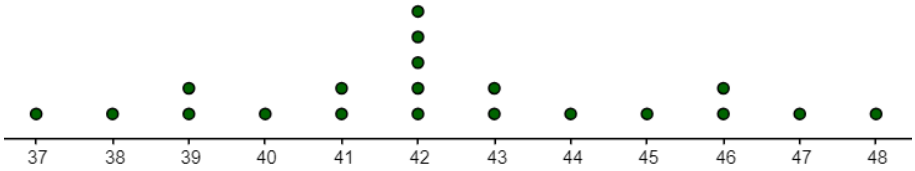
Mostly Harmless Elementary Statistics

Student Solutions Manual

Chapter 1 Exercises

1. The dotplot shows the height of some 5-year-old children measured in inches. Use the distribution of heights to find the approximate answer to the question, "How many inches tall are 5-year-olds?"



42

> "'Alright,' he said, 'but where do we start? How should I know?
> They say the Ultimate Answer or whatever is Forty-two, how am I supposed to know what the question is? It could be anything. I mean, what's six times seven?'
> Zaphod looked at him hard for a moment. Then his eyes blazed with excitement.
> 'Forty-two!' he cried."
> (Adams, 2002)

3. What are statistics? Answer c)
   a) A question with a variety of answers.
   b) A way to measure the entire population.
   c) The science of collecting, organizing, analyzing and interpreting data.
   d) A question from a survey.

5. Which of the following are statistical questions? Select all that apply. Answer b, c, e
   a) How old are you?
   b) What is the weight of a mouse?
   c) How tall are all 3-year-olds?
   d) How tall are you?
   e) What is the average blood pressure of adult men?

   The questions "How old are you?" and "How tall are you?" are not statistical questions because the answers to these questions will not vary. There is ONE correct answer, as opposed to many answers that can then be used to find an average, proportion, range, etc.

Questions 6-9:

Some helpful definitions:
Individual is a person or object that you are interested in finding out information about.
Variable is the measurement or observation of the individual.

Population is the total set of all the observations that are the subject of a study.
Sample a subset from the population.
Parameter a number calculated from the population.
Statistic a number calculated from the sample.

7. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1,000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.
   a) Percentage of all videos on YouTube that are cat videos. Population Parameter
   b) A video in your sample. Observation
   c) 2% Sample Statistic
   d) Whether a video is a cat video. Variable

9. The 2010 General Social Survey asked the question, "After an average workday, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an individual, a variable, a sample statistic, or a population parameter.
   a) Average number of hours all Americans spend relaxing after an average workday. Population Parameter
   b) 1.65 Sample Statistic
   c) An American in the sample. Individual
   d) Number of hours spent relaxing after an average workday. Variable

11. In a study, the sample is chosen by separating all cars by size, and selecting 10 of each size grouping. What is the sampling method? Stratified

13. In a study, the sample is chosen by asking people on the street. What is the sampling method? Convenience

15. In a study, the sample is chosen by surveying every 3rd driver coming through a tollbooth. What is the sampling method? Systematic

17. State whether each study is observational or experimental.
   a) You want to determine if cinnamon reduces a person's insulin sensitivity. You give patients who are insulin sensitive a certain amount of cinnamon and then measure their glucose levels. Experimental
   b) A researcher wants to evaluate whether countries with lower fertility rates have a higher life expectancy. They collect the fertility rates and the life expectancies of countries around the world. Observational
   c) A researcher wants to determine if diet and exercise together helps people lose weight over just exercising. The researcher solicits volunteers to be part of the study, and then randomly assigns the volunteers to be in the diet and exercise group or the exercise only group. Experimental
   d) You collect the weights of tagged fish in a tank. You then put an extra protein fish food in water for the fish and then measure their weight a month later. Experimental

An **observational study** is when the investigator collects data by observing, measuring, counting, watching, or asking questions. The investigator does not change anything.

An **experimental study** is when the investigator changes a variable or imposes a treatment to determine its effect.

19. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group, the probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating did not vary by age for boys, it decreased with age for girls. [Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children." In: Journal of Economic Psychology 32.1 (2011), pp. 73-78.] In this study, identify the variables. Select all that apply. Answers a, b, d, f, g
    a) Age
    b) Sex
    c) Paper Sheet
    d) Cheated or Not
    e) Reward for White Side of Coin
    f) White or Black Side of Coin
    g) Only Child or Not

A **variable** is the measurement or observation of an individual.

21. Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.

    a) Temperature in degrees Kelvin. Ratio
    b) Eye color. Nominal
    c) Year in school (freshman, sophomore, junior, senior). Ordinal
    d) The weight of a hummingbird. Ratio
    e) The height of a building. Ratio
    f) The amount of iron in a person's blood. Ratio
    g) A person's gender. Nominal

Nominal data is categorical data that has no order or rank, for example the color of your car, ethnicity, race, or gender.

Ordinal data is categorical data that has a natural order to it for example, year in school (freshman, sophomore, junior, senior), a letter grade (A, B, C, D, F), the size of a soft drink (small, medium, large) or **Likert scales**.

Interval data is numeric where there is a known difference between values, but zero does not mean "nothing." Interval data is ordinal, but you can now subtract one value from another and that subtraction makes sense. You can do arithmetic on this data. For example, Fahrenheit temperature, 0° is cold but it does not mean that no temperature exists. Time, dates and IQ scores are other examples.

Ratio data is numeric data that has a true zero, meaning when the variable is zero nothing is there. Most measurement data are ratio data. Some examples are height, weight, age, distance, or time running a race.

23. State which type of variable each is, qualitative or quantitative?

    a) The height of a giraffe. Quantitative
    b) A person's race. Qualitative
    c) Hair color. Qualitative
    d) A person's ethnicity. Qualitative
    e) Year in school (freshman, sophomore, junior, senior). Qualitative

    A **qualitative variable** is a word or name that describes a characteristic (quality) of the individual.

    A **quantitative or numerical variable** is a number (quantity), something that can be counted or measured from the individual.

25. State whether the variable is discrete or continuous.

    a) Temperature in degrees Celsius. Continuous
    b) The number of cars for sale at a car dealership. Discrete
    c) The time it takes to run a marathon. Continuous
    d) The amount of mercury in a tuna fish. Continuous
    e) The weight of a hummingbird. Continuous

    Discrete data can only take on particular values like integers. Discrete data are usually things you count.

    Continuous data can take on any value. Continuous data are usually things you measure.

27. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
    a) The quality control officer at a manufacturing plant needs to determine what percentage of items in a batch are defective. The officer chooses every 15^{th} batch off the line and counts the number of defective items in each chosen batch. Systematic
    b) The local grocery store lets you survey customers during lunch hour on their preference for a new bottle design for laundry detergent. Convenience
    c) Put all names in a hat and draw a certain number of names out. Random

d) The researcher randomly selects 5 hospitals in the U.S. then measures the cholesterol level of all the heart attack patients in each of those hospitals. Cluster

A **Simple Random Sample** is a sample collected from the population so that every sample of the same size has equal probability of being selected.

A **Systematic Sample** is a sample collected by organizing the population into a list, randomly selecting a starting point, and sampling every n^{th} value until the sample size is reached.

A **Stratified Sample** is a sample collected by first grouping the population into groups called strata and then selecting a random sample from each stratum.

A **Cluster Sample** is a sample collected by first grouping the population into groups called clusters and then sampling all individuals in one or more clusters that have been randomly selected.

A **Convenience Sample** is a sample collected at the researcher's convenience.

29. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
    a) In a research study, a list of all registered voters in a city is obtained, and a random sample of 500 voters is selected. Random
    b) A company wants to survey its employees about job satisfaction. They randomly select every 10th employee from the employee list to participate in the survey. Systematic
    c) A researcher wants to study the purchasing behavior of different age groups in a city. They divide the city into four age groups (18-25, 26-35, 36-45, 46 and above) and randomly select participants from each group. Stratified
    d) A university wants to conduct a survey among students. The university divides the campus into several sections and randomly selects a few sections. All students within the selected sections are surveyed. Cluster

31. Which type of sampling method is used for each scenario, Random, Systematic, Stratified, Cluster or Convenience?
    a) A school district wants to assess the performance of students in different grades. They randomly select a few schools from the district and test all students within those selected schools. Cluster
    b) A company wants to conduct a customer satisfaction survey. They select every 4^{th} customer who makes a purchase at the company's online store to participate in the survey. Systematic
    c) A survey is conducted at a shopping center, and shoppers passing by are asked to participate. The surveyors approach individuals who are readily available and willing to participate. Convenience
    d) A survey is conducted to determine the preferences of college students regarding online learning. The research team randomly selects 300 students from a list of all enrolled students to participate in the survey. Random

33. Which of the following best describes statistical ethics?
Answer: b) Protecting data privacy and confidentiality.
a) Ensuring the correct statistical analysis is used.
b) Protecting data privacy and confidentiality.
c) Using statistical methods to mislead others.
d) Ignoring ethical considerations in data collection.

35. Statistical results can have significant consequences because they:
Answer: b) Can impact decision-making in various fields
a) Are always accurate and reliable.
b) Can impact decision-making in various fields.
c) Have no practical implications in real-world applications.
d) Are unrelated to ethical considerations.

37. Which of the following is a potential negative consequence of conducting statistical analysis unethically?
Answer: b) Harmful effects on individuals or communities
a) Reliable and trustworthy results.
b) Harmful effects on individuals or communities.
c) Accurate interpretations of data.
d) Improved decision-making processes.

39. Ethical considerations in statistical analysis include:
Answer: a) Protecting data privacy and confidentiality
a) Protecting data privacy and confidentiality.
b) Ignoring the potential for biases.
c) Promoting discrimination and systemic biases.
d) Manipulating data to achieve desired results.

41. What is the purpose of Institutional Review Boards (IRBs)?
Answer: b) To protect the privacy and confidentiality of research participants
a) To oversee statistical analysis in research studies.
b) To protect the privacy and confidentiality of research participants.
c) To manipulate data to support specific conclusions.
d) To ignore ethical considerations in research studies.

Chapter 2 Exercises

1. Which types of graphs are used for quantitative data? Select all that apply. Answers a, c, d
   a) Ogive
   b) Pie Chart
   c) Histogram
   d) Stem-and-Leaf Plot
   e) Bar Graph

   Both **pie charts** and **bar graphs** are used to represent data that has been separated into categories. When data are separated into categories, it is qualitative data.

   Ogives, **histograms**, and **stem-and-leaf plots** are used to represent data that take on numeric values. When data are made up of numeric values, it is quantitative data.

3. The bars for a histogram should always touch, true or false? True

5. An instructor had the following grades recorded for an exam.

96	66	65	82	85
82	87	76	80	85
83	69	79	70	83
63	81	94	71	83
99	75	73	83	86

   a) Create a stem-and-leaf plot.

6	3 5 6 9
7	0 1 3 5 6 9
8	0 1 2 2 3 3 3 3 5 5 6 7
9	4 6 9

   b) Complete the following table.

Class	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
60 – 69	4	4	4/25 = 0.16	4/25 = 0.16
70 – 79	6	4+6 =10	6/25 = 0.24	10/25 = 0.4
80 – 89	12	10+12 = 22	12/25 = 0.48	22/25 = 0.88
90 – 99	3	22+3 = 25	3/25 = 0.12	25/25 = 1
Total	25		1	

   c) What should the relative frequencies always add up to? Answer = 1. Since relative frequencies can be converted into percentages, the total relative frequency should add up to 100%. Converting 100% back to decimal form gives us a total relative frequency of 1.
   d) What should the last value always be in the cumulative frequency column? The sample size.
   e) What is the frequency for students that were in the C range of 70-79? 6

f) What is the relative frequency for students that were in the C range of 70-79? 0.24
g) Which is the modal class? 80-89 is the most frequent
h) Which class has a relative frequency of 12%? 90-99
i) What is the cumulative frequency for students that were in the B range of 80-89? 0.88
j) Which class has a cumulative relative frequency of 40%? 70-79

7. The following table is from a sample of five hundred homes in Oregon asked the primary source of heating their residential home.

Type of Heat	Percent
Electricity	33
Heating Oil	4
Natural Gas	50
Firewood	8
Other	5

a) How many of the households heat their home with firewood? 40 households. The table above is giving a percentage for each household, 8% using Firewood. 8% of the 500 homes in the sample gives the following: $0.08 \cdot 500 = 40$
b) What percent of households heat their home with natural gas? 50%. Read this directly from the table.

9. A sample of heights of 20 people in cm is recorded below. Make a stem-and-leaf plot.

Height (cm)					
167	201	170	185	175	162
182	186	172	173	188	154
185	178	177	184	178	165
169	171	185	178	175	176

15	4
16	2 5 7 9
17	0 1 2 3 5 5 6 7 8 8 8
18	2 4 5 5 5 6 8
19	
20	1

Note: The smallest given data value is 154 and the largest is 201. The stems for each of these are 15 and 20 respectively, being that the stem includes all digits in the data value except the right most digit. The stems on the plot need to include every whole number from 15 to 20 (including 19, as you see above, despite there being no data values with a stem of 19).

11. The following is a sample of 25 temperatures in the summer for Portland, OR in ºF.

78	88	93	73	86
85	95	76	89	80
92	83	81	91	74

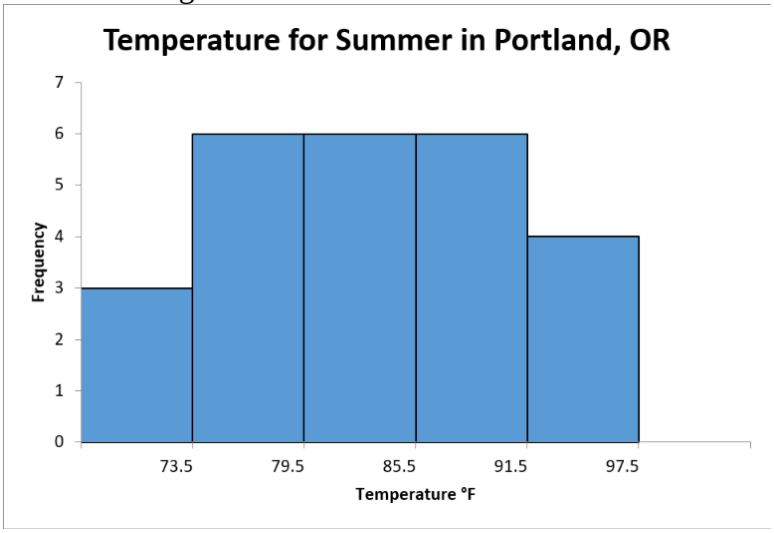
72	79	90	77	94
68	87	84	82	75

a) Create a frequency table with 5 classes.

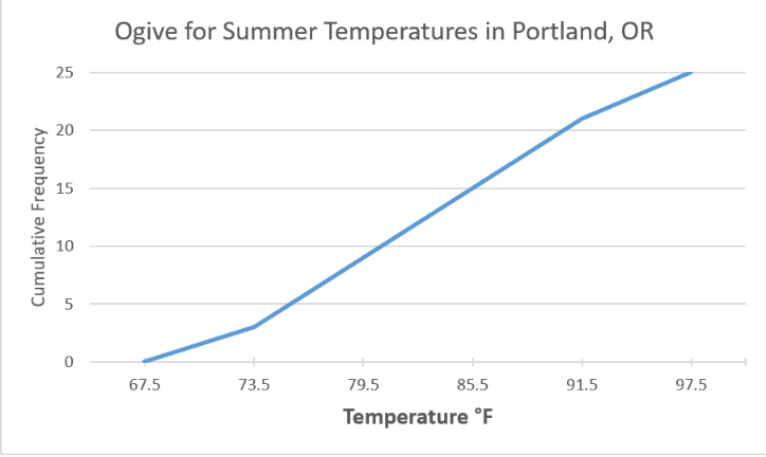
Find the range = max − min = 95 − 68 = 27. The question is asking for 5 classes so take range/5 = 27/5 = 5.4. Round up to the next integer and use a class length of 6 starting at the minimum value of 68. Then count the temperatures that fall into their respective classes and summarize in the following frequency table. Since we are creating an ogive in part c, also calculate the cumulative frequencies.

Class	Frequency	Cumulative Frequency
68 – 73	3	3
74 – 79	6	9
80 – 85	6	15
86 – 91	6	21
92 – 97	4	25

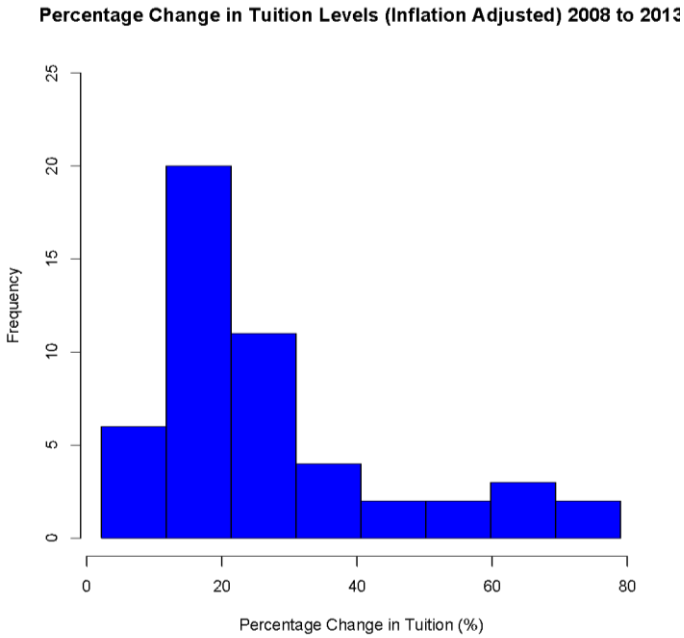
b) Create a histogram.



c) Create an ogive.



13. The following data represents the percent change in tuition levels at public, four-year colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013). Below is the frequency distribution and histogram.



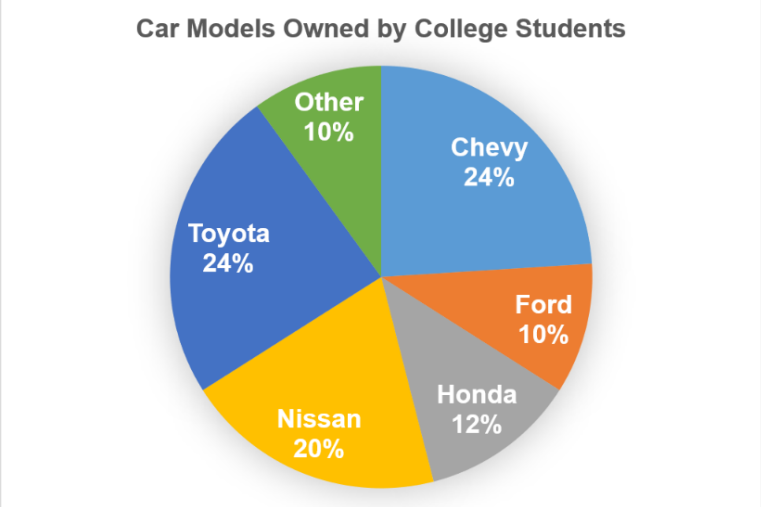
Class Limits	Class Midpoint	Frequency	Relative Frequency
2.2 – 11.7	6.95	6	0.12
11.8 – 21.3	16.55	20	0.40
21.4 – 30.9	26.15	11	0.22
31.0 – 40.5	35.75	4	0.08
40.6 – 50.1	45.35	2	0.04
50.2 – 59.7	54.95	2	0.04
59.8 – 69.3	64.55	3	0.06
69.4 – 78.9	74.15	2	0.04

a) How many colleges were sampled? 50. Find the total number sampled by adding up all the values in the frequency column of the frequency distribution.
6 + 20 + 11 + 4 + 2 + 2 + 3 + 2 = 50

b) What was the approximate value of the highest change in tuition? 78. We don't have the original data to know the exact highest value, but we can approximate the highest value from both the histogram and the classes in the frequency distribution. The right-most value on the x-axis of the histogram is just under 80, and the last class of the frequency distribution is 69.4-78.9. Given these two pieces of information, the answer of 78 is the best approximation for the highest change in tuition.

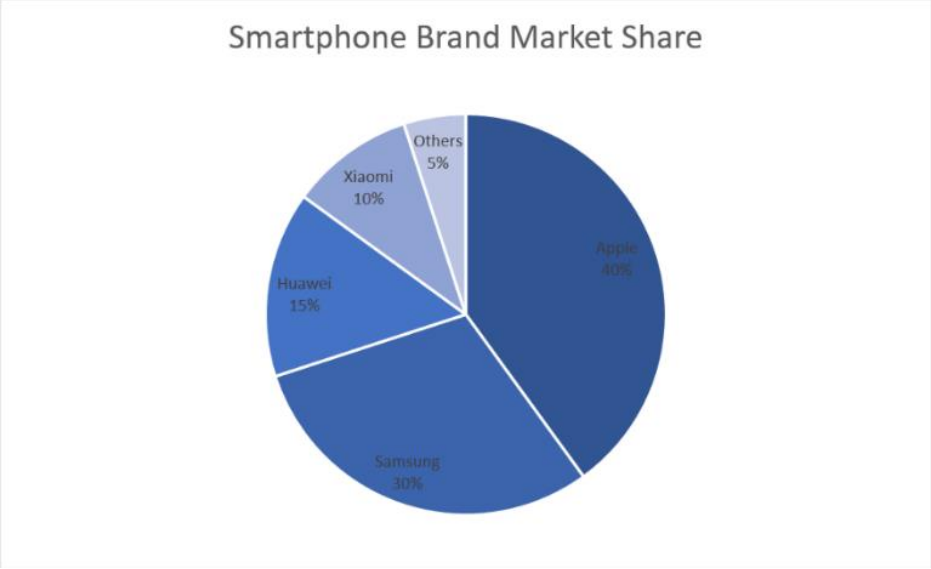
c) What was the approximate value of the most frequent change in tuition? $\frac{(11.8+21.3)}{2} = 16.55$
While we don't have the original data to know that answer for sure, we do know that the most frequent class is 11.8-21.3 by observing that the frequency for that class is 20 (the highest frequency on the table). To approximate the most frequent change in tuition, we simply take the class midpoint of that class.

15. The following graph represents a random sample of car models driven by college students. What percent of college students drove a Nissan? 20%. The sector of the pie graph that is labeled "Nissan" is given to be 20%.



17. The table below shows the percentage of market share held by each brand of smartphone. Create a pie graph for the market share of different smartphone brands.

Smartphone Brand	Market Share (%)
Apple	40%
Samsung	30%
Huawei	15%
Xiaomi	10%
Others	5%



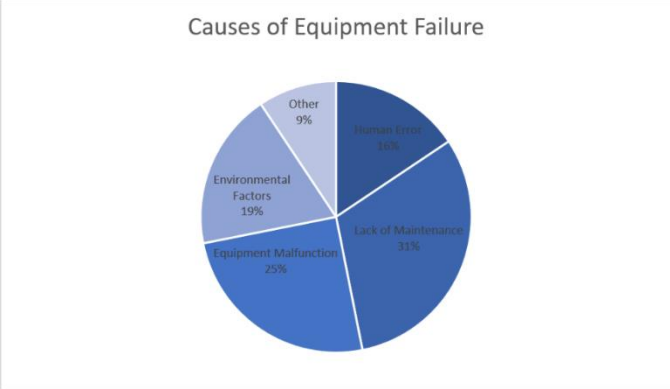
19. A manufacturing company wants to analyze the causes of equipment failures. The following table shows the number of failures caused by different factors in the past month.

Factor	Number of Failures
Human Error	5
Lack of Maintenance	10
Equipment Malfunction	8
Environmental Factors	6
Other	3

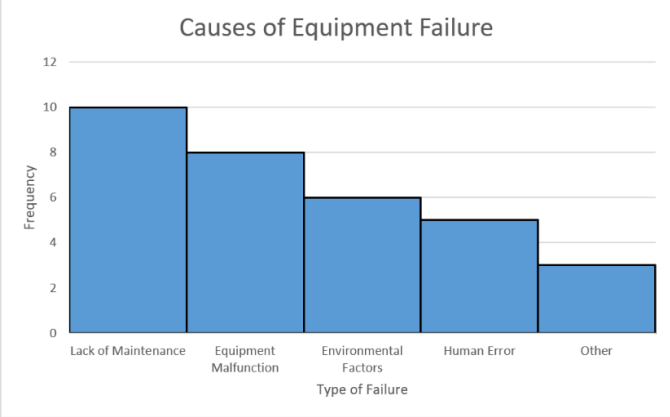
a) Create a relative frequency table.

Factor	Number of Failures	Relative Frequency
Human Error	5	5/32 = 0.15625
Lack of Maintenance	10	10/32 = 0.3125
Equipment Malfunction	8	8/32 = 0.25
Environmental Factors	6	6/32 = 0.1875
Other	3	3/32 = 0.09375
Total	32	1

b) Create a pie graph for the data.



c) Create a Pareto chart for the data.

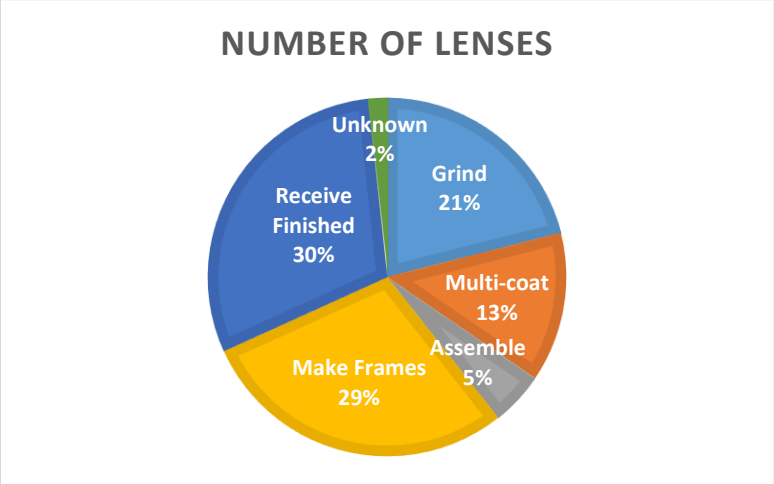


d) Identify the most significant factor contributing to the failures. Lack of Maintenance

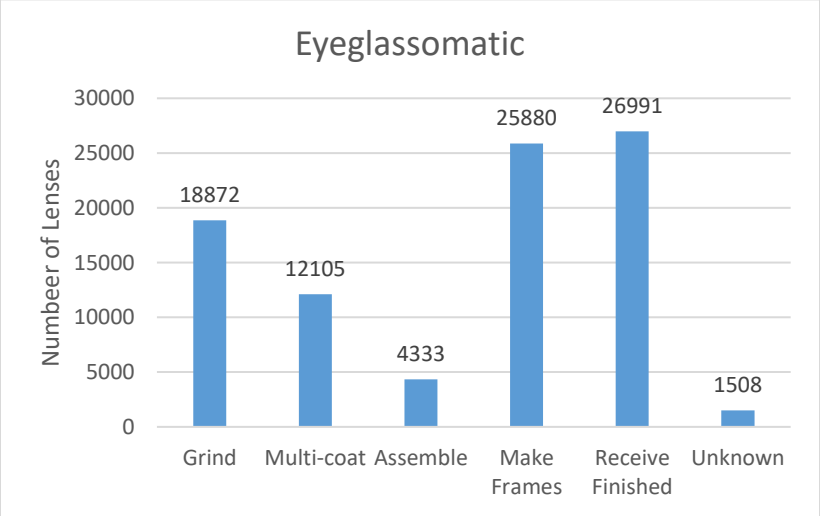
21. Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses for different activities is in table.

Activity	Grind	Multi-coat	Assemble	Make Frames	Receive Finished	Unknown
Number of lenses	18,872	12,105	4,333	25,880	26,991	1,508

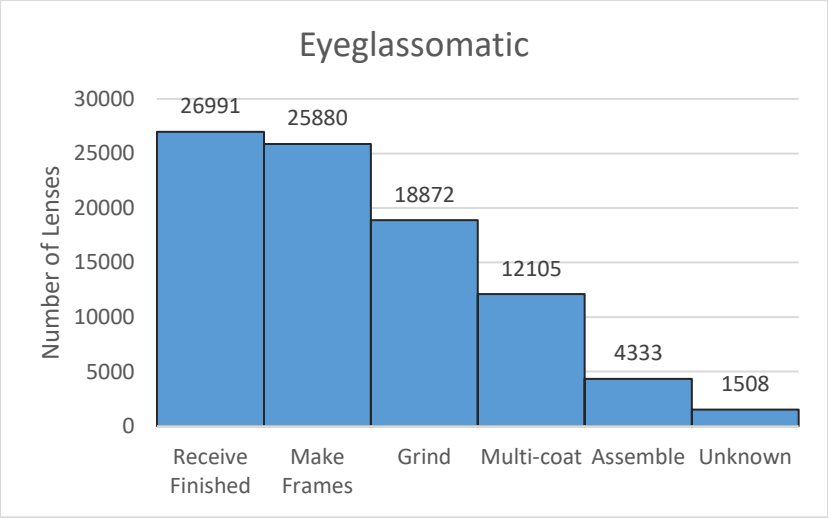
a) Make a pie chart.



b) Make a bar chart.



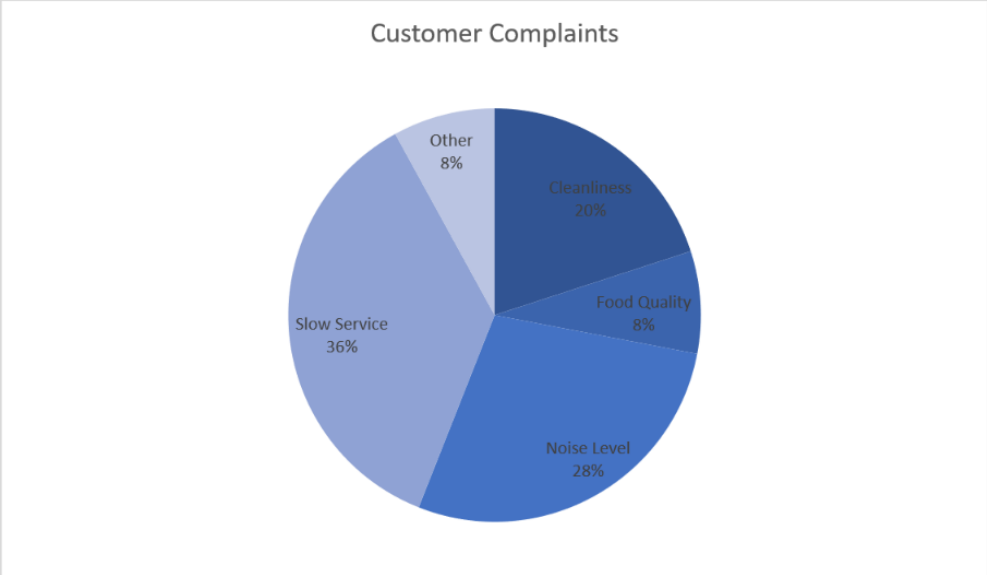
c) Make a Pareto chart.



23. A restaurant is interested in analyzing customer feedback to improve service quality. The following table shows the number of customer complaints received in the past month, categorized by the type of complaint.

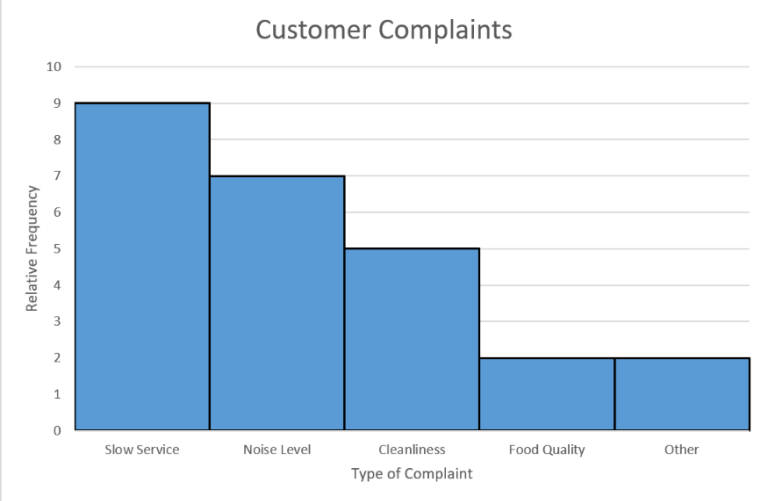
Complaint Category	Frequency
Cleanliness	5
Food Quality	2
Noise Level	7
Slow Service	9
Other	2

a) Create a pie graph for the data.

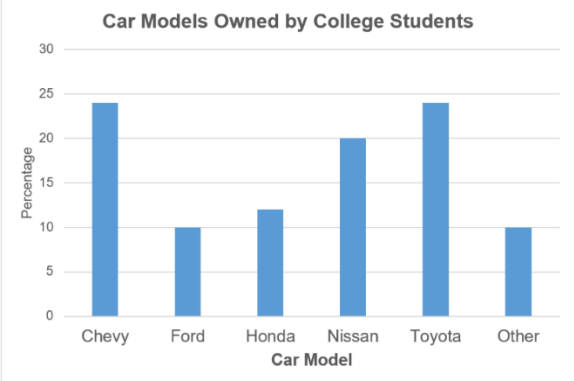


b) Create a bar graph using relative frequencies for the data.

c) Create a Pareto graph.



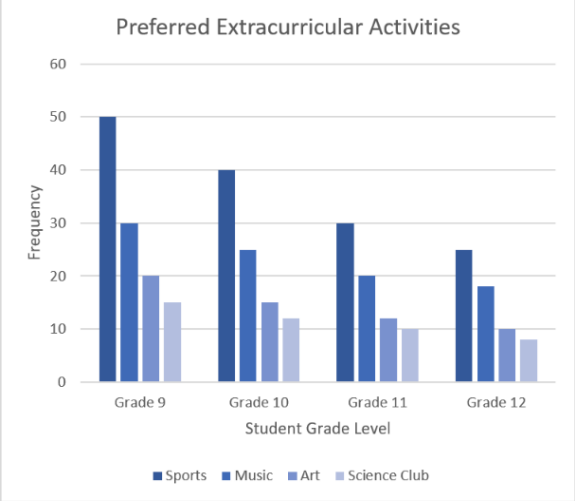
25. The following graph represents a random sample of car models driven by college students. What was the most common car model? Chevy & Toyota – they each have a frequency of around 24.



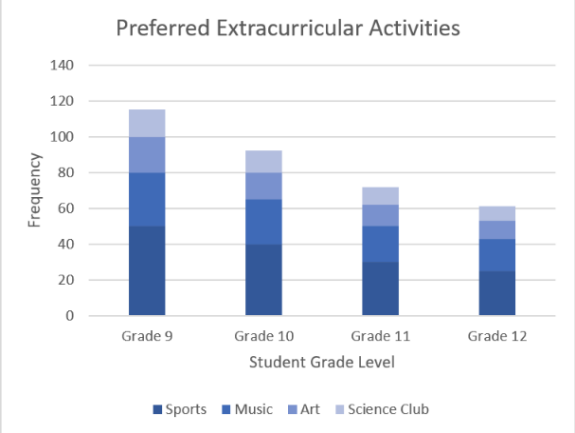
27. A high school principal conducted a survey to gather data on students' preferred extracurricular activities: Sports, Music, Art, and Science Club. The table below shows the number of students who chose each activity based on their grade level (Grade 9, Grade 10, Grade 11, Grade 12).

Activity	Grade 9	Grade 10	Grade 11	Grade 12
Sports	50	40	30	25
Music	30	25	20	18
Art	20	15	12	10
Science Club	15	12	10	8

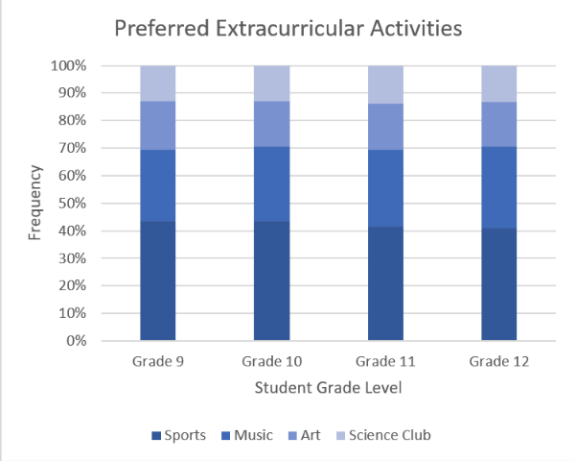
a) Create a side-by-side bar graph for the data by day of the week.



b) Create a stacked column graph for the data.



c) Create a 100% stacked column graph for the data.

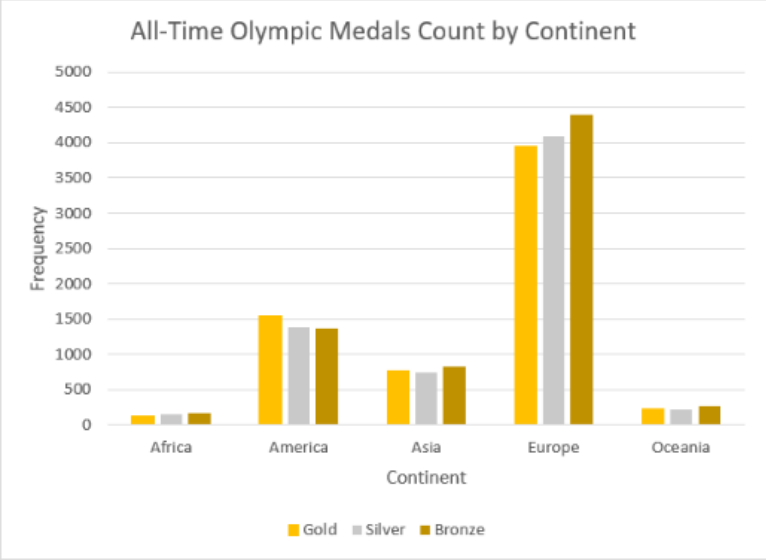


29. The following data show the all-time Olympic medals count by continent.

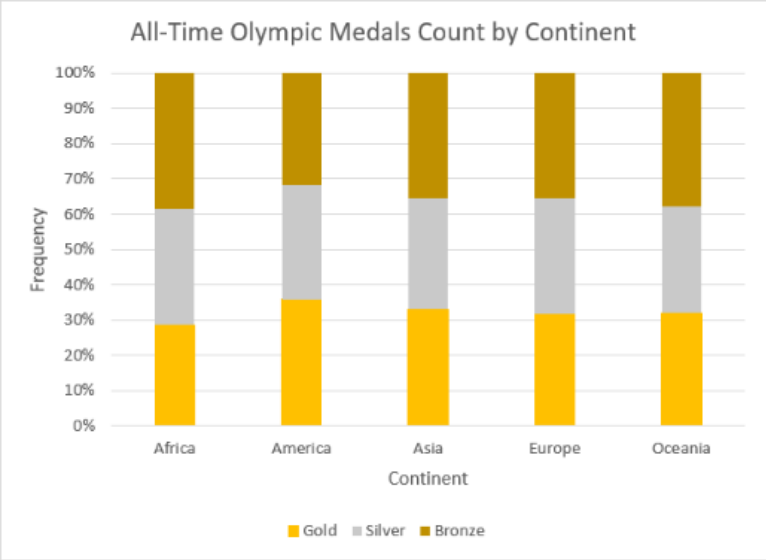
Continent	Gold	Silver	Bronze
Africa	126	146	169
America	1542	1390	1370
Asia	772	734	832
Europe	3949	4095	4393
Oceania	232	218	273

https://www.olympiandatabase.com/index.php?id=21633&L=1

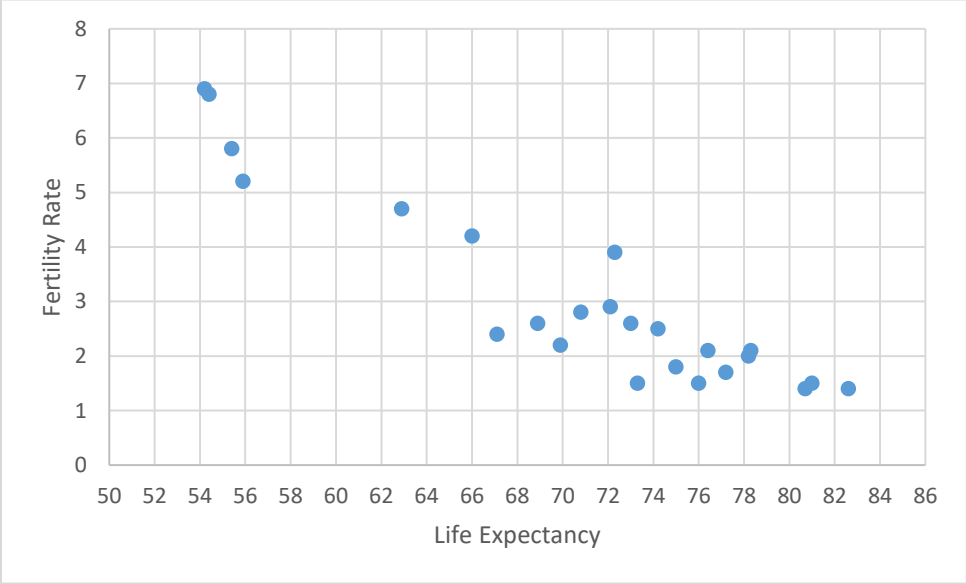
a) Create a side-by-side bar graph for the data by type of medal.



b) Create a 100% stacked column graph for the data where the continent is on the horizontal axis.

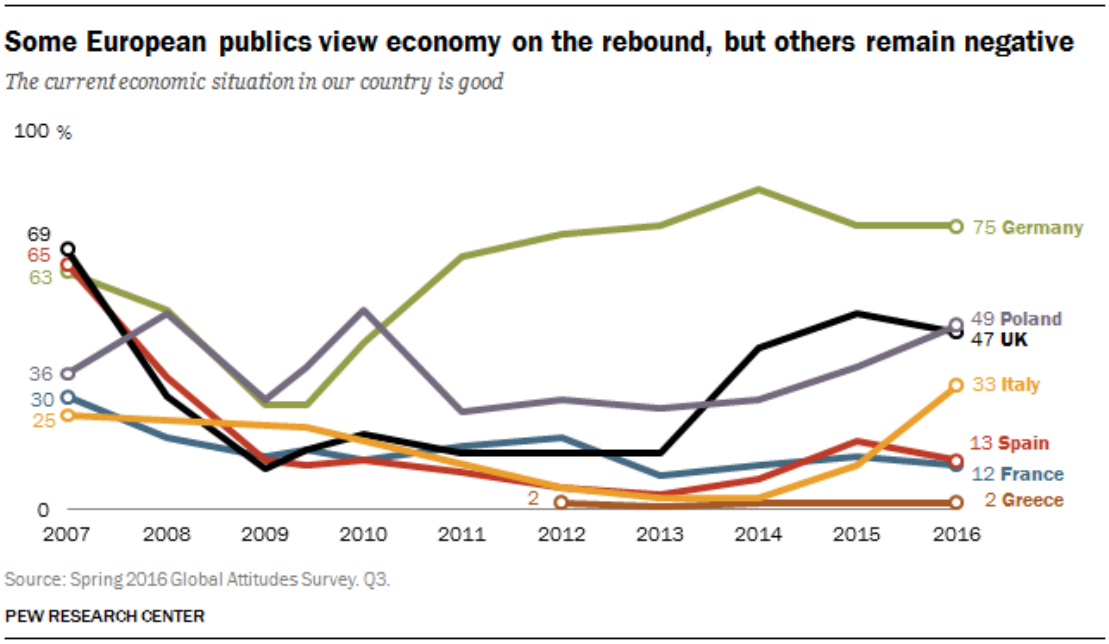


31. A scatter plot for a random sample of 24 countries shows the average life expectancy and the average number of births per woman (fertility rate). What is the approximate fertility rate for a country that has a life expectancy of 76 years? (2013, October 14). 1.5. The x-value of 76 has a corresponding y-value of 1.5.



Retrieved from http://data.worldbank.org/indicator/SP.DYN.TFRT.IN

33. A survey by the Pew Research Center, conducted in 16 countries among 20,132 respondents from April 4 to May 29, 2016, before the United Kingdom's so-called Brexit referendum to exit the EU. The following is a time series graph for the proportion of survey respondents by country that responded that the current economic situation is their country was good.

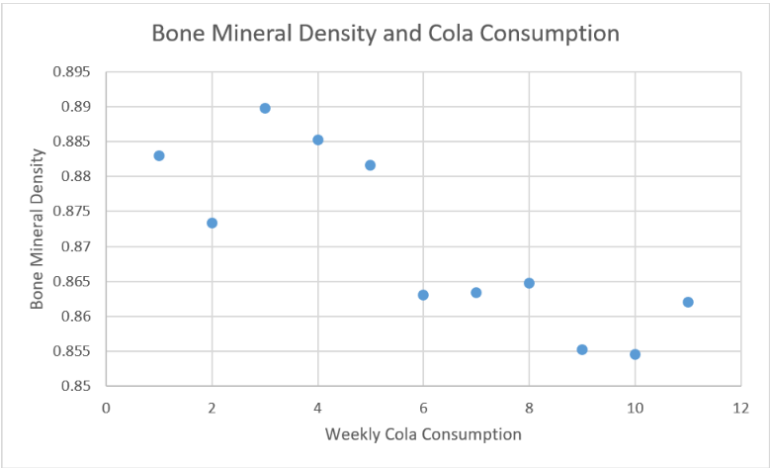


http://www.pewglobal.org/2016/08/09/views-on-national-economies-mixed-as-many-countries-continue-to-struggle/

a) Which country had the most favorable outlook of their country's economic situation in 2010? Poland, because in 2010 the highest line is the purple line, which represents Poland.
b) Which country had the least favorable outlook of their country's economic situation in 2016? Greece, because in 2016 the lowest line is the brown line, which represents Greece.

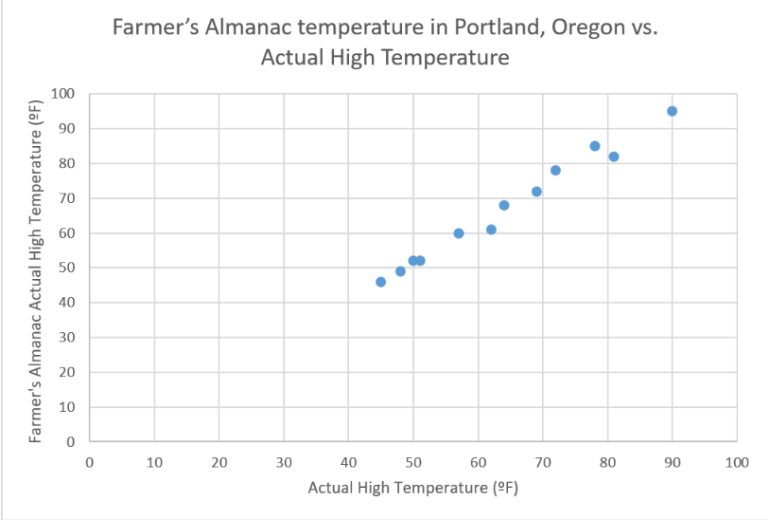
35. Bone mineral density and cola consumption have been recorded for a sample of patients. Let x represent the number of colas consumed per week and y the bone mineral density in grams per cubic centimeter. Create a scatter plot for the following data.

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862

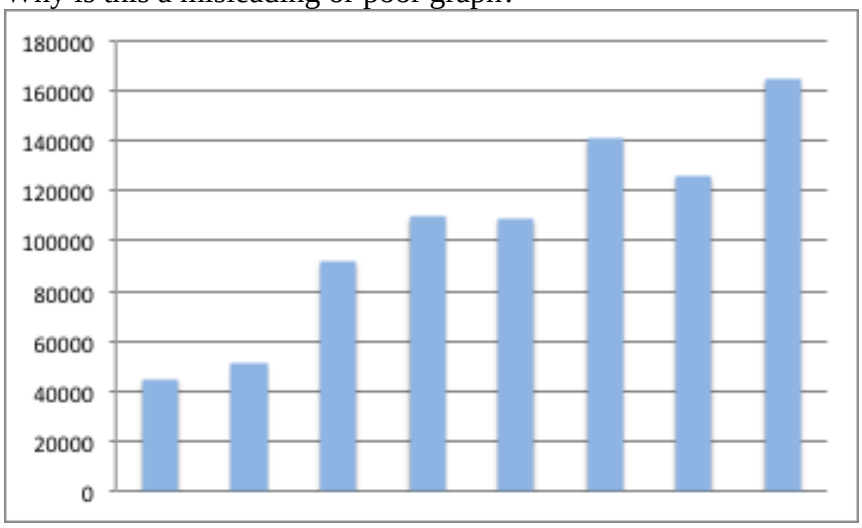


37. Create a scatter plot for the predicted average high temperature (ºF) per month by the Farmer's Almanac (x) in Portland, Oregon and the actual high (y) temperature per month that occurred.

Farmer's Almanac	45	50	57	62	69	72	81	90	78	64	51	48
Actual High	46	52	60	61	72	78	82	95	85	68	52	49

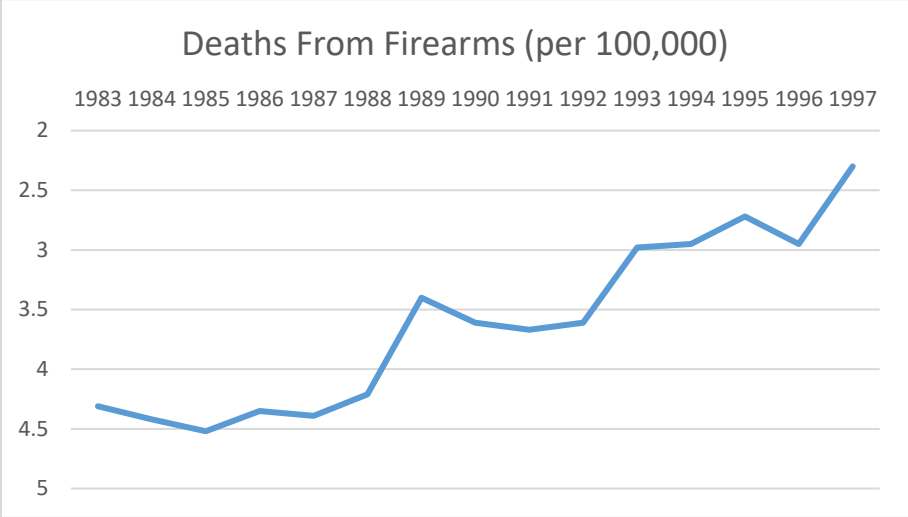


39. Why is this a misleading or poor graph?



There are no labels for both axis or categories.

41. The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997 (2013, September 26). Why is this a misleading or poor graph? The vertical axis is reversed, making the graph appear to increase when it is actually decreasing.



Retrieved from http://www.statsci.org/data/oz/firearms.html.

Chapter 3 Exercises

1. A sample of 8 cats found the following weights in kg. Compute the mean, median and mode.

4.0	4.1	3.2	4.0	3.8	3.6	3.7	3.4

   a) Compute the mode. Mode = 4.0 kg, because 4.0 occurs the most in the list of data.
   b) Compute the median. Median = 3.75 kg, because 3.75 is the average of the middle two values in the sorted data.
   Sorted list: 3.2 3.4 3.6 **3.7** **3.8** 4.0 4.0 4.1
   $\frac{3.7+3.8}{2} = 3.75$
   c) Compute the mean. $\bar{x} = \frac{4.0+4.1+3.2+4.0+3.8+3.6+3.7+3.4}{8} = 3.725\ kg$

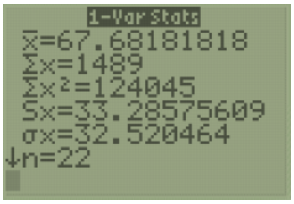
3. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed below.

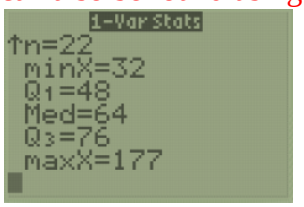
River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

Data from http://www.statsci.org/data/oz/nzrivers.html

   a) Compute the mode. 56 km & 64 km – both 56 and 64 occur three times, so there are two modes in this case.
   b) Compute the median. 64 km – find the median by sorting the data from lowest to highest and then taking the middle value. Since there are *two* middle values in this case, take the average of the two $\frac{64+64}{2} = 64$.
   c) Compute the mean. 67.6818 km – find the mean by adding up all the data values and dividing the result by the total number of data values.

   The mean and median can also be found using 1-Var Stats in your calculator.





5. A university assigns letter grades with the following 4-point scale: A = 4.00, A− = 3.67, B+ = 3.33, B = 3.00, B− = 2.67, C+ = 2.33, C = 2.00, C− = 1.67, D+ = 1.33, D = 1.00, D− = 0.67, F = 0.00. Calculate the grade point average (GPA) for a student who took in one term a 3-credit biology course and received a C+, a 1-credit lab course and received a B, a 4-credit engineering course and received an A− and a 4-credit chemistry course and received a C+.
Find the weighted mean by multiplying each score by the corresponding weights and adding up the results for the numerator. The denominator will be the sum of the weights.

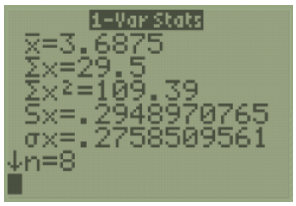
$$\frac{(3 \cdot 2.33 + 1 \cdot 3 + 4 \cdot 3.67 + 4 \cdot 2.33)}{3 + 1 + 4 + 4} = 2.833$$

7. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives an 85 on test 1, a 76 on test 2, an 83 on test 3, a 74 on the homework, a 65 on the project, and a 61 on the final, what grade did the student earn in the course? All the assignments were out of 100 points.
Find the weighted mean by multiplying each score by the corresponding weights and adding up the results for the numerator. The denominator will be the sum of the weights.

$$\frac{15 \cdot 85 + 15 \cdot 76 + 15 \cdot 83 + 10 \cdot 74 + 20 \cdot 65 + 25 \cdot 61}{15 + 15 + 15 + 10 + 20 + 25} = 72.25$$

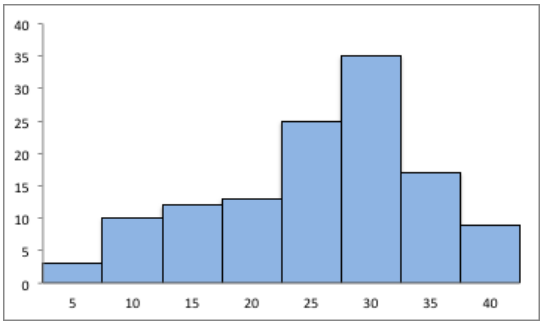
9. A sample of 8 cats found the following weights in kg.

3.7	4.1	3.2	4.0	3.8	3.6	3.7	3.4



a) Compute the range. Max − Min = 4.1 − 3.2 = 0.9
b) Compute the variance. $s^2 = 0.2949^2 = 0.087$
c) Compute the standard deviation. $s = 0.2949$

11. The following is a histogram of quiz grades.



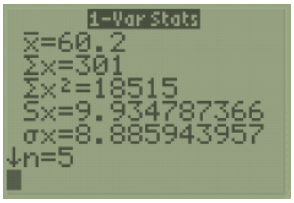
a) What is the shape of the distribution? Negatively skewed, since the peak of the graph is on the right side and there is a tail to the left.
b) Which is higher, the mean or the median? The mean is pulled in the direction of the tail, so the median would be higher.

13. Suppose that a manager wants to test two new training programs. The manager randomly selects 5 people for each training type and measures the time it takes to complete a task after the training. The times for both trainings are in table below. Which training method is more variable?

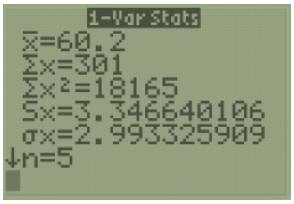
Training 1	56	75	48	63	59
Training 2	60	58	66	59	58

Since the scale and units are the same, we can compare standard deviations.

$$s_1 = \sqrt{\frac{(56-60.2)^2+(75-60.2)^2+(48-60.2)^2+(63-60.2)^2+(59-60.2)^2}{5-1}} = 9.9348$$



$$s_2 = \sqrt{\frac{(60-60.2)^2+(58-60.2)^2+(66-60.2)^2+(59-60.2)^2+(58-60.2)^2}{5-1}} = 3.3466$$



Training type 1 is more variable since it has a larger standard deviation.

15. Here are pulse rates before and after exercise. Which group has the larger range?

Before	Pulse Rates	After
9 8 8 7 6 5 2	6	
9 8 8 8 6 5 5 5 1 1 0 0	7	
8 8 7 5 4 2	8	5 6 6 7 8 9
4 0	9	0 1 1 2 3 4 5 5 6 8
4	10	0 1 4 6 7
	11	6 7
	12	4 5 7

Find the range of each using $Max - Min$.
Before range = 104 − 62 = 42, after range = 127 − 85 = 42. Both groups have the same range.

17. The following is a sample of quiz scores.

Score	17	44.5	16.1	37.2	42.8	37.5	19.5	28.2

a) Compute $\bar{x}$. 30.35
b) Compute s^2. 136.7
c) Compute the median. 32.7
d) Compute the coefficient of variation. 32.7 / 2035 = 38.52%
e) Compute the range. Max − Min = 28.4

19. The following is the height and weight of a random sample of baseball players.

Height (inches)	Weight (pounds)
76	212
76	224
72	180
74	210
75	215
71	200
77	235
78	235
77	194
76	185
72	180
72	170
75	220
74	228
73	210
72	180
70	185
73	190
71	186
74	200
74	200
75	210
78	240
72	208
75	180

a) Compute the coefficient of variation for both height and weight.

	Height (inches)	*Weight (pounds)*
Mean	74.08	203.08
Standard Deviation	2.253146	19.98733
Coefficient of Variation	$\frac{2.253146}{74.08} \cdot 100\% = 3.04\%$	$\frac{19.98733}{203.08} \cdot 100\% = 9.84\%$

b) Is there more variation in height or weight? Weight, because it has a higher coefficient of variation.

21. The length of a human pregnancy is normally distributed with a mean of 272 days with a standard deviation of 9.1 days. William Hunnicut was born in Portland, Oregon, at just 181 days into his gestation. What is the z-score for William Hunnicut's gestation?
Retrieved from: http://digitalcommons.georgefox.edu/cgi/viewcontent.cgi?article=1149&context=gfc_life

$$z = \frac{x-\bar{x}}{s} = \frac{181-272}{9.1} = -10,$$ this is 10 standard deviations below average.

23. The average time to run the Pikes Peak Marathon 2017 was 7.44 hours with a standard deviation of 1.34 hours. Rémi Bonnet won the Pikes Peak Marathon with a run time of 3.62 hours. Retrieved from: http://pikespeakmarathon.org/results/ppm/2017/.
The Tevis Cup 100-mile one day horse race for 2017 had an average finish time of 20.38 hours with a standard deviation of 1.77 hours. Tennessee Lane won the 2017 Tevis cup in a ride time of 14.75 hours. Retrieved from: https://aerc.org/rpts/RideResults.aspx.

a) Compute the z-score for Rémi Bonnet's time. $z = \frac{3.62-7.44}{1.34} = -2.8507$

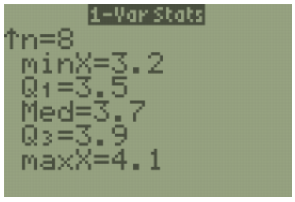
b) Compute the z-score for Tennessee Lane's time. $z = \frac{14.75-20.38}{1.77} = -3.1808$

c) Which competitor did better compared to their respective events? "Better" for race times would be the smaller of the two z-scores, so Tennessee Lane did better.

25. A sample of 8 cats found the following weights in kg. Compute the 5-number summary.

3.7	4.1	3.2	4.0	3.8	3.6	3.7	3.4

Excel	=QUARTILE.EXC	=QUARTILE.INC
Minimum	3.2	3.2
Q1	3.45	3.55
Q2	3.7	3.7
Q3	3.95	3.85
Maximum	4.1	4.1



27. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed below.

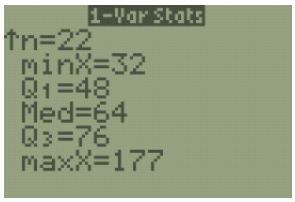
River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

Data from http://www.statsci.org/data/oz/nzrivers.html

a) Compute the 5-number summary.

Note: The values you find for the quartiles may differ depending on what form of technology you use to calculate them.

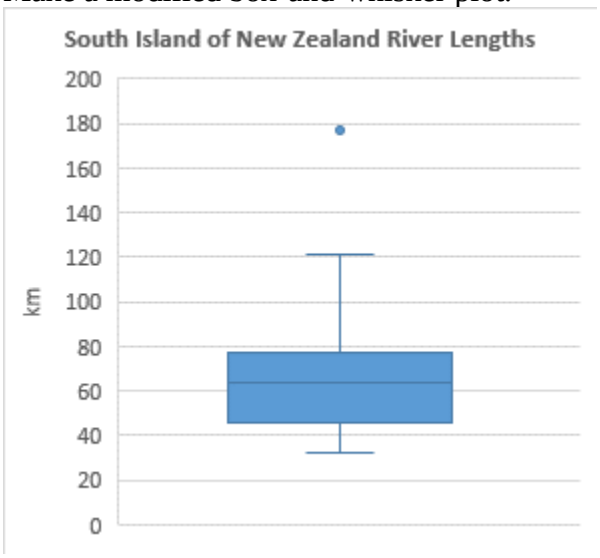
	Answers	Excel Formulas
Min	32	=MIN
Q_1	46 (Excel) or 48 (TI)	=QUARTILE.EXC
Q_2	64	=MEDIAN
Q_3	77 (Excel) or 76 (TI)	=QUARTILE.EXC
Max	177	=MAX

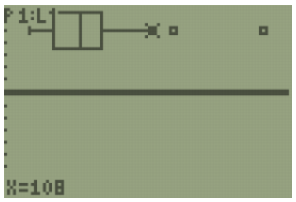


b) Compute the lower and upper limits and any outlier(s) if any exist.
Excel: IQR = $Q_3 - Q_1$ = 77 − 46 = 31, Lower Limit = Q_1 − 1.5*IQR = 46 − 1.5*31 = −0.5, Upper Limit = Q3 + 1.5*IQR = 77 + 1.5*31 = 123.5. Any data value outside the limits (−0.5, 123.5) is an outlier. Outlier = 177. Note the whiskers go out to the data values 32 and 121, not the limits.

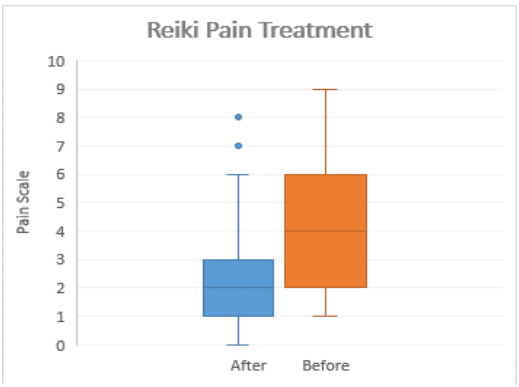
TI-Calculator: IQR = $Q_3 - Q_1$ = 76 − 48 = 28, Lower Limit = Q_1 − 1.5*IQR = 48 − 1.5*28 = 6, Upper Limit = Q3 + 1.5*IQR = 76 + 1.5*28 = 118. Any data value outside the limits (−0.5, 123.5) is an outlier. Outliers = 121 and 177. Note the whiskers go out to the data values 32 and 108, not the limits.

c) Make a modified box-and-whisker plot.





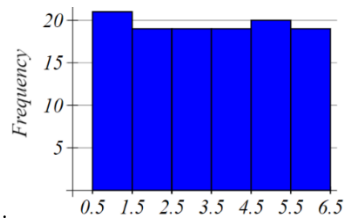
29. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997). Higher numbers mean the patients had more pain.

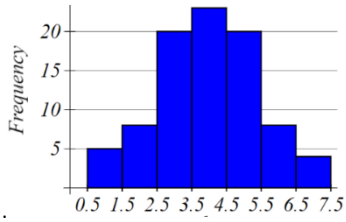


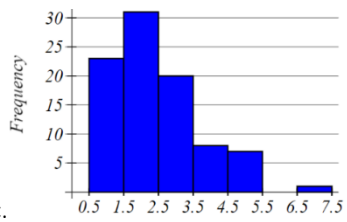
a) Use the box-and-whiskers plots to determine the IQR for the before treatment measurements. IQR = $Q_3 - Q_1 = 6 - 2 = 4$

b) Use the box-and-whiskers plots of the before and after VAS ratings to determine if the Reiki method was effective in reducing pain. Yes, the treatment was effective since all 3 quartiles for the after treatment measurements were smaller than the before treatment measurements.

31. Sort the following histograms from the smallest standard deviation to the largest and comment on the shape of each histogram.

a.


b.


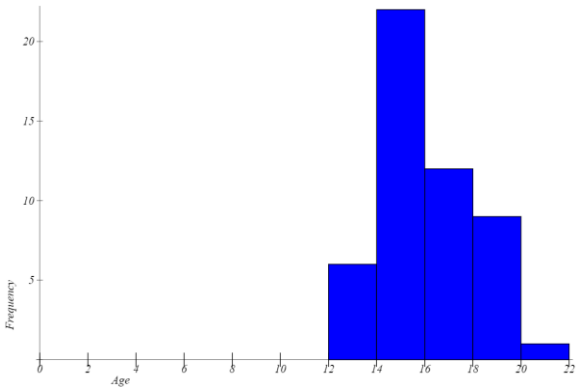
c.


When the taller bars are closer to the mean of a distribution, the standard deviation will be smaller. Answer: b, a, c

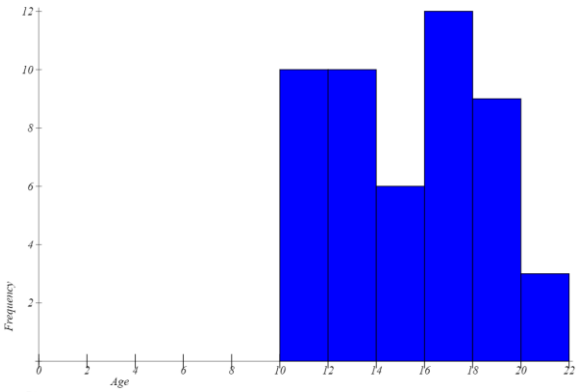
33. Match the correct descriptive statistics to the letter of the corresponding histogram and boxplot. Choose the correct letter for the corresponding histogram and Roman numeral for the corresponding boxplot. You should only use the visual representation, the definition of standard deviation and measures of central tendency to match the graphs with their respective descriptive statistics.

	Mean	Median	Standard Deviation	Histogram Letter	Boxplot Number
1	10.4	11	6.2	c	ii
2	16	15.8	1.9	a	iii
3	14.8	15	3.1	b	i

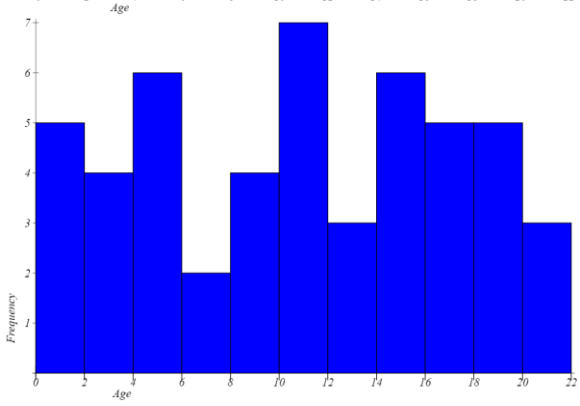
a)



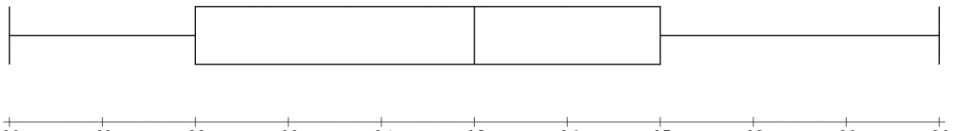
b)



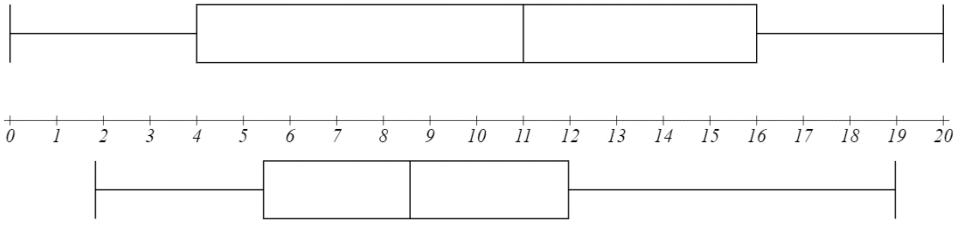
c)



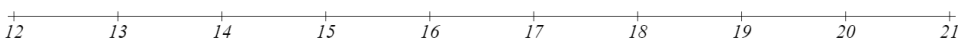
i.



ii.



iii.

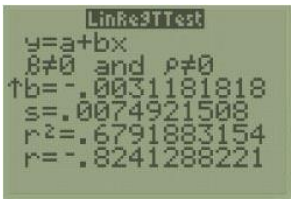


35. The correlation coefficient, r, is a number between _____. Answer: a

a) -1 and 1
b) -10 and 10
c) 0 and 10
d) 0 and ∞
e) 0 and 1
f) $-\infty$ and ∞

37. Bone mineral density and cola consumption have been recorded for a sample of patients. Let *x* represent the number of colas consumed per week and *y* the bone mineral density in grams per cubic centimeter. Assume the data is normally distributed. Calculate the correlation coefficient. $r = 0.8241$

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862



39. A teacher believes that the third homework assignment is a key predictor of how well students will do on the midterm. Let *x* represent the third homework score and *y* the midterm exam score. A random sample of last term's students were selected and their grades are shown below. Assume scores are normally distributed.

HW3	Midterm
13.1	59
21.9	87
8.8	53
24.3	95
5.4	39
13.2	66
20.9	89
18.5	78

HW3	Midterm
6.4	43
20.2	79
21.8	84
23.1	92
22	87
11.4	54
14.9	71
18.4	76

HW3	Midterm
20	86
15.4	73
25	93
9.7	52
15.1	70
15	65
16.8	77
20.1	78

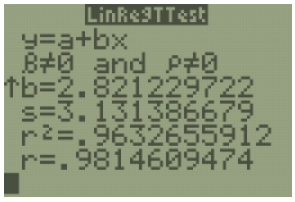
a) Compute the correlation coefficient.
Compute the 2-Var Stats and sum of squares.
$SS_{xx} = (n-1)s_x^2 = (24-1)5.558814322^2 = 710.7095833$

$SS_{yy} = (n-1)s_y^2 = (24-1)15.97892634^2 = 5872.5$

$SS_{xy} = \Sigma(xy) - n \cdot \bar{x} \cdot \bar{y} = 31156 - 24 \cdot 16.695833 \cdot 72.75 = 2005.075$

$$r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}} = \frac{2005.075}{\sqrt{(710.7095833 \cdot 5872.5)}} = 0.9815$$

$r = 0.9815$



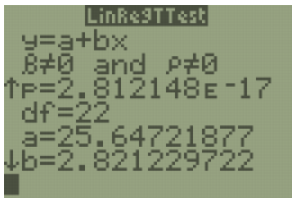
b) Compute the regression equation.

Calculate the slope: $b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{2005.075}{710.7095833} = 2.821229722.$

Then calculate the y-intercept: $b_0 = \bar{y} - b_1 \cdot \bar{x} = 72.75 - (2.821229722) \cdot 16.695833 = 25.64721877.$

Put the numbers back into the regression equation.

Write your answer as: $\hat{y}$ = 25.6472 + 2.8212x



c) Compute the predicted midterm score when the homework 3 score is 15.

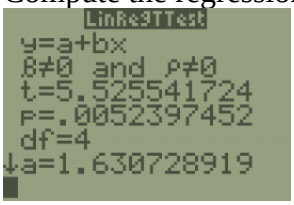
$\hat{y} = 25.6472 + 2.8212 * 15 = 67.96566$

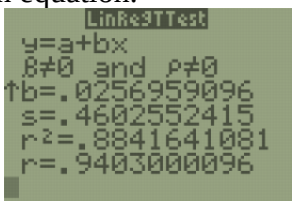
41. The following data represent the leaching rates (percent of lead extracted vs. time in minutes) for lead in solutions of magnesium chloride ($MgCl_2$).

Time (x)	4	8	16	30	60	120
Percent Extracted (y)	1.2	1.6	2.3	2.8	3.6	4.4

a) Compute the correlation coefficient. $r = 0.9403$

b) Compute the regression equation.





$\hat{y} = 1.630728919 - 0.0256959096x$

c) Predict the percent extracted for 100 minutes.

$\hat{y} = 1.630728919 - 0.0256959096 \cdot 100 = 4.20031988$

Chapter 4 Exercises

1. The number of M&M candies for each color found in a case were recorded in the table below.

Blue	Brown	Green	Orange	Red	Yellow	Total
481	371	483	544	372	369	2,620

What is the probability of selecting a red M&M? 372/2620 = 0.1420

3. An experiment is to flip a fair coin three times. What is the probability of getting exactly two heads? There are 3 outcomes with exactly two heads and a total of 8 outcomes in the sample space P(2 H) = 3/8 = 0.375

5. A raffle sells 1000 tickets for $35 each to win a new car. What is the probability of winning the car? There is only one winning ticket, 1/1000 = 0.001.

7. Compute the probability of rolling doubles when two 20-sided dice are rolled.
There are 20*20 = 400 possible outcomes, of which 20 of them would have matching numbers for doubles. P(D) = 20/400 = 0.05

9. Compute the probability of rolling a sum of two dice that is a 7 or a 12. There are 7 ways to get a sum of 7 or 12, P(7 or 12) = 7/36 = 0.1944.

+	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

11. The probability that a consumer entering a retail outlet for microcomputers and software packages will buy a computer of a certain type is 0.15. The probability that the consumer will buy a particular software package is 0.10. There is a 0.05 probability that the consumer will buy both the computer and the software package. What is the probability that the consumer will buy the computer or the software package?
P(Computer or SP) = P(Computer) + P(SP) − P(Computer and SP) = 0.15 + 0.10 − 0.05 = 0.2.

13. A poll showed that 48.7% of Americans say they believe that Marilyn Monroe had an affair with JFK. What is the probability of randomly selecting someone who does not believe that Marilyn Monroe had an affair with JFK. 1 − 0.487 = 0.513

15. Your favorite basketball player is an 81% free throw shooter. Find the probability that he does not make their next free throw shot. 1 − 0.81 = 0.19

17. Randomly pick a card from a standard deck.
    a) Compute the probability of selecting a card that shows a diamond.
    P(♦) = 13/52 = 0.25

b) Compute the probability of selecting a card that shows a diamond and an ace.
P(♦ ∩ Ace) = 1/52 = 0.0192

c) Compute the probability of selecting a card that shows a diamond or an ace.
P(♦ ∪ Ace) = P(♦) + P(Ace) − P(♦ ∩ Ace) = 13/52 + 4/52 − 1/52 = 16/52 = 0.3077

d) Compute the probability of selecting a card that shows a 7 or an ace.
P(7 ∪ Ace) = P(7) + P(Ace) − P(7 ∩ Ace) = 4/52 + 4/52 − 0/52 = 8/52 = 0.1538

e) Compute the probability of selecting a card that shows a 7 and an ace.
P(7 ∩ Ace) = 0/52 = 0

f) Are the events, selecting a 7, and selecting an ace, mutually exclusive? Why?
Yes, since you cannot get one card showing both a 7 and an ace, P(7 ∩ Ace) = 0.

19. The following table shows the frequency of items sold over several month by each item category and day of the week for a local restaurant. One item is randomly selected. Find the following. Find the marginal totals.

Item Category	Wednesday	Thursday	Friday	Saturday	Total
Appetizers	250	189	364	205	1008
Main Courses	483	407	672	591	2153
Desserts	128	103	238	196	665
Total	861	699	1274	992	3826

a) P(Main Course) = 2153/3826 = 0.5627

b) P(Main Course ∩ Friday) = 672/3826 = 0.1756

c) Compute the probability that a randomly selected item is a dessert, given that it was a Friday.
P(Dessert | Friday) = P(Dessert ∩ Friday)/P(Friday) = 238/1274 = 0.1868

d) Given that the item is a dessert, compute the probability that the item was sold on Friday.
P(Friday | Dessert) = P(Dessert ∩ Friday)/P(Dessert) = 238/665 = 0.3579

e) P(Saturday | Appetizer)
P(Saturday | Appetizer) = P(Saturday ∩ Appetizer)/P(Appetizer) = 205/1008 = 0.2034

21. Giving a test to a group of students, the grades and if they were business majors are summarized below. One student is chosen at random. Give your answer as a decimal out to at least 4 places.

	A	B	C	Total
Business Majors	4	5	13	22
Non-business Majors	18	10	19	47
Total	22	15	32	69

a) Compute the probability that the student was a non-business major or got a grade of C. P(NB or C) = P(NB) + P(C) − P(NB and C) = 47/69 + 32/69 − 19/69 = 60/69 = 0.8696

b) Compute the probability that the student was a non-business major and got a grade of C. P(NB and C) = 19/69 = 0.2754

c) Compute the probability that the student was a non-business major given they got a grade of C. P(NB|C) = P(NB and C)/P(C) = 19/32 = 0.5938

d) Compute the probability that the student did not get a B grade. $P(B^C) = 1 - P(B) = 1 - 15/69 = 54/69 = 0.7826$

e) Compute P(B ∪ Business Major). P(B) + P(Bus) − P(B and Bus) = 15/69 + 22/69 − 5/69 = 32/69 = 0.4638

f) Compute P(C | Business Major). P(C and Bus)/P(Bus) = 13/22 = 0.5909

23. The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.

	Inoculated	Not Inoculated	Total
Lived	238	5136	5374
Died	6	844	850
Total	244	5980	6224

Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

a) Compute the relative frequencies.

	Inoculated	Not Inoculated	Total
Lived	0.0382	0.8252	0.8634
Died	0.0010	0.1356	0.1366
Total	0.0392	0.9608	1

b) Compute the probability that a person was inoculated. 0.0392

c) Compute the probability that a person lived. 0.8634

d) Compute the probability that a person died or was inoculated. 0.1366 + 0.0392 − 0.0010 = 0.1748

e) Compute the probability that a person died if they were inoculated. P(Died | Inoculated) = P(Died and Inoculated)/P(Inoculated) = 0.001/0.0392 = 0.026

f) Given that a person was not inoculated, what is the probability that they died? P(Died | Not Inoculated) = P(Died and Not Inoculated)/P(Not Inoculated) = 0.1356/0.9608 = 0.141

25. A store purchases baseball hats from three different manufacturers. In manufacturer A's box there are 12 blue hats, 6 red hats, and 6 green hats. In manufacturer B's box there are 10 blue hats, 10 red hats, and 4 green hats. In manufacturer C's box, there are 8 blue hats, 8 red hats, and 8 green hats. A hat is randomly selected. Given that the hat selected is green, what is the probability that it came from manufacturer B's box? Hint: Make a table with the colors as the columns and the manufacturers as the rows.

Make a table:

	Blue	Red	Green	Total
A	12	6	6	24
B	10	10	4	24
C	8	8	8	24
Total	30	24	18	72

Find P(B | Green) = $\frac{P(B \cap Green)}{P(Green)} = \frac{4}{18} = 0.2222$

27. The probability of stock A rising is 0.3; and of stock B rising is 0.4. What is the probability that neither of the stocks rise, assuming that these two stocks are independent? Since A and B are independent then P(A and B) = P(A)*P(B) = 0.3*0.4 = 0.12. The probability that either stocks rise is P(A or B) = P(A) + P(B) − P(A and B) = 0.3 + 0.4 − 0.12 = 0.58. The probability of neither would be the complement to either which is 1 − 0.58 = 0.42.

29. The following data show the all-time Olympic medals count by continent. A medal is randomly selected. Find the marginal totals.

Continent	Gold	Silver	Bronze	Total
Africa	126	146	169	441
America	1542	1390	1370	4302
Asia	772	734	832	2338
Europe	3949	4095	4393	12437
Oceania	232	218	273	723
Total	6621	6583	7037	20241

https://www.olympiandatabase.com/index.php?id=21633&L=1

a) Compute the probability that the medal was gold.
P(Gold) = 6621/20241 = 0.3271

b) Compute the probability that the medal was gold and the athlete was from Africa.
P(Gold ∩ Africa) = 126/20241 = 0.0062

c) Compute the probability that the medal was gold or the athlete was from Africa.
P(Gold ∪ Africa) = P(Gold) + P(Africa) − P(Gold ∩ Africa) = 6621/20241 + 441/20241 − 126/20241 = 6936/20241 = 0.3427

d) Compute $P(Gold^C)$.
1 − P(Gold) = 1 − 6621/20241 = 0.6729

e) Compute P(Europe ∩ Bronze).
P(Europe ∩ Bronze) = 4393/20241 = 0.2170

f) Compute P(Europe ∪ Bronze).

P(Europe ∪ Bronze) = P(Europe) + P(Bronze) − P(Europe ∩ Bronze) = 12437/20241 + 7037/20241− 4393/20241 = 15081/20241 = 0.7451

g) Compute P(Europe | Bronze).
P(Europe ∩ Bronze)/P(Bronze) = 4393/7037 = 0.6243

31. How many different phone numbers are possible in the area code 503, if the first number cannot start with a 0 or 1? 8*10*10*10*10*10*10 = 8,000,000

33. The California license plate has one number followed by three letters followed by three numbers. How many different license plates are possible? 10*26*26*26*10*10*10 = 175,760,000

35. The PSU's Mixed Me club has 30 members. You need to pick a president, treasurer, and secretary from the 30. How many different ways can you do this? ${}_{30}P_3 = 24,360$

37. A baseball team has a 20-person roster. A batting order has nine people. How many different batting orders are there? ${}_{20}P_9 = 60,949,324,800$

39. A computer generates a random password for your account (the password is not case sensitive). The password must consist of 8 characters, each of which can be any letter or number. How many different passwords could be generated? 36^8 = 2,821,109,907,456

41. A typical PSU locker is opened with correct sequence of three numbers between 0 and 49 inclusive. A number can be used more than once, for example, 8-8-8 is valid. How many possible locker combinations are there? 50*50*50 = 125,000

Chapter 5 Exercises

1. Determine if the following tables are valid discrete probability distributions. If they are not state why.
   a) Yes, since all the probabilities are between 0 and 1 and the probabilities add up to one.

x	−5	−2.5	0	2.5	5
P(X = x)	0.15	0.25	0.32	0.18	0.1

   b) No, since the probabilities do not add up to 1.

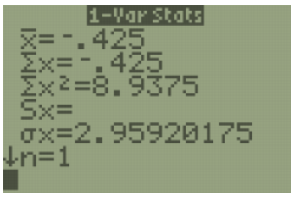
x	0	1	2	3	4
P(X = x)	0.111	0.214	0.312	0.163	0.159

   c) No, since not all the probabilities are between 0 and 1.

x	0	1	2	3	4
P(X = x)	0.2	−0.3	0.5	0.4	0.2

3. The following discrete probability distribution represents the amount of money won for a raffle game.

x	−5	−2.5	0	2.5	5
P(X = x)	0.15	0.25	0.32	0.18	0.1



   a) Compute μ. $\mu = \Sigma(x \cdot P(x)) = (-5) \cdot 0.15 + (-2.5) \cdot 0.25 + 0 \cdot 0.32 + 2.5 \cdot 0.18 + 5 \cdot 0.1 = -\$0.425$
   b) Compute σ.
   $\sigma = \sqrt{(\sum x^2 \cdot P(x)) - \mu^2} =$
   $\sqrt{((-5)^2 \cdot 0.15 + (-2.5)^2 \cdot 0.25 + 0^2 \cdot 0.32 + 2.5^2 \cdot 0.18 + 5^2 \cdot 0.1) - (-.425)^2} =$
   $\sqrt{8.756875} = \$2.9592$

5. The bookstore also offers a chemistry textbook for $159 and a book supplement for $41. From experience, they know about 25% of chemistry students just buy the textbook while 60% buy both the textbook and supplement, the remaining 15% of students do not buy either book. Find the standard deviation of the bookstore revenue.
   $\mu = \Sigma(x \cdot P(x)) = 159 \cdot 0.25 + 200 \cdot 0.6 + 0 \cdot 0.15 = \$159.75$
   $\sigma = \sqrt{(\sum x^2 \cdot P(x)) - \mu^2} = \sqrt{(159^2 \cdot 0.25 + 200^2 \cdot 0.6 + 0^2 \cdot 0.15) - (159.75)^2} =$
   $\sqrt{4800.1875} = \$69.283$

7. An LG Dishwasher, which costs $1000, has a 24% chance of needing to be replaced in the first 2 years of purchase. If the company has to replace the dishwasher within the two-year extended warranty, it will cost the company $112.10 to replace the dishwasher.

   a) Fill out the probability distribution for the value of the extended warranty from the perspective of the company.

x	-112.1	887.9
P(X = x)	0.76	0.24

   b) What is the expected value of the extended warranty? $\mu = \Sigma(x \cdot P(x) = (-112.1) \cdot 0.76 + 887.9 \cdot 0.24 = \$127.90$

c) Write a sentence interpreting the expected value of the warranty. For many of these extended warranties bought by customers, they can expect to gain 127.9 dollars per warranty on average.

9. The following table represents the probability of the number of pets owned by a college student.

x	0	1	2	3
P(X = x)	0.46	0.35	0.12	0.07

a) Is this a valid discrete probability distribution? Explain your answer. Yes, since Σ P(x) = 1 and $0 \le P(x) \le 1$

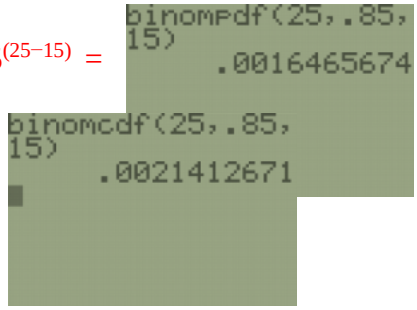
b) Find the mean number of pets owned. $\mu = \Sigma(x \cdot P(x)) = 0 \cdot 0.46 + 1 \cdot 0.35 + 2 \cdot 0.12 + 3 \cdot 0.07 = 0.8$

c) Find the standard deviation of the number of cars owned. $\sigma = \sqrt{(\sum x^2 \cdot P(x)) - \mu^2} = \sqrt{(0^2 \cdot 0.46 + 1^2 \cdot 0.35 + 2^2 \cdot 0.12 + 3^2 \cdot 0.07) - 0.8^2} = \sqrt{0.82} = 0.9055$

d) Find σ^2. 0.82

11. Suppose a random variable, X, arises from a binomial experiment. If $n = 25$, and $p = 0.85$, find the following probabilities.

a) P(X = 15)
binompdf(25, 0.85, 15), $P(X = 15) = {}_{25}C_{15} \cdot 0.85^{15} \cdot 0.15^{(25-15)} = 0.0016$



b) P(X ≤ 15)
binomcdf(25,0.85,15) = 0.0021

c) P(X < 15) = binomcdf(25,0.85,14) = 0.0005

d) P(X > 15) = 1−binomcdf(25, 0.85, 15) = 0.9979

e) P(X ≥ 15) = 1−binomcdf(25, 0.85, 14) = 0.9995

f) μ $\mu = n \cdot p = 25 \cdot 0.85 = 21.25$

g) σ $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{25 \cdot 0.85 \cdot 0.15} = 1.7854$

h) σ^2 $\sigma^2 = n \cdot p \cdot q = 25 \cdot 0.85 \cdot 0.15 = 3.1875$

13. A local county has an unemployment rate of 7.3%. A random sample of 20 employable people are picked at random from the county and are asked if they are employed. The distribution is a binomial. Round answers to 4 decimal places. Binomial, n = 20, p = 0.073

a) Find the probability that exactly 3 in the sample are unemployed. $P(X = 3) = {}_{20}C_3 \cdot 0.073^3 \cdot 0.927^{(20-3)} = 0.1222$

b) Find the probability that there are fewer than 4 in the sample are unemployed. $P(X < 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = {}_{20}C_0 \cdot 0.073^0 \cdot 0.927^{(20-0)} + {}_{20}C_1 \cdot 0.073^1 \cdot 0.927^{(20-1)} + {}_{20}C_2 \cdot 0.073^2 \cdot 0.927^{(20-2)} + {}_{20}C_3 \cdot 0.073^3 \cdot 0.927^{(20-3)} = 0.9464$

c) Find the probability that there are more than 2 in the sample are unemployed. $P(X > 2) = 1 - P(X \le 2) = 1 - ({}_{20}C_0 \cdot 0.073^0 \cdot 0.927^{(20-0)} + {}_{20}C_1 \cdot 0.073^1 \cdot 0.927^{(20-1)} + {}_{20}C_2 \cdot 0.073^2 \cdot 0.927^{(20-2)}) = 0.1759$

d) Find the probability that there are at most 4 in the sample are unemployed. $P(X \le 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = {}_{20}C_0 \cdot 0.073^0 \cdot 0.927^{(20-0)} +$

$_{20}C_1 \cdot 0.073^1 \cdot 0.927^{(20-1)} + {}_{20}C_2 \cdot 0.073^2 \cdot 0.927^{(20-2)} + {}_{20}C_3 \cdot 0.073^3 \cdot 0.927^{(20-3)} + {}_{20}C_4 \cdot 0.073^4 \cdot 0.927^{(20-4)} = 0.9873$

15. Approximately 10% of all people are left-handed. Out of a random sample of 15 people, find the following. Binomial, n = 15, p = 0.1
    a) What is the probability that 4 of them are left-handed?
    P(X = 4) = binompdf(15, 0.10, 4) = 0.0428
    b) What is the probability that less than 4 of them are left-handed? 0.9444
    P(X < 4) = binomcdf(15, 0.10, 3) = 0.9444
    c) What is the probability that at most 4 of them are left-handed? P(X ≤ 4) = binomcdf(15, 0.10, 4) = 0.9873
    d) What is the probability that at least 4 of them are left-handed? P(X ≥ 4) = 1 − binomcdf(15,0.10,3) = 0.0556
    e) What is the probability that more than 4 of them are left-handed? P(X > 4) = 1 − binomcdf(15,0.10,4) = 0.0127
    f) Compute μ. $\mu = n \cdot p = 15 \cdot 0.10 = 1.5$
    g) Compute σ. $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{15 \cdot 0.1 \cdot 0.9} = \sqrt{1.35} = 1.1619$
    h) Compute σ². $\sigma^2 = n \cdot p \cdot q = 15 \cdot 0.1 \cdot 0.9 = 1.35$

17. About 1% of the population has a particular genetic mutation. Find the standard deviation for the number of people with the genetic mutation in a group of 100 randomly selected people from the population. Binomial, n = 10, p = 0.01, $\sigma = \sqrt{10 \cdot 0.01 \cdot 0.99} = 0.995$

19. A poll is given, showing 72% are in favor of a new building project. Let X be the number of people who favor the new building project when 37 people are chosen at random. What is the probability that between 10 and 16 (including 10 and 16) people out of 37 favor the new building project? Binomial, n = 37, p = 0.72, P(10 ≤ X ≤ 16) = 0.0002

21. The Lee family had 6 children. Assuming that the probability of a child being a girl is 0.5, find the probability that the Smith family had at least 4 girls? Binomial, n = 6, p = 0.5, P(X ≥ 4) = P(X = 4) + P(X = 5) + P(X = 6) = $_6C_4 \cdot 0.5^4 \cdot 0.5^2 + {}_6C_5 \cdot 0.5^5 \cdot 0.5^1 + {}_6C_6 \cdot 0.5^6 \cdot 0.5^0 = 0.3438$

23. A manufacturing machine has a 6% defect rate. An inspector chooses 4 items at random. Binomial, n = 4, p = 0.06

    a) What is the probability that at least one will have a defect? P(X ≥ 1) = 1 − P(X = 0) = 1 − $_4C_0 \cdot 0.06^0 \cdot 0.94^{(4-0)} = 0.2193$
    b) What is the probability that exactly two will have a defect? P(X = 2) = $_4C_2 \cdot 0.06^2 \cdot 0.94^{(4-2)} = 0.0191$
    c) What is the probability that less than two will have a defect? P(X < 2) = P(X = 0) + P(X = 1) = $_4C_0 \cdot 0.06^0 \cdot 0.94^{(4-0)} + {}_4C_1 \cdot 0.06^1 \cdot 0.94^{(4-1)} = 0.9801$
    d) What is the probability that more than one will have a defect? P(X > 1) = 1 − P(X ≤ 1) = 1 − $(_4C_0 \cdot 0.06^0 \cdot 0.94^{(4-0)} + {}_4C_1 \cdot 0.06^1 \cdot 0.94^{(4-1)}) = 0.0199$

25. A small regional carrier accepted 20 reservations for a particular flight with 17 seats. 15 reservations went to regular customers who will arrive for the flight. Each of the remaining

passengers will arrive for the flight with a 60% chance, independently of each other. Binomial, n = 5, p = 0.6

a) Find the probability that overbooking occurs. $P(X > 2) = P(X = 3) + P(X = 4) + P(X = 5) = {}_5C_3 \cdot 0.6^3 \cdot 0.4^2 + {}_5C_4 \cdot 0.6^4 \cdot 0.4^1 + {}_5C_5 \cdot 0.6^5 \cdot 0.4^0 = 0.6826$

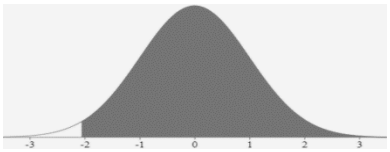
b) Find the probability that the flight has empty seats. $P(X < 2) = {}_5C_0 \cdot 0.6^0 \cdot 0.4^5 + {}_5C_1 \cdot 0.6^1 \cdot 0.4^4 = 0.087$

27. In a mid-size company, the distribution of the number of phone calls answered each day by each of the 12 employees is bell-shaped and has a mean of 59 and a standard deviation of 10. Using the empirical rule, what is the approximate percentage of daily phone calls numbering between 29 and 89? 99.7%

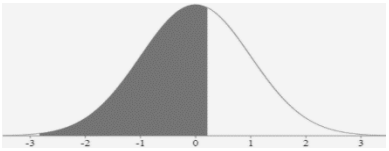
29. A company has a policy of retiring company cars; this policy looks at number of miles driven, purpose of trips, style of car and other features. The distribution of the number of months in service for the fleet of cars is bell-shaped and has a mean of 42 months and a standard deviation of 3 months. Using the Empirical Rule, what is the approximate percentage of cars that remain in service between 48 and 51 months? 2.35%

31. For a standard normal distribution, find the following probabilities.

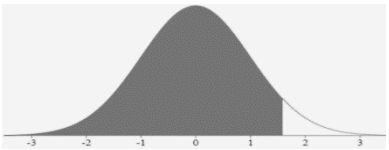
a) P(Z > −2.06)
Excel: =1−NORM.S.DIST(−2.06,TRUE) = 0.9803
TI-84: normalcdf(−2.06,1E99,0,1) = 0.9803



b) P(−2.83 < Z < 0.21)
Excel: =NORM.S.DIST(0.21,TRUE)−NORM.S.DIST(−2.83,TRUE) = 0.5809
TI-84: normalcdf(−2.83,0.21,0,1) = 0.5809

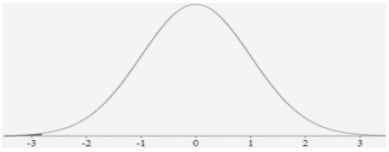


c) P(Z < 1.58)
Excel: =NORM.S.DIST(1.58,TRUE) = 0.9429
TI-84: normalcdf(−1E99,1.58,0,1) = 0.9429



d) P(Z ≥ 1.69)
Excel: =1−NORM.S.DIST(1.69,TRUE) = 0.0455
TI-84: normalcdf(1.69,1E99,0,1) = 0.0455

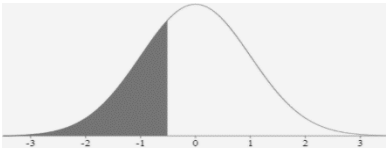
e) P(Z < −2.82)
Excel: =NORM.S.DIST(−2.82,TRUE) = 0.0024
TI-84: normalcdf(−1E99,−2.82,0,1) = 0.0024



f) P(Z > 2.14)
Excel: =1−NORM.S.DIST(2.14,TRUE) = 0.0162
TI-84: normalcdf(2.14,1E99,0,1) = 0.0162

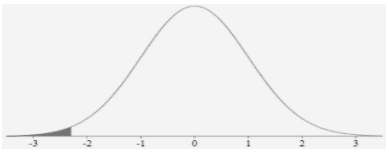
g) P(1.97 ≤ Z ≤ 2.93)
Excel: =NORM.S.DIST(2.93,TRUE)−NORM.S.DIST(1.97,TRUE) = 0.0227
TI-84: normalcdf(1.97,2.93,0,1) = 0.0227

h) P(Z ≤ −0.51)
Excel: =NORM.S.DIST(−0.51,TRUE) = 0.305
TI-84: normalcdf(−1E99,−0.51,0,1) = 0.305



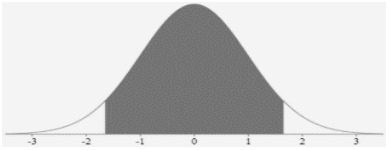
33. Compute the following probabilities where Z ~ *N*(0,1).

a) P(*Z* ≤ −2.03)
Excel: =NORM.S.DIST(−2.03,TRUE) = 0.0107
TI-84: normalcdf(−1E99,−2.03,0,1) = 0.0107

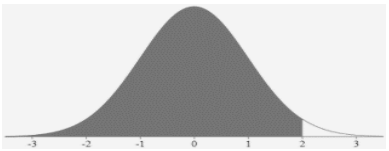


b) P(Z > 1.58)
Excel: =1−NORM.S.DIST(1.58,TRUE) = 0.0571
TI-84: normalcdf(1.58,1E99,0,1) = 0.0571

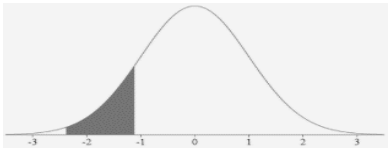
c) P(−1.645 ≤ *Z* ≤ 1.645)
Excel: =NORM.S.DIST(1.645,TRUE)−NORM.S.DIST(−1.645,TRUE) = 0.90
TI-84: normalcdf(−1.645,1.645,0,1) = 0.90



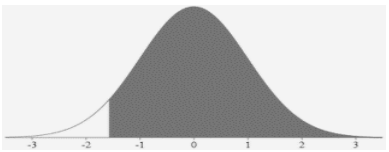
d) P(Z < 2)
Excel: =NORM.S.DIST(2,TRUE) = 0.9772
TI-84: normalcdf(−1E99,2,0,1) = 0.9772



e) P(−2.38 < Z < −1.12)
Excel: =NORM.S.DIST(−1.12,TRUE)−NORM.S.DIST(−2.38,TRUE) = 0.1227
TI-84: normalcdf(−2.38,−1.12,0,1) = 0.1227

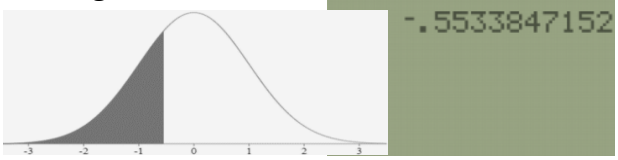


f) P(*Z* ≥ −1.75)
Excel: =1−NORM.S.DIST(−1.57,TRUE) = 0.9418
TI-84: normalcdf(−1.57,1E99,0,1) = 0.9418

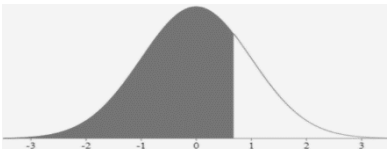


35. Use the standard normal distribution for the following.

a) Compute the z score that gives the 29th percentile.
Excel =NORM.S.INV(0.29) = −0.5534
TI-84: invNorm(0.29,0,1) = −0.5534

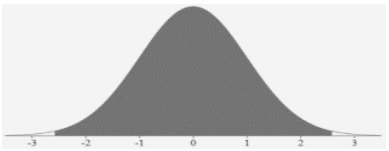


b) Compute the z score that gives the 75th percentile.
Excel =NORM.S.INV(0.29) = 0.6745
TI-84: invNorm(0.75,0,1) = 0.6745



c) Compute the two z-scores that give the middle 99% area.

Area between two unknown z-scores is 0.99, that leaves 1 − 0.99 = 0.01 area split between both tails. Half of 0.01 is 0.005. Use left tail areas 0.005 and 1 − 0.005 = 0.995.



Excel =NORM.S.INV(0.005) = −2.5758 and =NORM.S.INV(0.995) = 2.5758

TI-84: invNorm(0.005,0,1) = −2.5758 and invNorm(0.995,0,1) = 2.5758

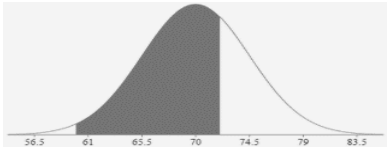
37. Arm span is the physical measurement of the length of an individual's arms from fingertip to fingertip. A man's arm span is approximately normally distributed with mean of 70 inches with a standard deviation of 4.5 inches.

a) Compute the probability that a randomly selected man has an arm span below 65 inches.

Excel: =NORM.DIST(65,70,4.5,TRUE) = 0.1333

TI-84: normalcdf(−1E99,65,70,4.5) = 0.1333

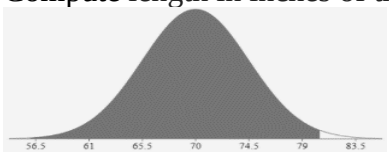
b) Compute the probability that a randomly selected man has an arm span between 60 and 72 inches.



Excel: =NORM.DIST(72,70,4.5,TRUE)−NORM.S.DIST(60,70,4.5,TRUE) = 0.6585

TI-84: normalcdf(60,72,70,4.5) = 0.6585

c) Compute length in inches of the 99th percentile for a man's arm span.



Excel =NORM.INV(0.99,70,4.5) = 80.4686

TI-84: invNorm(0.99,70,4.5) = 80.4686

39. A dishwasher has a mean life of 12 years with an estimated standard deviation of 1.25 years ("Appliance life expectancy," 2013). Assume the life of a dishwasher is normally distributed.

a) Compute the probability that a dishwasher will last less than 10 years.

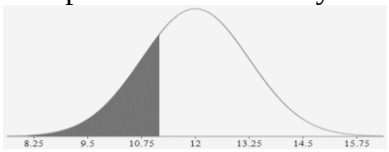
Excel: =NORM.DIST(10,12,1.25,TRUE) = 0.0548

TI-84: normalcdf(−1E99,10,12,1.25) = 0.0548

b) Compute the probability that a dishwasher will last between 8 and 10 years.

Excel: =NORM.DIST(10,12,1.25,TRUE)−NORM.S.DIST(8,12,1.25,TRUE) = 0.0541
TI-84: normalcdf(8,10,12,1.25) = 0.0541

c) Compute the number of years that the bottom 25% of dishwashers would last.



Excel =NORM.INV(0.25,12,1.25) = 11.1569
TI-84: invNorm(0.25,12,1.25) = 11.1569 years

41. Heights of 10-year-old children, regardless of sex, closely follow a normal distribution with mean 55.7 inches and standard deviation 6.8 inches.
    a) Compute the probability that a randomly chosen 10-year-old child is less than 50.4 inches.
    Excel: =NORM.DIST(50.4,55.7,6.8,TRUE) = 0.2179
    TI-84: normalcdf(−1E99,50.4,55.7,6.8) = 0.2179
    b) Compute the probability that a randomly chosen 10-year-old child is more than 59.2 inches.
    Excel: =1−NORM.DIST(59.2,55.7,6.8,TRUE) = 0.3034
    TI-84: normalcdf(59.2,1E99,55.7,6.8) = 0.3034
    c) What proportion of 10-year-old children are between 50.4 and 61.5 inches tall? 0.5853
    Excel: =NORM.DIST(61.5,55.7,6.8,TRUE)−NORM.S.DIST(50.4,55.7,6.8,TRUE) = 0.5853
    TI-84: normalcdf(50.4,61.5,55.7,6.8) = 0.0541
    d) Compute the 85th percentile for 10-year-old children.
    Excel =NORM.INV(0.85,55.7,6.8) = 62.7477 inches
    TI-84: invNorm(0.85,55.7,6.8) = 62.7477 inches

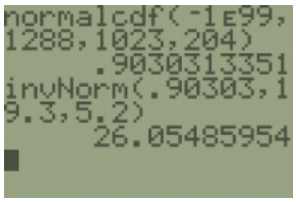
43. The mean daily milk production of a herd of cows is assumed to be normally distributed with a mean of 33 liters, and standard deviation of 10.3 liters. Compute the probability that daily production is more than 40.9 liters?
    Excel: =1−NORM.DIST(40.9,33,10.3,TRUE) = 0.2215
    TI-84: normalcdf(40.9,1E99,33,10.3) = 0.2215

45. A study was conducted on students from a particular high school over the last 8 years. The following information was found regarding standardized tests used for college admittance. Scores on the SAT test are normally distributed with a mean of 1023 and a standard deviation of 204. Scores on the ACT test are normally distributed with a mean of 19.3 and a standard deviation of 5.2. It is assumed that the two tests measure the same aptitude, but use different scales.
    a) Compute the SAT score that is the 50-percentile.
    Excel =NORM.INV(0.5,1023,204) = 1023

TI-84: invNorm(0.5,1023,204) = 1023

b) Compute the ACT score that is the 50-percentile.
Excel =NORM.INV(0.5,19.3,5.2) = 19.3
TI-84: invNorm(0.5,19.3,5.2) = 19.3

c) If a student gets an SAT score of 1288, find their equivalent ACT score. Go out at least 5 decimal places between steps.
$P(X \leq 1288)$ = normcdf(-1E99,1288,1023,204) = 0.90303
invNorm(0.90303,19.3,5.2) = 26.1



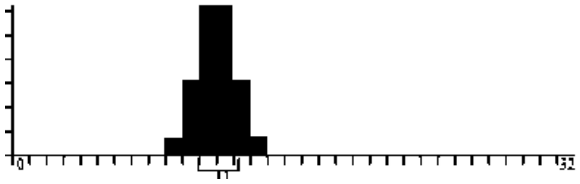
47. The MAX light rail in Portland, OR has a waiting time that is uniformly distributed with a mean waiting time of 5 minutes with a standard deviation of 2.9 minutes. A random sample of 40 wait times was selected. What is the probability the sample **mean** wait time is under 4 minutes?
Use the Central Limit Theorem to find the probability of a mean $P(\bar{x} < 4)$. Even though the distribution of the population is uniform, the sampling distribution will be normally distributed with a mean of 5 and a standard deviation of $\frac{2.9}{\sqrt{40}}$ since the sample size is over 30.
Excel: =NORM.DIST(4,5,2.9/SQRT(40),TRUE) = 0.0146
TI-84: normalcdf(−1E99,4,5,2.9/$\sqrt{40}$) = 0.0146

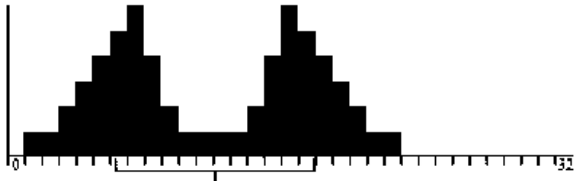
49. A certain brand of electric bulbs has an average life of 300 hours with a standard deviation of 45. A random sample of 100 bulbs is tested. What is the probability that the sample **mean** will be less than 295?
Use the Central Limit Theorem to find the probability of a mean $P(\bar{x} < 295)$. Even though the distribution of the population is unknown, the sample size is over 30. The sampling distribution will be normally distributed with a mean of 300 and a standard deviation of $\frac{45}{\sqrt{100}}$.
Excel: =NORM.DIST(295,300,45/SQRT(100),TRUE) = 0.1333
TI-84: normalcdf(−1E99,295,300,45/$\sqrt{100}$) = 0.1333

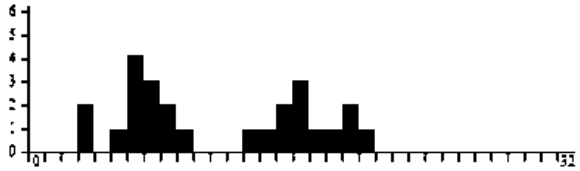
51. If the Central Limit Theorem is applicable, this means that the sampling distribution of a ____ _____ population can be treated as normal since the _____ is _____ .
a) symmetrical; variance; large
b) positively skewed; sample size; small
c) negatively skewed; standard deviation; large
d) non-normal; mean; large
e) negatively skewed; sample size; large

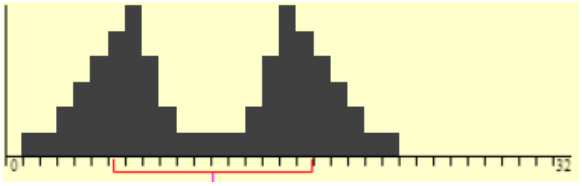
Answer e). If the Central Limit Theorem is applicable, this means that the sampling distribution of a negatively skewed population can be treated as normal since the sample size is large.

53. Match the following 3 graphs with the distribution of the population, the distribution of the sample, and the sampling distribution.
a) Distribution of the Population
b) Distribution of the Sample
c) Sampling Distribution

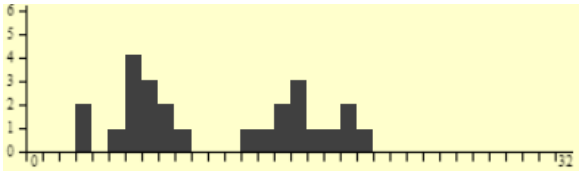
i. 

ii. 

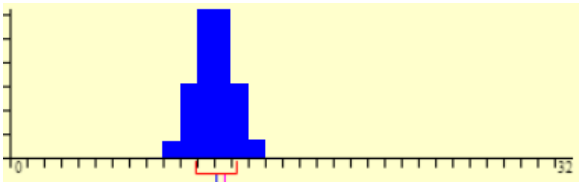
iii. 



Distribution of the Population

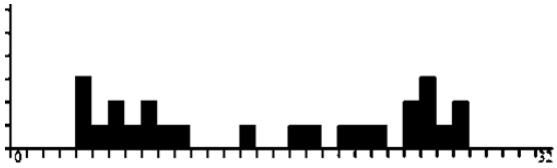


Distribution of the Sample

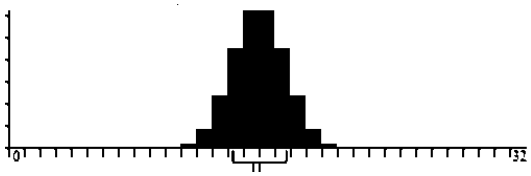


Sampling Distribution

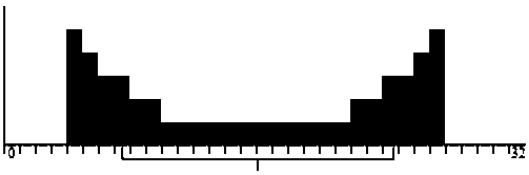
55. Match the following 3 graphs with the distribution of the population, the distribution of the sample, and the sampling distribution.
    a) Distribution of the Population
    b) Distribution of the Sample
    c) Sampling Distribution

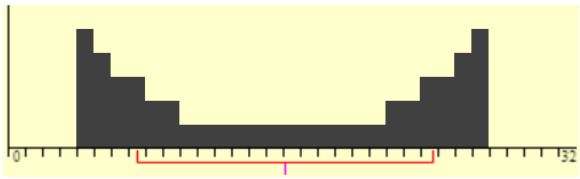
iv. 

v.

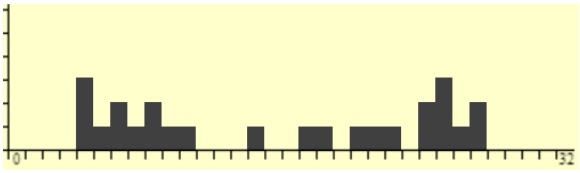


vi.

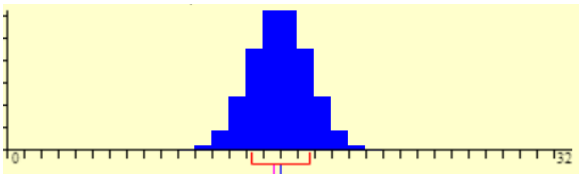




Distribution of the Population



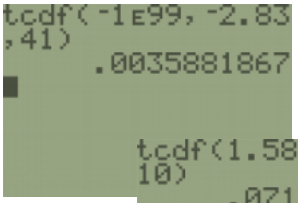
Distribution of the Sample



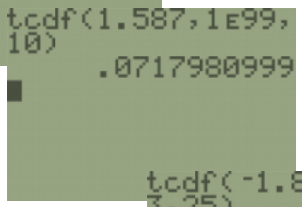
Sampling Distribution

57. For a t-distribution, find the following probabilities.

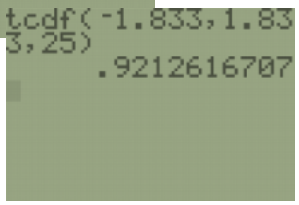
a) $P(t \leq -2.83)$, $n = 42$
TI: tcdf(−1E99,−2.83,41) = 0.0036
Excel: =T.DIST(−2.83,41,TRUE) = 0.0036



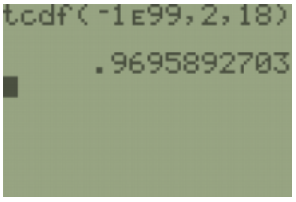
b) $P(t > 1.587)$, $n = 11$
TI: tcdf(1.587,1E99,10) = 0.0718
Excel: =1 − T.DIST(1.587,10,TRUE) = 0.0718



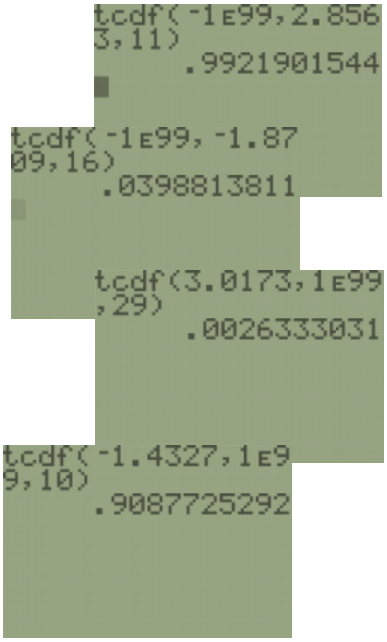
c) $P(-1.833 \leq t \leq 1.833)$, $df = 25$
TI: tcdf(−1.833,1.833,25) = 0.9213
Excel: =T.DIST(1.833,25,TRUE) − T.DIST(−1.833,25,TRUE) = 0.9213



d) $P(t < 2)$, $df = 18$
TI: tcdf(−1E99,2,18) = 0.9696
Excel: =T.DIST(2,18,TRUE) = 0.9696



59. Use the area under the t-distribution for the following.



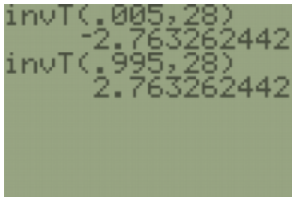
a) Compute the area to the left of t = 2.8563, when n = 12.
TI: tcdf(−1E99,2.8563,11) = 0.9922
Excel: =T.DIST(2.8563,11,TRUE) = 0.9922

b) Compute the area to the left of t = −1.8709, when df = 16.
TI: tcdf(−1E99,−1.8709,16) = 0.0399
Excel: =T.DIST(−1.8709,16,TRUE) = 0.0399

c) Compute the area to the right of t = 3.0173, when n = 30.
TI: tcdf(3.0173,1E99,29) = 0.0026
Excel: =1 − T.DIST(3.0173,29,TRUE) = 0.0026

d) Compute the area to the right of t = −1.4327, when df = 10.
TI: tcdf(−1.4327,1E99,10) = 0.9088
Excel: =1 − T.DIST(−1.4327,10,TRUE) = 0.9088

61. Compute the t-scores that give the middle 99% of the t-distribution for df = 28.



Subtract the middle area of .90 from the total area: 1 − 0.99 = 0.01. This is the area for both tails. The area in the left tail would be half this area, 0.01 / 2 = 0.005.
TI: invT(0.005,28) = −2.7633 and invT(0.995,28) = 2.7633
Excel: =T.INV(0.005,28) = −2.7633 and T.INV(0.995,28) = 2.7633
t = ±2.7633

Chapter 6 Exercises

1. Which confidence level would give the narrowest margin of error? Answer a) 80%
   a) 80%
   b) 90%
   c) 95%
   d) 99%

3. For a confidence level of 90% with a sample size of 35, find the critical z values. Use technology, invNorm(0.05,0,1), z = ±1.644853626

5. Which of the following would result in the widest confidence interval? Answer: d

   a) A sample size of 100 with 99% confidence.
   b) A sample size of 100 with 95% confidence.
   c) A sample size of 30 with 95% confidence.
   d) A sample size of 30 with 99% confidence.

7. In a random sample of 200 people, 135 said that they watched educational TV. Find and interpret the 95% confidence interval of the true proportion of people who watched educational TV.

   $\hat{p} = \frac{x}{n} = \frac{135}{200} = 0.675$ $\hat{q} = 1 - \hat{p} = 1 - 0.675 = 0.325$

   invNorm(0.025,0,1) = −1.959963986

   $\hat{p} \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ $0.675 \pm -1.959963986\sqrt{\left(\frac{0.675 \cdot 0.325}{200}\right)}$ $0.6101 < p < 0.7399$

9. A teacher wanted to estimate the proportion of students who take notes in her class. She used data from a random sample size of 82 and found that 50 of them took notes. The 99% confidence interval for the proportion of student that take notes is _____< p < _____.

   $\hat{p} = \frac{x}{n} = \frac{50}{82} = 0.6098$ $\hat{q} = 1 - \hat{p} = 1 - 0.6098 = 0.3902$

   invNorm(0.005,0,1) = −2.575829303

   $\hat{p} \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ $0.6098 \pm -2.575829303\sqrt{\left(\frac{0.6098 \cdot 0.3902}{82}\right)}$ $0.4710 < p < 0.749$

11. A 2019 survey of 2,000 adults, commissioned by the sleep-industry experts from Sleepopolis, reveals 34% sleep with a stuffed animal, blanket, or other anxiety-reducing item of sentimental value. Calculate the 90% confidence interval for the true proportion of adults that sleep with an anxiety-reducing item.

    $\hat{p} = 0.34$ $\hat{q} = 1 - \hat{p} = 1 - 0.34 = 0.66$ invNorm(0.05,0,1) = −1.644853626

    $\hat{p} \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ $0.34 \pm -1.644853626\sqrt{\left(\frac{0.34 \cdot 0.66}{2000}\right)}$ $0.3226 < p < 0.3574$

    We can be 90% confident that the population proportion of people that sleep with a stuffed animal, blanket, or other anxiety-reducing item of sentimental value is between 32.26% and 35.74%.

13. According to the 2023 survey of 30,300 U.S. households, 17% reported using a food pantry in the last year. Calculate the 95% confidence interval for the true proportion of U.S. households that used a food pantry in the last year.

$\hat{p} = 0.17$ $\hat{q} = 1 - \hat{p} = 1 - 0.17 = 0.83$ invNorm(0.025,0,1) = −1.959963986

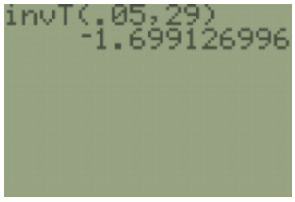
$\hat{p} \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)}$ $0.17 \pm -1.959963986\sqrt{\left(\frac{0.17 \cdot 0.83}{30300}\right)}$ 0.1658 < p < 0.1742

We can be 95% confident that the population proportion of U.S. households that used a food pantry in the last year is between 16.58% and 17.42%.

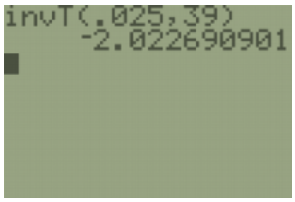
15. Gallup tracks daily the percentage of Americans who approve or disapprove of the job Donald Trump is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults. Margin of error is ±3 percentage points. On December 15, 2017, the gallop poll using a 95% confidence level showed that 34% approved of the job Donald Trump was doing. Which of the following the correct statistical interpretation of the confidence interval? Answer b)

a) As of December 15, 2017, 34% of American adults approve of the job Donald Trump is doing as president.
b) We are 95% confident that the interval $0.31 < p < 0.37$ contains the proportion of American adults who approve of the job Donald Trump is doing as president as of December 15, 2017.
c) As of December 15, 2017, 95% of American adults approve of the job Donald Trump is doing as president.
d) We are 95% confident that the proportion of adult Americans who approve of the job Donald Trump is doing as president is 0.34 as of December 15, 2017.

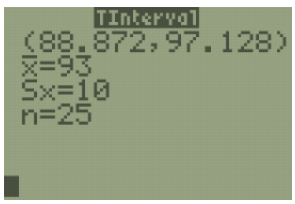
17. For a confidence level of 90% with a sample size of 30, find the critical t values. t = ±1.699127



19. For a confidence level of 95% with a sample size of 40, find the critical t values. t = ±2.023



21. A professor wants to estimate how long students stay connected during two-hour online lectures. From a random sample of 25 students, the mean stay time was 93 minutes with a standard deviation of 10 minutes. Assuming the population has a normal distribution, compute a 95% confidence interval estimate for the population mean.



Use a t-interval since we have a sample standard deviation. invT(0.025,24) = −2.063899

$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$ $93 \pm -2.063899\left(\frac{10}{\sqrt{25}}\right)$ $88.872 < \mu < 97.128$

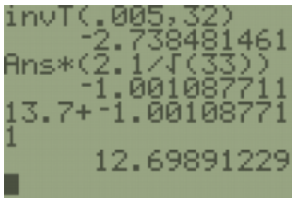
23. The age when smokers first start from previous studies is normally distributed with a mean of 13 years old. A survey of smokers of this generation was done to estimate if the mean age has changed. The sample of 33 smokers found that their mean starting age was 13.7 years old with a standard deviation of 2.1 years. Compute the 99% confidence interval of the mean.

invT(0.005,32) = −2.738481461 $\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

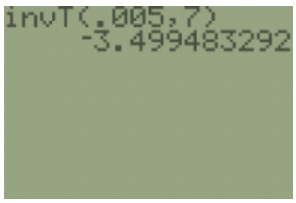
$13.7 \pm -2.738481461\left(\frac{2.1}{\sqrt{33}}\right)$

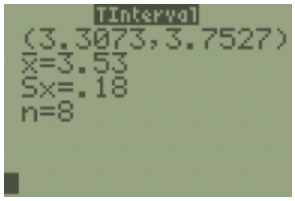
13.7 ± 1.001087711

12.6989 < μ < 14.7011



25. A college advisor wants to estimate the undergraduate grade point average (GPA) for students admitted to the top graduate business schools. The advisor randomly samples 8 students admitted to the top schools and found their GPA was 3.53 with a standard deviation of 0.18. Assume that the population is normally distributed. Calculate and interpret the 99% confidence interval estimate of the mean undergraduate GPA for all students admitted to the top graduate business schools.





invT(0.005,7) = −3.499483292

$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

$3.53 \pm -3.499483292\left(\frac{0.18}{\sqrt{8}}\right)$ $3.3073 < \mu < 3.7527$

We can be 99% confident that the population mean GPA for all graduate business students admitted into the top schools is between 3.3073 and 3.7527.

27. In a certain city, a random sample of executives finds the following personal monthly incomes (in thousands); 35, 43, 29, 55, 63, 72, 28, 33, 36, 41, 42, 57, 38, 30. Assume the population of incomes is normally distributed. Find and interpret the 95% confidence interval for the mean income.

invT(0.025,13) = −2.1604369 $\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ $43 \pm -2.1604369\left(\frac{13.626896}{\sqrt{14}}\right)$

35.13195 < μ < 50.86805 We can be 95% confident that the population mean personal monthly income is between $35,131.95 and $50,868.05.

29. Recorded here are the germination times (in days) for ten randomly chosen seeds of a new type of bean: 18, 12, 20, 17, 14, 15, 13, 11, 21, 17. Assume that the population germination time is normally distributed. Find and interpret the 99% confidence interval for the mean germination time.

invT(0.005,9) = −3.249836 $\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ $15.8 \pm -3.249836\left(\frac{3.359894}{\sqrt{10}}\right)$

12.3471 < μ < 19.2529

We can be 99% confident that the population mean germination time for the new bean is between 12.3 and 19.3 days.

31. Suppose you are a researcher in a hospital. You are experimenting with a new tranquilizer. You collect data from a random sample of 10 patients. The period of effectiveness of the tranquilizer for each patient (in hours) is as follows:

Hours	2	2.9	2.6	2.9	3	3	2	2.1	2.9	2.1

a) What is a point estimate for the population mean length of time? $\bar{x} = 2.55$

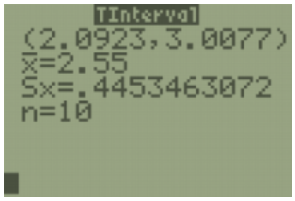
b) What must be true in order to construct a confidence interval for the population mean length of time in this situation? Choose the correct answer below. Answer: ii

i. The sample size must be greater than 30.
ii. The population must be normally distributed.
iii. The population standard deviation must be known.
iv. The population mean must be known.

c) Construct a 99% confidence interval for the population mean length of time.

invT(0.005,9) = −3.249836 $\quad \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

$2.55 \pm -3.249836\left(\frac{0.445346}{\sqrt{10}}\right)$ (2.0923, 3.0077)



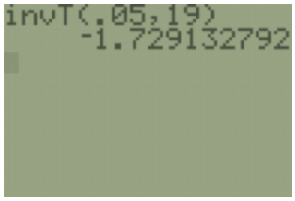
d) What does it mean to be "99% confident" in this problem? Choose the correct answer below. Answer: i

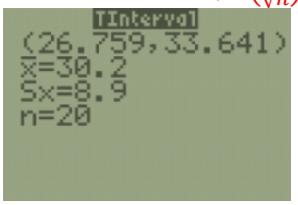
i. 99% of all confidence intervals found using this same sampling technique will contain the population mean time.
ii. There is a 99% chance that the confidence interval contains the sample mean time.
iii. The confidence interval contains 99% of all sample times.
iv. 99% of all times will fall within this interval.

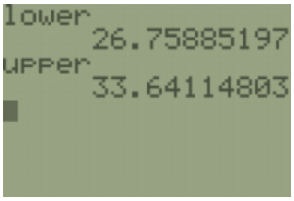
e) Suppose that the company releases a statement that the mean time for all patients is 2 hours. Is this possible? Is it likely? Answer: It is possible, but unlikely since 2 is not contained within the confidence interval boundaries.

33. The total of individual weights of garbage discarded by 20 households in one week is normally distributed with a mean of 30.2 lbs with a sample standard deviation of 8.9 lbs. Find the 90% confidence interval of the mean.

invT(0.05,19) = −1.729132792 $\quad \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ $\quad 30.2 \pm -1.729132792\left(\frac{8.9}{\sqrt{20}}\right)$







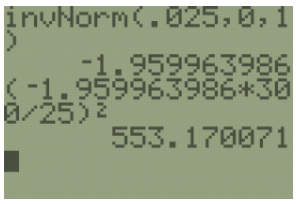
$26.7589 < \mu < 33.6411$

35. A researcher finds a 95% confidence interval for the average commute time in minutes using public transit is (15.75, 28.25). Which of the following is the correct interpretation of this interval? Answer d)
    a) We are 95% confident that all commute time in minutes for the population using public transit is between 15.75 and 28.25 minutes.
    b) There is a 95% chance commute time in minutes using public transit is between 15.75 and 28.25 minutes.
    c) We are 95% confident that the interval $15.75 < \mu < 28.25$ contains the sample mean commute time in minutes using public transportation.
    d) We are 95% confident that the interval $15.75 < \mu < 28.25$ contains the population mean commute time in minutes using public transportation.

37. A political candidate has asked you to conduct a poll to determine what percentage of people support her. If the candidate only wants a 9% margin of error at a 99% confidence level, what size of sample is needed? Since $\hat{p}$ is unknown, use p* = 0.5, $n = p^* \cdot q^* \left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.5)(0.5)\left(\frac{-2.575829303}{0.09}\right)^2 = 204.78$, always round up, so use n = 205.

39. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish that is consumed is allowed to contain 0.82 ppm of mercury or less. A laboratory is estimating the amount of mercury in tuna fish for a new company and needs to have a margin of error within 0.03 ppm of mercury with 95% confidence. Assume the population standard deviation is 0.138 ppm of mercury. What sample size is needed? Round up to the nearest integer. invNorm(0.025,0,1) = −1.959963986
$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{-1.959963986 \cdot 0.138}{0.03}\right)^2 = 81.29$ Always round up so use n = 82.

41. SAT scores are distributed with a mean of 1,500 and a standard deviation of 300. You are interested in estimating the average SAT score of first year students at your college. If you would like to limit the margin of error of your 95% confidence interval to 25 points, how many students should you sample? $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{-1.959963986 \cdot 300}{25}\right)^2 = 553.17$, round up to n = 554



Chapter 7 Exercises

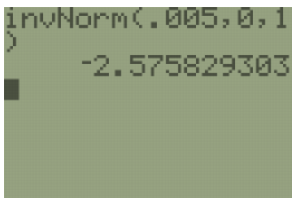
1. The plant-breeding department at a major university developed a new hybrid boysenberry plant called Stumptown Berry. Based on research data, the claim is made that from the time shoots are planted 90 days on average are required to obtain the first berry. A corporation that is interested in marketing the product tests 60 shoots by planting them and recording the number of days before each plant produces its first berry. The sample mean is 92.3 days. The corporation wants to know if the mean number of days is different from the 90 days claimed. Which one is the correct set of hypotheses?
   a) $H_0: p = 90\%$ $H_1: p \neq 90\%$
   b) $H_0: \mu = 90$ $H_1: \mu \neq 90$
   c) $H_0: p = 92.3\%$ $H_1: p \neq 92.3\%$
   d) $H_0: \mu = 92.3$ $H_1: \mu \neq 92.3$
   e) $H_0: \mu \neq 90$ $H_1: \mu = 90$

   Answer: b) $H_0: \mu = 90$ $H_1: \mu \neq 90$
   Note that your null hypotheses should always contain the signs $=, \leq, or \geq$ and your alternative hypotheses should always contain the signs $\neq, <, or >$. Since this problem is testing for an "average," we should use the variable μ in our hypotheses. The problem is specifying that we should test for a "difference" in the number of days, so we will choose $\neq$ for our alternative hypothesis.

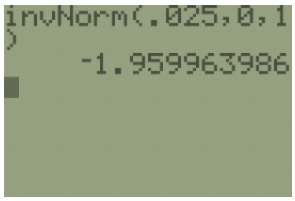
3. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints. Does this data provide enough evidence to show that Alaska had a lower proportion of identity theft than 23%? Which one is the correct set of hypotheses?
   Federal Trade Commission, (2008). *Consumer fraud and identity theft complaint data: January-December 2007.* Retrieved from website: http://www.ftc.gov/opa/2008/02/fraud.pdf.
   a) $H_0: p = 23\%$ $H_1: p < 23\%$
   b) $H_0: \mu = 23$ $H_1: \mu < 23$
   c) $H_0: p < 23\%$ $H_1: p \geq 23\%$
   d) $H_0: p = 0.224$ $H_1: p < 0.224$
   e) $H_0: \mu < 0.224$ $H_1: \mu \geq 0.224$

   Answer: a) $H_0: p = 23\%$ $H_1: p < 23\%$
   Note that your null hypotheses should always contain the signs $=, \leq, or \geq$ and your alternative hypotheses should always contain the signs $\neq, <, or >$. The problem pertains to proportions, so we will use the variable p in the hypotheses. It also specifies that we should test for a "lower" proportion, so we will select the sign of < for the alternative hypothesis.

5. Compute the z critical value for a two-tailed test when $\alpha = 0.01$.
   $z = invNorm(0.005, 0, 1) = \pm 2.5758$



7. Compute the z critical value for a two-tailed test when $\alpha = 0.05$.
   $z = invNorm(0.025, 0, 1) = \pm 1.96$



9. Match the phrase with the correct symbol.

a.	Sample Size	i.	α
b.	Population Mean	ii.	n
c.	Sample Variance	iii.	σ^2
d.	Sample Mean	iv.	s^2
e.	Population Standard Deviation	v.	s
f.	P(Type I Error)	vi.	$\bar{x}$
g.	Sample Standard Deviation	vii.	σ
h.	Population Variance	viii.	μ

n = Sample Size = ii
μ = Population Mean = viii
s^2 = Sample Variance = iv
$\bar{x}$ = Sample Mean = vi
σ = Population Standard Deviation = vii
α = P(Type I Error) = i
s = Sample Standard Deviation = v
σ^2 = Population Variance = iii

11. Match the symbol with the correct phrase.

$100(1-\alpha)\%$	Parameter
$1-\beta$	P(Type II Error)
β	Power
μ	Significance Level
α	Confidence Level

$100(1-\alpha)\%$	Confidence Level
$1-\beta$	Power
β	P(Type II Error)
μ	Parameter
α	Significance Level

13. The SAT exam in previous years is normally distributed with an average score of 1,000 points. The test writers for this upcoming year want to make sure that the new test does not have a significantly different mean score. A random sample of 20 students take the new SAT exam and their mean score was 1,050 points with a standard deviation of 150 points.

a) State the hypotheses to test to see if the mean time has significantly changed using a 5% level of significance.
H_0: μ = 1000
H_1: μ ≠ 1000 (claim)

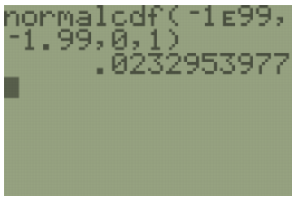
b) What is a type I error for this problem?
The test writers conclude that the average test score is not 1000 when it really was. They would need to change the exam when they really did not need to.

c) What is a type II error for this problem?
The test writers conclude that average test score is 1000 points when it really was not. Students taking the exam would have a higher or lower mean then the previous years. This could go either way, the students could have a harder test, score lower and hence not get into their choice of colleges. Or, on the flip side, students could have a higher mean score than previous years and get an unfair edge into colleges when they are not necessarily prepared.

15. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish contains at most 0.82 ppm of mercury. A scientist at the FDA believes the mean amount of mercury in tuna fish for a new company exceeds the ppm of mercury. The hypotheses are H_0: μ = 0.82 H_1: μ > 0.82. Which answer is the correct type II error in the context of this problem?
a) The fish is rejected by the FDA when in fact it had less than 0.82 ppm of mercury.
b) The fish is accepted by the FDA when in fact it had less than 0.82 ppm of mercury.
c) The fish is rejected by the FDA when in fact it had more than 0.82 ppm of mercury.
d) The fish is accepted by the FDA when in fact it had more than 0.82 ppm of mercury.

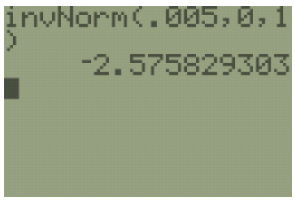
Answer: d)
A Type 2 Error occurs when the null hypothesis is not rejected, although it should have been. In this case, that means that it was determined the mercury level was less than or equal to 0.82 ppm, but in fact it was actually greater than 0.82 ppm.

17. A left-tailed z-test found a test statistic of z = −1.99. At a 5% level of significance, what would the correct decision be?
a) Do not reject H_0
b) Reject H_0
c) Accept H_0
d) Reject H_1
e) Do not reject H_1



Answer: b)
Find the p-value in your calculator using: normalcdf(−1E99, −1.99, 0, 1).
Since $p < \alpha$, reject H_0.

19. A two-tailed z-test found a test statistic of z = −2.19. At a 1% level of significance, which would the correct decision?
    a) Do not reject H_0
    b) Reject H_0
    c) Accept H_0
    d) Reject H_1
    e) Do not reject H_1



Answer: a)
The critical values for $\alpha = 0.01$ are $z = \pm 2.5757$. Do not reject H_0 since the test statistic is not in the critical rejecting region.

21. A hypothesis test was conducted during a clinical trial to see if a new COVID-19 vaccination reduces the risk of contracting the virus. What is the Type I and II errors in terms of approving the vaccine for use?

The implication of a Type I error from the clinical trial is that the vaccination will be approved when it indeed does not reduce the risk of contracting the virus.

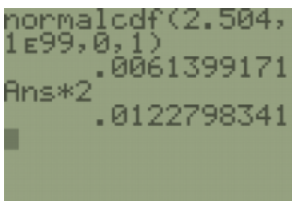
The implication of a Type II error from the clinical trial is that the vaccination will not be approved when it indeed does reduce the risk of contracting the virus.

23. A commonly cited standard for one-way length (duration) of school bus rides for elementary school children is 30 minutes. A local government office in a rural area conducts a study to determine if elementary schoolers in their district have a longer average one-way commute time. If they determine that the average commute time of students in their district is significantly higher than the commonly cited standard they will invest in increasing the number of school buses to help shorten commute time. What would a Type II error mean in this context?

The local government decides that the data do not provide convincing evidence of an average commute time higher than 30 minutes, when the true average commute time is in fact higher than 30 minutes.

25. You are conducting a study to see if the accuracy rate for fingerprint identification is significantly different from 0.34. Thus, you are performing a two-tailed test. Your sample data produce the test statistic z = 2.504. Use your calculator to find the p-value and state the correct decision and summary.

2*Normalcdf(2.504,1E99,0,1) = 0.0123

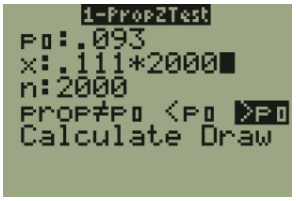


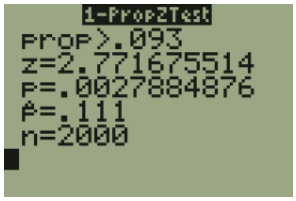
For exercises 26-31, show all 5 steps for hypothesis testing:

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

27. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.3% of American adults suffer from depression or a depressive illness. Suppose that in a survey of 2000 people in a certain city, 11.1% of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that city suffering from depression or a depressive illness is more than the 9.3% in the general adult American population. Test the relevant hypotheses using a 5% level of significance. Show all 5 steps using the p-value method.

$H_0: p = 0.093$
$H_1: p > 0.093$
$z = \frac{\hat{p}-p_0}{\sqrt{\left(\frac{p_0 q_0}{n}\right)}} = \frac{0.111-0.093}{\sqrt{\left(\frac{0.093 \cdot 0.907}{2000}\right)}} = 2.7116$

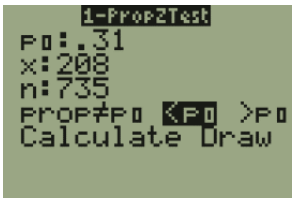


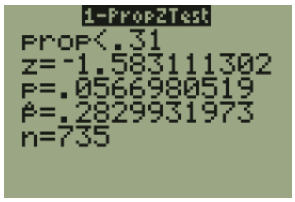


p-value = 0.0027
p-value = 0.0027 < α = 0.05 Reject H_0
There is enough evidence to support the claim the population proportion of American adults that suffer from depression or a depressive illness is more than 9.3%.

29. You are conducting a study to see if the proportion of men over the age of 50 who regularly have their prostate examined is significantly less than 0.31. A random sample of 735 men over the age of 50 found that 208 have their prostate regularly examined. Do the sample data provide convincing evidence to support the claim? Test the relevant hypotheses using a 5% level of significance.

$H_0: p = 0.31$
$H_1: p < 0.31$
Before finding the test statistic, find the sample proportion $\hat{p} = \frac{208}{735} = 0.282993$ and $q_0 = 1 - 0.31 = 0.69$.
$z = \frac{\hat{p}-p_0}{\sqrt{\left(\frac{p_0 q_0}{n}\right)}} = \frac{0.282993-0.31}{\sqrt{\left(\frac{0.31 \cdot 0.69}{735}\right)}} = -1.5831$



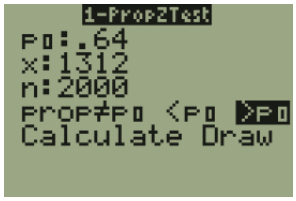


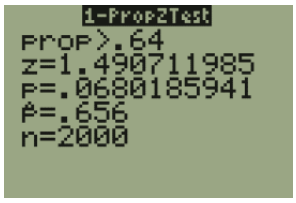
p-value = 0.0567
p-value = 0.0567 > α = 0.05
Fail to reject H_0
There is not enough evidence to support the claim the population proportion of men over the age of 50 who regularly have their prostate examined is significantly less than 0.31.

31. Nationally the percentage of adults that have their teeth cleaned by a dentist yearly is 64%. A dentist in Portland, Oregon believes that regionally the percent is higher. A sample of 2,000

Portlanders found that 1,312 had their teeth cleaned by a dentist in the last year. Test the relevant hypotheses using a 10% level of significance.

$H_0: p = 0.64$
$H_1: p > 0.64$
Before finding the test statistic, find the sample proportion $\hat{p} = \frac{1312}{2000} = 0.656$ and $q_0 = 1 - 0.64 = 0.36$.
$z = \frac{\hat{p}-p_0}{\sqrt{\left(\frac{p_0 q_0}{n}\right)}} = \frac{0.656-0.64}{\sqrt{\left(\frac{0.64 \cdot 0.36}{2000}\right)}} = 1.4907$





p-value = 0.0680
p-value = 0.0680 < α = 0.10
Reject H_0
There is enough evidence to support the claim the population proportion of adults that have their teeth cleaned by a dentist yearly is higher than 64%.

33. A student is interested in becoming an actuary. They know that becoming an actuary takes a lot of schooling and they will have to take out student loans. They want to make sure the starting salary will be higher than $55,000/year. They randomly sample 30 starting salaries for actuaries and find a p-value of 0.0392. Use α = 0.05.
    a) Choose the correct hypotheses.
        i. $H_0: \mu = 55{,}000$ $H_1: \mu < 55{,}000$
        ii. $H_0: \mu > 55{,}000$ $H_1: \mu \leq 55{,}000$
        iii. $H_0: \mu = 55{,}000$ $H_1: \mu > 55{,}000$
        iv. $H_0: \mu < 55{,}000$ $H_1: \mu \geq 55{,}000$
        v. $H_0: \mu = 55{,}000$ $H_1: \mu \neq 55{,}000$

    Answer: iii)
    The student is testing that the starting salary is higher than $55,000, so we will reflect that in the alternative hypothesis with a > sign.

    b) Should the student pursue an actuary career?
        i. Yes, since we reject the null hypothesis.
        ii. Yes, since we reject the claim.
        iii. No, since we reject the claim.
        iv. No, since we reject the null hypothesis.

    Answer: i)
    Since p < α, we reject the null hypothesis and conclude that the starting salary is higher than $55,000. Thus, the student should pursue the actuary career.

35. The Food & Drug Administration (FDA) regulates that fresh albacore tuna fish contains at most 0.82 ppm of mercury. A scientist at the FDA believes the mean amount of mercury in

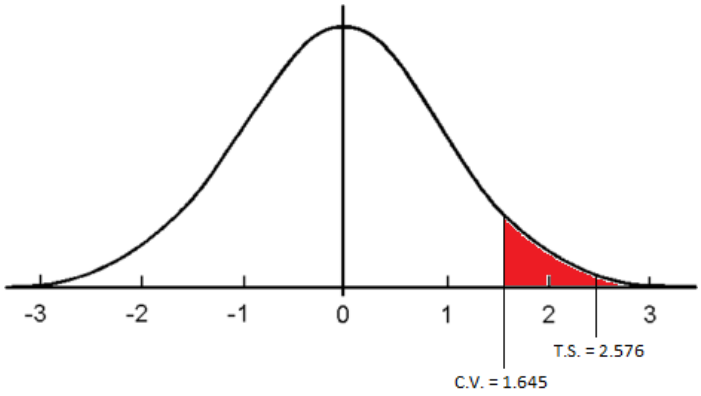
tuna fish for a new company exceeds the ppm of mercury. A test statistic was found to be 2.576 and a critical value was found to be 1.708, what is the correct decision and summary?

a) Reject H_0, there is enough evidence to support the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
b) Accept H_0, there is not enough evidence to reject the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
c) Reject H_1, there is not enough evidence to reject the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
d) Reject H_0, there is not enough evidence to support the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.
e) Do not reject H_0, there is not enough evidence to support the claim that the amount of mercury in the new company's tuna fish exceeds the FDA limit of 0.82 ppm.

Answer: a)

Plot the critical value on a graph and shade to the right of it since this is a right tailed test (see below). Plot the test statistic.

Since the test statistic lands in the shaded region, we reject the null hypothesis, and thus, we are supporting the claim in the alternative hypothesis – that the mercury level exceeds 0.82 ppm.



For exercises 37-42, show all 5 steps for hypothesis testing:

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

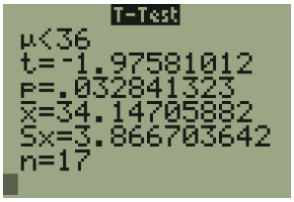
37. The total of individual pounds of garbage discarded by 17 households in one week is shown below. The current waste removal system company has a weekly maximum weight policy of 36 pounds. Test the claim that the average weekly household garbage weight is less than the company's weekly maximum. Use a 5% level of significance. Assume the population or garbage weights are approximately normally distributed.

Weight				
34.5	32.9	42.9	32.9	31.8
40	33.8	35.8	35.4	30.5
31.4	39.2	26.8	30.6	34.5
34.7	32.8			

$H_0: \mu = 36$, $H_1: \mu < 36$

t-Test: Paired Two Sample for Means

	Weight	*Dummy*
Mean	34.1471	0
Variance	14.9514	0
Observations	17	17
Pearson Correlation	#DIV/0!	
Hypothesized Mean Difference	36	
df	16	
t Stat	-1.9758	
P(T<=t) one-tail	0.0328	
t Critical one-tail	1.7459	
P(T<=t) two-tail	0.0657	
t Critical two-tail	2.1199	



$t = \frac{34.1471-36}{\left(3.8667/\sqrt{17}\right)} = -1.9758$ p-value = 0.0438

p-value = 0.0438 < α = 0.05 Reject H_0

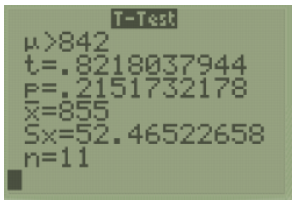
There is enough evidence to support the claim the average weekly household garbage weight is less than the company's weekly 36 lb. maximum.

39. The average number of calories from a fast-food meal for adults in the United States is 842 calories. A nutritionist believes that the average is higher than reported. They sample 11 meals that adults ordered and measure the calories for each meal shown below. Test the claim using a 5% level of significance. Assume that fast food calories are normally distributed.

Calories	855	854	785	854	952	860	853	760	862	851	919

$H_0: \mu = 842$ $H_1: \mu > 842$

$t = \frac{855-842}{\left(52.465227/\sqrt{11}\right)} = 0.8218$ p-value = 0.2152



t-Test: Paired Two Sample for Means

	Calories	*Dummy*
Mean	855	0
Variance	2752.6	0
Observations	11	11
Pearson Correlation	#DIV/0!	
Hypothesized Mean Difference	842	

df	10
t Stat	0.8218
P(T<=t) one-tail	0.2152
t Critical one-tail	1.8125
P(T<=t) two-tail	0.4303
t Critical two-tail	2.2281

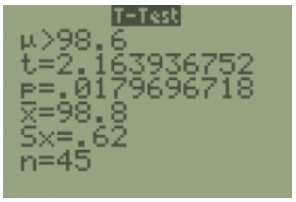
p-value = 0.2152 > α = 0.05 Do not reject H_0

We do not have evidence to support the claim the average calories from a fast food meal is higher than reported.

41. A sample of 45 body temperatures of athletes had a mean of 98.8˚F and a standard deviation of 0.62˚F. Test the claim that the mean body temperature for all athletes is more than 98.6˚F. Use a 1% level of significance.

H_0: μ ≤ 98.6

H_1: μ > 98.6 (claim)



$$t = \frac{\bar{x}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{98.8-98.6}{\left(\frac{0.62}{\sqrt{45}}\right)} = 2.1639$$

p-value = 0.018

The p-value is greater than α = 0.01, therefore the decision is to not reject H_0.

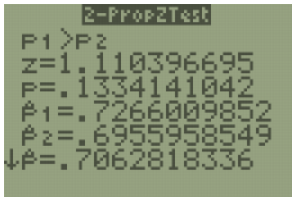
Summary: There is not enough evidence to support the claim that the mean body temperature for all athletes is more than 98.6˚F

Chapter 8 Exercises

For exercises 1-5, show all 5 steps for hypothesis testing:

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

1. A random sample of 406 college freshman found that 295 bought most of their textbooks from the college's bookstore. A random sample of 772 college seniors found that 537 bought their textbooks from the college's bookstore. You wish to test the claim that the proportion of all freshman that purchase most of their textbooks from the college's bookstore is greater than the proportion of all seniors at a significance level of $\alpha = 0.01$.



$H_0: p_1 = p_2$, $H_1: p_1 > p_2$

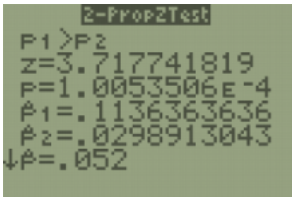
$\hat{p} = \frac{(x_1+x_2)}{(n_1+n_2)} = \frac{(295+537)}{(406+772)} = 0.706282$; $\hat{q} = 1 - 0.706282 = 0.293718$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{295}{406} = 0.726601$, $\hat{p}_2 = \frac{x_2}{n_2} = \frac{11}{368} = 0.695596$

Test Statistic $z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\left(\hat{p}\cdot\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)}} = \frac{(0.726601-0.695596)}{\sqrt{\left(0.706282*0.293718*\left(\frac{1}{406}+\frac{1}{772}\right)\right)}} = 1.1104$;

p-value = 0.1334; fail to reject H_0; There is not enough evidence to support the claim that the proportion of all freshman that purchase most of their textbooks from the college's bookstore is greater than the proportion of all seniors.

3. TDaP is a booster shot that prevents Diphtheria, Tetanus, and Pertussis in adults and adolescents. The shot should be administered every 8 years in order for it to remain effective. A random sample of 500 people living in a town that experienced a pertussis outbreak this year were divided into two groups. Group 1 was made up of 132 individuals who had not had the TDaP booster in the past 8 years, and Group 2 consisted of 368 individuals who had. In Group 1, 15 individuals caught pertussis during the outbreak, and in Group 2, 11 individuals caught pertussis. Is there evidence to suggest that the proportion of individuals who caught pertussis and were not up to date on their booster shot is significantly higher than those that were? Test at the 0.05 level of significance.



$H_0: p_1 = p_2$, $H_1: p_1 > p_2$

$\hat{p} = \frac{(x_1+x_2)}{(n_1+n_2)} = \frac{(15+11)}{(132+368)} = 0.052$; $\hat{q} = 1 - 0.052 = 0.948$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{15}{132} = 0.113636$, $\hat{p}_2 = \frac{x_2}{n_2} = \frac{11}{368} = 0.029891$

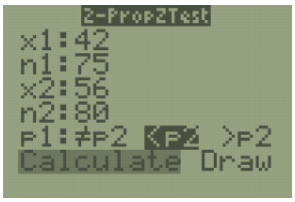
Test Statistic $z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\left(\hat{p}\cdot\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)}} = \frac{(0.113636-0.029891)}{\sqrt{\left(0.052*0.948*\left(\frac{1}{132}+\frac{1}{368}\right)\right)}} = 3.717742$

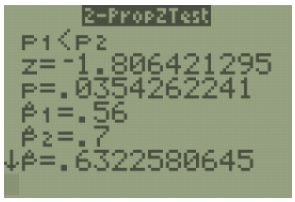
p-value = 0.00010051

Reject H_0. Yes, there is evidence that the proportion of those who caught pertussis is higher for those who were not up to date on their booster.

5. The makers of a smartphone have received complaints that the facial recognition tool often does not work, or takes multiple attempts to finally unlock the phone. The company upgraded to a new version and are claiming the tool has improved. To test the claim, a critic takes a random sample of 75 users of the old version (Group 1) and 80 users of the new version (Group 2). They find that the facial recognition tool works on the first try 56% of the time in the old version and 70% of the time in the new version. Can it be concluded that the new version is performing better? Test at α=0.10.

H_0: $p_1 = p_2$, H_1: $p_1 < p_2$
z = -1.8064; p-value = 0.0354
Do not reject H_0. There is not enough evidence that the new facial recognition tool is performing better.



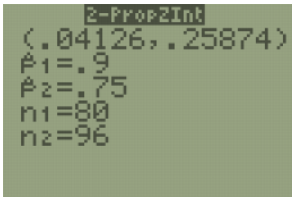


7. In a sample of 80 faculty from Portland State University, it was found that 90% were union members, while in a sample of 96 faculty at University of Oregon, 75% were union members. Find the 95% confidence interval for the difference in the proportions of faculty that belong to the union for the two universities.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}\right)}$$

$$(0.9 - 0.75) \pm 1.96\sqrt{\left(\frac{0.9 \cdot 0.1}{80} + \frac{0.75 \cdot 0.25}{96}\right)}$$

$0.04126 < p_1 - p_2 < 0.25874$



For exercises 8-9, show all 5 steps for hypothesis testing:

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

9. An adviser is testing out a new online learning module for a placement test. Test the claim that on average the new online learning module increased placement scores at a significance level of $\alpha = 0.05$. For the context of this problem, $\mu_D = \mu_{Before} - \mu_{After}$ where the first data set represents the after test scores and the second data set represents before test scores. Assume the population is normally distributed. You obtain the following paired sample of 19 students that took the placement test before and after the learning module.

Before	After
55.8	57.1
51.7	58.3
76.6	83.6
47.5	49.5
48.6	51.1

Before	After
30.6	35.2
53	46.7
21	22.5
58.5	47.7
42.6	51.5

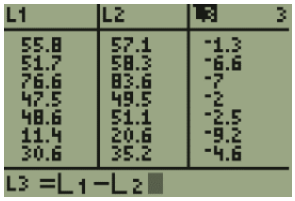
Before	After
26.8	28.6
11.4	14.5
56.3	43.7
46.1	57
72.8	66.1

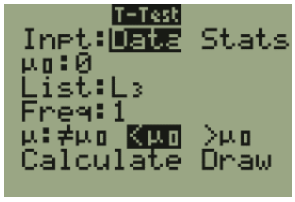
11.4	20.6	61.2	76.6	42.2	38.1
				51.3	42.4

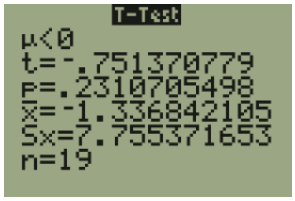
a) State the hypotheses. Before < After $H_0: \mu_D = 0$, $H_1: \mu_D < 0$

b) Compute the test statistic. $t = \frac{\bar{d}-\mu_d}{\left(s_d/\sqrt{n}\right)} = \frac{-1.3368-0}{\left(7.7553/\sqrt{19}\right)} = -0.7514$

c) Compute the p-value.







t-Test: Paired Two Sample for Means

	Before	*After*
Mean	45.5474	46.8842
Variance	333.8804	327.577
Observations	19	19
Pearson Correlation	0.9091	
Hypothesized Mean Difference	0	
df	18	
t Stat	-0.7514	
P(T<=t) one-tail	0.2311	
t Critical one-tail	1.7341	
P(T<=t) two-tail	0.4621	
t Critical two-tail	2.1009	

p-value = 0.2311

d) State the decision. Do not reject H_0

e) Write a summary. There is not enough evidence to support the claim on average the new online learning module increased placement scores.

11. Doctors developed an intensive intervention program for obese patients with heart disease. Subjects with a BMI of 30 kg/m² or more with heart disease were assigned to a three-month lifestyle change of diet and exercise. Patients' Left Ventricle Ejection Fraction (LVEF) are measured before and after intervention. Assume that LVEF measurements are normally distributed.

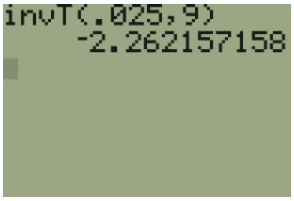
Before	**After**
44	56
49	58
50	64
49	60
57	63
62	71
39	49
41	51
52	60
42	55

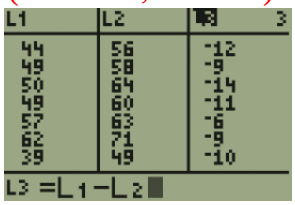
a) Find the 95% confidence interval for the mean of the differences.

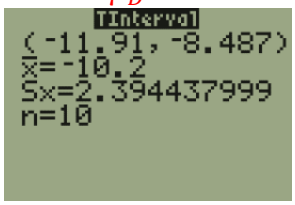
$\bar{D} \pm t_{\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right)$ where $t_{\alpha/2}$= invT(0.025,9) = −2.262157

$-10.2 \pm 2.262157\left(\frac{2.394438}{\sqrt{10}}\right)$ -10.2 ± 1.782878

Use interval notation (−11.9129, −8.4871) or standard notation $-11.9129 < \mu_D < -8.4871$







b) Using the confidence interval answer, did the intensive intervention program significantly increase the mean LVEF? Explain why.

Yes, since $\mu_D = 0$ is not captured in the interval (−11.9129, −8.4871).

13. A researcher is testing reaction times between the dominant and non-dominant hand. They randomly start with different hands for 20 subjects and their reaction times for both hands is recorded in milliseconds. Use the following computer output to test to see if the reaction time is faster for the dominant hand using a 5% level of significance.

t-Test: Paired Two Sample for Means		
	Non-Dominant	*Dominant*
Mean	63.33	56.28
Variance	218.9643158	128.7522105
Observations	20	20
Pearson Correlation	0.9067	
Hypothesized Mean Difference	0	
df	19	
t Stat	4.7951	
P(T<=t) one-tail	0.0001	
t Critical one-tail	1.7291	
P(T<=t) two-tail	0.0001	
t Critical two-tail	2.0930	

a) State the hypotheses. Non-Dominant > Dominant $H_0: \mu_D = 0$, $H_1: \mu_D > 0$

b) Compute the test statistic. t Stat = t = 4.7951

c) Compute p-value. P(T<=t) one-tail = 0.0001

d) State the decision. Reject H_0, since the p-value < α.

e) Write a summary. There is enough evidence to support the claim that the mean reaction time is significantly faster for a person's dominant hand.

For exercises 15-26, show all 5 steps for hypothesis testing:

a) State the hypotheses.

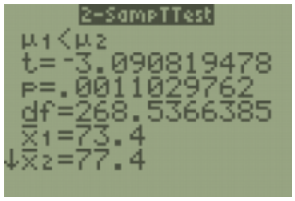
b) Compute the test statistic.

c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

15. A liberal arts college in New Hampshire implemented an online homework system for their introductory math courses and wanted to know whether the system improved test scores. In the Fall semester, homework was completed with pencil and paper, checking answers in the back of the book. In the Spring semester, homework was completed online – giving students instant feedback on their work. The results are summarized below. Is there evidence to suggest that the online system improves test scores? Use $\alpha = 0.05$. Assume the population variances are unequal.

	Fall Semester	**Spring Semester**
Number of Students	127	144
Mean Test Score	73.4	77.4
Sample Standard Deviation	10.2	11.1

Since the populations are independent, use the 2-Sample T-Test.



Fall Semester Score < Spring Semester Score

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 < \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{(73.4 - 77.4) - 0}{\sqrt{\left(\frac{10.2^2}{127} + \frac{11.1^2}{144}\right)}} = -3.0908$$

p-value = 0.0011

Reject H_0, since the p-value < α.

There is enough evidence to support the claim that the online homework system for introductory math courses improved student's average test scores.

17. In Major League Baseball, the American League (AL) allows a designated hitter (DH) to bat in place of the pitcher, but in the National League (NL), the pitcher has to bat. However, when an AL team is the visiting team for a game against an NL team, the AL team must abide by the home team's rules and thus, the pitcher must bat. A researcher is curious if an AL team would score more runs for games in which the DH was used. She samples 20 games for an AL team for which the DH was used, and 20 games for which there was no DH. The data are below. Assume the population is normally distributed with unequal population variances. Is there evidence to suggest that AL team would score more runs for games in which the DH was used? Use $\alpha = 0.10$.

With Designated Hitter			
0	5	4	7
1	2	7	6
6	4	2	10
1	2	7	5
8	4	11	0

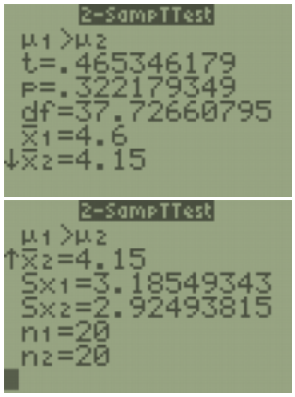
Without Designated Hitter			
3	6	5	2
12	4	0	1
6	3	7	8
4	0	5	1
2	4	6	4

Since the populations are independent, use the 2-Sample T-Test.

With DH > Without DH

4H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 > \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{(4.6 - 4.15) - 0}{\sqrt{\left(\frac{3.185493^2}{20} + \frac{2.924938^2}{20}\right)}} = 0.4653$$



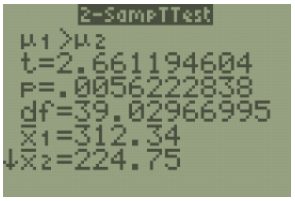
p-value = 0.3222

Do not reject H_0, since the p-value > α.

There is not enough evidence to support the claim that the American League team would score on average more runs for games in which the designated hitter was used.

19. A national food product company believes that it sells more frozen pizza during the winter months than during the summer months. Weekly samples of sales found the following statistics in volume of sales (in hundreds of pounds). Use α = 0.10. Use the p-value method to test the company's claim. Assume the populations are approximately normally distributed with equal variances

Season	n	$\bar{x}$	s
Winter	24	312.34	135
Summer	22	224.75	84.42



H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 > \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{312.34 - 224.75}{\sqrt{\left(\frac{135^2}{24} + \frac{84.42^2}{22}\right)}} = 2.6612 \qquad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2\left(\frac{1}{n_2 - 1}\right)} = 39$$

p-value = 0.0056 Reject H_0

There is enough evidence to support the claim that the mean number of frozen pizzas sold during the winter months is more than during the summer months.

21. "Durable press" cotton fabrics are treated to improve their recovery from wrinkles after washing. "Wrinkle recovery angle" measures how well a fabric recovers from wrinkles. Higher scores are better. Here are data on the wrinkle recovery angle (in degrees) for a random sample of fabric specimens. Assume the populations are approximately normally distributed with unequal variances. A manufacturer believes that the mean wrinkle recovery angle for Hylite is better. A random sample of 20 Permafresh (group 1) and 25 Hylite (group 2) were measured. Test the claim using a 10% level of significance.

Permafresh			
124	139	164	142
144	102	131	118
136	127	137	148
117	137	147	129
133	137	148	135

Hylite				
139	146	139	139	146
131	138	138	132	142
133	142	138	137	134
146	137	138	138	133
139	140	141	140	141

H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 < \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{134.75 - 138.68}{\sqrt{\left(\frac{180.197368}{20} + \frac{16.476667}{25}\right)}} = -1.2639$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)\right)} = 22$$

t-Test: Two-Sample Assuming Unequal Variances

	Permafresh	*Hylite*
Mean	134.75	138.68
Variance	180.1973684	16.47666667
Observations	20	25
Hypothesized Mean Difference	0	
df	22	
t Stat	-1.263872394	
P(T<=t) one-tail	0.109752065	
t Critical one-tail	1.321236742	
P(T<=t) two-tail	0.219504131	
t Critical two-tail	1.717144374	

p-value = 0.1098 Do not reject H_0

There is not enough evidence to support the claim that the mean wrinkle recovery angle for Hylite is better than Permafresh.

23. A new over-the-counter medicine to treat a sore throat is to be tested for effectiveness. The makers of the medicine take two random samples of 25 individuals showing symptoms of a sore throat. Group 1 receives the new medicine and Group 2 receives a placebo. After a few days on the medicine, each group is interviewed and asked how they would rate their comfort level 1-10 (1 being the most uncomfortable and 10 being no discomfort at all). The results are below. Is there sufficient evidence to conclude the mean scores from Group 1 is more than Group 2? Test at $\alpha = 0.01$. Assume the populations are normally distributed and have unequal variances.

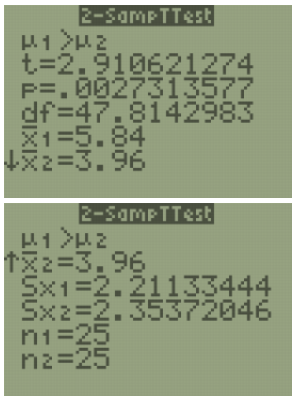
Group 1				
3	5	6	7	5
3	4	5	7	7
3	2	5	8	8
7	7	8	4	8
4	8	3	9	10

Group 2				
4	5	8	3	5
2	7	8	2	4
1	2	2	3	2
1	3	5	5	1
6	4	7	8	1

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{5.84 - 3.96}{\sqrt{\left(\frac{2.21133^2}{25} + \frac{2.35372^2}{25}\right)}} = 2.9106$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2\left(\frac{1}{n_2 - 1}\right)} = 47.814$$



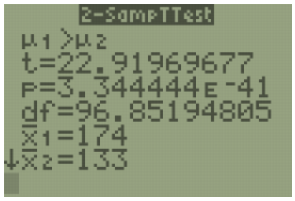
p-value = 0.0027 Reject H_0

There is enough evidence to support the claim that the new medicine is effective.

25. In a random sample of 60 pregnant women with preeclampsia, their systolic blood pressure was taken right before beginning to push during labor. The mean systolic blood pressure was 174 with a standard deviation of 12. In another random sample of 80 pregnant women without preeclampsia, there was a mean systolic blood pressure of 133 and a standard deviation of 8 when the blood pressure was also taken right before beginning to push. Is there sufficient evidence to conclude that women with preeclampsia have a higher mean blood pressure in the late stages of labor? Test at the 0.01 level of significance. Assume the population variances are unequal.

H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$; $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{174 - 133}{\sqrt{\left(\frac{12^2}{60} + \frac{8^2}{80}\right)}} = 22.9197$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2\left(\frac{1}{n_2 - 1}\right)} = 96.8519$$



p-value = 3.34E-41 Reject H_0

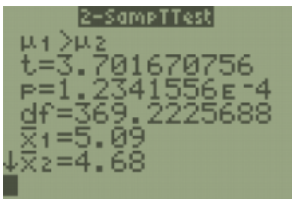
There is enough evidence to support the claim that women with preeclampsia have a higher mean blood pressure in the late stages of labor.

27. In a study that followed a group of students who graduated from high school in 2015, each was monitored in progress made toward earning a bachelor's degree. The group was divided in two – those who started at community college and later transferred to a four-year college, and those that started out in a four-year college as freshmen. That data below summarizes the findings. Is there evidence to suggest that community college transfer students take longer to earn a bachelor's degree? Use $\alpha = 0.05$. Assume the population variances are unequal.

	Community College Transfers	**Non-Transfers**
Number of Students	317	1,297
Mean Time to Graduate *(in years)*	5.09	4.68
Sample Standard Deviation	1.896	1.097

Since the populations are independent, use the 2-Sample T-Test.
CC Time > Non-Transfer Time
$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{(5.09 - 4.68) - 0}{\sqrt{\left(\frac{1.896^2}{317} + \frac{1.097^2}{1297}\right)}} = 3.7017$$



p-value = 0.000123 Reject H_0, since the p-value < α.
There is enough evidence to support the claim that community college transfer students take longer to earn a bachelor's degree.

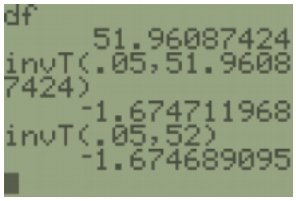
29. A researcher is curious what year in college students make use of the gym at a university. They take a random sample of 30 days and count the number of sophomores and seniors who use the gym each day. Is there evidence to suggest that a difference exists in gym usage based on year in college? Construct a 90% confidence interval for the data below to decide. Assume the population variances are unequal.

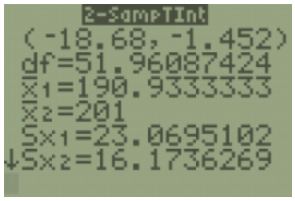
Sophomores				
189	203	167	154	217
209	198	143	208	220
188	197	165	207	231
201	177	186	193	201
190	165	180	245	200
199	155	165	188	187

Seniors				
209	199	186	210	221
204	214	230	170	197
190	201	165	183	235
187	199	189	194	197
192	195	200	211	205
200	190	218	210	229

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2\left(\frac{1}{n_2 - 1}\right)} = 51.96087$$



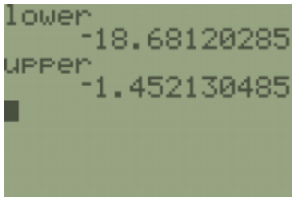


$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$(190.9333 - 201) \pm -1.6747\sqrt{\left(\frac{532.2023}{30} + \frac{261.5862}{30}\right)}$$

$$-10.06667 \pm -8.614536$$

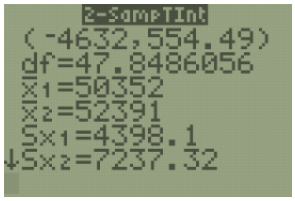
$$-18.6812 < \mu_1 - \mu_2 < -1.4521$$

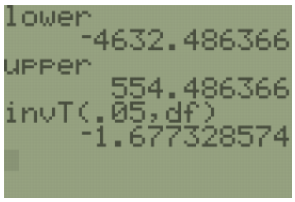


Reject H_0
There is enough evidence to support the claim that there is a difference in the average gym usage of sophomores and senior college students.

31. An employee at a large company is told that the mean starting salary at her company differs based on level of experience. The employee is skeptical and randomly samples 30 new employees with less than 5 years of experience and categorizes them as Group 1 and 30 new employees with 5 years of experience or more and categorizes them as Group 2. In Group 1,

she finds the sample mean starting salary to be $50,352 with a standard deviation of $4,398.10. Group 2 has a sample mean starting salary of $52,391 with a standard deviation of $7,237.32. Compute the 90% confidence interval for the difference in the mean starting salaries. Assume the populations are normally distributed with unequal variances.





$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$(50352 - 52391) \pm -1.677328574\sqrt{\left(\frac{4398.1^2}{30} + \frac{7237.32^2}{30}\right)}$$

$$-4632.49 < \mu_1 - \mu_2 < 554.49$$

Since $\mu_1 - \mu_2 = 0$ is between the endpoints of the confidence interval, we fail to reject H_0.
There is not enough evidence to support the claim that the mean starting salary at her company differs based on level of experience

For exercises 33-39, show all 5 steps for hypothesis testing:

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

33. A professor wants to know if there is a difference in comprehension of a lab assignment among students depending if the instructions are given all in text, or if they are given primarily with visual illustrations. She randomly divides her class into two groups of 15, gives one group instructions in text and the second group instructions with visual illustrations. The following data summarizes the scores the students received on a test given after the lab. Assume are populations are normally distributed with equal variances. Is there evidence to suggest that a difference exists in the comprehension of the lab based on the test scores? Use $\alpha = 0.10$.

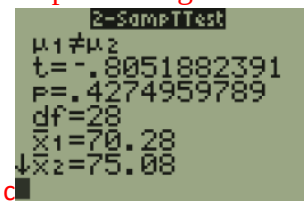
Text		
57.3	87.3	67.2
45.3	75.2	54.4
87.1	88.2	93.0
61.2	67.5	89.2
43.1	86.2	52.0

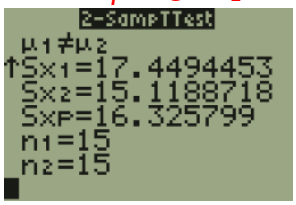
Visual Illustrations		
59.0	76.7	88.2
57.6	78.2	43.8
72.9	64.4	97.1
83.2	89.0	95.1
64.0	72.9	84.1

Since the population variances are unknown and equal, use the pooled 2-Sample T-Test.
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(70.28 - 75.08)}{\sqrt{\left(\left(\frac{(14 \cdot 17.449445^2 + 14 \cdot 15.118872^2)}{28}\right) \cdot \left(\left(\frac{1}{15}\right)+\left(\frac{1}{15}\right)\right)\right)}} = -0.8052$$

Use the t-distribution with pooled degrees of freedom $df = n_1 + n_2 - 2 = 28$.





p-value = 0.4275 Do not reject H_0, since the p-value > α.

There is not enough evidence to support the claim that there is a statistically significant difference in the mean comprehension score between text and visual illustrations.

35. The CEO of a large manufacturing company is curious if there is a difference in productivity level of her warehouse employees based on the region of the country the warehouse is located in. She randomly selects 35 employees who work in warehouses on the East Coast and 35 employees who work in warehouses in the Midwest and records the number of parts shipped out from each for a week. She finds that East Coast group ships an average of 1,287 parts and a standard deviation of 348. The Midwest group ships an average of 1,449 parts and a standard deviation of 298. Using a 0.01 level of significance, test if there is a difference in productivity level. Assume the population variances are equal.

Since the population variances are unknown and equal, use the pooled 2-Sample T-Test.

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

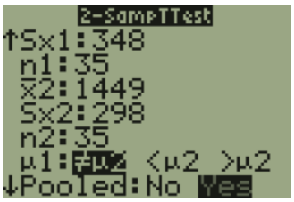
Test Statistic is $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(1287 - 1449)}{\sqrt{\left(\left(\frac{(34 \cdot 348^2 + 34 \cdot 298^2)}{68}\right) \cdot \left(\left(\frac{1}{35}\right)+\left(\frac{1}{35}\right)\right)\right)}} = -2.0919$

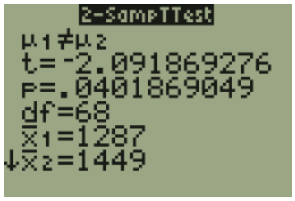
Use the t-distribution with pooled degrees of freedom $df = n_1 + n_2 - 2 = 68$.

p-value = 0.0402

Do not reject H_0, since the p-value > α.

There is not enough evidence to support the claim that there is a statistically significant difference in the mean productivity level between the two locations.





37. A large shoe company is interested in knowing if the amount of money a customer is willing to pay on a pair of shoes is different depending on location. They take a random sample of 50 single-pair purchases from Southern states and another random sample of 50 single-pair purchases from Midwestern states and record the cost for each. The results can be found below. At the 0.05 level of significance, is there evidence that the mean cost differs between the Midwest and the South? Assume the population variances are equal.

Midwest (cost in dollars)				
70	43	21	62	45
60	23	15	37	66
65	38	30	46	64
71	54	51	61	82
33	79	28	84	63
68	72	78	43	84
78	80	24	80	16
82	45	38	84	84
73	23	36	69	78
76	71	38	46	18

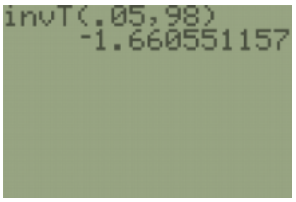
South (cost in dollars)				
73	75	34	59	81
80	17	18	65	27
62	32	60	60	56
30	28	31	17	34
36	54	48	85	54
46	55	30	41	53
80	16	67	36	39
22	16	38	46	50
16	49	43	54	27
83	50	57	51	68

t-Test: Two-Sample Assuming Equal Variances

	Midwest	*South*
Mean	55.5	46.98
Variance	480.9489796	388.1832653
Observations	50	50
Pooled Variance	434.5661224	
Hypothesized Mean Difference	0	
df	98	
t Stat	2.043533049	
P(T<=t) one-tail	0.021841632	
t Critical one-tail	1.660551217	
P(T<=t) two-tail	0.043683264	
t Critical two-tail	1.984467455	

H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$; t = 2.0435; p-value = 0.0437; reject H_0; There is enough evidence to support the claim that the mean cost for a pair of shoes in the Midwest and the South are different.

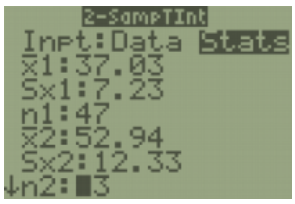
39. In a random sample of 100 college students, 47 were sophomores and 53 were seniors. The sophomores reported spending an average of \$37.03 per week going out for food and drinks with a standard deviation of \$7.23, while the seniors reported spending an average of \$52.94 per week going out for food and drinks with a standard deviation of \$12.33. Find the 90% confidence interval for difference in the mean amount spent on food and drinks between sophomores and seniors? Assume the population variances are equal.

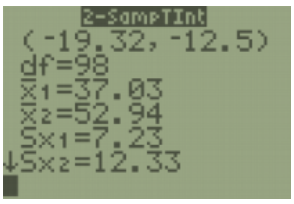


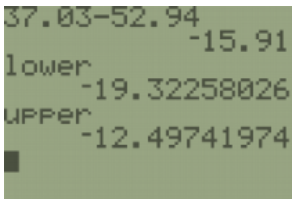
$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$(37.03 - 52.94) \pm -1.6605512\sqrt{\left(\frac{47*7.23^2+53*12.33^2}{98}\right)\left(\frac{1}{47}+\frac{1}{53}\right)}$$

$$-19.3226 < \mu_1 - \mu_2 < -12.4974$$







41. Two groups of students are given a problem-solving test, and the results are compared. Assume the populations are normally distributed with equal variances. Compute the 95% confidence interval for the difference of the mean scores.

Mathematics Majors	**Computer Science Majors**
$\bar{x}_1 = 83.6$	$\bar{x}_2 = 79.2$
$s_1 = 4.3$	$s_2 = 3.8$
$n_1 = 16$	$n_2 = 20$

Math Majors ≠ CS Majors

H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$

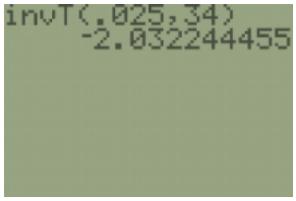
$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

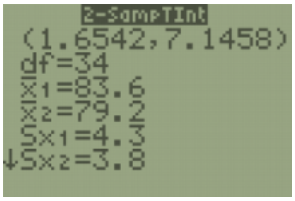
$$(83.6 - 79.2) \pm -2.032244455\sqrt{\left(\frac{15*4.3^2+19*3.8^2}{34}\right)\left(\frac{1}{16}+\frac{1}{20}\right)}$$

$$4.4 \pm -2.745795$$

$$1.6542 < \mu_1 - \mu_2 < 7.1458$$

Since $\mu_1 - \mu_2 = 0$ is not between the endpoints of the confidence interval, we reject H_0.





Reject H_0.

There is enough evidence to support the claim that there is a statistically significant difference in the mean scores on the problem-solving test for mathematics and computer science majors.

Chapter 9 Exercises

1. The shape of the χ^2 Distribution is usually: Answer d)
   a) Normal
   b) Bell-shaped
   c) Skewed Left
   d) Skewed Right
   e) Uniform

3. What are the requirements to be satisfied before using a Goodness of Fit test? Check all that apply. Answers b and c
   a) The data are obtained using systematic sampling.
   b) The data are obtained from a simple random sample.
   c) The expected frequency from each category is 5 or more.
   d) The observed frequency from each category is organized from largest to smallest.
   e) The degrees of freedom are less than 30.

5. Calculate the critical value for a right-tailed test using a χ^2-distribution with $\alpha = 0.05$ and $df = 17$.
   =CHISQ.INV.RT(0.05,17) = 27.5871

7. Calculate the critical value for a right-tailed test using a χ^2-distribution with $\alpha = 0.10$ and $df = 7$.
   =CHISQ.INV.RT(0.10,7) = 12.017

9. What is the mean of a Chi Square distribution with 6 degrees of freedom? 6

11. Suppose you are trying to determine whether a 20-sided dice is fair or if it tends to land on certain numbers more than others. You roll the dice 200 times, what is the expected value for each of the 20 sides?
    1/20 = 0.05, take n*p = 200*0.05 = 10

For exercises 12-27, show all 5 steps for hypothesis testing:

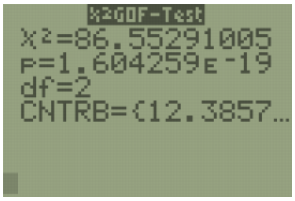
a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

13. A professor using an open-source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the term she asks her students to complete a survey where they indicate what format of the book they used. Of the 126 students, 45 said they bought a

hard copy of the book, 25 said they printed it out from the web, and 56 said they read it online. Run a Goodness of Fit test at α = 0.05 to see if the distribution is different than expected.

Format	HC	Print	Online	Total
Observed	45	25	56	126
Expected	126*0.6 = 75.6	126*0.25 = 31.5	126*0.15 = 18.9	126
$\frac{(O-E)^2}{E}$	$\frac{(45-75.6)^2}{75.6}$ = 12.3857	$\frac{(25-31.5)^2}{31.5}$ = 1.3413	$\frac{(56-18.9)^2}{18.9}$ = 72.8259	86.5529

H_0: $p_1 = 0.6$, $p_2 = 0.25$, $p_3 = 0.15$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum\frac{(O-E)^2}{E}$ = 86.5529; p-value = 1.604E-19 = 0
Reject H_0. There is enough evidence to support the claim that the distribution is different than expected. There were more students than expected that would read the text online.



15. You might think that if you looked at the first digit in randomly selected numbers that the distribution would be uniform. Actually, it is not! Simon Newcomb and later Frank Benford both discovered that the digits occur according to the following distribution.

Digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

A forensic accountant can use Benford's Law to detect fraudulent tax data. Suppose you work for the IRS and are investigating an individual suspected of embezzling. The first digit of 192 checks to a supposed company are as follows.

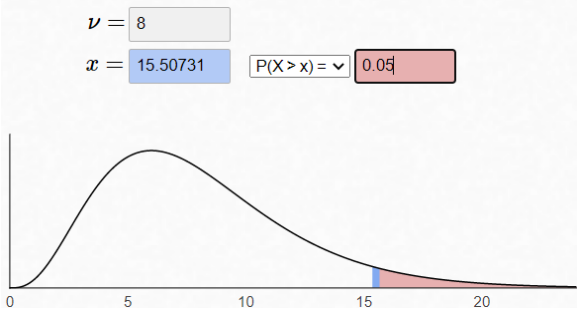
Digit	1	2	3	4	5	6	7	8	9
Observed Frequency	56	23	19	20	16	19	17	10	12

Run a complete Goodness of Fit test to see if the individual is likely to have committed tax fraud. Use α = 0.05. Should law enforcement officials pursue the case? Explain.

H_0: $p_1 = 0.301$, $p_2 = 0.176$, $p_3 = 0.125$, $p_4 = 0.097$, $p_5 = 0.079$, $p_6 = 0.067$, $p_7 = 0.058$, $p_8 = 0.051$, $p_9 = 0.046$
H_1: At least one proportion is different.

Digit	1	2	3	4	5	6	7	8	9	Total
O	56	23	19	20	16	19	17	10	12	192
E	57.792	33.792	24	18.624	15.168	12.864	11.136	9.792	8.832	192
$\frac{(O-E)^2}{E}$	0.0556	3.4466	1.0417	0.1017	0.0456	2.9268	3.0879	0.0044	1.1363	11.8466

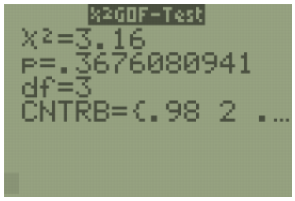
Test Statistic is $\chi^2 = \sum\frac{(O-E)^2}{E}$ = 11.8466 CV = CHISQ.INV.RT(0.05,8) = 15.5073



Do not reject H_0. There is no evidence of tax fraud so law enforcement officials should not pursue the case.

17. Consumer panel preferences for four store displays follow. Test to see whether there is a preference among the four display designs. Use $\alpha = 0.05$.

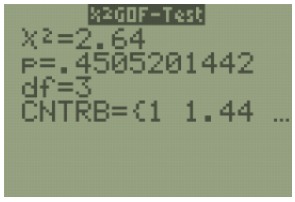
Display	A	B	C	D	Total
Observed	43	60	47	50	200
Expected	0.25*200 = 50	50	50	50	200
$\frac{(O-E)^2}{E}$	0.98	2	0.18	0	3.16



H_0: $p_1 = 0.25$, $p_2 = 0.25$, $p_3 = 0.25$, $p_4 = 0.25$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E} = 3.16$ p-value = 0.3676
Do not reject H_0. There is not enough evidence to support the claim that preference among the four display designs.

19. The director of a Driver's Ed program is curious if the time of year has an impact on number of car accidents in the United States. They assume that weather may have a significant impact on the ability of drivers to control their vehicles. They take a random sample of 100 car accidents and record the season each occurred in. They found that 20 occurred in the spring, 31 in the summer, 23 in the fall, and 26 in the winter. Can it be concluded at the 0.05 level of significance that car accidents are not equally distributed throughout the year?

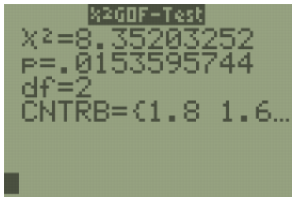
Season	Spring	Summer	Fall	Winter	Total
Observed	20	31	23	26	200
Expected	0.25*100 = 25	25	25	25	100
$\frac{(O-E)^2}{E}$	1	1.44	0.16	0.04	3.16



H_0: $p_1 = 0.25$, $p_2 = 0.25$, $p_3 = 0.25$, $p_4 = 0.25$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E} = 2.64$ p-value = 0.4505
Do not reject H_0. There is not enough evidence to support the claim that car accidents are not equally distributed throughout the year.

21. The permanent residence of adults aged 18-25 in the U.S. was examined in a survey from the year 2000. The survey revealed that 27% of these adults lived alone, 32% lived with a roommate(s), and 41% lived with their parents/guardians. In 2008, during an economic recession in the country, another such survey of 1,500 people revealed that 378 lived alone, 452 lived with a roommate(s), and 670 lived with their parents. Is there a significant difference in where young adults lived in 2000 versus 2008? Test with a Goodness of Fit test at $\alpha = 0.05$.

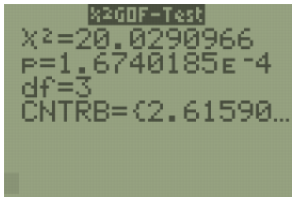
Residence	Alone	Roommate	Parent	Total
Observed	378	452	670	1500
Expected	1500*0.27 = 405	1500*0.32 = 480	1500*0.41 = 615	1500
$\frac{(O-E)^2}{E}$	1.8	1.63333	4.91869	8.352



H_0: $p_1 = 0.27$, $p_2 = 0.32$, $p_3 = 0.41$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E} = 8.352$ p-value = 0.0154
Reject H_0. There is enough evidence to support the claim that there a significant difference in where young adults lived in 2000 versus 2008. There are fewer young adults living at home than expected.

23. An urban economist is curious if the distribution in where Oregon residents live is different today than it was in 1990. She observes that today there are approximately 3,050 thousand residents in NW Oregon, 907 thousand residents in SW Oregon, 257 thousand in Central Oregon, and 106 thousand in Eastern Oregon. She knows that in 1990 the breakdown was as follows: 72.7% NW Oregon, 19.7% SW Oregon, 4.8% Central Oregon, and 2.8% Eastern Oregon. Can she conclude that the distribution in residence is different today at a 0.05 level of significance?

Area	NW	SW	C	E	Total
Observed	3050	907	257	106	4320
Expected	0.727*4320 = 3140.64	0.207*4320 = 851.04	0.048*4320 = 207.36	0.028*4320 = 120.96	4320
$\frac{(O-E)^2}{E}$	2.6159	3.6796	11.8833	1.8502	20.0291



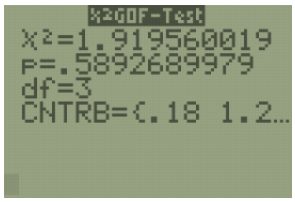
H_0: $p_1 = 0.727$, $p_2 = 0.197$, $p_3 = 0.048$, $p_4 = 0.028$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E} = 20.0291$; p-value = 0.0002
Reject H_0. There is enough evidence to support the claim that the distribution of Oregon residents is different now compared to 1990. There were more Oregonians in central Oregon than expected.

25. Students at a high school are asked to evaluate their experience in a class at the end of each school year. The courses are evaluated on a 1-4 scale − with 4 being the best experience possible. In the History Department, the courses typically are evaluated at 10% 1's, 15% 2's,

34% 3's, and 41% 4's. A new history teacher, Mr. Mendoza, sets a goal to outscore these numbers. At the end of the year, he takes a random sample of his evaluations and finds 11 1's, 14 2's, 47 3's, and 53 4's. At the 0.05 level of significance, can Mr. Mendoza claim that his evaluations are significantly different from the History Department's?

Score	1	2	3	4	Total
Observed	11	14	47	53	125
Expected	0.1*125 = 12.5	0.15*125 = 18.75	0.34*125 = 42.5	0.41*125 = 51.25	125
$\frac{(O-E)^2}{E}$	0.18	1.20333	0.47647	0.05975	1.9196

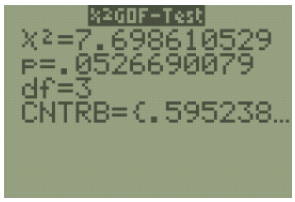
H_0: $p_1 = 0.1$, $p_2 = 0.15$, $p_3 = 0.34$, $p_4 = 0.41$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ = 1.9196 p-value = 0.5893
Do not reject H_0. There is not enough evidence to support the claim that the Mr. Mendoza's course evaluation scores are different compared to the rest of the History Department's evaluations.



27. A company that develops over-the-counter medicines is working on a new product that is meant to shorten the length of sore throats. To test their product for effectiveness, they take a random sample of 100 people and record how long it took for their symptoms to completely disappear. The results are in the table below. The company knows that on average (without medication) it takes a sore throat 6 days or less to heal 42% of the time, 7-9 days 31% of the time, 10-12 days 16% of the time, and 13 days or more 11% of the time. Can it be concluded at the 0.01 level of significance that the patients who took the medicine healed at a different rate than these percentages?

	6 days or less	7-9 days	10-12 days	13 or more days	Total
Duration of Sore Throat	47	38	10	5	100
Expected	42	31	16	11	100
$\frac{(O-E)^2}{E}$	0.5952	1.5806	2.25	3.2727	7.6986

H_0: $p_1 = 0.42$, $p_2 = 0.31$, $p_3 = 0.16$, $p_4 = 0.11$
H_1: At least one proportion is different.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ = 7.6986 p-value = 0.05267
Do not reject H_0. There is not enough evidence to support the claim that the patients who took the medicine healed at a different rate than these percentages.



29. The null hypothesis for the χ^2 Independence Test always states that _____.
    a) The two values are equal.
    b) One variable is dependent on another variable.
    c) One variable is independent of another variable.

d) The expected values and observed values are the same.

Answer: c

31. What are the degrees of freedom used in the χ^2 Independence Test?
    a) n − 1
    b) Rows + Columns
    c) n
    d) (Rows − 1)*(Columns − 1)
    e) n − 2

    Answer: d

For exercises 32-42, show all 5 steps for hypothesis testing:

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

33. A restaurant chain that has 3 locations in Portland is trying to determine which of their 3 locations they should keep open on New Year's Eve. They survey a random sample of customers at each location and ask each whether they plan to go out to eat on New Year's Eve. The results are below. Run a test for independence to decide if the proportion of customers who will go out to eat on New Year's Eve is dependent on location. Use α = 0.05.

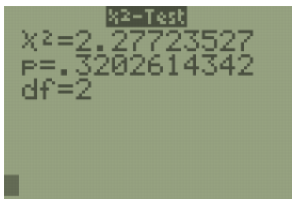
Observed	NW Location	NE Location	SE Location	Total
Will Go Out	45	33	36	114
Won't Go Out	23	29	25	77
Total	68	62	61	191

Expected	NW Location	NE Location	SE Location	Total
Will Go Out	40.5864	37.0052	36.4084	114
Won't Go Out	27.4136	24.9948	24.5916	77
Total	68	62	61	191

H_0: The proportion of customers who will go out to eat on New Year's Eve is independent of location.

H_1: The proportion of customers who will go out to eat on New Year's Eve is dependent on location.

Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ = 2.2772; p-value = 0.3203



Do not reject H_0. There is not enough evidence to support the claim that the proportion of customers who will go out to eat on New Year's Eve is dependent on location.

35. A high school offers math placement exams for incoming freshmen to place students into the appropriate math class during their freshman year. Three middle schools were sampled and the following pass/fail results were found. Test to see if the math placement exam and where students are placed are dependent at the 0.10 level of significance.

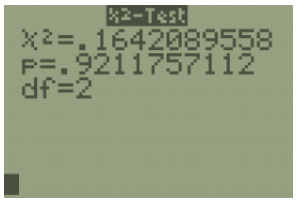
Observed	School A	School B	School C	Total
Pass	42	29	45	116
Fail	57	35	61	153
Total	99	64	106	269

Expected	School A	School B	School C	Total
Pass	42.69145	27.59851	45.71004	116
Fail	56.30855	36.401489	60.28996	153
Total	99	64	106	269

H_0: The math placement exam and where students are placed are independent.
H_1: The math placement exam and where students are placed are dependent.

Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ = 0.1642; p-value = 0.9212

Do not reject H_0. There is not enough evidence to support the claim that the math placement exam and where students are placed are dependent.

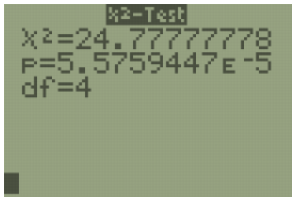


37. A university changed to a new learning management system (LMS) during the past school year. The school wants to find out how it is working for the different departments – the results in preference found from a survey are below. Test to see if the department and LMS preference are dependent at α = 0.05.

Observed	Prefers Old LMS	Prefers New LMS	No Preference	Total
School of Business	15	24	6	45
College of Liberal Arts & Science	34	7	19	60
College of Education	21	19	5	45
Total	70	50	30	150

Expected	Prefers Old LMS	Prefers New LMS	No Preference	Total
School of Business	21	15	9	45
College of Liberal Arts & Science	28	20	12	60
College of Education	21	15	9	45
Total	70	50	30	133

H_0: Department and LMS preference are independent.
H_1: Department and LMS preference are dependent.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ = 24.7778;
p-value = 0.000056
Reject H_0.
There is enough evidence to support the claim that department and LMS preference are dependent.

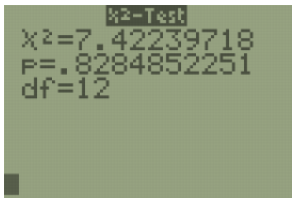


39. An electronics store has 4 branches in a large city. They are curious if sales in any particular department are different depending on location. They take a random sample of purchases throughout the 4 branches – the results are recorded below. Test to see if the type of electronic device and store branch are dependent at the 0.05 level of significance.

Observed	Appliances	TV	Computers	Cameras	Cellphones	Total
Branch 1	54	28	61	24	81	248
Branch 2	44	21	55	23	92	235
Branch 3	49	18	49	30	72	218
Branch 4	51	29	65	29	102	276
Total	198	96	230	106	347	977

Observed	App.	TV	Comp.	Cameras	Cellphones	Total
Branch 1	50.25998	24.36847	58.3828	26.90686	88.08188	248
Branch 2	47.62538	23.0911	55.32242	25.49642	83.46469	235
Branch 3	44.18014	21.42068	51.32037	23.652	77.42682	218
Branch 4	55.9345	27.11975	64.97441	29.94473	98.02661	276
Total	198	96	230	106	347	977

H_0: Type of electronic device and store branch are dependent.
H_1: Type of electronic device and store branch are dependent.
Test Statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ = 7.4224;
p-value = 0.8285
Do not reject H_0.
There is not enough evidence to support the claim that type of electronic device and store branch are dependent.



41. A manufacturing company knows that their machines produce parts that are defective on occasion. They have 4 machines producing parts and want to test if defective parts are dependent on the machine that produced it. They take a random sample of 300 parts and find the following results. Test at the 0.05 level of significance.

Observed	Machine 1	Machine 2	Machine 3	Machine 4	Total
Defective	9	12	15	6	42
Non-Defective	63	70	68	57	258
Total	72	82	83	63	300

Find the row and column totals. Compute the expected values by taking $\frac{\text{Row Total}\cdot\text{Column Total}}{\text{Grand Total}}$ for each of the 8 cells.

Expected	Machine 1	Machine 2	Machine 3	Machine 4	Total
Defective	10.08	11.48	11.62	8.82	42
Non-Defective	61.92	70.52	71.38	54.18	258
Total	72	82	83	63	300

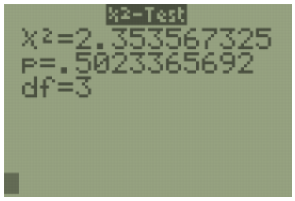
H_0: The number of defective parts is independent on the machine that produced it.
H_1: The number of defective parts is dependent on the machine that produced it.

Compute the test statistic.

$\frac{(O-E)^2}{E}$	Machine 1	Machine 2	Machine 3	Machine 4	Total
Defective	$\frac{(9-10.08)^2}{10.08} = 0.1157$	$\frac{(12-11.48)^2}{11.48} = 0.0236$	$\frac{(15-11.62)^2}{11.62} = 0.9832$	$\frac{(6-8.82)^2}{8.82} = 0.9016$	
Non-Defective	$\frac{(63-61.92)^2}{61.92} = 0.0188$	$\frac{(70-70.52)^2}{70.52} = 0.0038$	$\frac{(68-71.38)^2}{71.38} = 0.1601$	$\frac{(57-54.18)^2}{54.18} = 0.1468$	2.3536

Test Statistic is $\chi^2 = \sum\frac{(O-E)^2}{E} = 2.3536$ p-value = 0.5023

Do not reject H_0. There is not enough evidence to support the claim that the number of defective parts is dependent on the machine that produced it.



Chapter 10 Exercises

1. What is the primary purpose of conducting a one-way ANOVA? Answer b)

   a) To examine the relationship between two continuous variables.
   b) To compare the means of multiple independent groups.
   c) To analyze the difference between paired samples.
   d) To determine the correlation coefficient between two variables.

3. What does the acronym ANOVA stand for? Answer a)
   a) Analysis of Variance
   b) Analysis of Means
   c) Analyzing Various Means
   d) Anticipatory Nausea and Vomiting
   e) Average Noise Variance

5. What is the alternative hypothesis in a one-way ANOVA? Answer b)

   a) There is no difference between the group means.
   b) There is a difference between the group means.
   c) There is a linear relationship between the variables.
   d) There is no relationship between the variables.

7. Which assumption is required for conducting a one-way ANOVA? Answer d)

   a) The dependent variable is normally distributed.
   b) Equal variances across the groups.
   c) Independence of observations.
   d) All of the above

9. What is the purpose of post-hoc tests in a one-way ANOVA? Answer a)

   a) To identify the specific group means that are significantly different from each other.
   b) To determine the overall significance of the analysis.
   c) To estimate effect sizes for each group.
   d) To assess the normality assumption of the data.

11. What is the critical value for a right-tailed F-test with a 1% level of significance with $df_1 = 3$ and $df_2 = 55$? Round answer to 4 decimal places. =F.INV.RT(0.01,3,55) = 4.1591

13. What are the critical values for a two-tailed *F*-test with a 1% level of significance with $df_1 = 31$ and $df_2 = 10$? Round answer to 4 decimal places. =F.INV(0.005,31,10) = 0.3018

15. An ANOVA was run for the per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area. Assume tuition costs are normally distributed. At $\alpha = 0.05$, test to see if there is a difference in the means.

SUMMARY				
Groups	*n*	*Sum*	*Average*	*Variance*
Clackamas County	11	147215	13383.1818	36734231.36
Multnomah County	12	182365	15197.0833	33731956.63
Washington County	10	124555	12455.5	40409869.17

a) State the hypotheses.

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: At least one mean is different.

b) Fill out the ANOVA table to find the test statistic.

SSW = $\Sigma(n_i - 1)s_i^2$ = 10*36734231.36 + 11*33731956.63 + 9*40409869.17 = 1102082659.06
$\bar{x}_{GM} = \frac{\Sigma x_i}{N}$ = (147215 + 182365 + 124555)/33 = 13761.6666667
SSB = $\Sigma n_i(\bar{x}_i - \bar{x}_{GM})^2$ = 11(13383.1818 − 13761.6666667)² + 12(15197.0833 − 13761.6666667)² + 10(12455.5 − 13761.6666667)² = 43361523.2834
dfW = k − 1 = 2; dfB = N − k = 33 − 3 = 30; dfTotal = N − 1 = 32
MSB = SSB/dfB = 43361523.9/2 = 21680761.97
MSW = SSW/dfW = 1102082659/30 = 36736088.64
F = MSW/MSB = 21680761.97/36736088.64 = 0.590176112

ANOVA				
Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	43361523.2834	2	21680761.64	0.5901761
Within Groups	1102082659.06	30	36736088.64	
Total	1145444183.3434	32		

c) Find the p-value. =F.DIST.RT(0.5901761,2,30) = 0.5605
d) State the correct decision and summary. Do not reject H_0. There is not enough evidence to support the claim that there is a difference in the mean per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area.

17. What does the Bonferroni comparison test for? Answer c)
    a) The analysis of between and within variance.
    b) The difference between all the means at once.
    c) The difference between two pairs of mean.
    d) The sample size between the groups.

19. True or False: The Bonferroni post-hoc test compares each group mean to the grand mean. Answer: False.

21. When should you use a Bonferroni post-hoc test? Answer b)

    a) When the overall ANOVA test is not statistically significant.

b) When the overall ANOVA test is statistically significant.
c) When you want to compare group means to a control group.
d) When you have a large sample size.
e) When the sample sizes are unequal.

23. A manufacturing company wants to see if there is a significant difference in three types of plastic for a new product. They randomly sample prices for each of the three types of plastic and run an ANOVA. Use $\alpha = 0.05$ to see if there is a statistically significant difference in the mean prices. Part of the computer output is shown below.

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Plastic A	39	512	13.12821	15.48313
Plastic B	41	679	16.56098	1.302439
Plastic C	34	470	13.82353	22.08913

a) State the hypotheses.

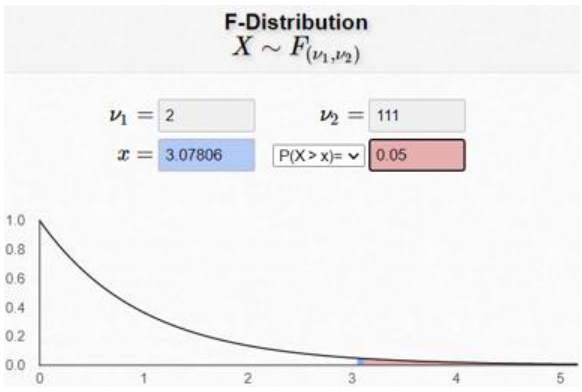
H_0: $\mu_A = \mu_B = \mu_C$
H_1: At least one mean is different.

b) Fill in the ANOVA table to find the test statistic.

$SSW = \sum(n_i - 1)s_i^2$ = 38*15.48313 + 40*1.302439 + 33*22.08913 = 1369.39779
$\bar{x}_{GM} = \frac{\sum x_i}{N}$ = (512 + 679 + 470)/114 = 14.570175
$SSB = \sum n_i(\bar{x}_i - \bar{x}_{GM})^2$ = 39(13.12821 − 14.570175)² + 41(16.56098 − 14.570175)² + 34(13.82353 − 14.570175)² = 262.5410236
dfW = k − 1 = 2; dfB = N − k = 114 − 3 = 111; dfTotal = N − 1 = 113
MSB = SSB/dfB = 262.5410236/2 = 131.2705118
MSW = SSW/dfW = 1369.39779/111 = 12.33691703
F = MSW/MSB = 131.2705118/12.33691703 = 10.64046

ANOVA				
Source of Variation	*SS*	*df*	*MS*	*F*
Between Groups	262.5410236	2	131.2705118	10.64046
Within Groups	1369.39779	111	12.33691703	
Total	1631.939	113		

c) Find the critical value. CV = F_α = 3.0781



d) State the decision and summary.

Reject H_0. There is enough evidence to support the claim that there is a difference in the mean price of the three types of plastic.

e) Which group(s) are significantly different based on the Bonferroni test?

Multiple Comparisons

Dependent Variable: Cost

Bonferroni

(I) Plastic Type	(J) Plastic Type	Mean Difference (i−j)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Plastic A	Plastic B	-3.43277*	.78564	.000	-5.3425	-1.5230
	Plastic C	-.69532	.82412	1.000	-2.6986	1.3080
Plastic B	Plastic A	3.43277*	.78564	.000	1.5230	5.3425
	Plastic C	2.73745*	.81471	.003	.7570	4.7178
Plastic C	Plastic A	.69532	.82412	1.000	-1.3080	2.6986
	Plastic B	-2.73745*	.81471	.003	-4.7178	-.7570

H_0: $\mu_A = \mu_B$
H_1: $\mu_A \neq \mu_B$
p-value = 0;
Reject H_0;
There is significant difference in the mean price between plastics A and B.

H_0: $\mu_A = \mu_C$
H_1: $\mu_A \neq \mu_C$
p-value = 1;
Do not reject H_0;
There is not a significant difference in the mean price between plastics A and C.

H_0: $\mu_B = \mu_C$

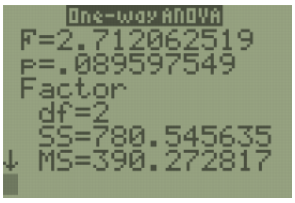
$H_1: \mu_B \neq \mu_C$
p-value = 0.003;
Reject H_0;
There is significant difference in the mean price between plastics B and C.

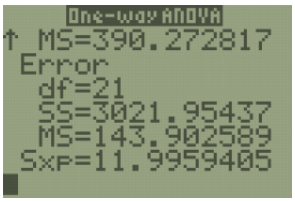
For exercises 25-32, Assume that all distributions are normal with equal population standard deviations, and the data was collected independently and randomly. Show all 5 steps for hypothesis testing. If there is a significant difference is found, run a Bonferroni test to see which means are different.

a) State the hypotheses.
b) Compute the test statistic.
c) Compute the critical value or p-value.
d) State the decision.
e) Write a summary.

25. Is a statistics class's delivery type a factor in how well students do on the final exam? The table below shows the average percent on final exams from several randomly selected classes that used the different delivery types. Assume that all distributions are normal with equal population standard deviations, and the data was collected independently and randomly. Use a level of significance of $\alpha = 0.10$.

Face-to-Face	Blended	Online
79	70	100
77	58	66
75	55	91
68	74	91
95	76	98
78	83	74
69	66	57
65		88
65		





$H_0: \mu_1 = \mu_2 = \mu_3$
H_1: At least one mean is different.

Source	SS	*df*	MS	F
Between	780.5456	2	390.2728	2.7121
Within	3021.9544	21	143.9026	
Total	3802.5	23		

$F = 2.7121$; p-value = 0.0896; Reject H_0. There is sufficient evidence to support the claim that course delivery type is a factor in final exam score.

27. The dependent variable is movie ticket prices, and the groups are the geographical regions where the theaters are located (suburban, rural, urban). A random sample of ticket prices were taken from randomly chosen states. Test to see if there is a significant difference in the means using $\alpha = 0.05$.

Suburb	Rural	Urban
11.25	11.75	11.25
11	9.5	11.25
11	11.25	12.25
12.25	10.5	9.75
11.25	10	10.75
10	10	11.75
8.75	11.5	12
11	10.75	12.5
10.75	10.25	11
10.75	9.25	10.75
11.5	10.75	12
9.75	10	12
12.25	13	10.75
9.75	11	10.5
9.25	12	12.75

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: At least one mean is different.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Suburb	15	160.5	10.7	1.0375
Rural	15	161.5	10.76667	1.022024
Urban	15	171.25	11.41667	0.71131

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.702778	2	2.351389	2.545865	0.090448	3.2199
Within Groups	38.79167	42	0.923611			
Total	43.49444	44				

$F = 2.5459$; p-value = 0.0904; Fail to reject H_0. There is not enough evidence to support the claim that there is a difference in the mean movie ticket prices by geographical regions where the theaters are located (suburban, rural, urban).

29. An ANOVA was run to test to see if there was a significant difference in the average cost between three different types of fabric for a new clothing company. Random samples for each of the three fabric types was collected from different manufacturers. Assume the costs are normally distributed. At $\alpha = 0.10$, run an ANOVA test to see if there is a difference in the means. If a difference is found, run a Bonferroni test to see which means are different. Based off the Bonferroni results which fabric type should you choose?

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	34	10608	312	204.1212
B	37	11655	315	97.5
C	32	9600	300	2019.548

H_0: $\mu_A = \mu_B = \mu_C$

H_1: At least one mean is different.

ANOVA

Price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4217.417	2	2108.709	2.895	.060
Within Groups	72852.000	100	728.520		
Total	77069.417	102			

$F = 2.895$; p-value = 0.06; Reject H_0.

There is sufficient evidence to support the claim that there is a difference in the mean cost between three different types of fabric.

Multiple Comparisons

Dependent Variable: Price

Bonferroni

(I) Fabric Type	(J) Fabric Type	Mean Difference (I–J)	Std. Error	Sig.	90% Confidence Interval	
					Lower Bound	Upper Bound
Type A	Type B	-3.00000	6.41224	1.000	-16.8367	10.8367
	Type C	12.00000	6.64780	.222	-2.3450	26.3450
Type B	Type A	3.00000	6.41224	1.000	-10.8367	16.8367
	Type C	15.00000*	6.51583	.070	.9398	29.0602
Type C	Type A	-12.00000	6.64780	.222	-26.3450	2.3450
	Type B	-15.00000*	6.51583	.070	-29.0602	-.9398

H_0: $\mu_A = \mu_B$

H_1: $\mu_A \neq \mu_B$

p-value = 1; Do not reject H_0; There is not a significant difference in the mean cost of fabrics A and B.

H_0: $\mu_A = \mu_C$

$H_1: \mu_A \neq \mu_C$
p-value = 0.222; Do not reject H_0; There is not a significant difference in the mean cost of fabrics A and C.

$H_0: \mu_B = \mu_C$
$H_1: \mu_B \neq \mu_C$
p-value = 0.07; Reject H_0; There is significant difference in the mean cost of fabrics B and C.

31. A researcher is testing to see if there is a difference in the average per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area. Assume tuition costs are normally distributed. The following table shows random samples for the per-pupil costs for private school tuition in thousands of dollars. Using a significance level of 5%, test to see if there is a significant difference.

Clackamas	Multnomah	Washington
15.74	14.97	14.77
14.66	12.28	15.2
14.6	12.94	15.13
15.17	11.33	15.35
14.83	13.27	14.8
15.06	13.38	14.45
14.54	12.95	14.64
15.13	11.86	13.97
14.76	13.83	15.12
15.3	12.48	15.42
15.19	10.68	
14.94	12.29	
14.96	11.25	
15.1	12.31	
14.96		
15.2		

$H_0: \mu_1 = \mu_2 = \mu_3$
H_1: At least one mean is different.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Clackamas	16	240.14	15.00875	0.090158
Multnomah	14	175.82	12.55857	1.259044
Washington	10	148.85	14.885	0.20265

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	52.6231	2	26.31155	49.8126029	3.19E-11	3.251923846
Within Groups	19.5438	37	0.528211			
Total	72.1669	39				

F = 49.8126; p-value = 0.000; Reject H_0.
There is sufficient evidence to support the claim that there is a difference in the average per-pupil costs for private school tuition for three counties in the Portland, Oregon, metro area.

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$
t = 9.2121695; p-value = 0.000, Reject H_0. There is significant difference in the mean per-pupil costs for private school tuition for Clackamas and Multnomah counties.

H_0: $\mu_1 = \mu_3$
H_1: $\mu_1 \neq \mu_3$
t = 0.4223955; p-value = 1, Fail to reject H_0. There is not a significant difference in the mean per-pupil costs for private school tuition for Clackamas and Washington counties.

H_0: $\mu_2 = \mu_3$
H_1: $\mu_2 \neq \mu_3$
t = -7.7312358; p-value = 0.000, Reject H_0. There is a significant difference in the mean per-pupil costs for private school tuition for Multnomah and Washington counties.

Chapter 11 Exercises

1. To test the significance of the correlation coefficient, we use the t-distribution with how many degrees of freedom? Answer: d)
   a) $n - 1$
   b) n
   c) $n + 1$
   d) $n - 2$
   e) $n_1 + n_2 - 2$

3. The coefficient of determination is a number between _____.
   a) -1 and 1
   b) -10 and 10
   c) 0 and 10
   d) 0 and ∞
   e) 0 and 1
   f) $-\infty$ and ∞

   Answer: e)
   The coefficient of determination is r^2, thus all the values of r between -1 and 0 are squared and turned into positive numbers between 0 and 1.

5. The standard error of estimate is a number between _____.
   a) -1 and 1
   b) -10 and 10
   c) 0 and 10
   d) 0 and ∞
   e) 0 and 1
   f) $-\infty$ and ∞

   Answer: d)
   The standard error of estimate s, is 0 if all the dots line up perfectly on the regression equation. As the dots get more dispersed away from the regression equation the s gets larger. Depending on the units of y, the value for s can become quite large. S would never be a negative number since standard deviations are never negative.

7. True or False: Correlation measures the strength and direction of the linear relationship between two variables. Answer: True

9. True or False: Correlation implies causation, meaning that if two variables are strongly correlated, one variable causes the other. Answer: False

11. True or False: A p-value less than the significance level indicates that there is a significant linear relationship between the independent and dependent variables. Answer: True

13. True or False: In simple linear regression, the y-intercept represents the predicted value of the dependent variable when the independent variable is zero. Answer: True

15. Which of the following is not the correct notation for a linear regression equation?

a) $\hat{y} = -5 + \frac{2}{9}x$

b) $\hat{y} = 3x + 2$

c) $\hat{y} = \frac{2}{9} - 5x$

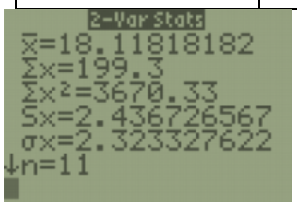
d) $\hat{y} = 5 + 0.4x$

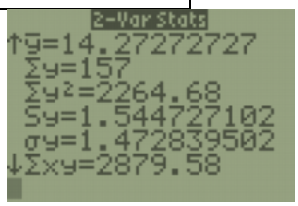
Answer: d)

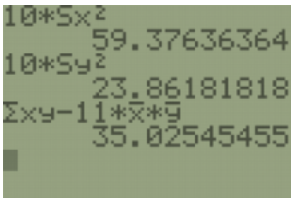
A regression equation needs the hat over the y for the predicted value $\hat{y}$.

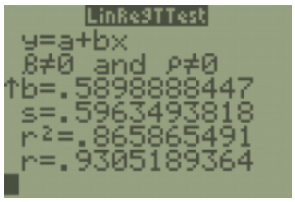
17. It has long been thought that the length of one's femur is positively correlated to the length of one's tibia. The following are data for a classroom of students who measured each (approximately) in inches. Use $\alpha = 0.10$ to test to see if there is a significant correlation between the two variables.

Femur Length	Tibia Length
18.7	14.2
20.5	15.9
16.2	13.1
15.0	12.4
19.0	16.2
21.3	15.8
21.0	16.2
14.3	12.1
15.8	13.0
18.8	14.3
18.7	13.8









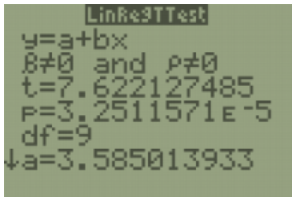
$H_0: \rho = 0$; $H_1: \rho \neq 0$; $t_{\alpha/2} = -1.833$,

$SS_{xx} = (n-1) \cdot s_x^2 = (11-1) \cdot 2.436726567^2 = 59.37636364$

$SS_{yy} = (n-1)s_y^2 = (11-1)1.544727102^2 = 23.861818$

$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y} = 2879.58 - 11 \cdot 18.11818182 \cdot 14.27272727 = 35.02545455$

$r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}} = \frac{35.02545455}{\sqrt{(59.37636364 \cdot 23.861818)}} = 0.9305189$



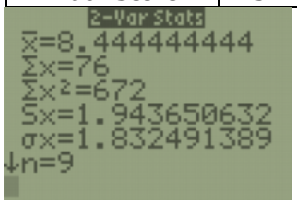
$t = r\sqrt{\left(\frac{n-2}{1-r^2}\right)} = 0.9305189\sqrt{\left(\frac{9}{1-0.865865491}\right)} = 7.6221$

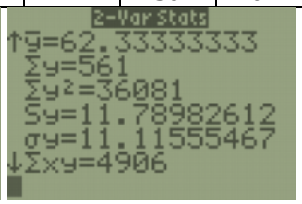
p-value = 0.0000325

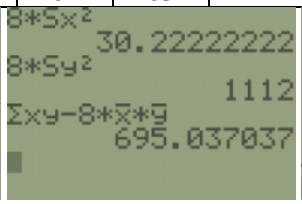
Reject H_0; There is a statistically significant correlation between a person's femur and tibia length.

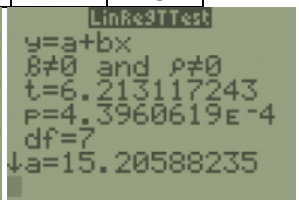
19. An elementary school uses the same system to test math skills at their school throughout the course of the 5 grades at their school. The age and score (out of 100) of several students is displayed below. Use $\alpha = 0.10$ to test to see if there is a significant correlation between the two variables.

Student Age	6	6	7	8	8	9	10	11	11
Math Score	54	42	50	61	67	65	71	72	79









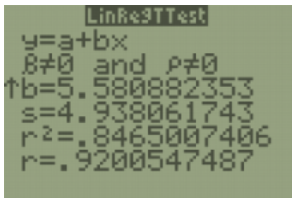
H_0: $\rho = 0$; H_1: $\rho \neq 0$; $t_{\alpha/2} = -1.895$

$SS_{xx} = (n-1) \cdot s_x^2 = (9-1) \cdot 1.943650632^2 = 30.22222222$

$SS_{yy} = (n-1)s_y^2 = (9-1)11.78982612^2 = 1112$

$SS_{xy} = \sum(xy) - n \cdot \bar{x} \cdot \bar{y} = 4906 - 9 \cdot 8.444444 \cdot 62.333333 = 695.037037$

$$r = \frac{SS_{xy}}{\sqrt{(SS_{xx} \cdot SS_{yy})}} = \frac{695.037037}{\sqrt{(30.222222 \cdot 1112)}} = 0.9201$$



$$t = r\sqrt{\left(\frac{n-2}{1-r^2}\right)} = 0.92005474879\sqrt{\left(\frac{7}{1-0.8465007406}\right)} = 6.2131$$

p-value = 0.0004

Reject H_0; There is a statistically significant correlation between the students age and their math score.

21. Body frame size is correlated with a person's wrist circumference in relation to height. A researcher measures the wrist circumference and height of a random sample of individuals. The data is displayed below.

Regression Statistics	
Multiple R	0.7938
R Square	0.6301
Adjusted R Square	0.6182
Standard Error	3.8648
Observations	33

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	31.6304	5.2538	6.0205	1.16E-06
x	5.4496	0.7499	7.2673	3.55E-08

a) What is the value of the test statistic to see if the correlation is statistically significant?

i. 6.0205
ii. 1.16E-06
iii. 3.55E-08

iv. 5.2538
v. 7.2673
vi. 0.7499
vii. 0.7938

Answer: v.

The correct test statistic can be found in the last row of the last table under "t Stat".

b) What is the correct p-value and conclusion for testing if there is a significant correlation?

i. 1.16E-06; There is a significant correlation.
ii. 3.55E-08, There is a significant correlation.
iii. 1.16E-06; There is not a significant correlation.
iv. 3.55E-08, There is not a significant correlation.
v. 0.7938, There is a significant correlation.
vi. 0.7938, There is not a significant correlation.

Answer: ii.

The correct p-value can be found in the last row of the last table under "P-value". $p < \alpha$, thus, there is a significant correlation.

c) Which number is the standard error of estimate? s = 3.8648

d) Which number is the coefficient of determination? $R^2 = 0.6301$

e) Compute the correlation coefficient. r = (sign of slope)* $\sqrt{0.6301} = 0.7938$

f) What is the correct test statistic for testing if the slope is significant H_1: $\beta_1 \neq 0$? t = 7.2673

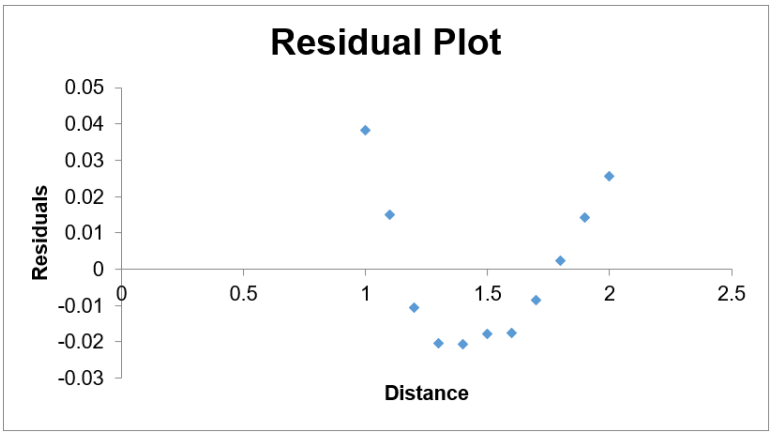
g) What is the correct p-value for testing if the slope is significant H_1: $\beta_1 \neq 0$? p = 3.55E-08

h) At the 5% level of significance, is there a significant relationship between wrist circumference and height? Reject H_0. Yes, there is a significant relationship between wrist circumference and height.

23. The intensity (in candelas) of a 100-watt light bulb was measured by a sensing device at various distances (in meters) from the light source. A linear regression was run and the following residual plot was found.

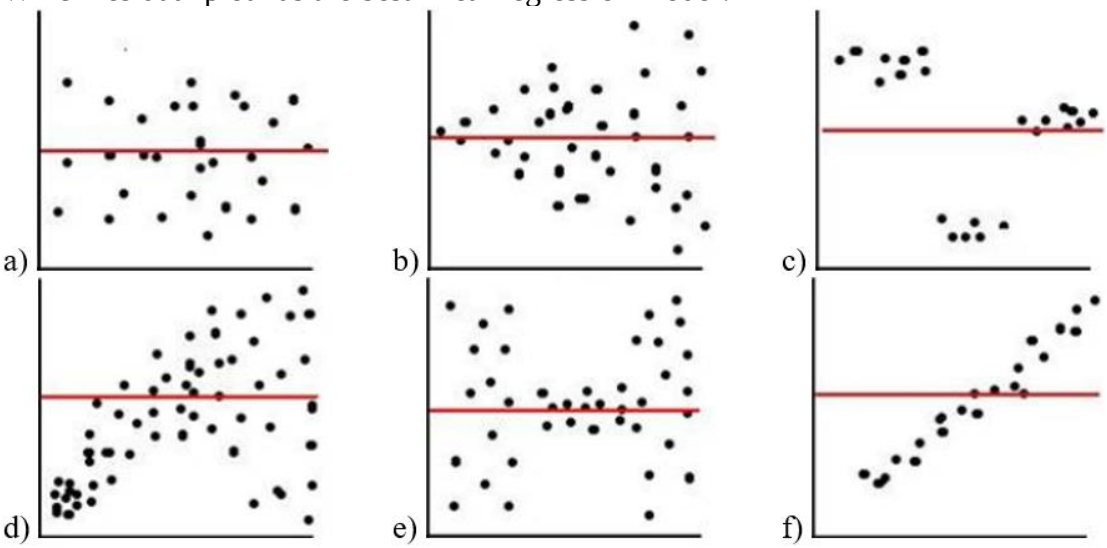
Regression Statistics	
Multiple R	0.95936
R Square	0.920371
Adjusted R Square	0.911523
Standard Error	0.021636
Observations	11

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.468618	0.031624	14.81856	1.25E-07
Distance	-0.2104	0.020629	-10.1992	3.04E-06



a) Is linear regression a good model to use? No, since there is a U shape in the residual plot.
b) Write a sentence explaining your answer. The p-value = 0.00000304 suggests that there is a significant linear relationship between intensity (in candelas) of a 100-watt light bulb was measured by a sensing device at various distances (in meters) from the light source. However, the residual plot clearly shows a nonlinear relationship. Even though we can fit a straight line through the points, we would get a better fit with a curve.

25. Which residual plot has the best linear regression model?

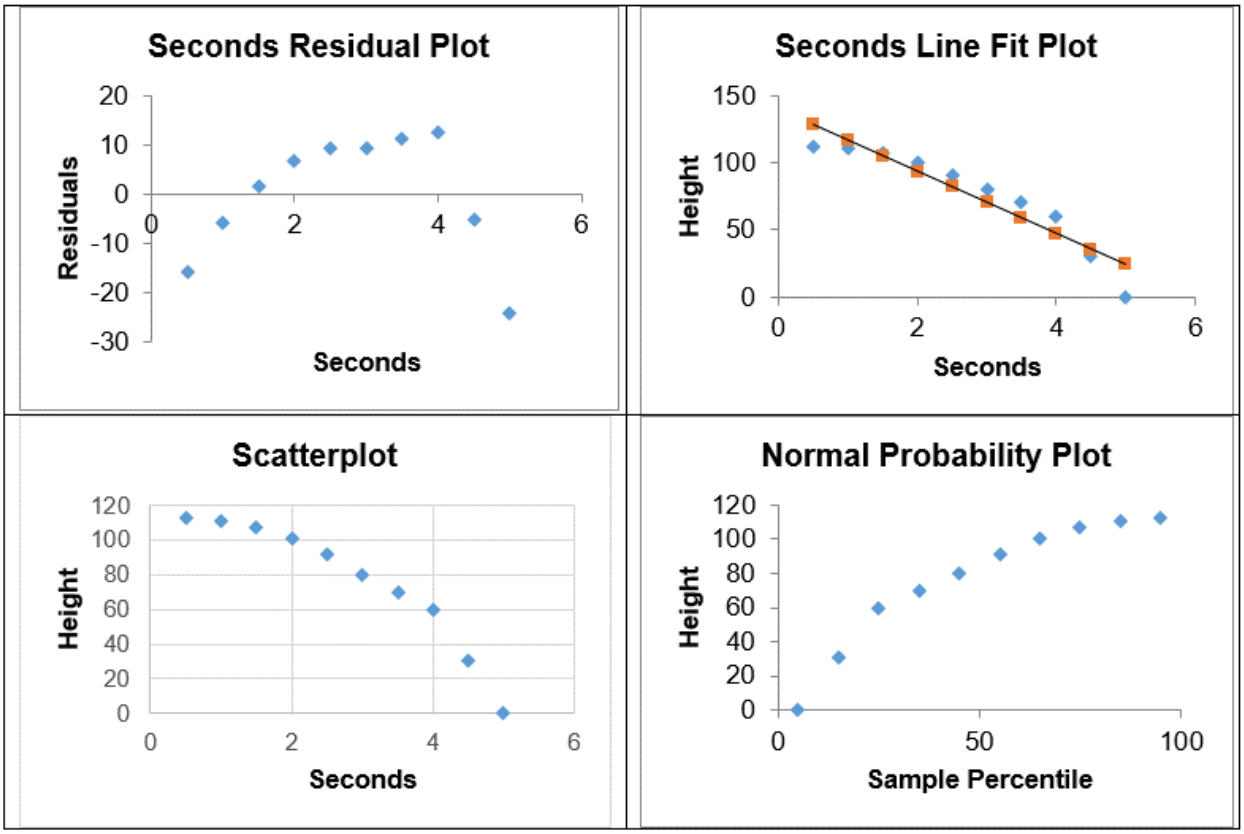


Answer: a)
A residual plot should have no pattern and the data points on the graph should vary in distance from the red line in the middle of the graph for it to represent a linear model.

27. An object is thrown from the top of a building. The following data measure the height of the object from the ground for a five-second period.

Seconds	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Height	112.5	110.875	106.8	100.275	91.3	79.875	70.083	59.83	30.65	0

The following four plots were part of the regression analysis.



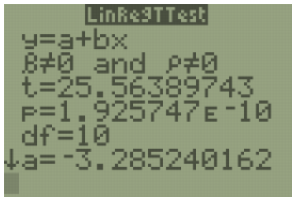
There is a statistically significant correlation between time and height, r = -0.942, p-value = 0.0000454. Should linear regression be used for this data? Why or why not? Choose the correct answer.

a) Yes, the p-value indicates that there is a significant correlation so we can use linear regression.
b) Yes, the normal probability plot has a nice curve to it.
c) Yes, there is a nice straight line in the line fit plot.
d) No, there is a curve in the residual plot, normal plot and the scatterplot.
Answer: d)
The residual plot should not have a pattern (curve) for linear data. You would also not find a curve in the scatterplot and normal plot.

29. The data below show the predicted average high temperature (ºF) per month by the Farmer's Almanac in Portland, Oregon alongside the actual high temperature per month that occurred.

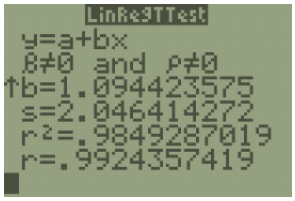
Farmer's Almanac	45	50	57	62	69	72	81	90	78	64	51	48
Actual High	46	52	60	61	72	78	82	95	85	68	52	49

a) Compute the regression equation.



$\hat{y} = -3.285240162 + 1.094423575x$

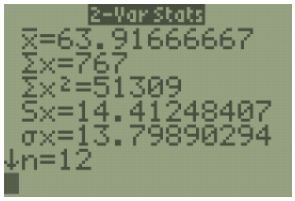
b) Test to see if the slope is significantly different from zero, use α = 0.01.



$H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$

$SS_{xx} = (n-1) \cdot s_x^2 = (12-1) \cdot 14.41248407^2 = 2284.916668$

$t = \frac{b_1}{\sqrt{\left(\frac{MSE}{SS_{xx}}\right)}} = \frac{1.094423575}{\sqrt{\left(\frac{2.046414272^2}{2284.916668}\right)}} = 25.5639$



p-value = 0.00000000019; Reject H_0

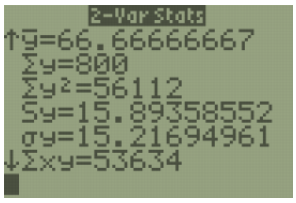
There is significant linear relationship between the predicted high temperature in the Farmer's Almanac and the actual high temperatures.

c) Predict the high temperature in the coming year, given that the Farmer's Almanac is predicting the high to be 58 ºF.

$\hat{y} = -3.285240162 + 1.094423575 \cdot 58 = 60.19132719$

d) Compute the 99% prediction interval for the actual high temperature in the coming year, given that the Farmer's Almanac is predicting the high to be 58 ºF.

Use $\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$



$t_{\alpha/2} = -3.169$,

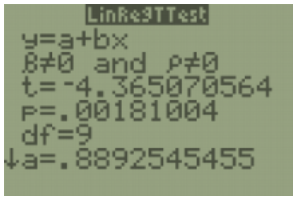
$SS_{xx} = (n-1) \cdot s_x^2 = (12-1) \cdot 14.41248407^2 = 2284.916668$

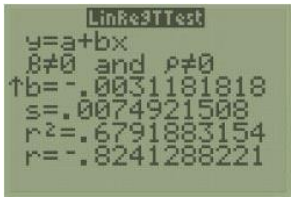
$60.19132719 \pm 3.169 \cdot 2.046414 \cdot \sqrt{1 + \frac{1}{12} + \frac{(58-63.91666667)^2}{2284.916668}}$

Answer: $53.394 < y < 66.989$

31. Bone mineral density and cola consumption has been recorded for a sample of patients. Let x represent the number of colas consumed per week and y the bone mineral density in grams per cubic centimeter. Assume the data is normally distributed.

x	y
1	0.883
2	0.8734
3	0.8898
4	0.8852
5	0.8816
6	0.863





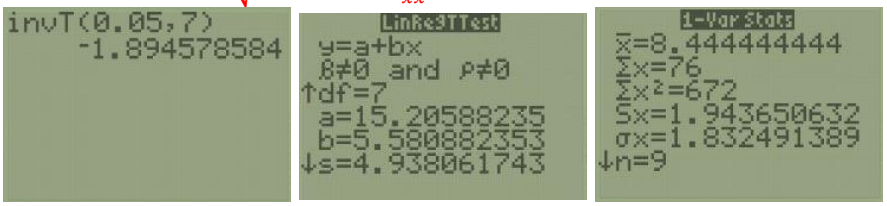
7	0.8634
8	0.8648
9	0.8552
10	0.8546
11	0.862

a) State the hypotheses to test for a significant linear relationship. $H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$
b) Compute the correlation coefficient. $r = 0.8241$
c) Compute the p-value to see if there is a significant linear relationship. p-value = 0018
d) State the correct decision, use $\alpha = 0.05$. Is there a significant linear relationship? Reject H_0; Yes
e) Compute the coefficient of determination. $r^2 = R^2 = 0.6792$
f) Compute the regression equation. $\hat{y} = 0.8893 - 0.0031x$
g) Interpret the slope coefficient. For every additional average weekly soda consumption, a person's bone density decreases by 0.0031 grams per cubic centimeter.
h) Compute the predicted bone mineral density for a person that consumes 7 colas per week. $\hat{y} = 0.8893 - 0.0031 \cdot 7 = 0.8674$
i) Compute the residual for the point (7, 0.8634). $y - \hat{y} = 0.8634 - 0.8674 = -0.004$

33. An elementary school uses the same system to test math skills at their school throughout the course of the 5 grades at their school. The age and score (out of 100) of several students is displayed below. A significant linear relationship is found between Student Age and Math Score. Compute a 90% prediction interval for the score a student would earn given that they are 5 years old.

Student Age	6	6	7	8	8	9	10	11	11
Math Score	54	42	50	61	67	65	71	72	79

Use $\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$



$t_{\alpha/2} = -1.895$, $SS_{xx} = (n-1) \cdot s_x^2 = (9-1) \cdot 1.943650632^2 = 30.22222223$

$\hat{y} = 15.20588235 + 5.580882353 \cdot 5 = 43.11029412$

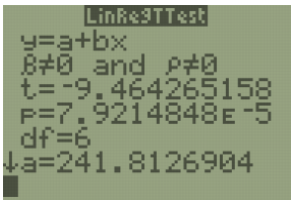
$43.11029412 \pm 1.894578584 \cdot 4.938061743 \cdot \sqrt{1 + \frac{1}{9} + \frac{(5-8.444444444)^2}{30.22222223}}$

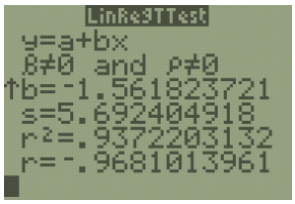
Answer: $31.636 < y < 54.585$

35. A study was conducted to determine if there was a linear relationship between a person's age and their peak heart rate. Use α=0.05.

Age (x)	16	26	32	37	42	53	48	21
Peak Heart Rate (y)	220	194	193	178	172	160	174	214

a) What is the estimated regression equation that relates number of hours worked and test scores for high school students.





$\hat{y} = 241.8127 - 1.5618x$

b) Interpret the slope coefficient for this problem. Every year a person ages, their peak heart rate decreases by an average of 1.5618.

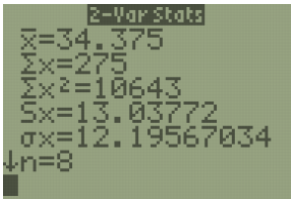
c) Compute and interpret the coefficient of determination. $R^2 = 0.93722$

d) Compute the coefficient of nondetermination. $1 - R^2 = 1 - 0.93722 = 0.06278$

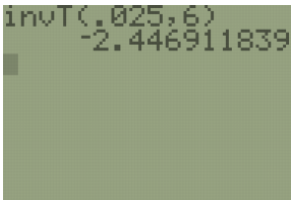
e) Compute the standard error of estimate. s = 5.692404918

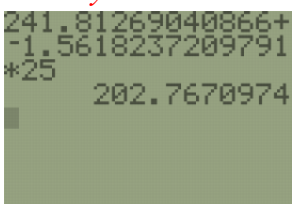
f) Compute the correlation coefficient. $r = -0.9681$

g) Find the 95% Prediction Interval for peak heart rate for someone who is 25 years old.



$\hat{y} = 241.8126904 - 1.561823721 \cdot 25 = 202.7670974$





$t_{\alpha/2} = -2.446911839$, $SS_{xx} = (n-1)*s^2 = 7*13.03772^2 = 1189.875$

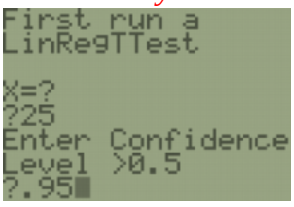
s = 5.692404918

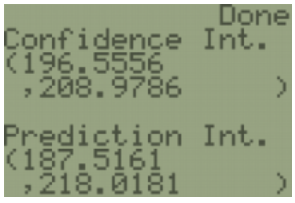
$$\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$$

$$202.7670974 \pm 2.446911839 \cdot 5.692404918 \cdot \sqrt{1 + \frac{1}{8} + \frac{(25-34.375)^2}{1189.875}}$$

$202.7670974 \pm 15.25103527$

$187.5161 < y < 218.0181$





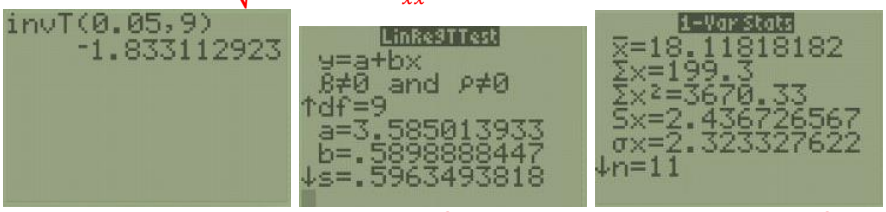
Program has some rounding issues (187.5161, 218.0181)

37. It has long been thought that the length of one's femur is positively correlated to the length of one's tibia. The following are data for a classroom of students who measured each (approximately) in inches. A significant linear correlation was found between the two

variables. Find the 90% prediction interval for the length of someone's tibia when it is known that their femur is 23 inches long.

Femur Length	Tibia Length
18.7	14.2
20.5	15.9
16.2	13.1
15.0	12.4
19.0	16.2
21.3	15.8
21.0	16.2
14.3	12.1
15.8	13.0
18.8	14.3
18.7	13.8

Use $\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$



$t_{\alpha/2} = -1.833$, $SS_{xx} = (n-1) \cdot s_x^2 = (11-1) \cdot 2.436726567^2 = 59.37636362$

$\hat{y} = 3.585013933 + 0.5898888447 \cdot 23 = 17.15245736$

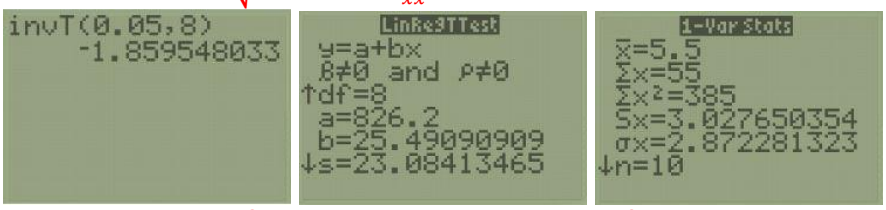
$17.15245736 \pm 1.833 \cdot 0.596349 \cdot \sqrt{1 + \frac{1}{11} + \frac{(23-18.11818182)^2}{59.37636362}}$

Answer: $15.817 < y < 18.488$

39. The following data represent the enrollment at a small college during its first 10 years of existence. A significant linear relationship is found between the two variables. Find a 90% prediction interval for the enrollment after the college has been open for 14 years.

Years	1	2	3	4	5	6	7	8	9	10
Enrollment	856	842	923	956	940	981	1025	996	1057	1088

Use $\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$



$SS_{xx} = (n-1) \cdot s_x^2 = (10-1) \cdot 3.027650354^2 = 82.4999807$

$\hat{y} = 826.2 + 25.49090909 \cdot 14 = 1183.072726$

$1183.072726 \pm 1.860 \cdot 23.084135 \cdot \sqrt{1 + \frac{1}{10} + \frac{(14-5.5)^2}{82.49998}}$

Answer: $1122.720 < y < 1243.425$

41. The data below represent the driving speed (mph) of a vehicle and the corresponding gas mileage (mpg) for several recorded instances.

Driving Speed	Gas Mileage
57	21.8
66	20.9
42	25.0
34	26.2
44	24.3
44	26.3
25	26.1
20	27.2
24	23.5
42	22.6
52	19.4
54	23.9
60	24.8
62	21.5
66	20.5
67	23.0
52	24.2
49	25.3
48	24.3
41	28.4
38	29.6
26	32.5
24	30.8
21	28.8
19	33.5
24	25.1

a) Do a hypothesis test to see if there is a significant correlation. Use $\alpha = 0.10$.

$H_0: \rho = 0$; $H_1: \rho \neq 0$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	32.40313	1.426646	22.7128	9.81E-18
Driving Speed	-0.1662	0.031648	-5.25143	2.2E-05

p-value = 0.000022, which is less than $\alpha = 0.10$, so reject H_0. There is a significant correlation between driving speed of a vehicle and the corresponding gas mileage.

b) Compute the standard error of estimate.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.731219
R Square	0.534681
Adjusted R Square	0.515293
Standard Error	2.49433
Observations	26

$s = 2.49433$

c) Compute the regression equation and use it to find the predicted gas mileage when a vehicle is driving at 77 mph.

	Coefficients
Intercept	32.40313
Driving Speed	-0.1662

$\hat{y} = 32.40313 - 0.1662x$

d) Compute the 90% prediction interval for gas mileage when a vehicle is driving at 77 mph.

$\hat{y} = 32.40313 - 0.1662 \cdot 77 = 19.60606$ mpg

Use $\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$

$t_{\alpha/2}$ = T.INV(0.05,24) = −1.71088

Driving Speed	
Mean	42.34615385
Standard Error	3.091398642
Median	43
Mode	24
Standard Deviation	15.763102
Sample Variance	248.4753846

$SS_{xx} = (n-1) \cdot s_x^2 = (26-1) \cdot 248.4753846 = 6211.884615$

$19.60606 \pm 1.71088 \cdot 2.49433 \cdot \sqrt{1 + \frac{1}{26} + \frac{(77-42.34615385)^2}{6211.884615}}$

Answer: $14.8697 < y < 24.3424$