. P(Red | Straight)

5. P(not Brown)

6. P(Wavy | Yellow)

7. P(Wavy or Red)

You Try It

Example 5.14. A large car dealership examined a sample of vehicles sold or leased in the past year. Data is classified by vehicle type (car, SUV, van, truck) and by type of sale (new vehicle sale, used vehicle sale or lease).

	Care	SUV	Van	Truck	Total
New vehicle sale	86	25	21	38	170
Used vehicle sale	39	13	4	22	78
Lease	34	12	6	0	52
Total	159	50	31	60	300

Suppose a vehicle in the sample is randomly selected to review its sales or lease papers.

1. Find the probability that the vehicle was leased.
2. Find the probability that a vehicle is a truck.
3. Find the probability that a vehicle is NOT a truck.
4. Find the probability that the vehicle was a car AND was leased.
5. Find the probability that the vehicle was used GIVEN THAT it was a van.
6. Find the probability that the vehicle was a van GIVEN THAT it was used.
7. Find the probability that the vehicle was used OR was a van.
8. Find the probability that the vehicle was leased OR was a truck.
9. Find the probability that the vehicle was a car GIVEN THAT it was new.
10. Find the probability that the vehicle was new IF it was a car.
11. Find the probability that a vehicle was new.
12. Find the probability that the vehicle was new AND was a car.
13. Find the probability that a vehicle was new OR a car.

Example 5.15. Answer the following using the table from Example 5.14.

1. Are the events New vehicle sale and Van independent?
2. Are the events SUV and Used vehicle sale independent?
3. Are the events New vehicle sale and Van mutually exclusive?
4. Are the events Lease and Truck mutually exclusive?