Math 140 Introductory Statistics Course Workbook

Fall 2017

DEPARTMENT OF MATHEMATICS M. SCHILLING

Table of Contents

Math 140 Class Questionnaire

Instructions: Your answers will become part of the class data base. Please answer each question honestly, but do not give your name. For numerical questions, give a single number answer, not a range (for example, answer 25 rather than 20-30).

1) Have you ever enrolled in a college or high school statistics course before? Yes No

2) Which of the following most closely reflects your attitude toward taking this course:

   a. Can't wait to start!
   b. I'm ambivalent
   c. I'd rather take a course on the history of garbage cans, but it's required for my major.

3) What is your class year? Freshman Sophomore Junior Senior Graduate student

4) Your gender: M F

5) Your month of birth:

6) Your height in feet and inches:

7) Your weight in pounds:

8) What is your perception of your own body? Do you feel that you are overweight, underweight, or about right?

9) Approximate number of hours you slept last night:

10) How many minutes do you typically spend in the shower? _____

11) Do you have a job?

   a. I work full time.
   b. I work part time.
   c. I currently don't have a job.

12) The <u>last digit</u> in your primary phone number:

13) Number of letters in your mother's first name:

14) The approximate distance in miles between your current residence and CSUN. Just make your best guess if you have to.

15) The amount of money you spent on your last haircut:

16) The amount of money you think your best friend would spend for your birthday gift:

17) The approximate number of pairs of shoes you own:

18) Pick a random number between 1 and 10.

19) A small ice cream shop has the following flavors. Which one would you pick?

a. Vanilla
b. Chocolate
c. Strawberry
d. Cherry
e. Chocolate chip
f. Rocky road
g. French vanilla
h. Raspberry
i. Banana nut
j. Pralines 'n cream
k. Rum raisin

20) How many hours would you prefer to spend alone on a typical day?

21) Do you believe in ghosts?

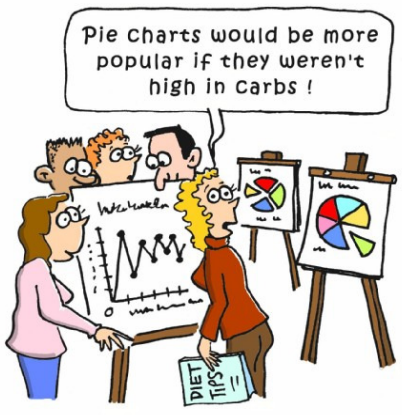
22) Have you ever had a personal encounter with a ghost?

23) Do you believe in love at first sight?

Numerical and Graphical Displays for a Categorical Variable

What you need to know:

- **Categorical variable; quantitative variable**
- **Distribution of a variable**
- **Numerical summaries for a categorical variable: category counts and percents**
- **Graphical displays for a categorical variable: bar graphs and pie charts**



Variables

Quantitative

- Numerical values
- "average" makes sense

Categorical

- Labeled values
- "average" doesn't make sense

1. For the Math 140 questionnaire, identify the variables for the following questions:

   #2:

   #10:

   #19:

2. Classify each of the 23 variables as categorical (C) or quantitative (Q):

   #1: #2: #3: #4: #5: #6: #7: #8: #9: #10: #11: #12:

   #13: #14: #15: #16: #17: #18: #19: #20: #21: #22: #23:

3. Here is the questionnaire data from another Math 140 class for Question #2: *Which of the following most closely reflects your attitude toward taking this course?* b, b, a, b, a, c, b, b, b, a, b, b, b, b, c, a, c, b, c, b, b, b, c, c, b, b, b, c, c, b, b, b, b, b, b, b, a, b, a

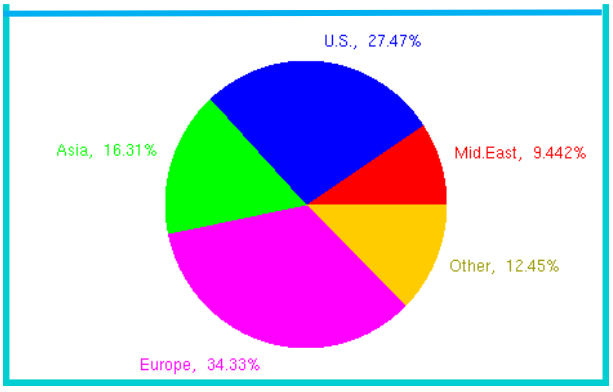
   a. Determine the percentages of students that fell into each category.

   b. What should your percentages add up to? Check this.

   c. Create an appropriate graphical display of the distribution of the variable. Make sure to label your graph appropriately.

   d. Summarize the results regarding the attitudes of this class toward taking this course. Use a well-constructed and interesting sentence or two.

4. Fortune magazine publishes the list of the world's billionaires annually. The 1992 list (Fortune, September 7, 1992. "The Billionaires." pp. 98-138) included 233 individuals or families. Their wealth, age and geographic location (Asia, Europe, Middle East, United States or Other) was reported. Look at the pie chart.

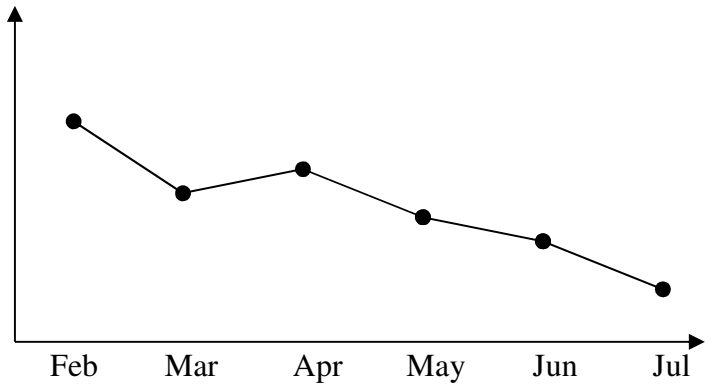


   a. What is the variable of interest?

   b. Summarize the information in the pie chart using a well-constructed and interesting sentence.

c. How could we check whether there might be any errors in the percentages given?

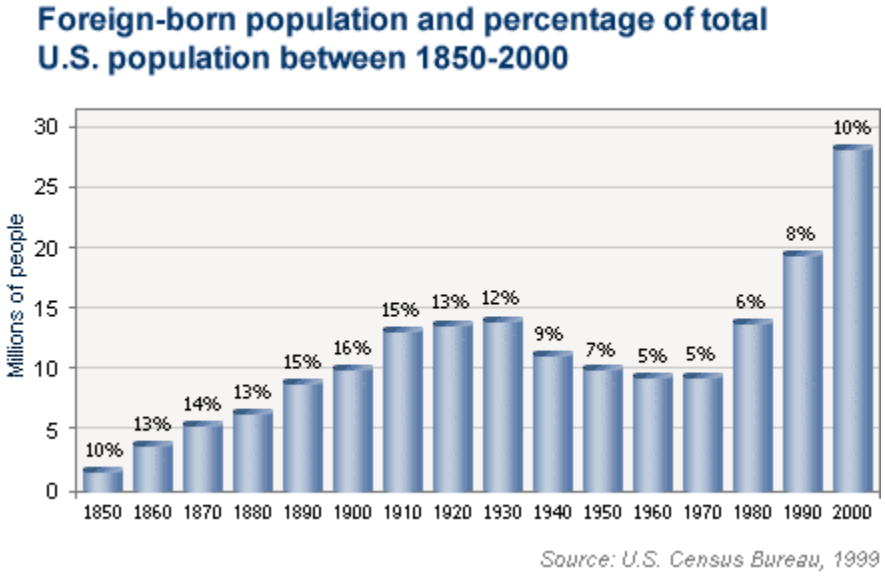
5. The graph below shows the number (in thousands) of SUV's sold by a certain company over the past six months:



a. Explain how this graph could be misleading.

b. How could the graph be improved?

6. Explain how a politician (for example) could use the information in the graph below to either argue that: (i) the number of foreign-born individuals in the U.S. has increased dramatically in the last 150 years, or (ii) the number of foreign-born individuals in the U.S. has *not* increased in the last 150 years.



(i)

(ii)

Additional Practice

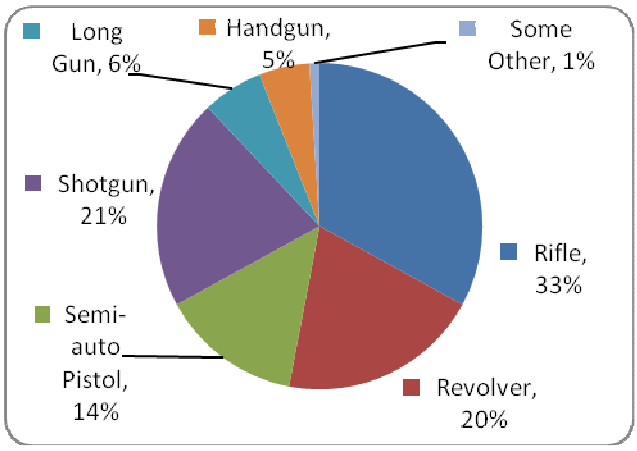
1. a. Explain what is meant by the *distribution* of a categorical variable.

2. List two ways that this distribution can be displayed visually.

3. For each of the following variables, indicate with Q or C whether it is a quantitative variable or a categorical variable.

   a. the color of an M&M candy
   b. the weight of an airplane
   c. how many miles a person walks in one day
   d. the age of a mother when her first child is born
   e. whether or not a student eats breakfast
   f. the length of a snake
   g. whether or not a car has automatic transmission or manual transmission/stick shift
   h. the number of calories in a pint of vanilla ice cream
   i. the running time of a Tom Cruise movie
   j. whether or not a state's name consists of one word
   k. the diameter of a pizza
   l. the number of dogs an animal shelter has
   m. the height of a sequoia tree
   n. the race of a person

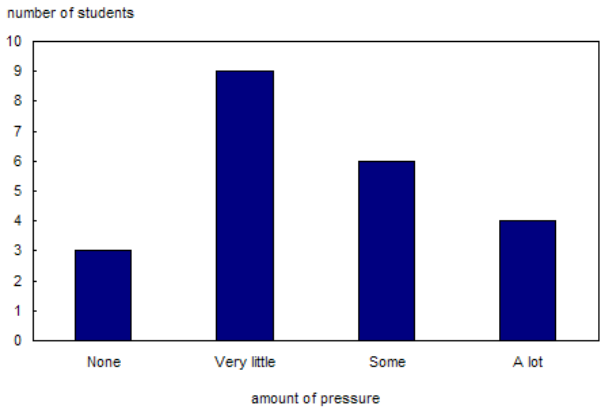
4. Consider the following pie chart:



a. What is the variable described in the pie chart?

b. Summarize what the pie chart shows.

5. The following bar graph shows the amount of pressure students experienced from schoolwork:

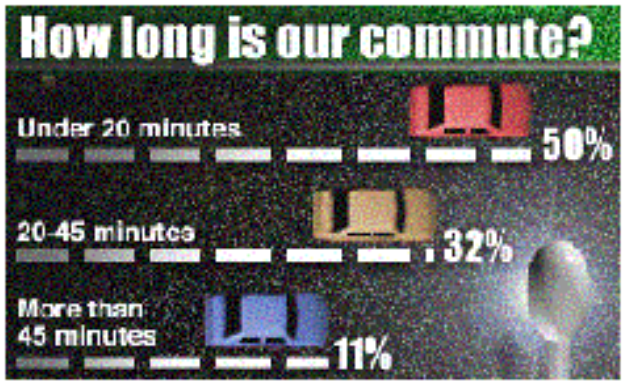


a. What is the variable described in the bar graph?

b. What percent of students feel no pressure from schoolwork?

6. In 2007, 5488 people were killed while working. Here is a breakdown of causes: transportation: 2234; contact with objects or equipment: 916; assaults or violent acts: 839; falls: 835; exposure to harmful substances or harmful environment: 488; fires or explosions:151; others: 25. (The data are from the Bureau of Labor Statistics.) Construct a bar graph.

6. The graph below came from the USA Today Snapshots: Commuting Time.



List two things that are wrong with this graph.

Graphical Displays for a Quantitative Variable

What you need to know:

- **How to create, describe and interpret graphical displays for a quantitative variable (histograms, stemplots)**

---

1. a. Using either the results for your class from #17 of the class questionnaire, *The approximate number of pairs of shoes you own*, or using the following data from another Math 140 class that answered the questionnaire, create histograms from the data: 8, 15, 15, 4, 8, 9, 200, 30, 8, 20, 30, 4, 45, 10, 15, 15, 6, 100, 12, 25, 20, 10, 15, 6, 7, 4, 10, 10, 15, 4, 12, 15, 15, 15, 15, 30, 10, 12, 9

   a. Fill in the empty columns of the table:

Approximate number of pairs of shoes	Count (Frequency)	Percent
[0-10)		
[10-20)		
[20-30)		
[30-40)		
[40-50)		
[50-60)		
[60-70)		
Total:		

   b. Create a histogram below using these intervals. Use the counts for the vertical axis.

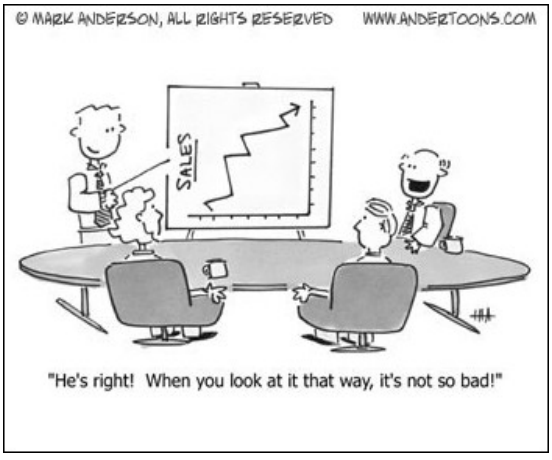
   c. Create a histogram using the same intervals as before, using percents for the vertical axis.

d. How does the shape of the histogram in (c) compare to the one you made in (b)?

2. Using either the results from the class questionnaire from either Question #7: *Your weight in pounds* or Question #10: *How many minutes do you typically spend in the shower?*, or using the following values from another Math 140 class that answered the questionnaire, make a stemplot of the data. Then briefly summarize its features and what it reveals about the distribution of that variable.

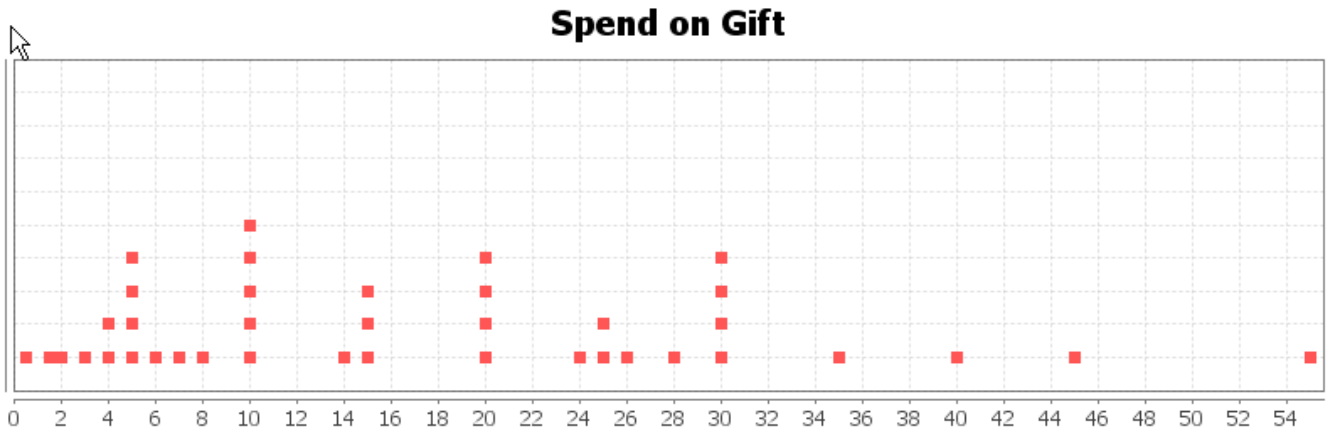
   #7 data (weights): 160, 200, 175, 175, 127, 112, 106, 130, 130, 95, 125, 160, 135, 137, 103, 140, 120, 120, 136, 190, 115, 120, 115, 110, 120, 155, 130, 135, 145, 138, 116, 160, 170, 170, 145, 125, 206, 145, 160

   #10 data (shower time): 20, 30, 10, 15, 15, 20, 25, 20, 15, 15, 20, 10, 15, 45, 15, 10, 12, 30, 10, 10, 15, 30, 20, 15, 15, 20, 10, 15, 30, 10, 15, 15, 20, 20, 20, 20, 25, 20, 7



Another Graphical Display for a Quantitative Variable: Dot plots

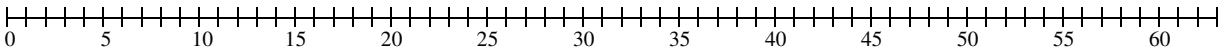
Another way to display the values of a quantitative variable is to use a ***dot plot***. In a dot plot, each data value is shown with a dot; when there are ties, data with the same value are stacked above each other. Here is a dot plot of the data for a Math 140 class for the question "The amount of money you think your best friend would spend for your birthday gift":



(a) Describe the shape of the distribution.

(b) What would you say is a typical amount that members of this class said they thought their best friend would spend on their birthday gift?

2. a. Using either the results for your class from #17 of the class questionnaire, *The approximate number of pairs of shoes you own*, or using the following values from another Math 140 class that answered the questionnaire, make a dot plot of the data: 8, 15, 15, 4, 8, 9, 200, 30, 8, 20, 30, 4, 45, 10, 15, 15, 6, 100, 12, 25, 20, 10, 15, 6, 7, 4, 10, 10, 15, 4, 12, 15, 15, 15, 15, 30, 10, 12, 9



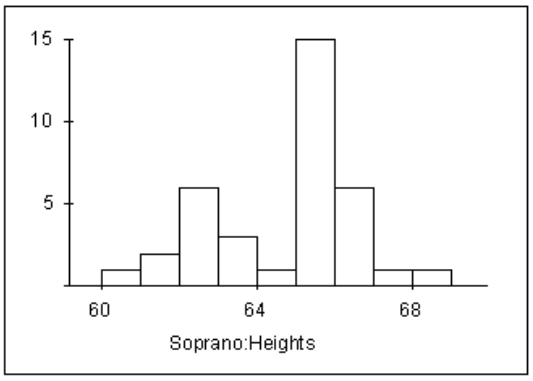
b. Circle your own value on the dot plot. Describe where you fall in relation to your classmates with regard to the number of pairs of shoes owned.

c. Write a few sentences describing the distribution. (Look at *center*, *variability*, *shape*, *outliers*, etc.)

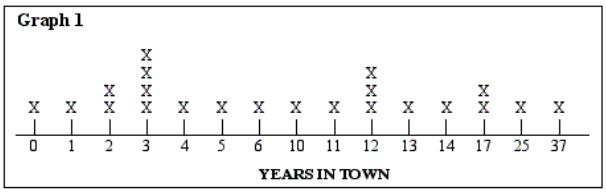
Additional Practice

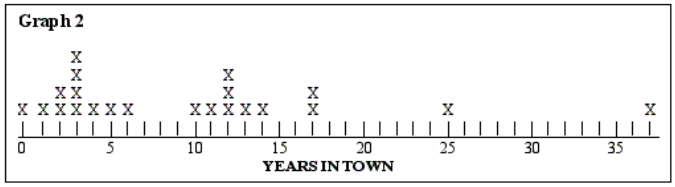
1. Using either the results for your class from Question #14 of the class questionnaire: *The approximate distance in miles between your current residence and CSUN* or using the following values from another Math 140 class that answered the questionnaire, make a stemplot of the data and summarize its features (center, variability, shape, outliers, etc): 30, 5, 10, 40, 6, 24, 7, 10, 35, 4, 5, 1.5, 30, 25, 14, 0.5, 55, 20, 10, 20, 8, 5, 30, 20, 45, 25, 20, 28, 3, 5, 4, 15, 15, 15, 10, 26, 30, 10, 2

2. The graph below shows the heights of singers in a large chorus. Write a description of the distribution of this data, indicating all the important features (center, variability, shape, outliers, etc).



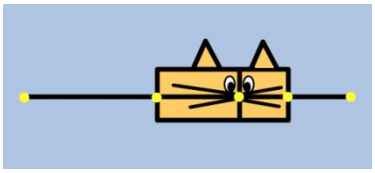
3. A class of students recorded the number of years their families had each lived in their town. Here are two graphs that the students drew to summarize the data. Which graph gives a correct representation of the data? Why?





Measures of Center and Spread; Five Number Summary, Boxplots

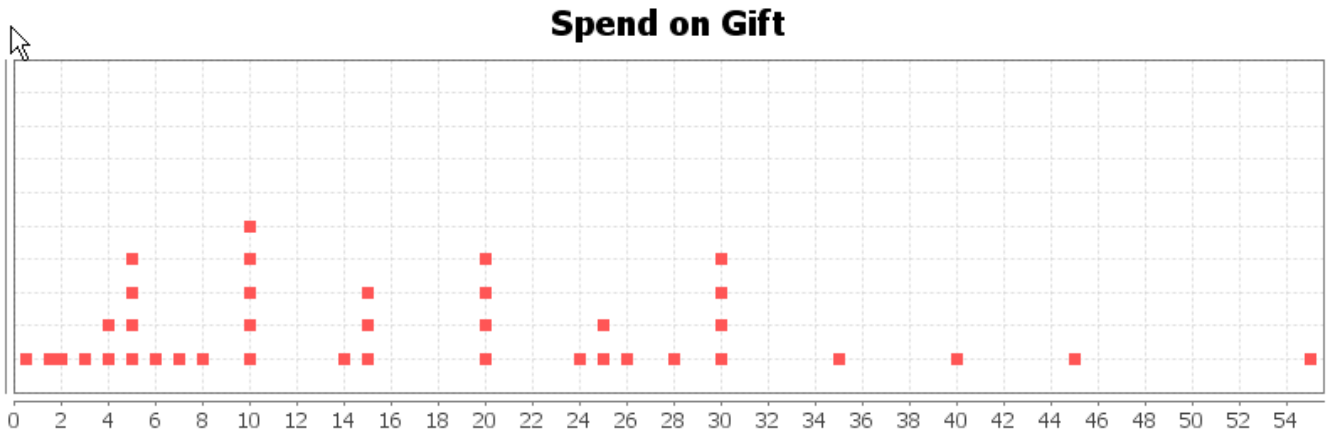
What you need to know:



- **How to quantify the center of a distribution: mean, median**
- **Behavior of mean and median for skewed distributions**
- **How to find the five-number summary and create a boxplot**
- **Interquartile range (IQR); identifying outliers using the 1.5(IQR) rule (optional)**

---

1. Here is a *dot plot* of the data for a Math 140 class for the question "The amount of money you think your best friend would spend for your birthday gift":



a. Without doing any calculations, choose a number that you feel represents a **typical value** for this data. Justify your choice.

b. Calculate the **mean** amount of money students felt their best friend would spend on their gift. To make this go faster, you can share the work with a partner. Split the data and have each person add up the values in their part. Then combine your results and find the mean. (Note: One student gave a value that was not a whole number of dollars, namely $1.50.) Mark and label this value at the correct location on the horizontal axis of the plot.

c. Determine the **median** of the distribution. Then mark and label this value on the horizontal axis of dot plot.

d. Were half of the students in the class above the mean number and half below? If not, does this mean you made an error in computing the mean? Explain.

2. The stemplot below shows the populations of the 50 states and Washington D.C. in the year 2000, in millions of people. (Source: U.S. Census Bureau, Current Population Reports, P25-1106; ‘‘Table CO-EST2001-12-00 - Time Series of Intercensal State Populaton Estimates: April 1, 1990 to April 1, 2000’’; published 11 April 2002; http://www.census.gov/popest/archives/2000s/vintage2001/CO-EST2001-12/CO-EST2001-12-00.html)

```
0|11111111111112222233333344444
0|5555666667888
1|0122
1|69
2|1
2|
3|4
```

a. California was the most populous state with a population of ___ million. How many states had a population of approximately one million in the year 2000? How many states had populations of at least ten million?

b. Which numerical measure of central tendency is more appropriate for this data, the mean or the median? Explain your choice.

c. Without any calculation, which must be larger, the mean or the median? Explain.

d. The average (mean) population of the 50 states and Washington D.C. in the year 2000 was 5.52 million people. Find the median. Was your prediction in part (c) correct?

e. Suppose the population of California were 340 million rather than 34 million (Heaven forbid!). Would the mean increase or decrease?

What about the median?

f. Find the interquartile range (IQR) for the data. What does it measure here?

3. The table below shows the team payrolls (total amount of salaries paid to all team members) for the a recent season in each of the major U.S. professional sports leagues:

Major League Baseball (MLB)		National Basketball Association (NBA)		National Football League (NFL)	
New York Yankees	$209,081,579	New York Knicks	$93,341,391	Washington Redskins	$123,408,019
Detroit Tigers	$138,685,197	Cleveland Cavaliers	$85,428,923	New England Patriots	$117,963,182
New York Mets	$138,293,378	Dallas Mavericks	$84,206,021	New Orleans Saints	$110,417,011
Boston Red Sox	$133,440,037	Portland Trail Blazers	$81,508,534	Buffalo Bills	$108,875,882
Chicago White Sox	$121,152,667	Boston Celtics	$78,019,509	Kansas City Chiefs	$108,482,459
Los Angeles Angels	$119,216,333	Los Angeles Lakers	$75,330,112	Dallas Cowboys	$107,376,072
Chicago Cubs	$118,595,833	Denver Nuggets	$73,412,428	San Francisco 49ers	$106,877,077
Los Angeles Dodgers	$118,536,038	Phoenix Suns	$73,222,550	Detroit Lions	$106,731,910
Seattle Mariners	$117,993,982	Houston Rockets	$71,286,452	Pittsburgh Steelers	$106,293,914
Atlanta Braves	$102,424,018	Milwaukee Bucks	$69,084,243	Baltimore Ravens	$104,997,764
St. Louis Cardinals	$100,624,450	Indiana Pacers	$68,898,240	Chicago Bears	$104,151,969
Toronto Blue Jays	$98,641,957	Utah Jazz	$65,632,827	Indianapolis Colts	$102,786,398
Philadelphia Phillies	$98,269,881	Detroit Pistons	$65,503,760	San Diego Chargers	$102,460,685
Houston Astros	$88,930,415	Sacramento Kings	$64,996,855	Cleveland Browns	$102,394,922
Milwaukee Brewers	$81,004,167	Minnesota Timberwolves	$64,728,383	Denver Broncos	$102,152,344
Cleveland Indians	$78,970,067	New Orleans Hornets	$64,547,807	Philadelphia Eagles	$100,807,309
San Francisco Giants	$76,904,500	Miami Heat	$63,888,839	St. Louis Rams	$100,340,467
Cincinnati Reds	$74,277,695	San Antonio Spurs	$63,559,368	New York Jets	$99,971,535
San Diego Padres	$73,677,617	Orlando Magic	$62,863,107	Seattle Seahawks	$99,567,188
Colorado Rockies	$68,655,500	Atlanta Hawks	$62,545,088	Arizona Cardinals	$98,694,817
Texas Rangers	$68,239,551	New Jersey Nets	$62,167,320	Cincinnati Bengals	$98,529,188
Baltimore Orioles	$67,196,248	Toronto Raptors	$58,784,702	Houston Texans	$98,154,775
Arizona Diamondbacks	$66,202,713	Washington Wizards	$57,803,157	Tampa Bay Buccaneers	$98,105,565
Minnesota Twins	$62,182,767	Seattle SuperSonics	$52,876,667	Green Bay Packers	$97,653,823
Kansas City Royals	$58,245,500	Chicago Bulls	$49,169,990	Tennessee Titans	$97,081,153
Washington Nationals	$54,961,000	Charlotte Bobcats	$45,309,481	Jacksonville Jaguars	$94,030,775
Pittsburgh Pirates	$49,365,283	Memphis Grizzlies	$44,706,020	Carolina Panthers	$93,944,262
Oakland Athletics	$47,967,126	Los Angeles Clippers	$43,892,696	Miami Dolphins	$92,573,123
Tampa Bay Rays	$43,820,598	Philadelphia 76ers	$35,166,714	Minnesota Vikings	$92,161,921
Florida Marlins	$21,836,500	Golden State Warriors	$34,582,090	Oakland Raiders	$90,869,865
				Atlanta Falcons	$83,845,371
				New York Giants	$75,755,388

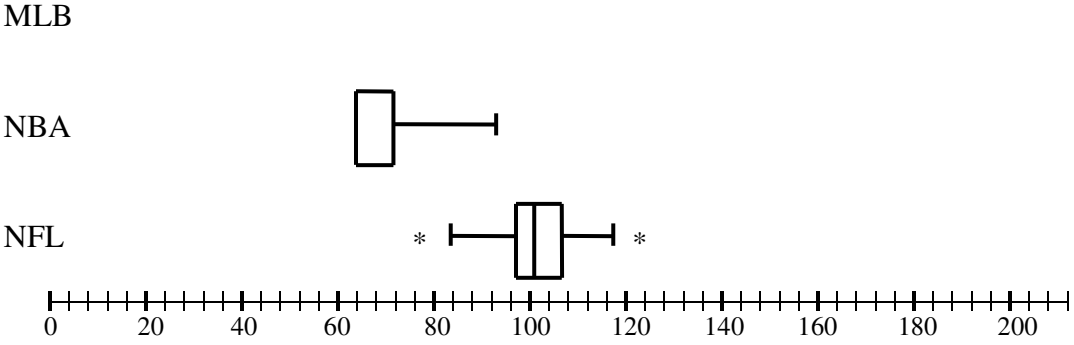
a. Which league has the greatest range of payrolls (maximum payroll minus minimum payroll)?

Which league has the smallest range?

b. Fill in the missing entries in the following table of five number summaries (show the figures in millions of dollars, using correct rounding)

	MIN	Q1	MEDIAN	Q3	MAX
MLB					
NBA	35	58	65	73	93
NFL	76	97	101	107	123

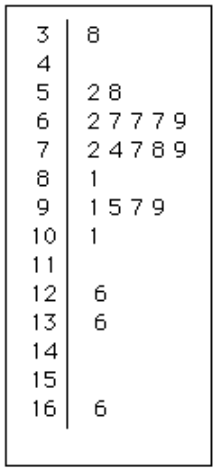
c. Complete the following display by completing the NBA boxplot and constructing the boxplot for Major League Baseball. (You'll need to determine the cutoff values for outliers for the MLB data):



d. Summarize your findings as if you were a sports reporter trying to write an interesting article about the similarities and differences between the three leagues.

Additional Practice

1. The following stemplot represents the yearly percentage increases in a college's comprehensive fees over the past 22 years. (3 | 8 means that one year had a percentage increase of 3.8%.)



3	8
4	
5	2 8
6	2 7 7 7 9
7	2 4 7 8 9
8	1
9	1 5 7 9
10	1
11	
12	6
13	6
14	
15	
16	6

   a. If one were asked to give the value of the center of this data, is there a single obvious correct answer? Explain.

   b. Find both the median and mean yearly percentage increases.

   c. Which is larger, the median or the mean? Is there a big difference between them? Are these answers what you would have expected based on the stemplot? Explain.

   d. Find the first and third quartiles, and the interquartile range.

   e. Use the IQR to fill in the blanks below:

   In about _____ of the past 22 years, the percentage increase in the college's comprehensive fees was _____ _____% and _____%.

   f. Determine if the data contains any outliers as judged by the 1.5(IQR) criterion.

g. Construct a boxplot of the data. Make sure to show outliers, if any, with asterisks.

2. Would it be more desirable for variability to be high or low for each of the following cases? Explain your decisions.

   a. Ages of trees in a national forest

   b. Diameters of new tires coming off one production line

   c. Actual weights of a particular size of cereal box

   d. Prices of cars available at a used car lot

3. For what kinds of variables is a histogram appropriate?

   a. categorical only
   b. quantitative only
   c. varies according to situation

4. What does it mean to say that the median is a RESISTANT measure of center?

5. Pretend you are constructing a histogram for describing the distribution of salaries for all employed people in California.

   a. What goes on the Y-axis?

b. What goes on the X-axis?

c. What would be the probable shape of the salary distribution? Explain why.

6. In 1798 the English scientist Henry Cavendish measured the density of the earth. Since it was difficult at that time to measure this density accurately, he repeated his calculation 30 times. Here are the results (the data represent the ratio of the earth's weight to an equivalent volume of water):

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.51	5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27
5.39	5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85

a. Make a stem and leaf plot of the values.

b. Determine the five-number summary for the data. Use three decimal places in your answers.

c. Construct a boxplot, showing outliers if any.

d. Write a complete description of the distribution.

7. The table below shows the average monthly temperatures for Albuquerque, New Mexico and Green Bay, Wisconsin.

	Albuquerque	Green Bay
Jan.	34	14
Feb.	40	18
Mar.	46	30
Apr.	55	44
May.	64	55
Jun.	74	64
Jul.	78	69
Aug.	75	67
Sep.	68	59
Oct.	57	48
Nov.	44	34
Dec.	35	20

a. Make side-by-side boxplots above (as in Figure 2.1, p. 52) of the two cities' temperature distributions.

b. Summarize the key differences and similarities in the temperature distributions of the two cities, using statistical terminology.

The Standard Deviation

What you need to know:

- **What the standard deviation measures**
- **When it is appropriate to summarize a distribution with the mean and standard deviation as opposed to the five-number summary.**

The tattoo parlor near campus got busy when the professor required hand calculation of the standard deviation.

---

1. The data below shows the ammonia levels (in parts per million) in air samples taken near the exit ramp of a San Francisco freeway tunnel for eight days during the afternoon drive.
(Source: Environmental Science and Technology, Sept.1, 2000).

Day	1	2	3	4	5	6	7	8
Ammonia Level	1.53	1.50	1.37	1.51	1.55	1.42	1.41	1.48

a. Complete the table below. The mean of the eight ammonia levels above is $\bar{x} = 1.47$.

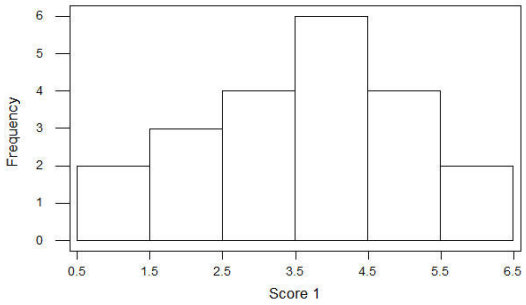
Ammonia level	Deviation from mean	Squared deviation
1.53	0.06	0.0036
1.50	0.03	0.0009
1.37	-0.10	0.0100
1.51	0.04	0.0016
1.55		
1.42	-0.05	0.0025
1.41		
1.48		
Sum		

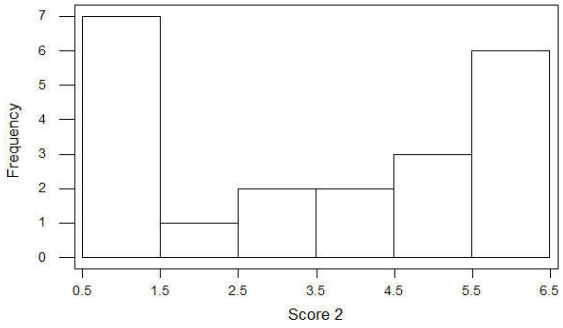
a. What is the sum of the deviations from mean? Is this a coincidence? Explain.

b. Find the sum of the squared deviations and use it to determine the standard deviation.

2. For each of the following pairs of histograms, indicate how the standard deviations compare:

a.



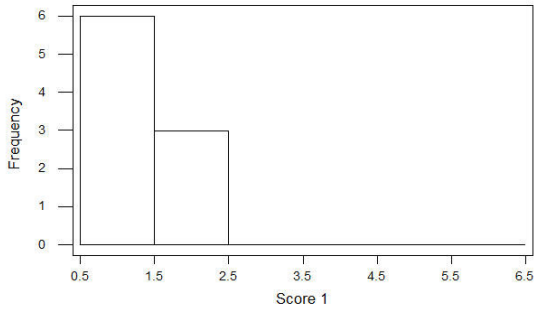


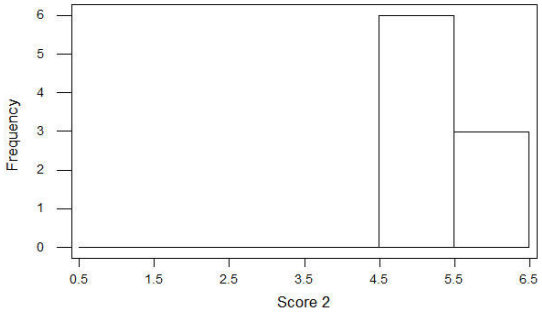
☐ The Score 1 data has a larger standard deviation than the Score 2 data

☐ The Score 2 data has a larger standard deviation than the Score 1 data

☐ Both data sets have about the same standard deviation

b.





☐ The Score 1 data has a larger standard deviation than the Score 2 data

☐ The Score 2 data has a larger standard deviation than the Score 1 data

☐ Both data sets have the same standard deviation

3. Consider the following samples of student scores on a seven point quiz:

Sample 1: 0, 1, 2, 3, 4, 5, 6, 7
Sample 2: 0, 0, 0, 0, 7, 7, 7,7
Sample 3: 0, 3.5, 3.5, 3.5, 3.5, 3.5, 7

Without doing any calculations answer the following questions:

a. Which sample has the greatest standard deviation? Why?

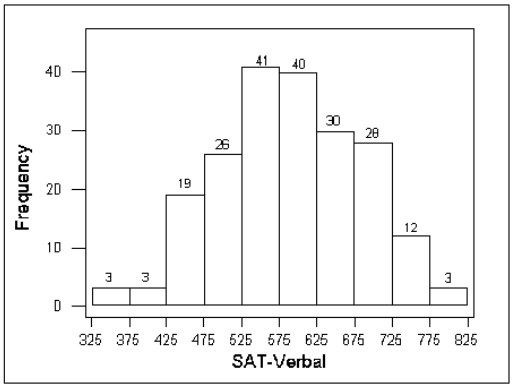
b. Which sample has the smallest standard deviation? Why?

Additional Practice

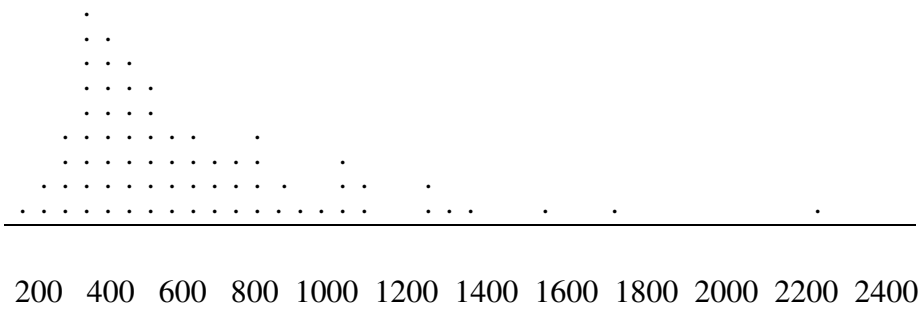
1. What if the diameters of a sample of new tires coming off one production line turned out to have a standard deviation of 0. Would the manufacturer be happy or unhappy, assuming the average diameter was correct? Explain.

2. For each of the following cases, indicate which would give a better summary of the data: (i) the five-number summary (min, Q1, median, Q3, max) or (ii) the mean and standard deviation? Explain your choice.

   a. Verbal SAT scores for 205 students entering a local college in the fall of 2002:



   b. Recent sales prices of homes in a local neighborhood (in thousands of dollars):

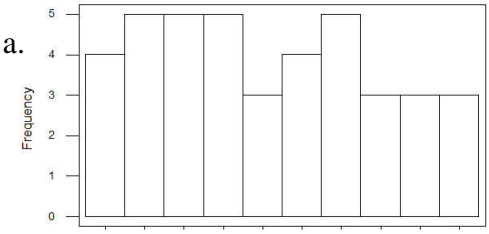


3. I have data sets for four different variables, and I made a histogram and a boxplot and found the summary statistics for each. Somehow they all got mixed up. Match each variable to the correct histogram, boxplot, and set of summary statistics. Use the diagram on the following page to show your answers.

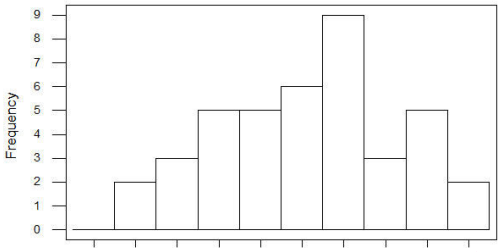
Variables: A. Age at death of a sample of people
B. Heights of a class of college students
C. Number of gold medals won by medal-winning countries in the 2008 Olympics
D. Random digits between 0 and 9 generated by a computer

Histograms:

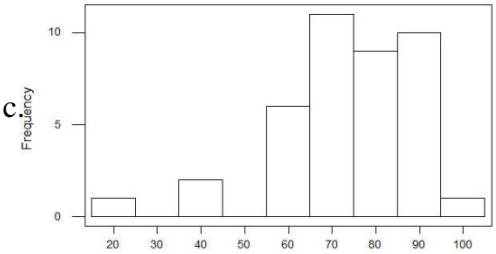
a.



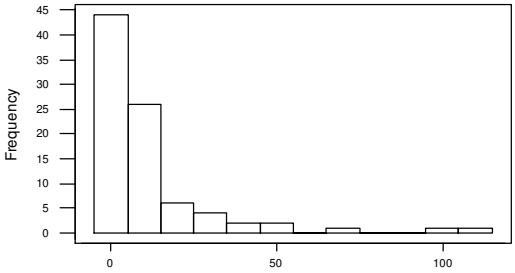
b.



c.

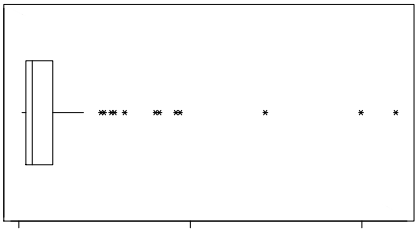


d.

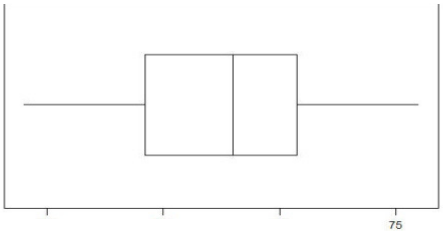


Boxplots:

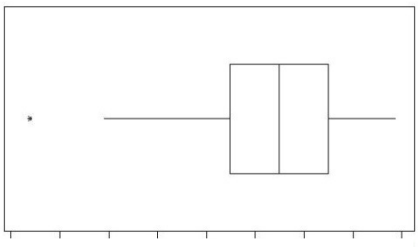
i.



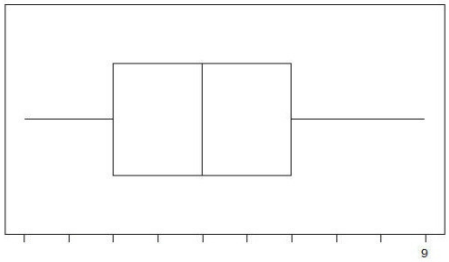
ii.



iii.



iv.



Summary statistics (each row goes with one set of data):

SS1:	Mean: 72.82	Median: 75	Standard deviation: 15.51	IQR: 20
SS2:	Mean: 11.01	Median: 4	Standard deviation: 19.08	IQR: 8
SS3:	Mean: 4.1	Median: 4	Standard deviation: 2.808	IQR: 4
SS4:	Mean: 67.8	Median: 68	Standard deviation: 4.22	IQR: 6.5

Show your answers here, or on a copy of the array below, by drawing lines connecting each set of data to the correct histogram, then to the correct boxplot, then to the correct summary statistics. For example, you could connect A---c---ii---SS4 if that happened to be one of the correct matches.

A	**a**	**i**	**SS1**
B	**b**	**ii**	**SS2**
C	**c**	**iii**	**SS3**
D	**d**	**iv**	**SS4**

Normal Distributions; 68-95-99.7 Rule

What you need to know:

- **How the mean and s.d. affect the appearance of a normal density curve.**
- **How to draw a normal curve with change-of-curvature points and label the mean and s.d.**
- **How to estimate the mean and s.d. of a normal distribution from its graph**
- **How to use the 68-96-99.7 Rule**

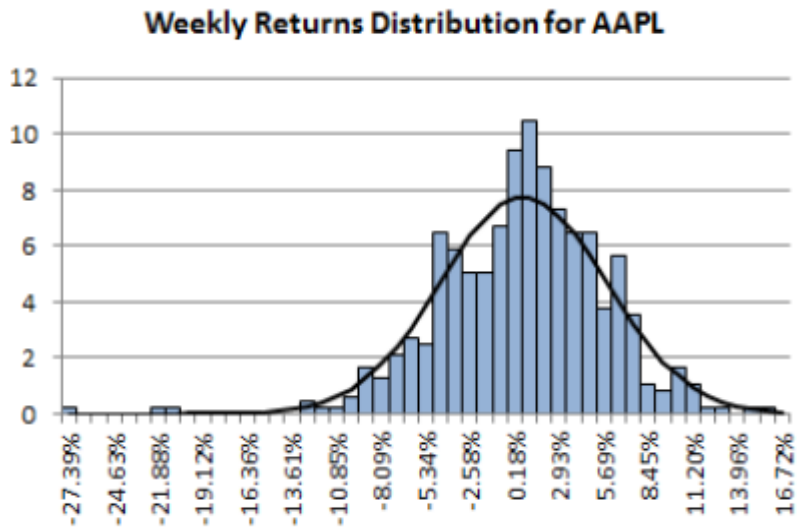
---

1. a. Draw two normal curves on the same axis below that have the same standard deviation but different means. Describe the similarities and differences in their graphs.

b. Draw two normal curves on the same axis below that have the same mean but different standard deviations. Describe the similarities and differences in their graphs.

2. A famous test of human intelligence is the Wechsler IQ test. The IQ scores closely follow a normal curve with mean 100 and s.d. 15. Carefully draw this normal density below. Label the axis with the numerical values of the mean and the points one, two and three s.d.s above and below the mean:

3. The histogram below shows the distribution of the weekly changes in the price of *Apple* stock for the ten-year period ending March 20, 2014. A normal density curve has been fitted to the histogram.



a. Use the normal curve to estimate the mean weekly return on Apple stock during this period.

$$\mu \approx \_\_\_\_\_$$

b. As accurately as possible, mark the inflection points (change-of-curvature points) on the normal curve. (Keep in mind that they should be at equal distances from the mean.)

c. Estimate the standard deviation of the mean weekly return on Apple stock.

$$\sigma \approx \_\_\_\_\_$$

4. Suppose the scores on an achievement test follow an approximately normal distribution with mean 500, min = 350, and max = 650. Which of the following values could be the standard deviation?

10 50 100 150

Justify your answer. A sketch may help.

5. Student GPA's at a certain university are approximately normally distributed with a mean of 2.5 and an s.d. of 0.5 grade points.

    a. About what percentage of students have a GPA between 2.0 and 3.0?

    b. About what percentage of students have a GPA between 1.5 and 3.5?

    c. About what percentage of students have a GPA below 2.0?

    d. A GPA of 3.5 or higher is required to make the Dean's List. About what proportion of students at the university qualify for the Dean's List?

Additional Practice

1. The distribution of adult human height is approximately normal for both males and females. The summary statistics below are for American adults ages 20-29. Using these values, carefully sketch the female and male height distributions on the same labeled axis.

	Mean	S.D.
Females	64.1	2.75
Males	69.3	2.92

---

2. The mean Wechsler IQ score is 100 with a standard deviation of 15 points. IQ scores are roughly symmetrical, mound-shaped distribution.

   a. About what percent of adults have IQ scores between 85 and 115?

   b. About what proportion of adults have IQ scores between 70 and 130?

   c. About what percent of adults have IQ scores between 55 and 145?

   d. How high would it be necessary to score to fall in the highest 2.5% of IQs?

   e. What are the IQ scores of adults in the lowest 16%?

Working with Normal Distributions

What you need to know:

- **The Standard Normal Distribution**
- **How to calculate and interpret z-scores**
- **How to find proportions for a normal distribution ("forwards"):**
  - **Draw the normal curve and shade the area of interest**
  - **Standardize the variable value x (i.e. find its z-score)**
  - **Find the desired proportion using the z-table or your calculator**
- **How to find the value of a variable x corresponding to a given proportion ("backwards"):**
  - **Draw the normal curve and shade the area of interest**
  - **Find the z-score from the given proportion, using the z-table or your calculator**
  - **Find x from z by "unstandardizing"**

---

1. Explain what a z-score of -1.8 means: *The value of the variable x is* _____

   _____ .

2. The weights of adult male rhesus monkeys are approximately normally distributed with a mean of 15 pounds and a standard deviation of 3 pounds.

   a. What's the probability that a randomly selected monkey weighs less than 15 pounds? Draw a sketch and shade and write the answer.

   b. What proportion of monkeys weigh less than 13 pounds? Draw a sketch, find the z-score, and determine the proportion from either the normal table or your calculator.

   c. What proportion of monkeys weigh more than 20 pounds? Draw a sketch, find the z-score, and determine the proportion from either the normal table or your calculator.

d. What proportion of monkeys weigh between 11 and 19 pounds? Draw a sketch, find the two z-scores, and determine the proportion from either the normal table or your calculator.

e. What are the weights of the lightest 10% of the monkeys? Draw a sketch, find the z-score from either the normal table or your calculator, then determine the value of *x* that has 10% of the area below it.

f. What are the weights of the heaviest 15% of the monkeys? Draw a sketch, find the z-score from either the normal table or your calculator, then determine the borderline value of *x*.

3. Here are the mean and s.d. of the heights of young (ages 20-29) adult Americans:

	Mean	S.D.
Female	64.1”	2.75”
Male	69.3”	2.92”

a. Find your *percentile rank* in the distribution of height for your gender.

b. Find your percentile rank in the distribution of height for the opposite gender. What proportion of the opposite sex are taller than you?

c. How tall are the tallest 1% of men?

Additional Practice

1. Suppose your Statistics professor reports test grades as z-scores, and you got a z-score of 2.20 on the midterm exam. Are you happy? Write a sentence explaining what your score means, giving as much detail as possible.

2. A town's January high temperatures average 36˚F with a standard deviation of 10˚F, while in July the mean high temperature is 74˚F and the standard deviation is 8˚F. In which month is it more unusual to have a day with a high temperature of 55˚F? Explain.

3. The time that it takes me to make dinner is approximately normal with mean 40 minutes and standard deviation 15 minutes.

   a. *Carefully* draw a normal distribution curve on which this mean and standard deviation are correctly indicated.

   b. What proportion of the time will I make dinner in less than 20 minutes?

   c. On a randomly chosen night, what is the chance that I spend more than an hour in the kitchen preparing dinner?

   d. How often will I spend between 30 and 45 minutes at the stove making dinner?

4. John Beale of Stanford, CA, recorded the speeds of cars driving past his house, where the speed limit read 20 mph. Suppose that these speeds follow an approximately normal distribution. The mean of 100 readings was 23.84 mph, with a standard deviation of 3.56 mph.

   a. How many standard deviations from the mean would a car going at the speed limit be?

   b. Which would be more unusual, a car traveling 34 mph or one going 10 mph?

   c. What proportion of cars travelling past his house were going 15 mph or less?

   d. What proportion of cars travelling past his house were going at least 30 mph?

5. In a large section of a statistics class, the points for the final exam are approximately normally distributed with a mean of 68 and a standard deviation of 9. Grades are assigned according to the following rules:

   - The top 10% receive As
   - The next 20% receive Bs
   - The middle 40% receive Cs
   - The next 20% receive Ds
   - The bottom 10% receive Fs

   Find the scores on the final exam that would qualify a student for an A, a B, a C, a D, and an F.

End of *Exploring Data: Data and Distributions.* Check your knowledge:

1. In statistics, what is meant by a ***variable***?
2. What is the difference between a ***categorical*** *variable* and a ***quantitative*** *variable*? Also give an example of each.
3. What is meant by the ***distribution*** of a variable?
4. What two types of graphs are most appropriate for displaying data for a ***categorical*** variable?
5. Should there be spaces between the bars of a ***bar graph***?
6. When describing the overall pattern of a distribution of a quantitative variable, name ***three features*** you should mention.
7. What two measures describe the ***center*** of a distribution of a quantitative variable?
8. How do you describe the ***spread*** of a distribution of a quantitative variable?
9. Informally define an ***outlier***.
10. List at least three graphs that are used to display *quantitative* data.
11. What information is lost when you choose a ***histogram*** over a *stemplot* or a *dot plot?*
12. Should there be spaces between the bars of a *histogram*?
13. List the steps for constructing a *histogram.*
14. If a distribution is ***skewed right***, what does its shape look like?
15. If a distribution is ***skewed left***, what does its shape look like?
16. Explain how to calculate the ***mean*** $\bar{x}$ of a set of data**.**
17. Explain how to find the ***median*** M of a set of data**.**
18. Explain why the median is ***resistant*** to extreme observations but the mean is *not resistant*.
19. The *mean* and *median* are close together if the distribution is what?
20. In a skewed distribution, which will be farther towards the long tail—the mean or the median?
21. Which measure is most appropriate for a highly skewed distribution—the mean or the median?
22. What is the ***five-number summary***?
23. Explain how to calculate the ***first quartile*** Q_1 and the ***third quartile*** Q_3.
24. What type of graph gives a picture of the *five-number summary*?
25. The "box" in a ***boxplot*** represents what percentage of the data?
26. The middle line of a *boxplot* represents the _____.
27. Can the value of the *mean* be identified from a *boxplot*?
28. What is the ***interquartile range (IQR)***?
29. What is the *IQR*-based "rule of thumb" for defining ***outliers***?
30. What does ***standard deviation*** measure?
31. The sum of all the **deviations** of the observations from their mean will always be ___. Explain why.
32. When does the ***standard deviation*** equal zero?
33. Can the ***interquartile range*** or the ***standard deviation*** ever be negative?
34. Is the ***standard deviation*** resistant or nonresistant to extreme observations? Explain.
35. When is it better to use the ***five-number summary*** rather than the ***mean and standard deviation****?*
36. How would you describe the shape of a *normal curve*? Draw several examples.
37. Draw and describe the Standard Normal Distribution, labeling the axis with several values.

38. Where on the normal curve are the *inflection points* located?
39. Explain how to *standardize* a variable.
40. What information does the *standard normal table* (z-table) give?
41. What do z-scores mean? Make sure you know how to interpret z-scores.
42. Make sure you know how to find probabilities for a normal random variable ("forwards problems").
43. Make sure you know how to find the value of a variable (x) corresponding to a given probability ("backwards problems").

Examining Relationships: Scatterplots and Correlation

What you need to know:

- **How scatterplots are used to show a relationship between two quantitative variables**
- **How to describe a scatterplot**
  - **Overall pattern**
    - **Direction**
    - **Form**
    - **Strength**
  - **Deviations from the pattern**
- **That *r* is a numerical measure for assessing the direction and strength of linear relationships between quantitative variables**
- **The properties of *r***

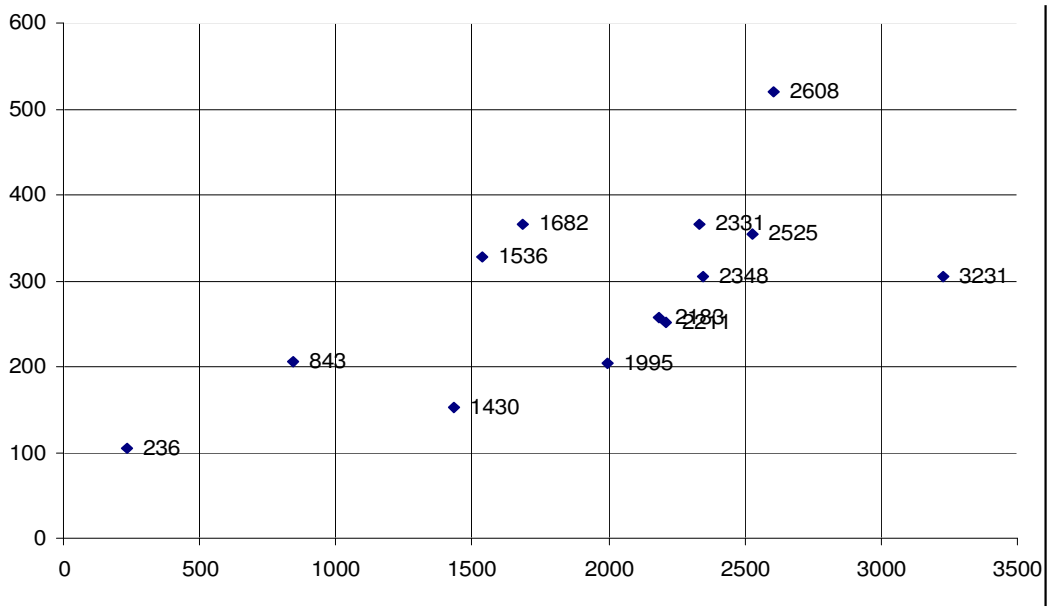
---

1. The table below shows flight distance (in miles), and the least expensive non-stop airfare (in dollars) from Los Angeles to selected cities (round trip). The prices are based on the following dates: leaving Sept.16, returning on Sept.23, 2008 (Sources: prices: www.orbitz.com, flight distance:www.convertunits.com/distance/)

City	Flight distance (in miles)	Non-stop airfare (in dollars)
Atlanta:	3231	305
Baltimore:	2331	365
Boston:	2608	520
Chicago:	2183	258
Dallas:	1430	152
Denver:	843	205
Detroit:	1995	203
Las Vegas:	236	105
Miami:	2348	305
Minneapolis:	1536	327
New Orleans:	1682	365
New York:	2525	355
Orlando:	2211	252
Salt Lake City:	591	179
San Antonio:	1482	247
Seattle:	954	178
St.Louis:	1601	285

a. Identify the explanatory variable: _____ and the response variable:_____. What type (categorical or quantitative) is each variable?

b. Complete the scatterplot below by adding the points for the last four cities and labeling the axes. The explanatory variable is on the x-axis and the response variable is on the y-axis.



c. Describe the relationship between flight distance and airfare using the scatterplot. Make sure you mention direction, form, strength, and outliers, if any.

d. Guess the value of the correlation coefficient *r*.

e. What are the units of *r*?

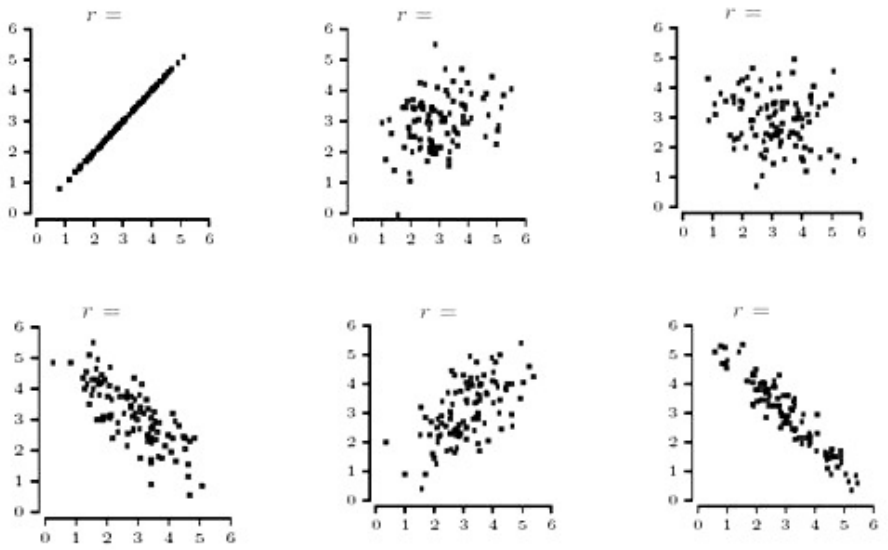
f. The actual value of the correlation coefficient is one of the following: 0.3 0.5 0.7 0.9. Decide with your classmates which is the correct value. Or you can enter the data into your calculator and find it that way. Circle the correct value above.

Was your estimate close to the calculated value?

g. If we change the units of flight distance from miles to kilometers, how would that change the scatterplot? How would it change the correlation coefficient?

h. If we interchange the *x* and *y* variables, how would that change the scatterplot? How would it change the correlation coefficient?

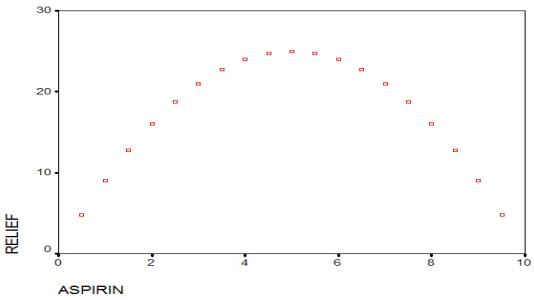
2.



Match the diagrams with the following correlations:

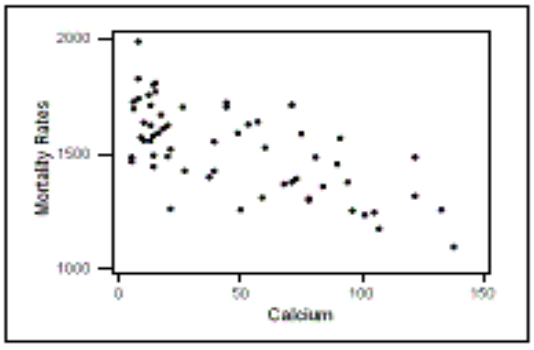
– 0.93 – 0.75 –0.20 0.27 0.63 1.0

3. Does the scatterplot at the right show a relationship between the two variables?



The correlation coefficient $r = 0$. Explain why.

4. The scatterplot below shows the annual mortality rate (number of deaths per 100,000 per year) for the years 1958-1964 and the calcium concentration (in parts per million) in the drinking water supply for a random sample of 61 large towns in England and Wales.



a. Each point represents a _____.

b. Summarize the key features (direction, form, strength) of the relationship between calcium concentration and mortality rate.

c. What does the plot tell us about the relationship between calcium concentration and mortality?

Additional Practice

1. For each of the following questions, suppose you want to collect data and make a scatterplot to help answer the question. (i) Which variable would you use as the explanatory variable and which would you use as a response variable? Why? (ii) What do you expect to see in the scatterplot? Discuss the likely direction (positive, negative, or neither), form (linear, nonlinear, no pattern), and strength.

   a. Do people who spend more on haircuts tend to own more pairs of shoes?

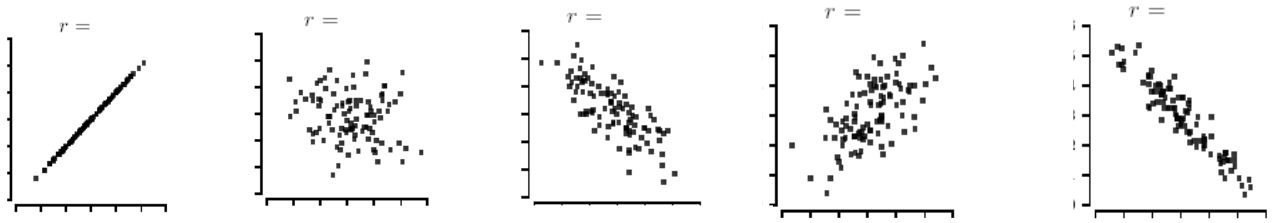
   b. Do magazines with more pages also use more ink?

   c. Do students who live far from CSUN tend to get less sleep?

   d. Do students who miss class less often get better test scores?

   e. Do professional baseball teams with younger players tend to hit more home runs?

2. a. Match the following scatterplots with the given correlation coefficients.



Correlation coefficients:	− 0.93	− 0.75	0.13	0.63	1.0

b. Which of these scatterplots could show the relationship between:

- shoe size and GPA of college students?
- weights of people in kilograms and weights of those same people in pounds?

3. A recent article in an educational research journal describes a study in which tests of basic math skills, math anxiety, and overall math achievement were administered to a large sample of students. The article reports a correlation of +0.8 between their basic math skills and math achievement scores. It also reports a correlation of -0.8 between math achievement and a math anxiety scores. Which of the following interpretations is the most correct? Circle the correct letter and justify your choice.

a. The correlation of +.8 indicates a stronger relationship than the correlation of -.8
b. The correlation of +.8 is just as strong as the correlation of -.8
c. It is impossible to tell which correlation is stronger without more information

4. Jose's cell phone plan requires him to pay a monthly base amount, $10, plus $0.30 per minute that the phone is in use.

a. Ignoring taxes and other fees, what would a scatterplot of Jose's monthly bills look like?

b. Why is it best to put Monthly Bill on the vertical axis and Minutes on the horizontal, rather than the other way around?

c. What would the correlation coefficient be? Explain.

5. If females of a certain species of lizard always mate with males that are half a year younger than they are, what would the correlation between the ages of the male and female lizards be?

   a. 0.5
   b. -0.5
   c. 0 (males and females are not related)
   d. 1
   e. -1
   f. Not enough information to tell

> **Hint:** Make up a few data values that follow the description above, where x is the male's age and y is the female's age, and draw the scatterplot.

6. Suppose that a study of a certain species of tree found a correlation of 0.78 between x = tree circumference, measured 3 ft above its base, and y = tree height. What would the correlation be between x = tree height and y = tree circumference?

   a. -0.78
   b. 1 - 0.78 = 0.22
   c. 0.78
   d. Not enough information to tell

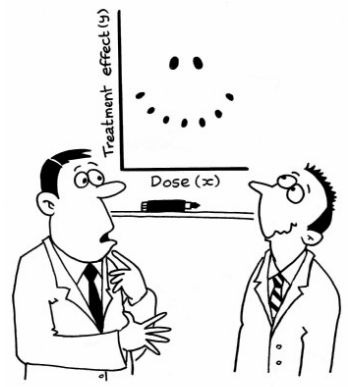
7. Suppose that the correlation between educational level attained and yearly income is 0.72. Which one of the following must be true?

   a. Income is one result of higher educational level.
   b. Lower income is associated with higher educational level.
   c. People with lower educational levels tend to have lower incomes.
   d. People with higher educational levels tend to have lower incomes.

Least Squares Regression

What you need to know:

- **What is the least squares regression line.**
- **How to find the equation of the regression line either from raw data or from summary statistics.**
- **How to interpret the slope of the least squares regression line in context.**
- **How to predict values of the response variable using the equation of the regression line.**
- **What extrapolation is, and that it is very unreliable.**



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

---

1. Jose's cell phone plan requires him to pay a monthly base amount, $10, plus $0.30 per minute that the phone is in use. In a scatterplot of Jose's monthly bills (ignoring taxes and other fees), all of the points would fall exactly on a line. Plot this line, using Minutes for *x* and Monthly Bill for *y*.

2. Recall the flight distance and airfare example from last time. We will now use this data set to determine an equation that will predict the least expensive non-stop airfare (in dollars) from Los Angeles to any other city (*not* just those included in the original list).

   a. What variable does *y* represent in this example?

   b. What variable does *x* represent in this example?

   c. What does *b* represent graphically for a regression line?

   d. What does *a* represent graphically for a regression line?

e. The correlation between flight distance and airfare is $r = 0.7$. Here are the other statistics:

$$s_y = 100 \qquad s_x = 792$$
$$\overline{Y} = 270.9 \qquad \overline{X} = 1752$$

Use the following formulas to calculate b and a:

$$b = r\frac{s_y}{s_x} = \qquad\qquad a = y - b\bar{x} =$$

Now plug the values of a and b into $\hat{y} = a + bx$ to find the equation of the least squares regression line:

f. Sketch the regression line on the scatterplot from last class.

g. Interpret the slope of the line in context, i.e., in terms of airfare and distance.

h. Interpret the y-intercept in context. Does it make practical sense here?

i. Using the equation of the least squares regression line, predict the airfare to Cincinnati, Ohio. The flight distance from LAX to Cincinnati is 1911 miles.

j. Predict the airfare from L.A. to Kansas City, Missouri (1369 miles).

k. Predict the airfare from L.A. to Tokyo, Japan (5484 miles). Is this prediction reliable?

l. Fill in the blank: Predictions for values of the explanatory variable that fall outside the range in the data are not reliable and should be avoided. This type of prediction is called

_____.

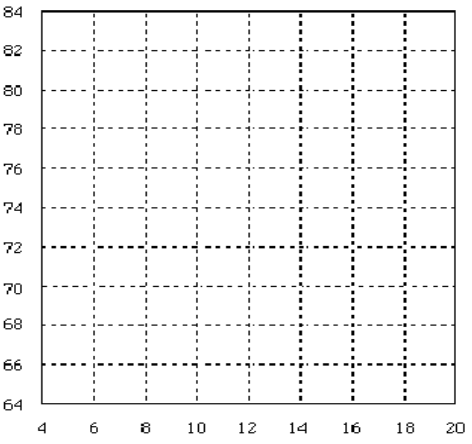
Additional Practice

1. Ten states were randomly selected from among the 50 United States. The data set below presents the percentage of households in each state that were below the poverty level (Poverty Rate) and the percentage of adults in the state who had earned a at least high school degree (HS and Above).

   a. Draw a scatterplot of the two variables using the grid provided. Clearly label the graph to indicate the variable being represented on both the horizontal and the vertical axes.

Variable	Mean	Standard Deviation	Correlation
Poverty Rate	12.33	3.117	-.796
% HS and Above	76.11	6.176	

Poverty Rate	HS and Above
8.7	77.5
19.0	64.7
13.1	76.7
10.2	82.5
10.2	78.8
13.0	70.0
12.5	75.6
11.9	83.1



   b. Briefly describe the relationship between the poverty rate and the percentage of adults with a high school degree or above, for these states.

   c. Determine the equation for the least squares regression line for predicting the percentage of adults in a state that have at least a high school degree (HS and Above) from the poverty rate for that state.

   d. Draw the regression line on your scatterplot.

e. Predict the percentage of adults that would have at least a high school degree in a state where 12.3% of the households are below the poverty level.

f. Provide an interpretation in context of the slope *b*.

g. Would it be OK to use the regression line to predict the percentage of adults that would have at least a high school degree in Mississippi, a state where 25.2% of the households are below the poverty level? Explain.

2. The following data represents the yield y of a chemical reaction at various temperatures x:

$x(F^\circ)$	y
150	77.4
150	76.7
160	78.2
180	84.5
200	83.9
200	83.7
225	85.6
250	88.9
250	90.3
280	94.8

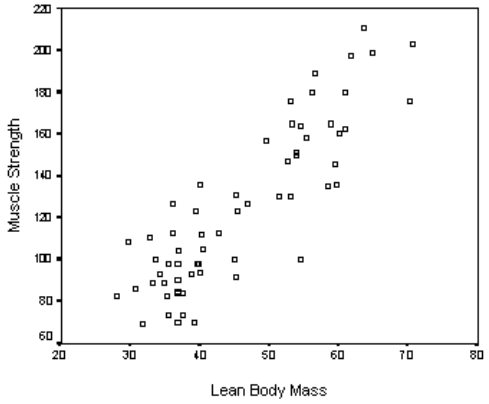
a. Find the summary statistics for the data--the sample mean and s.d. of the *x*'s, the sample mean and s.d of the *y*'s, and the correlation coefficient. Use calculator or software to do the computations.

b. Find the equation of the regression line of *y* on *x*.

c. What yield is predicted for a temperature of 190 degrees?

d. The yield is predicted to increase by _____ for every 1^o increase in temperature.

3. The scatterplot below shows the relationship between a certain measure of muscle strength and a person's lean body mass (in kg).



The summary statistics are as follows:

Variable	mean	s.d.
Muscle strength	122	31.2
Lean body mass	45	9.0

$r = .87$

a. Describe the nature, direction and strength of the association between muscle strength and lean body mass.

b. Find the regression equation that predicts a person's muscle strength from their lean body mass.

c. Predict the muscle strength score of a person with a lean body mass of 60 kg (about 132 lbs).

d. Simon's lean body mass is 10 kg more than Arnold's. How much more muscle strength is Simon predicted to have than Arnold?

Causation

What you need to know:

- **Association, however strong, does NOT imply causation. Only properly designed experimentation can show causation.**
- **What a lurking variable is.**

A ***lurking variable*** is a variable that has an important effect on the relationship among the variables in a study but is not one of the variables being studied. For example, lemonade consumption and crime rates are highly correlated. Now, does lemonade incite crime or does crime increase the demand for lemonade? Neither, of course—they are joint effects of a common cause, or lurking variable, namely hot weather.

1. A strong negative correlation exists between the horsepower of a car's engine and the fuel economy of the car. What is the hidden (lurking) variable? _____

2. Each pair of variables shown here has either a positive or negative association. Is it likely that item I influences item II, item II influences item I, or is there a lurking variable responsible for both?

   a. I. Wearing a hearing aid
      II. Dying within the next ten years

   b. I. The amount of milk a person drinks
      II. The strength of a person's bones

   c. I. A town's high school basketball gymnasium capacity
      II. Number of churches (or bars) in the same town

Additional Practice

1. Each pair of variables shown here has either a positive or negative association. Is it likely that item I influences item II, item II influences item I, or is there a lurking variable responsible for both?

    a. I. Shirt size of an American child
    II. Number of states an American child can name from memory

    b. I. The percentage of people in a country who own cars
    II. Life expectancy in that country

    c. I. Number firefighters at a fire site
    II. The amount of damage a fire does

2. Studies have shown that students who excel in algebra in high school do better overall in college. Does this show that mastering algebra is helpful in achieving success college? Answer based on what you have learned in this section.

Examining Relationships: Two-Way Tables for Categorical Variables

What you need to know:

- **Data display: two-way table**
- **Numerical summaries: marginal distributions, conditional distributions.**
- **Simpson's paradox**

---

1. For each of the potential studies described below, (i) identify the explanatory variable and indicate if it is quantitative (Q) or categorical (C), (ii) identify the response variable and indicate if it is quantitative or categorical. The first problem has been done for you:

a. You want to explore the relationship between high school students' gender and SAT score.

EV: C *RV: Q*

b. You want to explore the relationship between adult workers' years of education and income.

c. You want to explore the relationship between age and employment status (employed, unemployed).

d. You want to explore the relationship between color of soft drink and rating of taste (awful, poor, OK, good, excellent).

2. A clinical trial compared the effectiveness of two drugs, Lithium and Imipramine, in preventing a recurrence of depression among patients who were hospitalized with depression. Here are the results:

Imipramine	Recurrence
Lithium	No Recurrence
Lithium	Recurrence
Lithium	No Recurrence
Imipramine	No Recurrence
Lithium	Recurrence
Imipramine	Recurrence
Lithium	No Recurrence
Imipramine	No Recurrence
Lithium	No Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Lithium	Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Lithium	Recurrence
Lithium	Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Lithium	Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence
Imipramine	Recurrence
Imipramine	No Recurrence
Lithium	No Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence
Lithium	Recurrence

Lithium	No Recurrence
Imipramine	Recurrence
Imipramine	Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence
Lithium	No Recurrence
Imipramine	No Recurrence
Imipramine	No Recurrence

Imipramine	No Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Lithium	Recurrence
Imipramine	Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Imipramine	Recurrence
Lithium	Recurrence
Imipramine	No Recurrence
Lithium	Recurrence

a. Construct a two-way table that classifies drug against the recurrence variable. Use the explanatory variable for the rows and the response variable for the columns.

b. Determine the marginal distribution of the Drug variable. (This will show the percentage of patients that were/were not assigned to each drug.)

c. Determine the marginal distribution of the Recurrence variable (recurrence/no recurrence). (This will show the percentages of patients that had/did not have a recurrence of their depression.)

d. Find the conditional distribution of the Recurrence variable for patients assigned to Lithium. Then find the conditional distribution of the Recurrence variable for patients assigned to Imipramine.

e. Draw a segmented bar graph that compares these conditional row distributions.

f. Summarize what your analysis shows about the relationship between drug and the likelihood of recurrence.

g. Based on the number of subjects used in this study, are you reasonably convinced that one of the two drugs is better than the other at preventing a recurrence of depression? (Note: This question simply calls for a judgment call on your part, there is no "correct answer" at this point. We will need the methods of inference that will be covered later in this course in order to answer it.)

3. In a study conducted several years ago, researchers sent 2600 resumes in response to want ads in *The Boston Globe* and *The Chicago Tribune*. All resumes described a well-qualified applicant. The researchers randomly assigned half of the resumes to have fictitious "white-sounding" first names (e.g., Neil, Brett, Greg, Emily, Anne, Jill) and the other half to have "black-sounding" names (e.g., Tamika, Ebony, Aisha, Rasheed, Kareem and Tyron). The table below shows how many of each type of applicant name received a response (call, letter or email):

	"White" name	"Black" name	Total
Received response	131	87	218
Did not receive response	1169	1213	2382
Total	1300	1300	2600

a. Assuming one wants to study whether the type of name has an effect on whether an applicant receives a response, which would be the best summary to facilitate this analysis: conditional row percents, or conditional column percents?

b. Compute the conditional distributions you chose in part (a) and make a segmented bar graph. Then use these things to address the question of interest.

4. Don Maddeningly, manager of the Fresno Mudhens baseball team, has a decision to make. He has two third basemen, both equally good defensively. Which one he should choose to get the most playing time for the remainder of the season comes down to who he thinks is the better hitter. The batting statistics for two players for the first 100 games of the season are given below:

Alvin Rodney's batting average: .265
Blake Casey's batting average: .250

Here is how they have done against each type of pitcher:

Against right-handed pitchers:

	AB	Hits
Rodney	198	57
Casey	80	24

Against left-handed pitchers:

	AB	Hits
Rodney	70	14
Casey	100	21

a. Based on their overall batting averages, who should play the most?

b. Compute each player's batting average against right-handed pitchers by dividing each player's number of hits by the number of "at bats" (AB) they have had. Who has done better against right-handed pitchers so far this season?

c. Compute each player's batting average against left-handed pitchers. Who has done better against left-handed pitchers?

d. If you were the manager of the Mudhens, who would you play based on the two players' batting statistics, all other things being equal? What's going on here?

Additional Practice

1. Here is a two-way table showing data collected from a Math 140 class concerning body image, where students' possible answer choices were: "about right", 'overweight", and "underweight":

Body Image	Female	Male
About right	38	19
Overweight	13	4
Underweight	0	2

   a. What is the explanatory variable?

   b. What is the response variable?

   c. Determine the conditional distributions of body image for men and for women.

   d. Draw a segmented bar graph that compares these conditional distributions.

   e. Are the men and women in our class about equally likely to think their weight is about right? Compare the percentages.

   f. Among those students who do *not* think their weight is about right, is there a difference between the genders in their feelings about body image? Explain.

2. For each of the situations described below, (i) identify the explanatory variable and indicate if it is quantitative or categorical, and (ii) identify the response variable and indicate if it is quantitative or categorical:

   a. A student was curious about which route would get her to school faster, so she collected data on how long the trip took for a freeway route and for a non-freeway route, taking each route ten times.

   b. A psychologist is studying the effect of electroshock therapy on a subject's ability to solve simple tasks. The number of tasks completed in a 10-minute period is recorded for subjects, half of whom received electroshock and half of whom did not receive electroshock.

   c. You want to determine whether students' expected grades at the beginning of an introduction to statistics course are positively related to their final course grades.

   d. A study is made to analyze how students' SAT scores are related to whether they graduate from college.

3. The government Office of Vital Statistics studies a sample of married couples, measuring the heights of each husband and each wife. Is there a clear choice of explanatory variable and response variable here? Explain.

4. **True or False:** If the most severely ill patients in Mercy Hospital have a higher death rate than similar patients in General Hospital, and the least severely ill patients in Mercy Hospital also have a higher death rate than similar patients in General Hospital, then we can safely conclude that the overall death rate for Mercy Hospital is higher than that for General Hospital. Explain.

5. A company has just fired a total of 300 employees. The gender breakdown is given in the top table below. The other two tables show a further breakdown according to the type of work the employees were doing, either professional or clerical. For each of the three tables, determine and compare the proportion of men being fired to the proportion of women being fired. Do you notice something strange? Explain, and give a possible reason for the phenomenon that occurs here.

overall	retained	fired	total
men	300	200	500
women	400	100	500
total	700	300	1000

professional	retained	fired	total
men	255	195	450
women	25	25	50
total	280	220	500
clerical	retained	fired	total
men	45	5	50
women	375	75	450
total	420	80	500

6. Sketch a scatterplot that has a negative overall correlation, but is made of up groups of points that each have a positive correlation. What is this an example of?

End of *Exploring Data: Relationships*. Check your knowledge:

1. What is the difference between a ***response variable*** and an ***explanatory variable***?
2. What is another set of terms for "response and explanatory variables"?
3. For quantitative variables, which one is usually represented as X and which one is usually represented as Y?
4. Suppose you have a list of the values of two categorical variables for a number of individual cases. How should the data be organized in order to study the variables?
5. Explain how to find the conditional column distributions from a two-way table of counts.
6. Describe how to show the relationship between two categorical variables graphically.
7. **True or false:** Simpson's paradox involves a possible relationship between three variables.
8. Explain clearly what Simpson's paradox is and give an example.
9. **True or false:** If in each of the three schools in a certain school district, a higher percentage of girls than boys gave out Valentines, then it must be true for the district as a whole that a higher percentage of girls than boys gave out Valentines.
10. A ***scatterplot*** shows the relationship between two _____ variables.
11. Which variable goes on the *horizontal axis* of a scatterplot?
12. When describing a *scatterplot*, what are the three most important aspects of the pattern to assess?
13. **True or false**: In describing the form of a scatterplot, it is important to say whether the graph appears to be linear or not.
14. In describing the ***direction*** of a scatterplot, when there is a positive or negative slope, we say that the variables are positively or negatively _____.
15. **True or false**: In describing the ***strength*** of a scatterplot, we look at the amount of "scatter" in the data points—how close the points lie to a simple form such as a line or a curve.
16. What is the best method for judging the strength of a linear relationship aside from simply looking at the scatterplot?
17. What does the ***positive*** or ***negative sign*** associated with the correlation coefficient indicate?
18. **True or false**: The correlation depends on which variable is explanatory and which is response.
19. **True or false**: A correlation coefficient has the same units as the variables.
20. What is true about the ***relationship*** between two variables if r is:
    a. Near 0?
    b. Near 1?
    c. Near -1?
    d. Exactly 1?
    e. Exactly -1?
21. What sort of correlation coefficient do you find when two variables have a very strong linear relationship, and large values of one variable are associated with small values of the other?
22. Suppose that for each of the days of 2015 we knew the number of words Barack Obama spoke on that day (variable 1) and the peak barometric pressure for that day in Caracas, Venezuela (variable 2). About what would you guess the correlation between these two variables to be? Why?

23. Suppose there are two variables which, when graphed in a scatterplot, form an almost perfect U-shaped parabola. Would the strong relationship between these variables result in a high correlation coefficient (meaning close to 1 or -1)? Why or why not?
24. Does the correlation coefficient resemble the median and IQR in being fairly ***resistant*** to outliers, or does it resemble the mean and standard deviation in being heavily influenced by outliers (i.e., ***non-resistant***)?
25. A ***least squares regression line*** is a straight line that is used to _____ *y* from *x*.
26. To determine the ***equation*** for a regression line, one must determine _____ and _____.
27. The slope is the amount of change in _____ when _____ increases by one unit.
28. Once you have a regression line, how do you find a predicted value of *Y* for a given value of *X*?
29. Suppose that someone measures height and weight for a bunch of human adults and then calculates a regression equation predicting height from weight. Why does the *y*-intercept of the equation not have a meaningful interpretation?
30. **True or false**: With a *regression line*, as for a correlation coefficient, you get the same numbers (slopes and intercepts) no matter which variable is considered the explanatory variable and which is considered the response.
31. Every regression line passes through the point (_____ , _____)
32. What is *extrapolation* and what is the problem with it?
33. Define *lurking variable*. Why is it such an important concept?
34. If two variables have a strong positive association, then the larger the value of one variable, the larger the value of the other variable. Is it safe to say that an increase in one variable *causes* an increase in the other variable? Explain.
35. **True or false**: The stronger the association between two variables, the more confident we can be that there is a cause and effect relationship between them.

Producing Data: Sampling

What you need to know:

- **The goal of sampling: representativeness of the population**
- **The problem of bias**
- **Sampling methods**
  - **Voluntary response sample**
  - **Simple random sample (SRS)**
  - **Stratified random sample**
- **Sample surveys**
  - **Undercoverage**
  - **Nonresponse**
  - **Response bias**
  - **Wording of questions**

---

1. Let's assume I have a crazy idea: I will give an A as a final grade to a few randomly selected students in this class. I have a few ways to select these lucky students. Identify my sampling methods for each:

   a. I randomly pick 5 students from the roster

   b. I randomly pick 5 females and 5 males

   c. I randomly select 1 freshman, 1 sophomore, 1 junior, and 1 senior from the female students, and 1 freshman, 1 sophomore, 1 junior, and 1 senior from the male students

2. For the following situations identify the sampling method, including whether or not random sampling was employed, and potential sources of bias or any other problems.

   a. An online magazine asks its readers to vote on whether they thought that a 4-day workweek would hurt the economy.

b. Bureau of Labor Statistics researchers randomly selected 500 individuals from each of nine geographic regions in the United States and asked each one if he or she was employed full time.

c. Pollsters called 1000 randomly selected phone numbers to ask the callers' opinions about radio stations.

3. You need to gather students' opinions about the cost of parking for students on campus. It's not practical to ask all the students at CSUN as there are approximately 40,000 students in all.

   a. Give an example of a way to choose a sample of students that is a poor design because it depends on voluntary response.

   b. Give another example that doesn't use voluntary response but is still a poor way of sampling .

   c. Give at least two valid ways of sampling. That is, describe at least two different *unbiased* methods for sampling students' opinions.

Additional Practice

1. We need to survey a random sample of the 300 passengers on a flight from Los Angeles to New York. Identify each sampling method described below:

   a. From the boarding list, randomly select 5 people flying first class and 25 of the other passengers.

   b. Randomly pick 30 seat numbers, and survey the passengers who sit there.

2. In a large city school system with 20 elementary schools, the school board is considering the adoption of a new policy that would require elementary students to pass a test in order to be promoted to the next grade. The PTA wants to find out whether parents agree with this plan. Listed below are some ideas proposed for gathering data. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.

   a. Put a big ad in the newspaper asking people to log their opinions on the PTA website.

   b. Send a survey home with every student, and ask parents to fill it out and return it the next day.

   c. Randomly select 20 parents from each elementary school. Send them a survey, and follow up with a phone call if they do not return the survey within a week.

   d. Run a poll on the local TV news, asking people to dial one of two phone numbers to indicate whether they favor or oppose the plan.

3. For the situation described in #2, two members of the PTA committee proposed different questions to ask in seeking parents' opinions:

   Question 1: *Should elementary school-age children have to pass high-stakes tests in order to remain with their classmates the following year?*

   Question 2: *Should schools and students be held accountable for meeting yearly learning goals by testing students before they advance to the next grade?*

   a. Do you think responses to these two questions might differ? How? What kind of bias is this?

   b. Propose a question with more neutral wording that might better assess parental opinion.

Producing Data: Experiments

What you need to know:

- **What an observational study is**
- **The definition of an experiment**
- **That observational studies typically contain lurking variables**
- **What we mean by treatment, treatment group, control group**
- **What is a randomized controlled experiment**
- **What is *random assignment*, and that it is different from *random sampling***
- **What we mean by blind and double blind experiments**
- **What a placebo is, and the placebo effect**
- **That determining whether the explanatory variable actually *causes* changes in the response variable requires a randomized controlled experiment**
- **Block designs, including matched pairs designs**

"I couldn't afford a control group so I decided to go with an out-of-control group."

"Your doctor will be here in a minute, I'm a placebo."

---

1. a. What is the difference between an observational study and an experiment?

   b. What is the critical advantage that randomized controlled experiments have over other experiments and observational studies?

2. A food company assesses the nutritional quality of a new "instant breakfast" product by feeding it to newly weaned male white rats. In this experiment they used 30 rats, and compared the new product with the standard products. After 28 days they measured the rats' weight gain.

   a. What is the explanatory variable?

   b. What is the response variable?

   c. Use a diagram to outline the design of a randomized comparative experiment for this study.

3. We would like to study the effect that a certain medication has on the hours of sleep that people with insomnia get at night.

   a. Do we need an experiment or an observational study? Explain.

   b. What features should our experiment have?

   c. Draw the outline of the experiment.

Additional Practice

1. A pharmaceutical company has developed a new pain-relief medication for arthritis. Sixty patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, "About what percentage of your pain has gone away?"

   a. Why should this company not simply administer the new drug and record the patients' responses?

   b. Draw the outline of an experiment to compare the drug's effectiveness with that of aspirin and of a placebo.

   c. Should patients be told which drug they are receiving? How would this knowledge probably affect their reaction?

   d. Should this experiment be double blind? Explain.

2. An expert on worker performance is interested in the effect of room temperature on the performance of tasks requiring the use of both hands simultaneously. Twenty subjects are available.

   a. Outline a design to compare performance at 70˚ and 80˚.

   b. Because individuals differ greatly in performance, the wide variation in individual scores may hide the systematic effect of temperature unless there are many subjects in each group. Describe in detail the design of a matched pairs experiment in which each subject serves as his or her own control. Explain where you use randomization in this design.

Introducing Probability

What you need to know:

- **The concept of randomness**
- **The idea of probability**
- **What a sample space *S* is**
- **Continuous probability models**
- **What a random variable is**
- **Continuous random variables and their density graphs**

1. a. If you flip a fair coin, what is the approximate probability that it lands with the "heads" side face up?

b. Does this mean that if you flip a coin 10 times you will definitely have exactly 5 heads and 5 tails?

c. Does this mean that if you flip a coin 10 times you will *probably* have exactly 5 heads and 5 tails?

d. Does this mean that if you flip a coin 1000 times, you'll probably have *around* 500 heads and 500 tails?

2. Probability is important in **statistics** primarily because many statistical studies involve a *sample* drawn randomly from a *population*—thus chance is involved.

a. Among Americans age 25 or older, 32% have a bachelor's degree or higher. Suppose a simple random sample of Americans 25 years old or older is taken. What is the probability that a particular member of this sample has a bachelor's degree or higher?

b. A SRS of 50 companies listed on the New York Stock Exchange is taken. Suppose that 62% of all stocks listed on the New York Stock Exchange gained value in the past month. What is the chance that any particular company in the sample has gained value?

c. A SRS of 50 companies listed on the New York Stock Exchange is taken. Eight of these companies announced a stock split in the past year. If this is all you know, estimate the probability that *Watsco Inc*—which was not one of the companies in the sample—announced a stock split in the past year.

3. a. What is the probability that you will earn a grade of K (rather than A, B, C, D or F) in this class? Write your answer in probability notation.

b. What is the probability that you are going to take a breath in the next five minutes? Write your answer in probability notation.

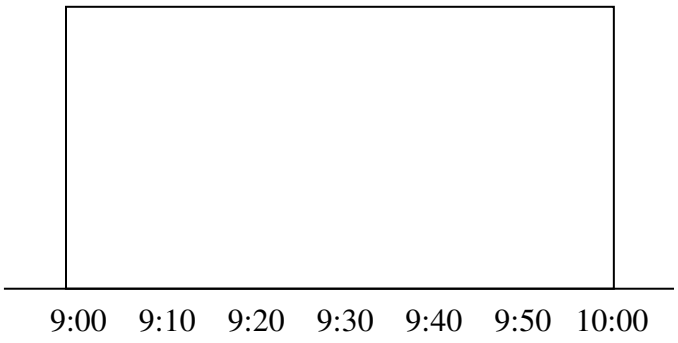
4. Write the sample space for the experiment of flipping two coins. Use H for Heads, T for Tails, and remember to list *all* possible specific outcomes.

$S = \{ \qquad \}$

a. What is the probability that if you flip two ordinary coins you will get either TT, HH, or HT?

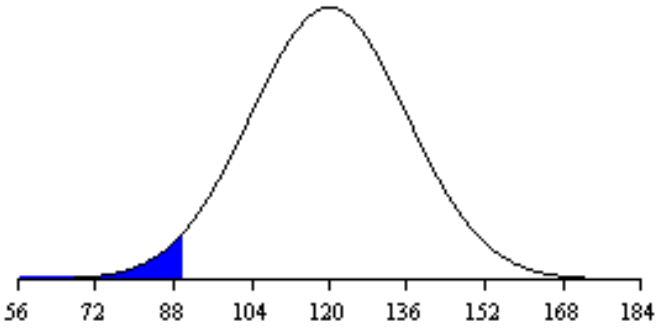
b. What is the probability that if you roll a die once, the outcome *won't* be a six?

5. Rachel is at a party. Rachel's boyfriend William has promised to show up at the party sometime between 9pm and 10pm. If William keeps his promise but the time he arrives during that hour is completely random, then the appropriate probability model for this situation is a **uniform distribution** having the following density curve:



Shade in the event "William arrives between 9:50 and 10:00" and find its probability using areas.

6. Assume that non-fasting glucose levels in females aged 20 to 29 years old who do not have diabetes follow a normal distribution, as shown below. The right edge of the shaded region is at 90.



a. The shaded region represents the probability that _____

_____ .

b. Estimate this probability by eye: ____ %

c. Let X be the non-fasting glucose level of a random selected female aged 20 to 29 years old who do not have diabetes. Is X a continuous random variable? Explain.

d. Express your answer to part (b) in mathematical notation:

$$P(\qquad) =$$

Additional Practice

1. a. According to the latest available figures, the percentage of registered California voters that are registered Republicans is 28.4%. What is the probability that a randomly chosen registered California voter is a registered Republican?

   b. If a person is randomly selected from the population of all registered California voters, the chance that they are a registered Democrat is 43.4%. Therefore, what proportion of all registered California voters are registered Democrats?

2. The members of a class of 50 students were asked their favorite type of movie. The survey data is displayed in the two-way table below:

	Females	Males
Romantic	7	2
Drama	6	4
Sci-fi	1	7
Comedy	6	8
Horror	3	5
Biographical	1	0
Total	**24**	**26**

   a. What percent of the males prefers comedies?

   b. What is the probability that a random male student in this class prefers comedies?

   c. What is the chance that a random female in this class prefers either comedies or romantic movies?

   d. A random student prefers sci-figure movies. What is the probability that the student is male?

End of *Producing Data* and *Probability.* Check your knowledge:

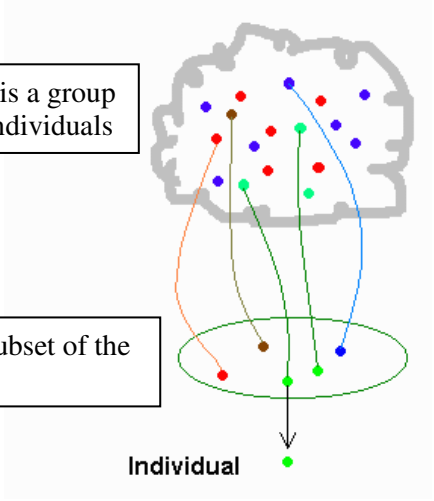
1. Why are *voluntary response samples* unreliable?
2. Why is *convenience sampling* unreliable?
3. What is a *biased* study?
4. Define *simple random sample*.
5. What is a *stratified random sample*?
6. How can the wording of questions cause *bias* in a sample survey?
7. Explain the difference between an *observational study* and an *experiment*. **Important!**
8. Define *treatment*.
9. Describe the *placebo effect*.
10. What is the significance of using a *control group*?
11. What do we mean be *random assignment*? **Important!**
12. What are the advantages of a *double-blind study*?
13. Describe a *block design*.
14. Describe a *matched pairs design*.
15. What is a *random phenomenon*?
16. What is a *sample space*?
17. What are the possible values of a probability?
18. How does one find the chance that an event *won't* happen from the chance that it will?
19. What is a *random variable*?
20. What is a *uniform distribution*?
21. What is a *continuous probability model*?
22. When using a *density curve*, how are probabilities determined?
23. What is the relationship between the proportion of a population that has a certain feature and the probability that a random individual from that population has that feature?

Introduction to Statistical Inference: Sampling Distributions

What you need to know:

- **What is a *parameter* and what is a *statistic* ***ONE OF THE MOST CRUCIAL THINGS TO KNOW IN THIS COURSE**
- **A *statistic* computed from a sample is a *point estimate* of the corresponding population *parameter***
- **The proper notation for the population proportion and the sample proportion**
- **The meaning of *sampling variability***
- **The effect of sample size on sampling variability**
- **What a *sampling distribution* is, and why they are the backbone of inference**
- **How sampling distributions can be studied through simulation**
- **The three properties of a sampling distribution: shape, center, spread**

A **population** is a group of objects or individuals



A **parameter** is a ***number*** that measures something about the population. Its value is usually unknown.

The **sample** is a subset of the population.

A **statistic** is a ***number*** calculated from the observations in the sample. We use the statistic to make *inferences* about the population parameter.

---

1. Consider the population of all the M&M Milk Chocolate candies manufactured by the Mars Company. Suppose that you want to learn about the distribution of colors of these candies but that you can only afford to take a sample of 50 candies. Suppose that at first you will only focus on a particular color, say, green.

   Take 50 M&Ms, disregarding the colors. The number of M&Ms you now have is the sample size, ***n***. Count the number of green candies ("successes") in your sample.

   n = # M&Ms = 50 number of green candies ("successes") = _____

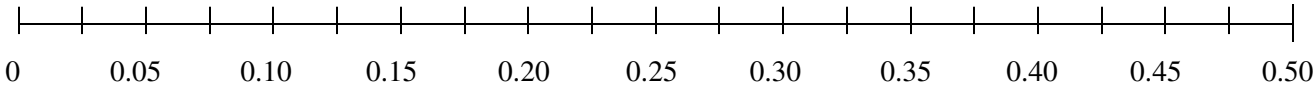
   a. Is the *proportion* of green candies in your sample a parameter or a statistic? What symbol represents it? Calculate this proportion for your sample.

b. Is the proportion of green M&Ms among all Milk Chocolate M&Ms manufactured by Mars a parameter or a statistic? What symbol represents it?

c. Do you <u>know</u> the value of the proportion of green M&Ms manufactured by Mars?

d. Clearly you and the other students in this class will get many different values for their sample proportions. What important statistical term refers to this phenomenon?

e. Construct a dot plot of the sample proportions of green candies obtained by you and the other students in the class. (We'll do this part together.)



The distribution of all the sample proportions from all possible samples of size *n* from the population is called the Sampling Distribution of the Sample Proportion.

f. Based on the plot above, the standard deviation of the sampling distribution is approximately:

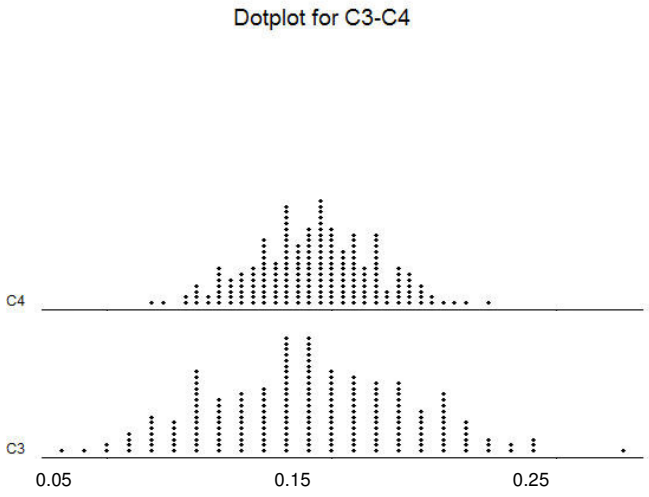
0.20 0.05 0.001 0.40

g. If every student were to estimate the population proportion of green candies (the proportion of Milk Chocolate M&M's the Mars Company makes that are green) by the proportion of green candies in his/her sample, would everyone arrive at the same estimate?

h. Even though sample values vary depending on the sample you happen to pick, there is a <u>pattern</u> to this variation. We need more samples to investigate this pattern more thoroughly. Since it is time-consuming (and possibly fattening) to <u>literally</u> sample more candies, we can instead use a computer to <u>simulate</u> the process of sampling.

Below are graphs of the sampling distributions of the sample proportion for many simulated samples of 100 imaginary M&Ms (n=100), and also for simulated samples of 200 imaginary

M&Ms (n=200). The C3 information is for the sampling distribution for n=100, and C4 is for the sampling distribution for n=200.



		Mean	S.D.
C4	n=200	0.16255	0.02567

		Mean	S.D.
C3	n=100	0.16095	0.03665

i. Do the sampling distributions appear approximately bell-shaped?

What can you say about the means of the sampling distributions?

How do the standard deviations of the two sampling distributions compare?

The Mars Company Milk Chocolate M&Ms formula specifies that the proportion p of green candies they manufacture is _____ . (Your instructor will provide the answer, or you find out by Googling "M and M Color Distribution".)

j. Assume the company is telling the truth and determine the mean and standard deviation of the sampling distribution of $\hat{p}$. Then on the following page, draw a normal curve showing the sampling distribution of $\hat{p}$ for samples of size 50. Be sure to label the axis with several values.

Mean of the sampling distribution of $\hat{p}$: _____.

Standard deviation of the sampling distribution of $\hat{p}$: _____ .

Does the value of the s.d. as determined from the formula agree well with your answer to part (f)?

Normal curve showing the sampling distribution of $\hat{p}$:

k. On the horizontal axis of dot plot, find your particular $\hat{p}$ value. How many standard deviations (roughly) is your value of $\hat{p}$ from p*?* Recall that the z-score answers this question exactly.

l. Does it look like your data is consistent with the M&M Company's claim? A common notion of what is "extreme" or "unusual" is an observation that is more than two standard deviations from the mean. Is your sample proportion within two standard deviations of p?

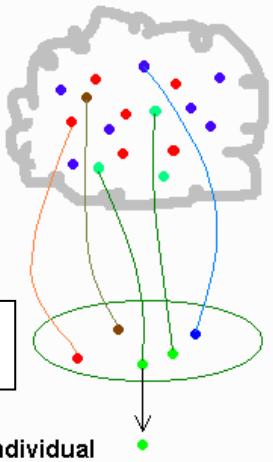
m. Find the approximate probability that in a random sample of 50 M&M's, less than 10% will be green.

Introduction to Statistical Inference: Sampling Distributions

What you need to know:

- **What is a *parameter* and what is a *statistic* ***ONE OF THE MOST CRUCIAL THINGS TO KNOW IN THIS COURSE**
- **A *statistic* computed from a sample is a *point estimate* of the corresponding population *parameter***
- **The proper notation for the population proportion and the sample proportion**
- **The meaning of *sampling variability***
- **The effect of sample size on sampling variability**
- **What a *sampling distribution* is, and why they are the backbone of inference**
- **How sampling distributions can be studied through simulation**
- **The three properties of a sampling distribution: shape, center, spread**

A **population** is a group of objects or individuals



A **parameter** is a ***number*** that measures something about the population. Its value is usually unknown.

The **sample** is a subset of the population.

A **statistic** is a ***number*** calculated from the observations in the sample. We use the statistic to make *inferences* about the population parameter.

---

1. You are attending a large party at a Las Vegas casino. At the front of the room there is a large jar filled with exactly 10,000 poker chips. (This is the *population* for this problem.) Some of them are red and the rest are blue. The host announces a contest to guess the proportion of red chips in the jar. Anyone whose guess is exactly right wins $500. You *could* make a wild guess, but to increase your chances of winning you can purchase a *sample* of chips from the jar and see what proportion are red in your sample. (Your chips are then returned to the jar.)

   a. Suppose you buy a sample of 100 chips for a cost of $25, and find that 32 of them are red. Estimate the *proportion* of chips in the jar that are red.

      Estimate of the proportion of chips in the jar that are red: $\hat{p}$ =

   This is your sample *statistic.*

b. Before a person purchases a sample of chips, the jar is twisted, turned and shaken many times. Why is this important?

c. As it turns out, many other people at the party buy their own samples of 100 chips. Needless to say, they don't all get the same proportion of red chips as you did. What is the important term that describes the fact that these different samples will produce different results?

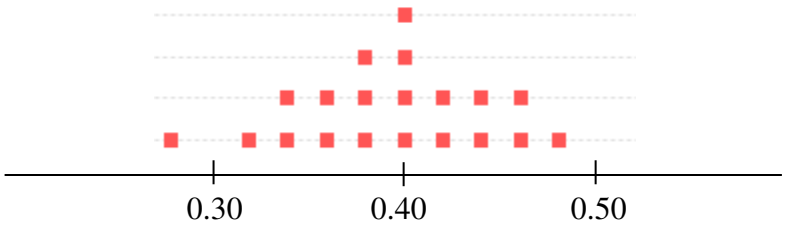
d. The 10,000 chips in the jar represent the population here. What is the population *parameter*?

The _____ of these 10,000 chips that are _____ .

e. Suppose that the jar actually contains exactly 4000 red chips. What is the *value* of the population *parameter*? Write an equation with the symbol for the population parameter on the left and its value on the right:

=

f. Twenty partygoers in all bought 100 chip samples. Here is a dot plot of the various proportions of red chips they obtained:



Use the information in the textbook to determine the mean, standard deviation and shape of the **sampling distribution of** $\hat{p}$. Round the s.d. off to two decimal digits.

Mean: S.d.: Shape:

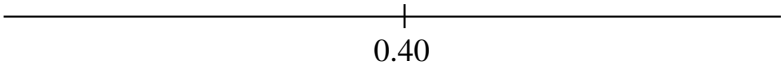
g. Does the dotplot above follow this distribution well? Explain, addressing shape, center and spread.

2. Suppose that the casino announces that, in addition to awarding the $500 prize, anyone whose guess is within 5% of the actual value will get $25.

   a. About what percent of the people who buy samples of 100 chips would you expect to get their money back? (Use the 68-95-99.7 Rule.)

   b. Several other people bought larger samples, containing 400 chips (for a cost of $100). Why did they do this?

   c. With samples of 400 chips, the sampling distribution of $\hat{p}$ will have the same center and shape as before, but *less* variability. Determine the standard deviation for $n = 400$, then add appropriate additional values to the axis and sketch a normal density curve below that represents the sampling distribution of $\hat{p}$.



   d. What percent of the people who bought 400 chips are likely to get $25 back? Shade the corresponding area on your plot and give its value.

3. Explain why the sampling distributions involved in #1 and #2 would look essentially the same if the jar had, say, 25,000 chips rather than 10,000 (assuming the same 40% were red in both cases).

Additional Practice

1. For each of the following, indicate whether what is described is a *parameter* or a *population*:

   a. All four-inch ham sandwiches sold at Quizno's.

   b. The average weight in ounces of all four-inch ham sandwiches sold at Quizno's.

   c. The proportion of registered drivers in California who had an accident in 2008.

   d. All apartment units in New York that are larger than 2000 square feet.

   e. The percentage of dogs and cats in Los Angeles that have been spayed or neutered.

   f. All 100 members of the United States Senate.

2. For each of the following, indicate whether what is described is a *parameter* or a *statistic*:

   a. The fraction of all Americans who have never seen an ocean in person.

   b. The mean number of spots that a sample of 100 ladybugs have.

   c. The proportion of 100 randomly chosen single-family houses in Orange County that have a swimming pool

   d. The percent of all defective iPod Touches made by Apple.

   e. The mean height of all kindergarten children in California.

3. For each of the following, indicate whether what is described is a *sample* or a *statistic*:

   a. 1000 tax returns randomly selected for audit.

   b. The proportion of 1000 tax returns randomly selected for audit that have serious errors.

   c. The percentage of people in a poll who favor withdrawing all troops immediately from Iraq.

   d. A list of 20 beaches along the Gulf Coast inspected to determine the degree of oil contamination.

   e. The number of patients in a clinical trial who experienced complications from the treatment.

4. A poll is administered to a random sample of 250 students at a certain university to determine the percentage that favor a new fee that will go towards enhancing the campus Recreation Center. Only 24% of those polled are in favor of the fee. The standard deviation of the sampling distribution for such polls is 2.7%.

   a. What is the population here?

   b. What is the parameter?

   c. Give the value of the statistic.

   d. Does the size of the university student population play a role in the accuracy of estimation?

5. For the following situations identify (i) the population, (ii) the parameter of interest, (iii) the sample, and (iv) the sample statistic. Your answers to all four should be in *words*; however also give the numerical value for the statistic.

   a. State police set up a roadblock to estimate the percentage of cars with up-to-date registration and insurance. They found problems with 10% of the cars they stopped.

   b. A company packaging snack foods maintains quality control by randomly selecting 10 cases from each day's production and weighing the cases. The weight of a case should be 2 lbs on average. One day they found that the total weight of the 10 cases was 20.5 lbs.

c. A magazine asked all subscribers whether they had used alternative medical treatments and, if so, whether they had benefited from them. 22% of those responding reported cures or substantial improvement in their condition.

6. Explain why sampling distributions are important in statistical inference.

7. A simple random sample of 1000 Americans found that 61% were satisfied with the service provided by the dealer from whom they bought their car. A simple random sample of 1000 Canadians found that 58% were satisfied with the service provided by the dealer from whom they bought their car. The sampling variability associated with these statistics (circle the correct answer):

   a. is about the same
   b. is smaller for the sample of Canadians since the population of Canada is smaller than that of the United States, hence the sample is a larger proportion of the population
   c. is smaller for the sample of Canadians since the percentage satisfied was smaller than that for the Americans
   d. is larger for the Canadians, since Canadian citizens are more widely dispersed throughout the country than in the United States, hence have more variable views

8. Suppose that studies were made to estimate the mean number of televisions owned by families in various California cities. Circle the choice a, b, c or d below that indicates which option would be likely to give the most accurate estimate of that city's mean, and which would be likely to give the least accurate estimate:

   - A random sample of 500 families from Bakersfield (population 330,000)
   - A random sample of 500 families from San Diego (population 1,350,000)
   - A survey form distributed in various locations (shopping malls, libraries, post offices, etc.) of Santa Barbara (population 95,000), with 7,285 responses received

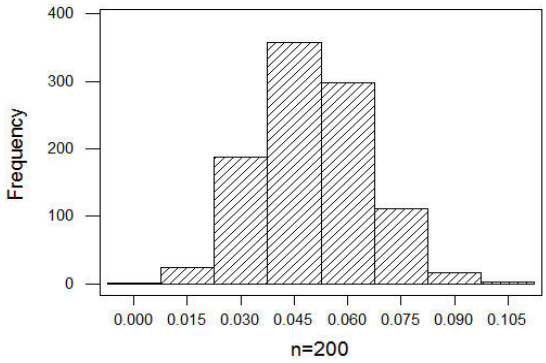
   (a) Most accurate: Bakersfield
   Least accurate: Santa Barbara

   (b) Most accurate: Santa Barbara
   Least accurate: Bakersfield, San Diego

   (c) Most accurate: Santa Barbara
   Least accurate: San Diego

   (d) Most accurate: Bakersfield, San Diego
   Least accurate: Santa Barbara

9. A fundraising organization typically gets a donation from about 5% of the people on their mailing list. However, this response rate may vary according to what the mailing is asking for as well as seasonal and economic factors. Suppose that the organization does a test mailing for a new appeal they plan to send out soon. The test mailing involves sending the appeal to 200 names randomly selected from its mailing list. They receive a donation from 18 of them, for a response rate of 9%. Does this give good evidence that if they do a mass mailing to their entire mail list they will receive donations from more than the typical 5% of the recipients? To find out, they consulted a statistician, who *simulated* 1000 "mailings" to 200 "individuals" from a population with a population response rate of 5% and constructed the histogram below of resulting sample proportions:



a. Show that the histogram reflects what the information in Chapter 15 says about the sampling distribution for a sample proportion. Consider shape, center and spread. Be sure to check the rule of thumb given on page 341.

b. Based on the histogram, if the response rate for the entire population will actually be 5%, would it be unusual to obtain a sample response rate of 9% in 200 appeals? Explain.

c. Now calculate the probability of obtain a sample response rate of 9% in 200 appeals if the response rate for the entire population is 5%. (Use a z-score.)

d. Do the test mailing results give convincing evidence that the response rate for this appeal will be higher than 5% when it is sent out to everyone on the mailing list (the population)? Explain. Base your answer on your answer to parts (b) and (c).

10. Assume that 30% of the students at a university wear contact lenses. We randomly pick 100 students. Let $\hat{p}$ represent the proportion of students in this sample who wear contacts.

a. Describe and draw the sampling distribution model for the proportion of students in such a random sample who wear contacts. Make sure to mention shape, center, and standard deviation in your description. Also, be sure to verify that the rule of thumb on page 341 is satisfied.

b. Find the approximate probability that less than 20% of this sample wear contacts.

c. Find the approximate probability that more than one third of this sample wear contacts.

Confidence Intervals

What you need to know:

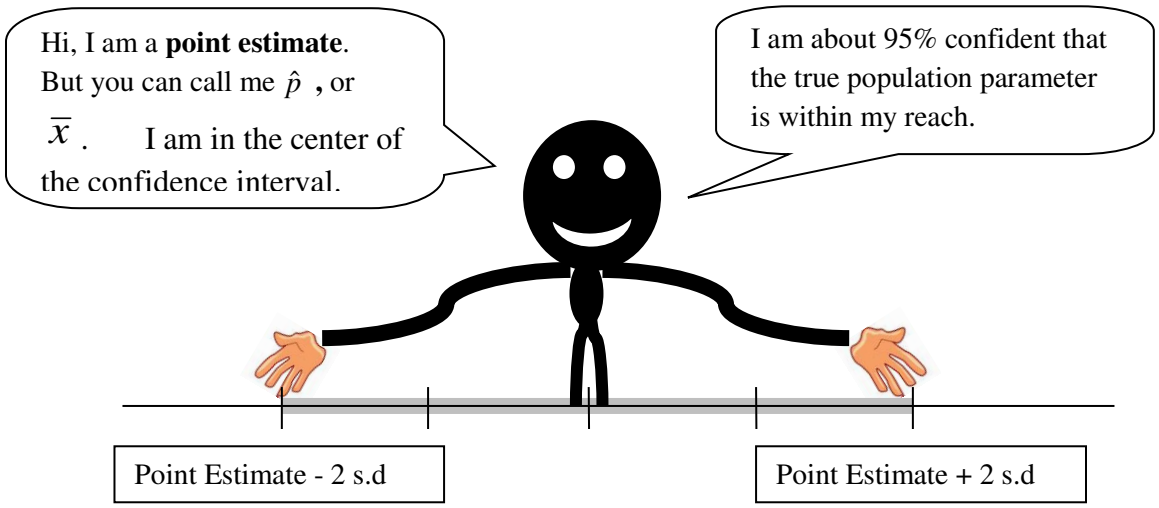
- **The purpose of confidence intervals**
- **The form of confidence intervals for a population proportion**
- **How margin of error depends on the confidence level**
- **How margin of error depends on the sample size**
- **Determining sample size for a desired margin of error**

" I got the instructions from my Statistics Professor. He was 80% confident that the true location of the restaurant was in this neighborhood."

Due to *sampling variability*, we cannot be sure exactly how close a point estimate is to the true value of the population parameter. The purpose of a confidence interval is to give an interval for which any value in the interval is a *plausible* value of the parameter, while values outside of the interval are *implausible.*

Think back to how sampling distributions for a sample proportion work. We know that if the sample size is large enough, the sampling distribution of $\hat{p}$ is approximately normal, and therefore we know that for 95% of random samples our point estimate will be no more than two standard deviations away from the population parameter. Let's look at this from the point estimate's point of view:

"If I am a sample proportion $\hat{p}$, there is about a 95% chance that I am no more than two standard deviations away from the population proportion *p.* That means that *p* is no more than two standard deviations away from *me*. If I reach out two standard deviations away from myself to both sides, I am 95% confident that *p* will be within my grasp."



So to determine the margin of error of a *95% confidence interval for a population proportion*, we simply need to determine the standard deviation of the sampling distribution and double it. Recall that the standard deviation of the sampling distribution of the sample proportions is $\sqrt{\frac{p(1-p)}{n}}$. But we have a problem here: we don't *know p*. After all, that's exactly what we want to estimate! So what should we do? The logical thing is to replace it in the formula above by its point estimate, $\hat{p}$.

The estimated standard deviation of the sampling distribution of the sample proportion we have just obtained is called the _____ _____ of $\hat{p}$:

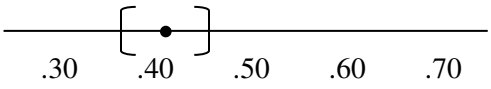
$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

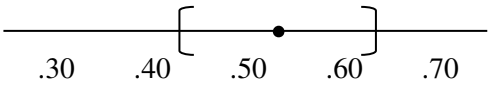
1. A Gallup poll surveyed a random sample of 1016 households in the U.S about their pets. Of those surveyed, 599 said they had at least one dog or cat as a pet.

   a. Check the conditions for inference about a population proportion.

   b. Find and interpret a 95% confidence interval for the population proportion of households that owned a dog or cat. First write it in the form *point estimate* ± *margin of error*, then write it in standard interval notation (,). ***The exact value of z^* for a 95% confidence interval is 1.96, very close to the value of 2 that the 68-95-99.7 Rule uses for 95%. Check with your instructor as to which value (s)he wants you to use when finding 95% confidence intervals.***

2. Two interval estimates of a proportion *p* are shown below. One is based on a sample size of 350 and the other is based on a sample size of 1200. Both samples were taken randomly. Indicate which interval estimate goes with which sample size, and indicate the locations of the two point estimates:





3. For the situation in Problem #1, suppose the Gallup organization had wanted a result with a margin of error of 1%. How many households would they have needed to survey? Use the value of $\hat{p}$ from #1 as your guess p^* of the actual value of p in order to determine the approximate sample size that would have been required.

4. The instructor of a nutrition course surveyed her students to see how many of them were vegetarians. None of her 24 students were vegetarians.

a. Treating her class as a random sample of all students at the instructor's university, find a 90% confidence interval for the proportion of the entire student body that is vegetarian. Does the answer make any sense?

b. Now find a 90% confidence interval using the **plus four confidence interval**. Is the answer more reasonable?

c. Do you believe it is reasonable to treat the students from this instructor's class as roughly equivalent to a simple random sample of the entire student body? Why or why not?

Additional Practice

1. What is the goal when using a confidence interval?

2. Explain carefully what we mean by the phrase "95% confidence".

3. Which level of confidence would produce the widest confidence interval from the same data?

   a. 90%
   b. 99%
   c. 95%

4. What is the probable effect on the width of the confidence interval when the sample size is increased?

   a. The width increases
   b. The width decreases
   c. No predictable effect

5. What value is in the exact middle of the confidence interval, the statistic or the parameter?

6. Suppose you wish to estimate from a sample, with 95% confidence, the proportion of computers that need repairs or have problems by the time they are three years old. Your estimate must be accurate within 3% of the true (population) proportion.

   a. If no preliminary estimate is available, find the minimum sample size required to guarantee that the margin of error will be no larger than specified above.

   b. Now suppose a prior study involving less than 100 computers found that 19% of these computers needed repairs or had problems by the time the product was three years old. Find the minimum sample size needed to achieve the specified margin of error.

7. In your biology lab you need to estimate the proportion of dead cells in a suspension. You need to pick a random sample from the suspension, so you stir it very well, then using a pipette you take a sample from it and put some special dye in it. The dye turns the dead cells blue. As you look in a microscope, six of the 85 cells you look at are dead.

a. What is the parameter of interest? Use both a symbol and words.

b. Find the value of the sample statistic.

c. Estimate the parameter of interest with 99% confidence.

d. Pretend you need to write a report in your lab book. Write an interpretation of the results.

8. Adverse drug reactions to legally prescribed medicines are among the leading causes of drug-related death in the U.S. Suppose in a random sample of 60 drug-related deaths, 53 were caused by legally prescribed drugs and the rest were result of illicit drug use. Based on these results, the reported 95% confidence interval is 86% ± 8.5%. Their interpretation is as follows:

> "*We are 95% confident that the true percent of all drug-related deaths caused by legally prescribed drugs in this sample is between 77.5% and 94.5%.*"

Is this interpretation correct? If not, correct the part that's wrong.

Tests of Significance for a Population Proportion

What you need to know:

- **The goals of a test of significance**
- **The reasoning of tests of significance**
- **The terminology of test of significance**
- **One-sided and two-sided alternatives**
- **How to carry out a test of significance for a population proportion,**
- **from stating the hypotheses to interpreting the results correctly**
- **Interpretation of the *p*-value**

---

1. In Question #18 of the Math 140 class questionnaire, each student in the class was asked to pick a number at random between one and ten. It is suspected that people are more likely to pick 7 than any other number.

   a. If people were really picking purely at random, about what percent of the class should have picked 7 as their number?

   b. Suppose we want to test the hypothesis that Math 140 students are more likely to pick 7 than they would be if they were choosing a number at random. State the null and alternative hypotheses in both symbols and in words:

   c. In a certain Math 140 class, ten of the 41 students picked the number 7. Give the value of the sample proportion, using the correct symbol:

   ____ = ____

   d. Treat the class as a random sample of all students who take Math 140, and check the other conditions for inference. Are they both satisfied?

e. Draw the sampling distribution of the sample proportion on the axis below, assuming that the true proportion of all Math 140 students who would pick 7 is 10%. What value is the center of the distribution? What is the standard deviation of the distribution? Mark the value of the sample proportion on the axis. Then shade the area corresponding to the *p*-value.

---

f. Compute the value of the test statistic.

g. Determine the *p*-value.

h. Mark the value of the test statistic underneath the value of the sample proportion on the graph you drew on the previous page. Does the *p*-value you found in part (g) seem to match the amount you shaded before?

i. Write a conclusion. Include a statement about the strength of evidence against the null hypothesis.

2. In the 1980s it was generally believed that congenital abnormalities affected about 5.2% of the nation's children. Some people believe that the increase in the number of chemicals in the environment has led to an increase in the incidence of abnormalities. A recent study examined 384 children and found that 28 of them showed signs of abnormality.

   a. Is the appropriate alternative here one-sided or two sided? Explain.

   b. Follow the steps of a significance test and state your conclusions by answering in context to the following questions: At the $\alpha = 5\%$ significance level, does this data give strong evidence that the risk has increased? At the 1% level? Perform the test of significance to find out.

3. A nutritionist estimates that 40% of U.S. adults eat breakfast every day. We want to test whether we can reject the nutritionist's claim. In a random sample of 300 U.S. adults, 45% say they eat breakfast every day.

   a. Is the appropriate alternative here one-sided or two sided? Explain.

   b. At the 10% significance level, is there enough evidence to reject the nutritionist's claim? What about at $\alpha = 5\%$? Carry out the test to find out.

Additional Practice

1. For each of the following situations, define the parameter (using a symbol and words) and set up appropriate null and alternative hypotheses about it. The parameter has been defined for you in the first problem:

   a. Under normal conditions, 64% of the seeds of the rare plant *Botanica statistica* germinate. Scientists at an agricultural research station believe that application of a certain vitamin formulation may increase the germination rate.

      p = the proportion of all seeds of *Botanica statistica* that will germinate if the vitamin formulation is applied

      H_0:

      H_a:

   b. A government agency reports that the proportion of automobiles that come off of assembly lines without significant manufacturing flaws is 96%. A consumer organization suspects that the actual figure is lower than this.

      p =

      H_0:

      H_a:

   c. Nationally, 8% of homeowners get their house painted in any given year. A random sample of California homeowners is taken to see if the percentage is different from the national figure.

      p =

      H_0:

      H_a:

2. Suppose we select a random sample of 100 Math 140 students and find that 43% said they believe in love at first sight. Mark each statement as True or False:

   a. There were 43 students in the sample who said they believe in love at first sight.
   b. Based on the information in the sample, we cannot determine exactly what proportion of the population of all Math 140 students would say they believe in love at first sight.
   c. $\hat{p} = 0.43$
   d. $p = 0.43$

3. In a recent Gallup Poll, about 50% of the respondents said they believed in love at first sight. Let *p* be the proportion of *all college students* who believe in love at first sight.

   a. What null and alternative hypotheses do you think would be appropriate?

   b. Using either the information from Problem #2 or the responses to Question #23 of the class questionnaire from your own class, perform a test of significance using a significance level of $\alpha = 10\%$.

   c. State your conclusion in context.

   d. Are you completely comfortable with your conclusion? (Hint: Was the sample drawn from the indicated population?)

4. Currently, about 10% of marriages in the United States end in divorce during the first five years of marriage. A sociologist is studying the effect on the divorce rate of having children within the first two years of marriage. Using hospital birth records, she selected a random sample of 100 couples who had a child within the first two years. Following up on these couples, she finds that 17 of these couples had divorced within the first five years.

   a. Set up appropriate null and alternative hypotheses to test whether having children within the first two years of marriage affects the divorce rate. Assume having children could either increase *or* decrease the divorce rate.

   b. Calculate the value of the test statistic.

   c. Determine the *p*-value of the test.

   d. Suppose a test is performed using a significance level of $\alpha = .05$. Based on your answer to (c), would we reject the null hypothesis? What could we conclude, if anything, about the effect on the divorce rate of having children within the first two years of marriage?

5. ***Understanding the p-value*:** A university administrator guesses that about 20% of the students at her school had sent a text message during class at least once. The head of the Faculty Senate believed that the percentage is higher than that, so he conducted a survey of randomly selected students. 28% of the students surveyed admitted to having sent a text message during class at least once. Using the survey data, a test of significance of H_0: $p = .20$ versus H_a: $p > .20$ gave a p-value of 3%.

   a. Give a correct interpretation of this p-value by filling in the blanks in the statement below:

   "If the true proportion of students who have texted during class is 20%, then the _____ that a survey like this would result in 28% or _____ of the students admitting to having texted during class is ____."

   b. Does this give *strong evidence* or *weak evidence* that the head of the Faculty Senate was right in thinking that more than 20% of students at her school have texted during class at least once? Carefully explain the reasoning behind your answer.

Inference for Comparing Two Population Proportions

What you need to know:

- **How to calculate and interpret a confidence interval for the difference of two population proportions**
- **How to carry out a test of significance to compare two population proportions**

---

1. A sociologist wants to know whether there is a difference between the proportions of male and female college students at his college who work over the summer. A random survey of students at the college gave the following results:

	n	***Worked in Summer***
Males	125	47
Females	153	52

   a. State the parameters of interest in both symbols and words.

   b. State the null and alternative hypotheses in symbol form.

   c. Are the conditions for inference satisfied?

   d. The 90% confidence interval for the difference between the two population proportions is (-0.059, 0.131). State your conclusion based on this interval.

e. The value of the test statistic is $z = 0.63$. Sketch a z curve and mark the location of the given value. Explain what this indicates about whether there is evidence for a difference in the proportions of male and female college students at the college who work over the summer. Do this without computing the *p*-value.

f. The *p*-value is 0.53. Is this consistent with what you concluded in part (d)? Explain.

g. Shade the *p*-value on the curve you drew in part (e).

2. Do more women or more men believe in love at first sight? Using either the results from #23 of the class questionnaire, or use the following information from another Math 140 class that answered the questionnaire to address this question. Pretend that the students who participated in the survey are a random sample from the population of interest.

	Females	Males
Yes	9	5
No	19	7

a. Create a segmented bar graph from the data and briefly describe what it shows.

b. Describe the parameters of interest in both symbols and words.

c. Compute the pooled sample proportion $\hat{p}$.

d. Do the data indicate that there is a difference in the proportion of men and women who believe in love at first sight? Perform a test of significance and discuss what you find.

3. If you have time, verify the confidence interval and the value of the test statistic given in #1.

Additional Practice

1. The Pew Research Center found in a recent survey that 61% of teenage girls surveyed use Instagram, compared to 44% of the boys. The report on this survey states that the difference is statistically significant. Explain what they meant by that.

2. The Centers for Disease Control and Prevention reported from a survey of randomly selected U.S. seniors age 65 of older that 411 of 1012 men and 535 of 1062 women suffered from some form of arthritis.

   a. Are the conditions for inference satisfied?

   b. Do these data indicate that a higher proportion of women of age 65 and older are affected by arthritis than men in the same age group? Carry out a significance test to answer the question.

3. In 1997 a random sample of 200 low income families was taken and it was found that 43 of them did not have health insurance. In 2003 a similar survey of 270 families was taken and 42 did not to have health insurance. Follow the steps below to determine whether there is statistically significant evidence from these samples that the proportion of low income families without insurance changed over that time. Use a 10% significance level.

   a. State the null and alternative hypotheses. Be careful to base the alternative on the wording of the question.

   b. Are the conditions satisfied for inference?

   c. State your conclusion in context using the results from the confidence interval and the z-test shown in the computer output below:

```
Test and CI for Two Proportions

Sample       X      N   Sample P
1997        43    200   0.2150
2003        42    270   0.1555

Difference = p(1997) - p(2003)
Estimate for difference: 0.3706
90% CI: (-0.0006, 0.1194)
Test for difference = 0 (vs ≠ 0):  z = 1.655    p-value: 0.0978
```

4. Here is a two-way table comparing the effectiveness of two drugs, Lithium and Imipramine, in preventing a recurrence of depression among patients who were hospitalized with depression:

	Recurrence	No Recurrence
Imipramine	8	21
Lithium	18	9

You may have previously used exploratory data analysis (conditional distributions and a segmented bar graph) to investigate the relationship between drug and the frequency of recurrence. Now use a test of significance to determine whether the apparent advantage of Imipramine shown in the data above could simply be due to chance. Follow the steps below:

a. Find the sample proportions $\hat{p}_{Lithiun}$ and $\hat{p}_{Imipramine}$ that had a recurrence. Then find the pooled proportion.

b. State the null and alternative hypotheses, check the conditions for inference, and then compute the test statistic and *p*-value for the data.

c. Does the data give convincing evidence that Imipramine is better than Lithium at preventing a recurrence of depression?

End of *Inference About Categorical Variables.* Check your knowledge:

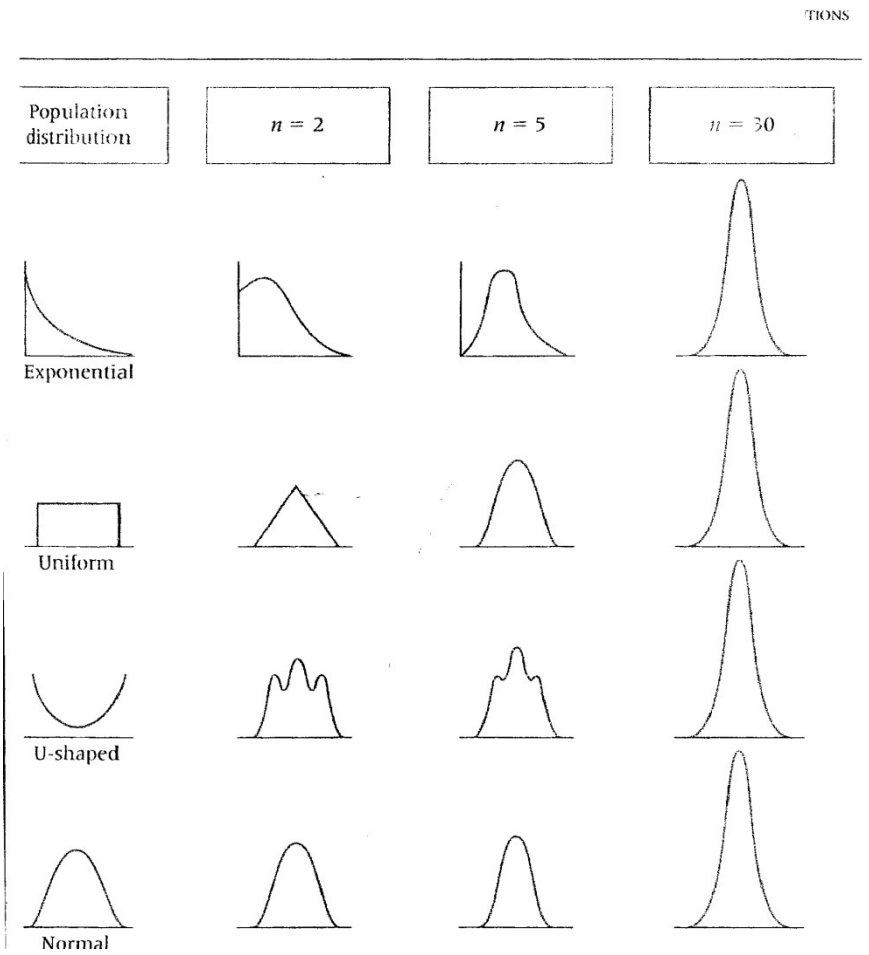
1. Define the term *statistical inference.*
2. Explain the difference between a *population* and a *sample*.
3. Be able to define and explain the difference between a *parameter* and a *statistic*. **Important!**
4. What are the two types of parameters we will do inference about in this class?
5. What is the meaning of the term *sampling variability*? **Important!**
6. What is meant by the *sampling distribution* of a statistic? **Important!**
7. How is the size of a sample related to the *variability* of the sampling distribution?
8. How is the probable accuracy of an estimate based on a sample related to the size of the sample?
9. How is the probable accuracy of an estimate based on a sample related to the size of the population?
10. **True or False:** We cannot predict the likely accuracy of an estimate obtained from a sample if the sample is not taken randomly.
11. **True or False:** Usually the parameter value will fall within the interval specified by the point estimate plus or minus its margin of error, but this is not guaranteed to happen.
12. What is meant by the terms *margin of error* and *interval estimate?*
13. Explain how simulation allows us to study sampling distributions.
14. What are the conditions for the shape of the sampling distribution of a sample proportion to be approximately normal?
15. Explain the concept behind the two sd/95% margin of error.
16. **True or False:** The margin of error of an estimate will typically be smaller if a large sample size is used than if a small sample size is used.
17. Explain the difference between the meaning of the symbols p and $\hat{p}$.
18. For a SRS of size n, what happens to the shape of the sampling distribution of $\hat{p}$ as the sample size n increases?
19. For a SRS of size n, what is the mean of the sampling distribution of $\hat{p}$?
20. For a SRS of size n, what is the standard deviation of the sampling distribution of $\hat{p}$?
21. What happens to the standard deviation of $\hat{p}$ as the sample size n increases?
22. **True or False:** A confidence interval is a set of plausible values for the parameter.
23. **True or False:** A confidence interval is centered at the population parameter.
24. Know how to find the point estimate and the margin of error if the confidence interval is given.
25. Know how the confidence interval behaves as we increase the confidence level, or change the sample size.
26. Name the two kinds of hypotheses involved in tests of significance.
27. **True or False:** The null hypothesis will probably be rejected if the point estimate is in the tail of the sampling distribution and its value supports the alternative.
28. **True or False:** The smaller the p-value, the stronger the evidence against the null hypothesis.
29. Be able to give the meaning of the *p-value* in the context of a problem.
30. Know the degree of evidence against H_0 that specific *p-values* represent.
31. Explain the term *significance level*.
32. What does the term *statistically significant* mean?
33. **True or False:** The null hypothesis will be rejected if the *p-value* is larger than α.

Sampling Distribution For a Mean

What you need to know:

- **The proper notation for the population mean and the sample mean**
- **What a sampling distribution for a sample mean is**
- **How the sampling distribution of $\overline{x}$ can be generated.**
- **The Central Limit Theorem**

Means Problems:	parameter = population mean:	μ
	statistic = sample mean:	$\overline{x}$



Shapes of the distributions of sample means for three sample sizes drawn from four different population distributions

1. A jar of Jif Peanut Butter (labeled as a 32 oz. jar) is selected randomly off of the assembly line, and the weight of the jar is measured. Suppose that jar weights are normally distributed with a mean weight of 32.3 oz. and a standard deviation of 0.40 oz.

   a. Identify the variable of interest. Is it categorical or quantitative?

   b. Identify the two population parameters. Give both their symbols and their values.

   c. What is the average weight of all jars of *Jif*? Is this value a parameter or a statistic?

   d. What is the probability a randomly selected jar is underweight (filled with less than 32.0 fl oz)?

   e. What proportion of all jars of *Jif* are underweight?

   f. Consider a six-pack containing six jars of *Jif*. Assume these jars represent a simple random sample of all jars produced. Let $\bar{x}$ be the sample mean weight for these six jars. Is $\bar{x}$ a parameter or a statistic?

   g. How does the variability in the mean weights of jars in different six-packs compare to the variability in the weights of individual jars?

h. Find the mean and standard deviation of the (sampling) distribution of the sample mean weight $\overline{x}$ of jars in a six-pack. Compare the standard deviation to the one given for individual jars at the beginning of this problem. Does this support your answer to (g)?

i. What is the shape of the distribution of sample means?

j. Find the probability that the average weight of the jars in a six-pack is less than 32.0 fl oz.

k. Now consider a *carton* containing 24 jars of *Jif*. Treat these jars as if they are a simple random sample from all jars of *Jif.* What are the mean and standard deviation of the distribution of $\overline{x}$ now? Note that the sample size has been multiplied by 4 when compared to a six-pack. How has the standard deviation changed with the quadrupling of the sample size?

m. Do you think it will be more likely or less likely that the mean weight of a carton is less than 32.0 fl oz. than it was for a six-pack? Take into account how the variability of the sampling distribution for a sample of size 24 compares to the variability for a sample of size 6.

n. Find the probability the sample mean weight of the 24 jars is less than 32.0 fl oz. Does your answer confirm your guess in part (l)?

2. The per capita consumption of processed fruits by people in the U.S. in a recent year was normally distributed with a mean of 152.7 lb and a standard deviation of 51.6 lb.

   a. Find the mean and standard deviation of the sampling distribution of $\overline{x}$ for random samples of size 9. Then determine whether it would be unusual for a random of 9 people to have a sample mean consumption of processed fruits as high as 175 lb.

   b. Find the mean and standard deviation of the sampling distribution of $\overline{x}$ for random samples of size 36. Then determine whether it would be unusual for a random of 36 people to have a sample mean consumption of processed fruits as high as 175 lb.

   c. Sketch the sampling distributions from parts (a) and (b) on the same scale. Label more values on each side of 152.7.

   152.7

   d. We might speculate that people over the age of 60 tend to eat more processed fruits than the rest of the country, on average. Suppose that a simple random sample of people over the age of 60 was taken, and their mean consumption of processed fruits for the year turned out to be 175 lb. Would this give good evidence that people over the age of 60 *do* tend to eat more processed fruits than the rest of the country? Answer this question first assuming the sample has $n = 9$ people, and then again assuming the sample has $n = 36$ people. Use the results of parts (a) & (b).

e. Thinking of the situation described in part (d) as a test of significance, what would be:

the null hypothesis, in words?

the alternative hypothesis, in words?

the p-value if $n = 9$?

the p-value if $n = 36$?

Additional Practice

1. Place an X by any of the following statements that are NOT true:

   ___ An increase in sample size from $n = 16$ to $n = 25$ will reduce the standard deviation of the sampling distribution for $\bar{x}$.

   ___ The mean of the sampling distribution of $\bar{x}$ is equal to the population mean divided by the square root of the sample size.

   ___ The larger the sample size, the more the sampling distribution of $\bar{x}$ resembles the shape of the population distribution.

   ___ The mean of the sampling distribution of $\bar{x}$ for samples of size $n = 15$ will be the same as the mean of the sampling distribution for samples of size $n = 100$.

   ___ The larger the sample size, the more the sampling distribution of $\bar{x}$ will resemble a normal distribution, regardless of the shape of the population distribution.

   ___ If the shape of the population distribution is itself normal, then the sampling distribution of sample means will resemble a normal distribution for *any* sample size.

2. The heights of American adult women are approximately normally distributed. The mean height of all adult American women is 63.5 inches with a standard deviation of 2.5 inches. Imagine that all possible random samples of size 25 ($n = 25$) are taken from the population of American adult women's heights, and then the means from each sample are graphed to form the sampling distribution of sample means.

   a. Using the Central Limit Theorem, draw and label this sampling distribution. Show the mean and standard deviation on your drawing.

   b. Find the probability that the mean height of a random sample of 25 women is less than 62.5 inches.

c. Find the probability that the mean height of a random sample of 25 women is more than 64 inches.

3. The amount of time it takes to complete an exam has a skewed-to-the-left distribution with a mean of 65 minutes and an s.d. of 8 minutes. A sample of 64 students is selected at random. Describe and draw a picture of the sampling distribution of the sample mean ($\bar{x}$) for samples of size $n = 64$.

4. A brake pad manufacturer claims its brake pads will last for 38,000 miles, on average. Assume that the lifespans of the brake pads are normally distributed. Past analyses indicate that $\sigma = 5000$ miles. You work for a consumer protection agency and you are testing this manufacturer's brake pads using a random sample of 30 brake pads. In your tests, the mean lifespan of the brake pads you sample is 35,700 miles.

   a. Would it be unusual to have an *individual* brake pad last for only 35,700 miles? Why or why not?

   b. Assuming the manufacturer's claim is correct, what is the probability that the *mean lifespan of the sample* is as low as 35,700 miles? (That is, 35,700 miles or lower.)

   c. Using your answer from (b), what do you think of the manufacturer's claim?

Inference for a Population Mean

What you need to know:

- **Using the *t* distribution**
- **How to calculate and interpret a confidence interval for a population mean**
- **How to carry out a test of significance for a population mean, from stating the hypotheses to the conclusion**
- **Matched pairs procedures**

"Air Traffic Control, I'm going 400 miles an hour and need to land this thing on a floating narrow runway, so I'd prefer something better than 95% confidence in those coordinates."

f

---

1. The 1960 Census results indicated that the age at which American men first married had a mean of 23.3 years. It is widely suspected that young people today are waiting longer to get married. We select a random sample of 20 men who married for the first time last year and find they married at an average age of 24.2 years, with a standard deviation of 5.1 years. Find the errors in the attempt below to test an appropriate hypothesis. How many mistakes can you find in the work below and in the conclusion? Cross out each error and write a correction.

$H_0: \bar{x} \geq 23.3$

$H_a: \bar{x} < 23.3$

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{23.3 - 24.2}{5.1/\sqrt{20}} = -0.789 \qquad p\text{-value} = 0.2151$$

Conclusion: There is a 21% chance that the mean age men first get married is 23.3. Thus, there is strong evidence that the average age at which men first get married has not increased since 1960.

2. Humerus bones (no, these are not the "funny bones") from different members of the same species of animal tend to have approximately the same length-to-width ratio, which is different from the ratio for other species. When fossils of humerus bones are discovered, therefore, archeologists can often determine the species of animal by measuring the length-to-width ratios of the bones. Forty-one fossils of humerus bones were unearthed at an archeological site in East Africa, in a location where the species *Ancientius Bonius* may have lived, and the length-to-width ratios of these bones were calculated.

   It is known from previous studies that the population mean ratio for *Ancientius Bonius* is 8.5. Therefore in order to judge whether the fossils found could be from this species, the archeologists plan to test at the 5% significance level whether the population mean ratio for all bones of the species that lived at this site is 8.5 or differs from 8.5.

a. State the two hypotheses in symbols.

b. Using the computer output below, make an appropriate conclusion in context.

T-Test of the Mean

```
Test of mu = 8.500 vs mu not = 8.500

Variable     N      Mean    StDev   SE Mean        T          P

RATIO       41     9.258    1.204     0.188     4.03     0.0002
```

c. Explain how the output below is consistent with the results of the test of significance.

T Confidence Intervals

```
Variable     N      Mean    StDev  SE Mean       95.0 % CI

RATIO       41     9.258    1.204    0.188   (8.878,   9.637)
```

3. Industrial pollution often makes water supplies either acidic (pH < 7.0) or basic (pH >7.0). An industrial company that dumps waste water into a nearby river claims that the mean pH level of the water in the river near the site of the company's discharge is 7.0, so it is neutral. You, as an environmentalist, randomly select 19 water samples and measure the pH of each. The sample mean and standard deviation are 6.9 and 0.24, respectively. Assume the population of pH levels in all possible water samples is normally distributed.

   a. Is there enough evidence to reject the company's claim at the 5% significance level? How about at the 10% level?

   b. Compute 90% and 95% confidence intervals for the mean pH level. Are the results consistent with what you found in part (a)?

4. Can a baseball player hit a baseball farther with an aluminum bat? A contest was held in which each player was given five swings against a pitching machine with a wooden bat and five swings with an aluminum bat. The longest distance the ball was hit was recorded for each participant using each type of bat. Here are the results:

Player	1	2	3	4	5	6	7	8	9	10
Wooden bat	280	315	308	340	295	369	325	330	307	355
Aluminum bat	305	320	340	330	318	385	340	330	312	330

a. Explain, step-by-step, how to test whether there is the difference in maximum distance for the two types of bat. Show the hypotheses and indicate the remaining steps, but do not actually carry out the test. Treat the players in this study as a SRS from the population of all players at their level.

b. Explain how to compute a confidence interval for the difference in maximum distance for the two types of bat. Do not actually compute the interval.

c. Carry out the steps for both the test and the confidence interval.

5. Here are the periods of revolution around the Sun (in Earth days) for the eight planets in our solar system:

Planet	**Period**
Mercury	88
Venus	225
Earth	365
Mars	687
Jupiter	4,332
Saturn	10,760
Uranus	30,684
Neptune	60,188

Does it make sense to calculate a 95% confidence interval for the mean period of revolution around the Sun (in Earth days) for the eight planets? Explain.

Additional Practice

1. A consumer advocacy agency plans to study a new gasoline additive that is being advertised as increasing a car's gas mileage. They plan to collect data from a sample of cars and do a test of significance. Either the product increases gas mileage or it has no effect at all. Which of these two possibilities is the null hypothesis, and which is the alternative? Justify your answer.

2. A SRS of $n = 100$ residents of Columbia, South Carolina yielded a 99% confidence interval for the mean age (in years) of Columbia residents of (29.8, 38.5). What is the correct interpretation attached to this interval?

   a. We are 99% confident that the mean age of all Columbia residents is between 29.8 and 38.5.
   b. 99% of the residents in our sample had ages between 29.8 and 38.5.
   c. We are 99 % confident that the mean age of the Columbia residents in our sample is between 29.8 and 38.5.
   d. All of the above are valid interpretations.

3. The 95% confidence interval for the mean number of latex gloves used per week by all health-care workers in a certain hospital is (15.86, 22.74).

   a. Write a correct interpretation of this confidence interval.

   b. What is in the exact middle of the interval? Circle all the correct answers.

Sample mean	*Population mean*	*19.30*	*Margin of error*
Sample statistic	*Population parameter*	*3.44*	*95%*

4. In Question 2, how could one decrease the width of the confidence interval?

   a. Increase the sample size
   b. Use a lower confidence level
   c. Both (a) and (b) are correct
   d. Neither (a) nor (b) are correct

5. Based on a random sample of 100 employees at a large firm, a 95% confidence interval is calculated for the mean age of all of the firm's employees. The interval is (34.5 years, 47.2 years).

   a. What was the sample mean? (Hint: Where does the point estimate fall within a confidence interval ?)

   b. Determine the margin of error.

6. It has long been reported that normal human body temperature follows a normal distribution with a mean of 98.6 degrees Fahrenheit. But is the mean really 98.6^{o}F? A group of medical researchers measured the body temperature of a random sample of 18 healthy individuals. After the researchers made their measurements, they entered the data into a statistical software package and ran the appropriate statistical test. The output generated from this test is printed below. Based on this output, what should they conclude and why?

T-Test of the Mean

Test of mu = 98.60 vs mu not = 98.60

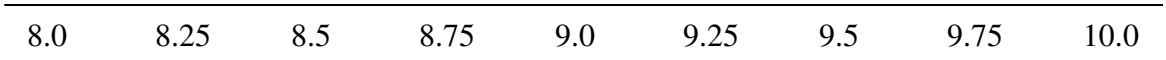
Variable	N	Mean	StDev	SE Mean	T	P
TEMP	18	8.2167	0.68363	0.16113	-2.39	0.029

7. A newspaper publisher wants to estimate the mean length of time all subscribers spend reading newspapers. To determine this estimate, the publisher takes a random sample of 15 people and obtains the following times (in minutes):

11, 9, 8, 10, 10, 9, 7, 11, 11, 7, 6, 9, 10, 8, 10

a. Construct and interpret the 90% and 95% confidence intervals.

b. Graph both confidence intervals on the same number line below. Which one is wider?



c. Do we have to assume that the population of reading times is normally distributed?

8. Ecologists periodically measure the concentration of coliform bacteria at an urban reservoir. Each time, water samples are taken from various random locations within the reservoir. The distribution of bacteria counts in the past has followed a normal distribution with a mean of 2.8 ppm (parts per million). In the latest set of readings, the mean count for the 27 water samples was 2.98 ppm with an s.d. of 0.455 ppm. Thus there is some evidence that the bacteria concentration in the reservoir has increased.

   a. Set up hypotheses for testing whether the bacteria concentration in the reservoir has increased.

   b. If the mean bacteria concentration in the reservoir has not changed, show that the chance of seeing evidence this strong that bacteria levels have increased is only 2.5%. (Compute the *t*-statistic for the given sample data and find the chance of obtaining a *t* value that high.)

   c. What is the term for what the 2.5% from part (a) represents?

   d. The fact that 2.5% is a small probability means that it is unlikely to see a sample mean concentration as large as 2.98 ppm just by chance. What is the strength of the evidence, therefore, that the reservoir is more contaminated than before?

      (i) little or none (ii) some, but weak (iii) fairly strong (iv) very strong

   e. What would you conclude at significance level 5%?

   f. What would you conclude at significance level 1%?

9. Educators are interested in knowing whether a new experimental program has an impact on the reading readiness of first graders. Assume that with the existing program the mean Reading Readiness Score for all first graders is 100. A random sample of 41 first graders who have been through the new program had a mean score of 104 and a standard deviation of 24.5.

   a. Draw and label a graph of the sampling distribution of $\bar{x}$, a mean score for a random sample of 41 first graders under the existing program. (Since $n = 41$ is fairly large, the Central Limit Theorem applies, so you know the shape that the sampling distribution of $\bar{x}$ will have.) Use the sample s.d. $s = 24.5$ in place of the unknown population s.d. σ.

   b. Mark the location of $\bar{x} = 104$ on your graph and shade in the region corresponding to the *p*-value of a test of H_0: $\mu = 100$ vs. H_a: $\mu > 100$, where μ is the mean Reading Readiness score that would be achieved if <u>all</u> first graders used the new experimental program.

   c. Based only on your sketch, does there appear to be statistically significant evidence that the new program is better than the existing one? Explain briefly.

   d. Determine the *p*-value using your calculator or approximate it using the *t* table, and see whether the improvement in mean score for the sample group is statistically significant at level $\alpha = 5\%$.

10. The Food and Drug Administration (FDA) has a maximum upper limit of 12 mg for the mean nicotine content in cigarettes. An FDA evaluator took a random sample of 10 cigarettes from a new brand and found the mean nicotine content of the sample to be 13 mg, with a standard deviation of 2 mg.

    a. Based on this sample, should the FDA conclude that the average nicotine level of the new brand higher than the maximum allowable upper limit of 12 mg? Perform the appropriate test of significance—set up hypotheses, determine the *p*-value, and draw an appropriate conclusion.

    b. Do the results of part (a) give convincing evidence that the mean nicotine level of the new cigarette brand is unacceptable, that is, higher than 12 mg?

    c. Suppose the FDA had taken a larger sample, and still found the mean nicotine content of the sample to be 13 mg. How would the *p*-value compare to the one you found in part (a)? How would the FDA's conclusion be likely to change as a result? (Hint: Think of what the sampling distribution would look like for a larger sample compared to what it would look like for the actual sample size used.)

11. A new treatment for depression is under investigation. Patients are given a psychological test before and after treatment in order to determine the effectiveness of the treatment. Researchers hope to find a difference in evaluation scores before and after treatment. A 90% confidence interval for the population average difference in evaluation scores (before-after) is given by (0.25, 3.75), which was computed based on a random sample of 22 patients.

    a. What is the population parameter of interest?

    b. Based on the confidence interval, what is the value of the sample statistic (point estimate)?

    c. Decide if the following statement is true or false: *We are 90% confident that the mean difference in sample evaluation scores (before-after) of patients undergoing this treatment lies in the interval (0.25, 3.75).*

    True ***False***

    d. State the hypotheses in symbols:

    H_0: H_a:

    e. Based on the confidence interval, would 0 be a reasonable value of the population parameter? Why or why not?

    f. Would a 95% confidence interval be wider or narrower than the 90% confidence interval?

    g. Assume the *p*-value for a two-sided test was 0.09, with a *t* statistic of 1.78. Determine if the following statement is a correct interpretation of the *p*-value.

    If we repeated this experiment many times (*taking random samples of size 22 each time*)*, we would expect to obtain a t-test statistic value of 1.78 or larger or -1.78 or smaller in only 9% of the repetitions, if in fact there is no difference, on average, in the population evaluation scores before and after psychological treatment.*

    True ***False***

12. Which of the following research situations would most probably involve independent samples, and which would likely involve paired data?

    a. Twenty-five people have their cholesterol measured before eating a Big Mac and again after eating it. On average, does eating a Big Mac increase cholesterol?

    b. What is the difference in average salaries for high school graduates and college graduates?

    c. In fifty married couples, the husband and wife each separately take the same test of marital satisfaction. Is there a difference, on average, between the scores of husbands and wives?

    d. Do college grade point averages differ for athletes in major sports (e.g., football) compared with those in minor sports (e.g., swimming)?

13. Automobile insurance appraisers examine cars that have been in accidents in order to assess the cost of repairs. An insurance executive is concerned that some appraisers may tend to make higher assessments than others. In one experiment**,** 10 cars that had recently been in accidents were shown to two appraisers, which we will call "A1" and "A2". Each appraiser assessed the estimated repair cost for each car.

    a. State appropriate hypotheses for comparing these two appraisers in both symbols and in words:

    H_0:

    H_a:

    b. The data and summary statistics are given below. Show how the value 54 was obtained.

Car:	1	2	3	4	5	6	7	8	9	10
A1	$2150	$860	$1140	$1510	$1390	$1250	$940	$1710	$1020	$1190
A2	$1900	$880	$1100	$1420	$1430	$1150	$910	$1580	$980	$1270

Difference	Mean of Diff.	S.D. of Diff.	DF	T-Stat	P-value	95% Lower Limit	95% Upper Limit
A1 - A2	54	94.77	9	1.80	0.1051	-13.8	121.8

c. Is there evidence that the two appraisers give significantly different assessments on average? Use $\alpha = 0.05$. In your conclusion use both the given *p*-value and confidence interval.

14. An herbal medicine is tested on 16 randomly selected patients with sleep disorders. Each patient's amount of sleep (in hours) is measured for one night with the herbal medicine and for one night without the herbal medicine. Researchers claim that the patients' condition will improve using the herbal medicine. At the 1% significance level, does the data support the researchers' claim?

Patient	Without	With
1	1.8	3.0
2	2.0	3.6
3	3.4	4.0
4	3.5	4.4
5	3.7	4.5
6	3.8	5.2
7	3.9	5.5
8	3.9	5.7
9	4.0	6.2
10	4.9	6.3
11	5.1	6.6
12	5.2	7.8
13	5.0	7.2
14	4.5	6.5
15	4.2	5.6
16	4.7	5.9

Mean (Without–With) = –1.525

S.d. (Without–With) = 0.542

15. Do women have a better vocabulary than men? Ten married couples are chosen at random and a vocabulary test is given to each of the husbands and each of the wives. Here are the scores:

Men	68	76	74	91	84	70	77	82	92	69
Women	73	81	73	89	88	79	81	88	92	80

(cont. on next page)

Compute the mean and standard deviation for the differences, defined by Difference = Women – Men. Then perform a test of significance to help answer the question above. Give the null and alternative hypotheses, calculate the test statistic, sketch a *t*-curve and mark the location of the test statistic, and make a decision at the 10% significance level. Write your conclusion in context.

16. A study of asthmatics measured the peak expiratory flow rate (basically, a person's maximum ability to exhale) before and after a walk on a cold winter's day for a random sample of nine asthmatics. Use the data to test whether there is a difference between the peak expiratory flow rate before vs. after a walk on a cold winter's day for asthmatics. Give null and alternative hypotheses, compute the test statistic (hint: there is more information above than you need), determine the approximate *p*-value, and assess the strength of the evidence for the claim that there is a difference between peak expiratory flow rates before and after the walk.

Subject	Before	After	Difference
1	312	300	12
2	242	201	41
3	340	232	108
4	388	312	76
5	296	220	76
6	254	256	-2
7	391	328	63
8	402	330	72
9	290	231	59
mean	**323.89**	**267.78**	**56.11**
s.d.	**59.83**	**50.01**	**34.17**

Inference for Comparing Two Means

What you need to know:

- **The difference between independent samples and matched pairs designs**
- **How to calculate and interpret a confidence interval for comparing two means**
- **How to carry out a test of significance for comparing two means, from stating the hypotheses to the conclusion**

---

1. Determine whether the following are independent samples or matched pairs situations:

   a. The effectiveness of Zantac for treating heartburn is tested by measuring gastric acid secretion in a group of patients treated with Zantac and another group of patients given a placebo.

   independent samples matched pairs

   b. The effectiveness of Zantac for treating heartburn is tested by measuring gastric acid secretion in patients before and after the drug is administered.

   independent samples matched pairs

   c. Comparing vitamin content of bread immediately after baking versus 3 days later (the same loaves are used on day one and 3 days later).

   independent samples matched pairs

   d. The effectiveness of a marketing campaign for a certain product is tested by comparing sales in one metropolitan area where the campaign was used to sales in another metropolitan area where the campaign was not used.

   independent samples matched pairs

   e. The effectiveness of a tartar control toothpaste is tested in an experiment involving several families; one child in each family used the regular toothpaste, and another child in that same family used the tartar control toothpaste.

   independent samples matched pairs

2. A study published in the *Journal of the American Academy of Business* examined whether guests' perception of the quality of service at five-star hotels in Jamaica differed by gender. It was suspected that in general, female guests are not as satisfied with the quality of service as male guests. Hotel guests were randomly selected from the lobby and restaurant areas and asked to rate 10 service-related items (e.g. "the personal attention you received from our employees"). Each item was rated on a five-point scale (from 1 = "much worse than I expected" to 5 = " much better than I expected") and the sum of the item ratings for each guest was determined. A summary of the resulting guest scores is provided in the following table:

Gender	Sample size	Mean score	Standard deviation
Males	127	39.08	6.73
Females	114	37.79	6.94

Carry out all the steps of a test of significance: State the hypothesis, check conditions for inference, find the test statistic, and determine the *p*-value. Write your conclusion in context.

3. Researchers randomly assigned participants either a tall, thin "highball" glass or a short, wide "tumbler," each of which held 355 ml. Participants were asked to pour a shot (1.5 oz = 44.3 ml) of liquor into their glass. Did the shape of glass make a difference in how much liquor they poured?

	Highball	Tumbler
no. of participants	99	99
mean	42.2 ml	47.9 ml
s.d.	16.2 ml	17.9 ml

Compute a 90% confidence interval for the difference in the mean amount poured. Then use this interval to address the question above.

Additional Practice

1. For each of the following instances say whether the study uses an independent samples or a matched pairs design.

   a. A survey is conducted of teens from inner city schools to estimate the proportion who have tried drugs. A similar survey is conducted of teens from suburban schools.

   b. A psychologist measures the response times of subjects under two stimuli; each subject is observed under both of the stimuli, in a random order.

   c. Lung cancer patients admitted in a hospital over a 12 month period are each matched with a non-cancer patient by age, sex, and race. To determine how much of a risk factor smoking is for lung cancer, it is noted for each patient whether he or she is a smoker.

   d. An advertising agency has come up with two different TV commercials for a household detergent. To determine which one is more effective, a test is conducted in which a sample of 100 adults is randomly divided into two groups. Each group is shown a different commercial, and their reactions are then assessed by the agency.

2. A mathematics test was given to a random sample of 1000 17-year-old students in 2004, and again to another random sample of 1000 students in 2014. The mean score in 2004 was 300.4. In 2014 it was 305.4. Is this five point change evidence of a real difference in means for the two years, or is it plausibly just due to chance variation?

   a. State the null and alternative hypotheses.

   b. The sample s.d.s were 34.9 for 2004 and 30.1 for 2014. Compute the test statistic.

c. Sketch the relevant t distribution and mark the location of the test statistic. Then shade in the area representing the p-value, and estimate its value from the graph.

d. Compute the p-value using technology or estimate it from the t table.

e. What can you conclude? Is there a statistically significant difference between average math scores in 2004 vs. 2014? Explain.

f. Compute the 95% confidence interval for the difference between the population mean scores. Write clear sentence interpreting the result:

"We are 95% confident that _____

_____ .

3. A study was carried out to investigate the effectiveness of a treatment. 1000 subjects participated in the study, with 500 being randomly assigned to the "treatment group" and the other 500 to the "control (or placebo) group". A statistically significant difference was reported between the responses of the two groups ($p < .005$). Thus, we can conclude that

   a. there is a large difference between the effects of the treatment and the placebo.
   b. there is strong evidence that the treatment is very effective.
   c. there is strong evidence that there is some difference in effect between the treatment and the placebo.
   d. there is little evidence that the treatment has any effect.
   e. there is evidence of a strong treatment effect.

4. The Excellent Drug Company claims its aspirin tablets will relieve headaches faster than those made by the Simple Drug Company. To determine whether Excellent's claim is valid, an aspirin is given to each of 30 randomly selected persons suffering from headaches and the number of minutes required for each to recover from their headache is recorded. 15 people were given aspirins made by the Excellent Drug Company and the other 15 were given aspirins made by Simple Drug Company.

A 5% significance level test will be performed to determine whether Excellent's (E) aspirin cures headaches significantly faster than Simple's (S) aspirin.

a. How should the choice be made of which people get the Excellent aspirins and which get the Simple aspirins?

b. The appropriate set of hypotheses to be tested is:

(a) H_0: $\mu_E - \mu_S = 0$ H_a: $\mu_E - \mu_S > 0$
(b) H_0: $\mu_E - \mu_S = 0$ H_a: $\mu_E - \mu_S \neq 0$
(c) H_0: $\mu_E - \mu_S = 0$ H_a: $\mu_E - \mu_S < 0$
(d) H_0: $\mu_E - \mu_S < 0$ H_a: $\mu_E - \mu_S = 0$
(e) H_0: $\mu_E - \mu_S > 0$ H_a: $\mu_E - \mu_S = 0$

Hint: Think carefully about what μ_E and μ_S represent.

5. The superintendent of the local public school district is concerned that boys in grades 1 through 5 are less mathematically competent than girls in grades 1 through 5. If this is the case, changes might need to be made in the mathematics curriculum delivered in these grades. To determine if such changes need to be made, the superintendent takes a random sample of boys and a random sample of girls in grades 1 through 5 from her district. She then compares the mean score for boys and the mean score for girls on a standardized math competency exam. Here are the results:

Gender	N	Mean	Standard deviation
Male	60	84.05	12.96
Female	34	90.21	14.42

Does the data above give good evidence that, if all boys and all girls in the district took the exam, the mean score for boys would be lower than the mean score for girls? Perform a test of significance to find out:

a. Give the null and alternative hypotheses using symbols.

b. Your hypotheses in part (a) should have used two parameters. Give the meaning of each.

c. Check the conditions for inference.

d. Compute the test statistic.

e. Determine the *p*-value.

f. Based on the *p*-value, how strong is the evidence that population of boys would achieve a lower mean score than the population of girls?

Little or no evidence Weak evidence Good evidence Very strong evidence

g. What conclusion would you make at significance level 5%?

h. What conclusion would you make at significance level 1%?

i. Find and interpret the 99% confidence interval for the difference between the mean score that all boys would achieve and the mean score that all girls would achieve.

6. Some of the situations below involve means, and some involve proportions. Write the null and alternative hypotheses in symbols that you would use for a test of each, and determine whether the right alternative is one-sided or two-sided.

   a. Is a coin fair (equally likely to land heads or tails when flipped)?

   b. Only 34% of people who try to quit smoking succeed. A company claims that their chewing gum can help people quit.

   c. In the 1950s only 40% of high school graduates went on to college. Has the percentage changed?

   d. A large city's DMV claimed that 80% of candidates pass driving tests, but a newspaper reporter claims that the actual passage rate is lower than this.

   e. The FDA claims that the mean mercury level in halibut is 0.26 parts per million. An environmental organization believes that the level is higher than this, and that people should avoid eating it.

   f. The average time it takes for a person to experience pain relief from aspirin is 25 minutes. A new ingredient is added to help speed up relief, and an experiment is conducted to verify that the new product is better.

7. What would a Type I error be for the situation described in the following cases?

   Problem #6a:

   Problem #6c:

8. What would a Type II error be in the following cases?

   Problem #6a:

   Problem #6e:

9. What would the chance of a Type I error be if a 5% significance level is used for a test of significance?

10. Consider again Problem #6b, and make the appropriate conclusions based on the *p*-value (shown as *p*) in each of the scenarios below:

(i) In random sample of 100 smokers, 38% quit. A test of significance of this data gave $p = .20$.

There is _____ evidence that the chewing gum increases the proportion of smokers who quit.

(ii) In random sample of 100 smokers, 41% quit. A test of significance of this data gave $p = .07$.

There is _____ evidence that the chewing gum increases the proportion of smokers who quit.

(iii) In random sample of 400 smokers, 40% quit. A test of significance of this data gave $p = .006$.

There is _____ evidence that the chewing gum increases the proportion of smokers who quit.

End of *Inference About Quantitative Variables.* Check your knowledge:

1. Explain the difference between the meaning of the symbols $\bar{x}$ and μ.
2. What are the conditions for the shape of the sampling distribution of a sample mean to be approximately normal?
3. For a SRS of size *n*, what happens to the shape of the sampling distribution of $\bar{x}$ as the sample size *n* increases?
4. For a SRS of size *n*, what is the mean of the sampling distribution of $\bar{x}$?
5. For a SRS of size *n*, what is the standard deviation of the sampling distribution of $\bar{x}$?
6. What happens to the standard deviation of the sampling distribution of $\bar{x}$ as the sample size *n* increases?
7. **True or False:** If a random sample of size 100 is taken from a population with mean 25 and s.d. 5, then the chance that $\bar{x}$will be between 15 and 35 is 95%.
8. Know how to find a confidence interval for the mean (conditions, formula, how to interpret it)
9. Know how to find the margin of error of a confidence interval for a mean.
10. Know how to carry out a test of significance for the mean (check conditions for inference, state the hypotheses, find the test statistic, find the *p*-value, give the conclusion in context: strength of the evidence against H_0, decision to reject or not reject if a significance level α is given).
11. **True or False:** A confidence interval is a set of plausible values for the parameter.
12. **True or False:** A confidence interval is centered at the population parameter.
13. Know how to find the point estimate and the margin of error if the confidence interval is given.
14. Know how the confidence interval behaves as we increase the confidence level, or change the sample size.
15. Name the two kinds of hypotheses involved in tests of significance.
16. **True or False:** The null hypothesis will probably be rejected if the point estimate is in the tail of the sampling distribution and its value supports the alternative.
17. **True or False:** The smaller the *p*-value, the stronger the evidence against the null hypothesis.
18. Be able to give the meaning of the *p-value* in the context of a problem.
19. Know the degree of evidence against H_0 that specific *p-values* represent.
20. Explain the term *significance level*.
21. What does the term *statistically significant* mean?
22. **True or False:** The null hypothesis will be rejected if the *p-value* is larger than α.
23. In a test of significance, what is a Type I error?
24. In a test of significance, what is a Type II error?
25. What is the *power* of a test of significance against a specific alternative value of the parameter?

Inference for Categorical Relationships

What you need to know:

- **How to assess a relationship between two categorical variables numerically and graphically**
- **How to carry out a chi-square test of independence/homogeneity**

---

1. A clinical trial compared the effectiveness of two drugs, Lithium and Imipramine, in preventing a recurrence of depression among patients who were hospitalized with depression. Here is a two-way table of the results:

	Lithium	Imipramine
Recurrence	18	8
No Recurrence	9	21

The columns of the following table show the conditional distributions of the Recurrence variable for the two drugs:

	Lithium	Imipramine
Recurrence	66.7%	27.6%
No Recurrence	33.3%	72.4%
Total	100.0%	100.0%

The data seem to indicate that Imipramine is far more effective than Lithium in preventing a recurrence of depression, but the number of patients in the study is small. Could the difference in recurrence rates in this study be simply due to chance? Or is this enough data to convince us that Imipramine is actually better than Lithium? Use a chi-square test of significance to help answer this question, following the steps below:

a. State the null and alternative hypotheses in words, in the context of the study.

b. Compute the expected counts for each cell under H_0 and enter them in the table below:

	Recurrence	No Recurrence
Imipramine		
Lithium		

c. Check that the row and column sums of the expected counts are the same as for the original data table. Also check that the expected counts are large enough to meet the requirement for the chi-square test.

d. Determine the value of the chi-square test statistic, and give the degrees of freedom.

e. Compute the *p*-value using technology or estimate it from the chi-square table, and use it to make an appropriate conclusion regarding the original question.

f. Is the result statistically significant at the 5% significance level? At the 1% level? At the 0.1% level? Justify your answer.

2. a. 100 men and 100 women were asked whether they believed in ghosts. Out of all 200 people interviewed, 45% said they believed in ghosts. Fill in the data for the two-way table below in the way that most strongly indicates that whether a respondent believed in ghosts was completely independent of their gender:

	Female	Male
Believe in ghosts		
Don't believe in ghosts		
Total	100	100

b. Now suppose that 100 men and 200 women were asked whether they believed in ghosts. Out of all 300 people interviewed, 45% said they believed in ghosts. Fill in the data for the two-way table below in the way that most strongly indicates that whether a respondent believed in ghosts was completely independent of their gender:

	Female	Male
Believe in ghosts		
Don't believe in ghosts		
Total	200	100

c. Compute the female and male conditional distributions and make a segmented bar graph to compare them.

d. Finally, suppose the actual data are as follows:

	Female	Male
Believe in ghosts	96	39
Don't believe in ghosts	104	61
Total	200	100

Compute the female and male conditional distributions and make a segmented bar graph to compare them. Based on this data, does there seem to be a relationship between gender and the likelihood a person believes in ghosts?

e. Compute the table of expected counts for the chi-square test. What do you notice?

f. Are the expected counts fairly close to the actual counts in (d), or are they quite different?

g. Based on your answer to part (f), do think the data in (d) will provide statistically significant evidence of a relationship between gender and the likelihood a person believes in ghosts, or do you think the discrepancy between the actual counts and the expected counts under the assumption of independence of these two factors could be due simply to chance? Explain.

h. Carry out the chi-square test and find out the real answer to the question in part (g). Did you guess right?

Additional Practice

1. An on-line music service company wants to know if customer age is a factor in whether a person decides to subscribe. The company has gathered a random sample of 1000 people from the population, and asked each person whether he or she would be likely to subscribe to the service. They also asked which age group the person falls into: under 18 years of age, 18 - 34 years of age, and 35 years or older.

   a. Explain why the company could use a chi-square test to answer its question.

   b. The answer choices for each person interviewed were “Yes”, “No” and “Maybe”. How many degrees of freedom would the chi-square test have?

2. There are many "indicators" that investors use to predict the stock market. One of these is the "January indicator", which says that if the market is up in January, then it will be up over the rest of the year. If the market is down in January, then it will be down over the rest of the year. Results for a 72 year period (as measured by the Dow Jones average) are shown below:

	January	
February-December	Up	Down
Up	33	13
Down	13	13

a. State the null hypothesis in words for chi-square test of independence.

b. Does the alternative hypothesis of the chi-square test represent the investors' belief?

c. Perform the test and make an appropriate conclusion.

3. A marketing firm is interested in studying consumer behavior would like to know whether the income level of consumers is related to their choice of brand of detergent. A stratified random sampling procedure sampled 600 consumers; the results are shown below:

	Brand1	Brand2	Brand3	Brand4	Total
Income:					
Lower	25	15	55	65	160
Middle	30	25	35	30	120
Upper Middle	50	55	20	22	147
Upper	60	80	15	18	173
Total	165	175	125	135	600

a. For the purposes of a chi-square test, how many rows and how many columns of data does this table have? (Be careful here!)

b. Carry out the chi-square test and make an appropriate conclusion. You are welcome to use technology.