Student Solution Manual

With An Emphasis On The TI-83

Data Analysis

An Applied Approach To Statistics With Technology

Second Edition

(2nd Printing)

Brian Jean
David Meyers
Rene' Sporer

Printed in the United States of America

ISBN 1-930090-03-X

Preface

The student solution manual to Data Analysis – An Applied Approach To Statistics With Technology contains detailed solutions to all odd problems contained in the textbook. When appropriate, the solutions emphasize the use of technology, specifically the Texas Instruments TI-83 graphing calculator.

For additional information/comments, contact the publisher at:

http://www.3RingPublishing.com

1 Introduction

Review Exercise Solutions

1.1 a) Ratio b) Nominal c) Interval d) Ratio e) Ratio f) Nominal
g) Nominal h) Ordinal i) Nominal

1.3 There are many ways to approach this problem. The key is to present language that is void of statistical jargon and emphasize that the sample is a smaller portion of the whole.

1.5 a) All students on campus b) 825 c) Weight

d) Proportion of students who fall into the categories of skinny, slender, appropriate, chunky, and obese.

e) Ordinal

1.7 a) Population: Elected representatives Variable:How a representative will vote on the bill

b) Population: Registered voters Variable: Opinion regarding candidate or important issues

c) Part (a) was a census. It is reasonable to contact all members of congress or all members of the house of representatives and poll them regarding an upcoming bill. Part (b) was a sample. It is not reasonable to expect we could contact every voter within a specific district and obtain their opinion.

1.9 Answers will vary. Possible solutions include: Name at least three (3) qualitative variables that could be measured for this container of milk.

a) Grade, homogenized, pasteurized, type (1%, 2%, whole milk, chocolate milk)

b) Weight, proportion of daily recommended amounts of various vitamins, calories, fat in grams.

c) Answers will vary depending on the variables chosen.

1.11 a) descriptive b) inferential c) inferential d) descriptive

1.13 a) Time interval between successive births. b) ratio c) continuous

1.15 a) All cell phone users. b) The sample is those who answered a survey in 12 metro areas in the U.S..

c) The true proportion of cell phone users that experience service problems. The true proportion of cell phone users that found their carrier's response helpful. The true proportion of cell phone users that have had an overcharge of $10 or more.

1.17 a) All persons and companies that might use their services.

b) 1. Are you planning any landscaping in near future?

2. If so, how far in the future.

c) Answers will vary according to questions in part b.

d) Answers will vary according to questions in part c.

e) You may report descriptive statistics when you look at the summary of the values calculated from the survey results. Then when you use the numbers to make broad statements about your population of interest you would be using inferential statistics.

1.19 a) All prescription drugs in the U.S..

b) The drugs or drug companies surveyed by the Federal government.

c) The true proportion of growth of prescription drug costs.

d) It is 15.3%.

e) It would be inferential. The sample statistic was reported and then a statement was made about the population.

f)The name of the drugs, the cost of each drug, the use of each drug, the amount produced each year. These answers will vary.

g) For the above answers:

Name: qualitative, nominal, discrete

Cost: quantitative, ratio, continuous

Use: qualitative, nominal, discrete

Amount Produced: quantitative, ratio, discrete

2 Experimental Design and Data Collection

Review Exercise Solutions

2.1 Answers will vary. Regardless of the sampling method chosen, the process is a survey, not an experiment because data is being collected without modifying the environment in any way.

2.3 Answers will vary, however, the basic idea behind Junk Science is that it is the use of faulty data or faulty analytical processes.

2.5 Answers will vary. The key to any solution will be to use simple language avoiding statistical jargon and discuss the idea that a survey is recording information that already exists whereas an experiment modifies the environment then records the results.

2.7 Chance error is the result of randomness in the sampling process whereas a bias is a systematic error inherent to the way you are taking your sample.

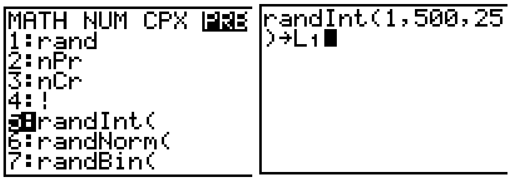
2.9 a) Controlled experiment. You, the experimenter, are controlling the environment by selecting the type of strawberry to be planted.

b) Observational study. You, the experimenter, are simply recording what has already taken place. You are not doing anything to manipulate the environment.

c) Controlled experiment. You, the experimenter, have selected the area to introduce the burger and will compare it to a control group, that possibly being sales in the same area prior to introduction of the new burger.

d) Observational study. You, the experimenter, are not doing anything to manipulate the environment. Rather, you are simply recording an opinion that already exists.

2.11 a) There are a total of 500 grid squares. To generate 25 random numbers between 1 and 500 on your TI-83, go to **MATH>PRB** and use the command **5: randInt(**. The command sequence will be **randInt(1, 500, 25) → L$_1$**.



```
randInt(1,500,25
)→L1
```

By storing the random numbers in L$_1$, you will have easy access to the values to sort and manipulate them as needed.

b) The number of samples required is 25, so we have 500/25 = 20. Next, we need a random start between 1 and 20. Use the same command as in part (a) only specify one random number. randInt(1, 20, 1). There is no need to store the random value in a list because you will only be receiving one number.

c) Answers will vary. d) Answers will vary.

2.13 This is directly tied to the placebo effect. The idea is if people believe they are getting "the real treatment" then there is a natural tendency to "become better" even though no real change occurs. This is related to controlled experiments by using a placebo group and comparing the results of the placebo group with the experimental group in a blind experiment.

2.15 The results are invalid for many reasons. First, there is no control over how many times people can vote. Second, the population is limited to those who frequent this web site, so any inference beyond that population would clearly be inappropriate. In addition, the wording of the question is very suggestive. It asks if you support animal testing *if it saves human lives*. The question is justifying animal testing by the wording, so is argumentative to start with. Better wording would be “Do you support animal testing for medical research.” The mention of saving human lives in the original wording may invoke an inappropriate emotional response.

2.17 Mail surveys are easy to conduct and can cover a wide population relatively inexpensively. They are also very biased due to the fact that typically, only persons who have a personal interest in the question(s) asked respond making the results biased.

2.19 a) The answers will vary for this problem as there are many ways to gather this information. An observational study would be most appropriate. Cell phone users could be systematically sampled through user lists from the companies.

b) Bias can enter when there is a strong feeling about the subject. If a user has had a bad experience, they might be more inclined to answer the survey than someone that has no problems at all. So, a higher proportion of users with problems could end up in the survey.

c) In my systematic survey, I would contact the users rather than relying on the users to return a voluntary survey.

2.21 a) Answers will vary for this question. A stratified sample based on the zip code or some other natural division in the population may be appropriate.

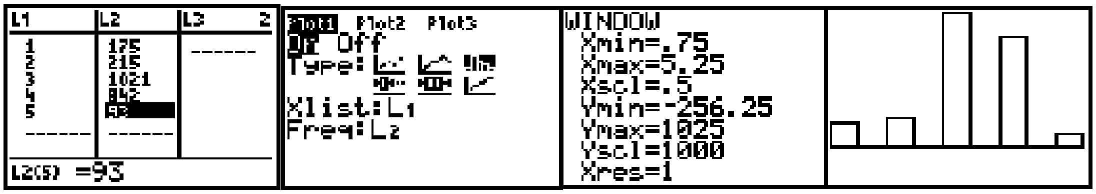
b) Answers will vary.

3 Graphical Displays of Univariate Data

Review Exercise Solutions

3.1 The variable is weight and the measurement scale is ordinal. Weight is typically thought of as being ratio, but the way the weights are being recorded in this example - skinny, slender, appropriate, chunk, and obese - make the measurement scale ordinal.

A bar graph will be used to display the data. Skinny will be coded as 1, slender as 2, appropriate as 3, chunky as 4 and obese as 5. By entering the coded data in L_1 and the frequencies in L_2, we can quickly generate an appropriate bar graph. Values can be read by pressing the TRACE button.



```
L1      L2      L3     2
 1      175     ------
 2      215
 3      1021
 4      842
 5      93
------  ------
L2(5) =93
```

```
Plot1  Plot2  Plot3
On Off
Type:
Xlist:L1
Freq:L2
```

```
WINDOW
 Xmin=.75
 Xmax=5.25
 Xscl=.5
 Ymin=-256.25
 Ymax=1025
 Yscl=1000
 Xres=1
```

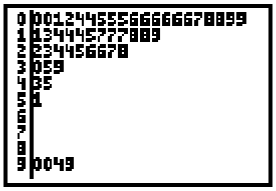
3.3 A histogram is used for quantitative data whereas a bar graph is used for qualitative data.

3.5 False. Cumulative frequency has no meaning for nominal data. If you had a frequency table that consisted of the eye color of everyone in your class, what would it mean to say 75% of everyone in the class has brown or less colored eyes? Cumulative frequency only has meaning for at least ordinal scaled variables.

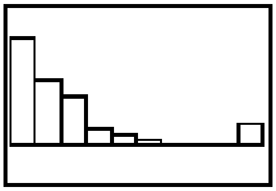
3.7 False. Stem-and-leaf displays have no meaning for qualitative data.

3.9 a) The variable is the percent of schools in compliance with the NCEE requirements. The measurement scale is ratio.

b) The stem-and-leaf display are from the TI-83 program STEMPLOT.



```
0|00124455566666678899
1|134445777889
2|234456678
3|059
4|35
5|1
6|
7|
8|
9|0049
```



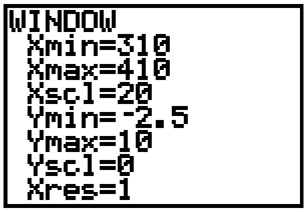
c) The distribution is skewed right.

d) Yes. If you rotate the stem-and-leaf display 90 degrees counter clockwise, the general shape matches that of the histogram.

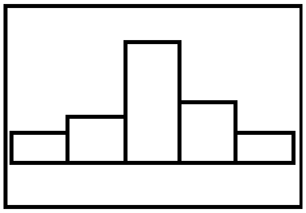
e) The compliance is very low. The vast majority are less than 50% in compliance.

3.11 a) Variable: Number of cases produced daily. Measurement scale: ratio.

b) The distribution is approximately symmetric.



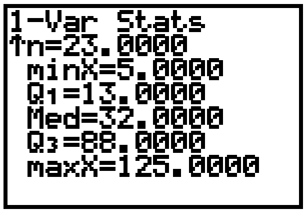
```
WINDOW
 Xmin=310
 Xmax=410
 Xscl=20
 Ymin=-2.5
 Ymax=10
 Yscl=0
 Xres=1
```



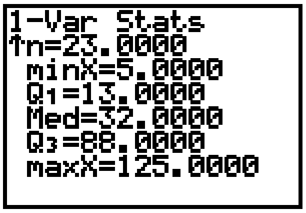
A student's graphical display should indicate class widths along with a y-axis scale. The window settings are provided here for your reference.

3.13 a) Ratio

b) The data is skewed right. A student's response should have a properly labeled graphical display. The needed values for the box-plot are provided for your reference.



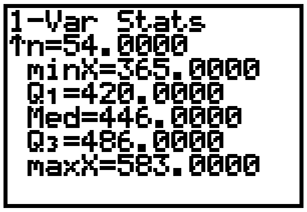
```
1-Var Stats
↑n=23.0000
 minX=5.0000
 Q1=13.0000
 Med=32.0000
 Q3=88.0000
 maxX=125.0000
```



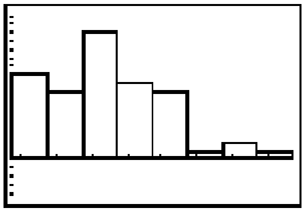
```
1-Var Stats
↑n=23.0000
 minX=5.0000
 Q1=13.0000
 Med=32.0000
 Q3=88.0000
 maxX=125.0000
```

c) The data value 6 means there was a country in the study that reported a mortality rate of 6 per 1000 births. Similarly for the data value 125.

3.15 a) Ratio b) The data is slightly skewed right.



```
1-Var Stats
↑n=54.0000
 minX=365.0000
 Q1=420.0000
 Med=446.0000
 Q3=486.0000
 maxX=583.0000
```

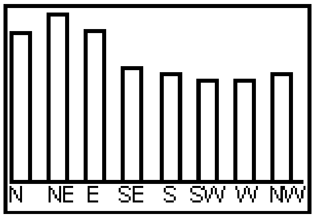


An appropriate response should have a properly labeled graphical display. The needed values for the box-plot are provided for your reference

3.17 a) Ordinal. Regardless of which direction you start, the next category is predetermined due to the obvious natural order.

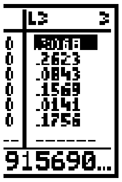
b) Yes. A histogram and a box-plot, as two examples, are not appropriate for this data.

c) The following bar graph should have a y-axis scale and a graph label. The categories along the x-axis have been added.

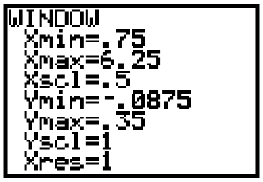


3.19 a) Both variables, gender and type of test, are nominal

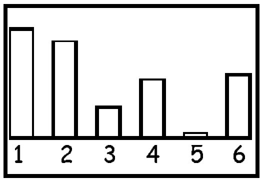
b) Data coding: 1 = Social Studies, 2 = English, 3 = Foreign Language, 4 = Calculus, 5 = Computer Science, 6 = Science.



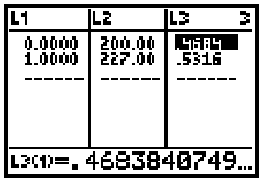
```
    L3      3
0   .3068
0   .2623
0   .0843
0   .1569
0   .0141
0   .1756
--  -------
915690…
```



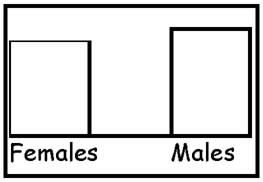
```
WINDOW
 Xmin=.75
 Xmax=6.25
 Xscl=.5
 Ymin=-.0875
 Ymax=.35
 Yscl=1
 Xres=1
```



c) Data coding: 0 = Female, 1 = Male.

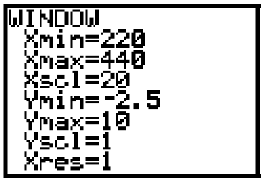


```
L1      L2      L3     3
 0.0000  200.00  .4684
 1.0000  227.00  .5316
 ------  ------  ------
L3(1)=.4683840749…
```

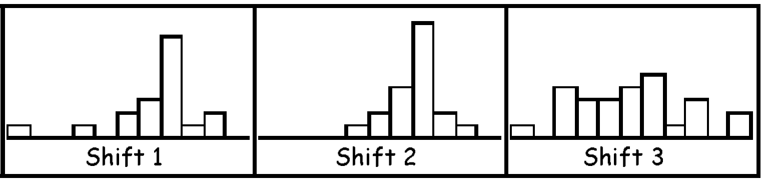


3.21 a) Ratio.

b)



```
WINDOW
 Xmin=220
 Xmax=440
 Xscl=20
 Ymin=-2.5
 Ymax=10
 Yscl=1
 Xres=1
```



c) It is difficult to see a difference between shift 1 and shift 2 in the overall production, although shift 1 clearly has at least one month of very low production. In general, the histogram of shift 2 appears to have greater production, but this is unclear without numerical summaries which will come in future chapters. The distribution of shift 3 appears to be relatively uniform covering a much larger range of shifts 1 and 2. Of the three shifts, shift 2 appears to be more consistent in their production.

3.23 a. WDS - Number of words in each advertisement. Discrete, Quantitative, Ratio

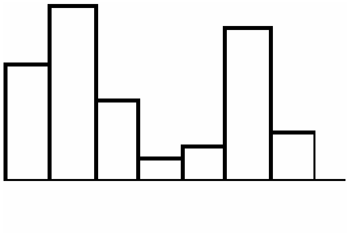
SEN - Number of sentences in each advertisement Discrete, Quantitative, Ratio

3SYL - Number of 3+ syllable words in each advertisement Discrete, Quantitative, Ratio

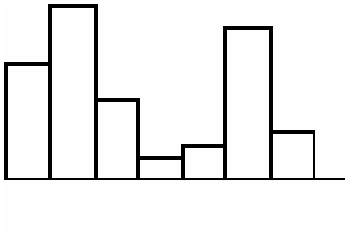
MAG - Which magazine in the sample. Discrete, Qualitative, Nominal

GROUP - Educational level of the magazine. Discrete, Qualitative, Ordinal b.

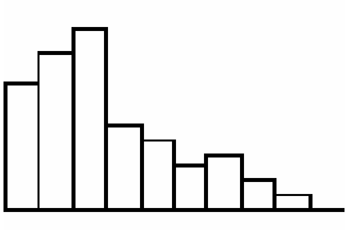
Number of Words in Ad - Bimodal



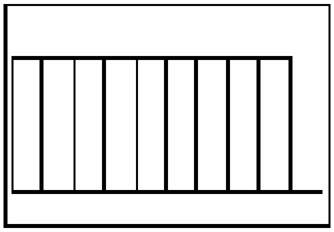
Number of Sentences in Ad
Approximately Bell Shaped



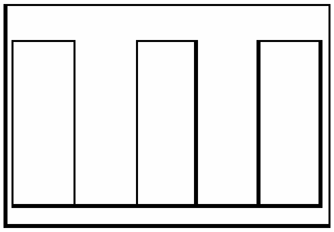
Number of 3 Syllable Words in Ad
Skewed Right



Magazines Used in Study
Uniform



Education Level
Uniform



3.25 a. January Temperature and July Temperature are both Continuous, Quantitative and Interval.

c. The temperatures in July are typically higher than the January temperatures.

CHAPTER 3

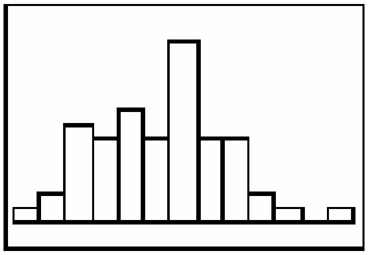
3.27 a. Mortality Rate - Ratio.

b.

Mortality Rate for U.S. Cities

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
800 = x < 825	1	1/59	1	1/59
825 = x < 850	2	2/59	3	3/59
850 = x < 875	7	7/59	10	10/59
875 = x < 900	6	6/59	16	16/59
900 = x < 925	8	8/59	24	24/59
925 = x < 950	6	6/59	30	30/59
950 = x < 975	13	13/59	43	43/59
975 = x < 1000	6	6/59	49	49/59
1000 = x < 1025	6	6/59	55	55/59
1025 = x < 1050	2	2/59	57	57/59
1050 = x < 1075	1	1/59	58	58/59
1075 = x < 1100	0	0	58	58/59
1100 = x < 1125	1	1/59	59	59/59
Total	**59**	**59/59**		

c. Annual Mortality in U.S. Cities



3.29 a. The level of measurement is nominal.

b. A bar graph would be best used for this data since the measurement scale is nominal.

CHAPTER 3

4 Measurements of Location and Position

Review Exercise Solutions

4.1 a) Both the mean and the median are measurements of the center of the data.

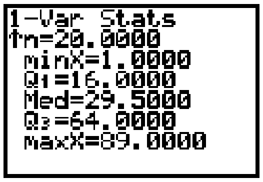
b) When the data is symmetric, the mean and the median are equal to each other. We could use either measure of center but the mean is preferred. When the data is not symmetric. The mean is influenced by extreme values whereas the median is always the value physically in the middle of the data.

c) The biggest advantage the median has over the mean is that the median is not influenced by extreme values.

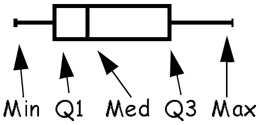
4.3 Yes, the new average is: $\frac{37(86) + 76 + 81 + 97}{37 + 3} = 85.9$.

4.5 You would ask that your client receive the mean salary. The distribution of salaries is clearly skewed right so that the mean salary will be higher than the median salary.

4.7 a) The values for the five-number summary are minimum = 1, Q_1=16, median = 29.5, Q_3 = 64, maximum = 89.



```
1-Var Stats
↑n=20.0000
 minX=1.0000
 Q1=16.0000
 Med=29.5000
 Q3=64.0000
 maxX=89.0000
```



b) This distribution is skewed right.

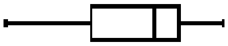
c) $P_{15} = \frac{x_3 + x_4}{2} = \frac{13 + 14}{2} = 13.5$

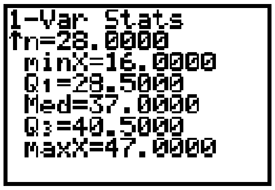
d) $P_{90} = \frac{x_{18} + x_{19}}{2} = \frac{84 + 85}{2} = 84.5$

e) $P_{85} = \frac{x_{17} + x_{18}}{2} = \frac{66 + 84}{2} = 75$

4.9 a) The variable of interest is nurturing tendency and the measurement scale is interval.

b) The values for the five-number summary are minimum = 16, Q_1 = 28.5, median = 37, Q_3 = 40.5, maximum = 47. See problem 4.7 b) for the labeling.





```
1-Var Stats
↑n=28.0000
 minX=16.0000
 Q1=28.5000
 Med=37.0000
 Q3=40.5000
 maxX=47.0000
```

c) $P_{80} = x_{23} = 42, \quad P_{90} = x_{26} = 45$

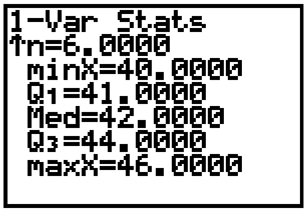
d) $\bar{x} = 34.5$ and M = 37. The median would be a better choice for measuring center since the distribution of the data set is skewed left.

4.11 Yes the biologist can use this information to estimate the total number of squirrels in the breeding ground. A box plot of the data shows that the distribution is skewed left. Using the median number of squirrels in each grid and multiplying by the total number of grids, a reasonable estimate for the total number of squirrels in the breeding ground is (71)(1478) = 104,938 squirrels.

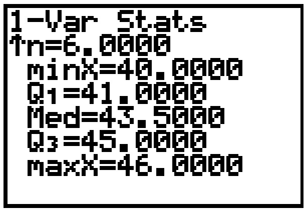
4.13 The mean and median are similar if the distributional shape of the data is symmetric. By knowing how close the mean and median are to one another and if the mean is greater than or less than the median, we know if the data has a small or large degree of skewness or is symmetric.

4.15 a) Variable: Sales, in thousands of dollars. Scale: Ratio.

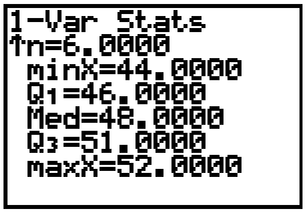
b) Campaign #1 Campaign #2 Campaign #3



```
1-Var Stats
↑n=6.0000
 minX=40.0000
 Q1=41.0000
 Med=42.0000
 Q3=44.0000
 maxX=46.0000
```



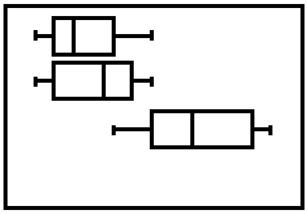
```
1-Var Stats
↑n=6.0000
 minX=40.0000
 Q1=41.0000
 Med=43.5000
 Q3=45.0000
 maxX=46.0000
```



```
1-Var Stats
↑n=6.0000
 minX=44.0000
 Q1=46.0000
 Med=48.0000
 Q3=51.0000
 maxX=52.0000
```

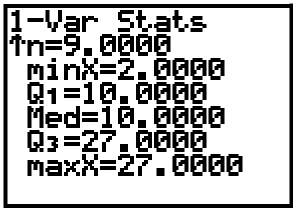
c) The side-by-side box-plot order shown has Campaign #1 on the top, then Campaign #2 followed by Campaign #3.

d) Based on the summary statistics and graphical displays, it appears that Campaign #3 is doing a better job.

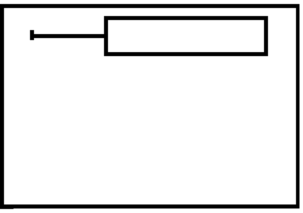


4.17 Answers will vary. The basic idea is that this was a silly statement, as worded. It is not possible for everyone to be above the 50th percentile. By definition, 50% are above and 50% are below.

4.19 Yes. Consider the following data: 2, 10, 10, 10, 10, 10, 27, 27, 27, 27. This will result in the five-number-summary shown below. The box plot for this data is also shown.



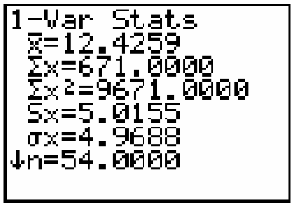
```
1-Var Stats
↑n=9.0000
 minX=2.0000
 Q1=10.0000
 Med=10.0000
 Q3=27.0000
 maxX=27.0000
```



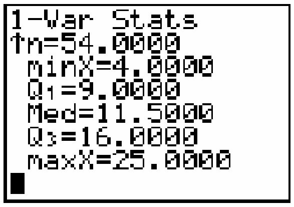
4.21 Answers will vary.

4.23 The data is skewed left. This can be easily observed once a box plot is drawn

4.25 a) Sentences: Mean = 12.4259, Median = 11.5, Mode = 9

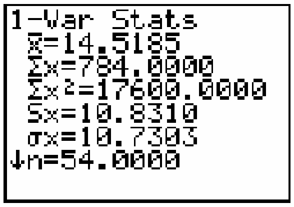


```
1-Var Stats
 x̄=12.4259
 Σx=671.0000
 Σx^2=9671.0000
 Sx=5.0155
 σx=4.9688
↓n=54.0000
```

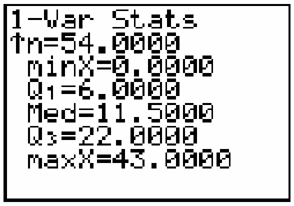


```
1-Var Stats
↑n=54.0000
 minX=4.0000
 Q1=9.0000
 Med=11.5000
 Q3=16.0000
 maxX=25.0000
█
```

3 Syllable : Mean = 14.5185, Median = 11.5, Mode = 6

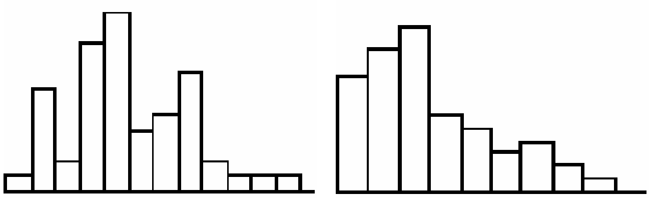


```
1-Var Stats
 x̄=14.5185
 Σx=784.0000
 Σx^2=17600.0000
 Sx=10.8310
 σx=10.7303
↓n=54.0000
```



```
1-Var Stats
↑n=54.0000
 minX=0.0000
 Q1=6.0000
 Med=11.5000
 Q3=22.0000
 maxX=43.0000
```

b. Number of Sentences in Ad, approximately Bell Shaped. Number of 3 Syllable Words in Ad, Right Skewed



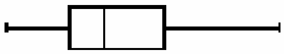
c. We would use the mean for the data "Sentences" since the data set is approximately bell shaped symmetric. We would use the median for the data set "3 Syllable" since the data set is right skewed.

d. No more than 50% of the ads have below 11.5 sentences and no more than 50% of the adds have more than 11.5 sentences.

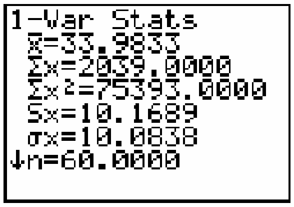
e. The average number of sentences in the ads is 14.5185.

f. Minimum = 0, 1st Quartile = 6 , Median = 11.5, 3rd Quartile = 22, Maximum = 43

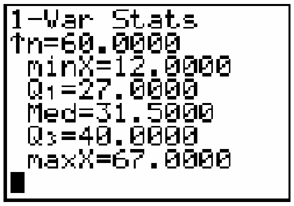
Both graphs represent the data well. The shape of the distribution is visible in either graph.



4.27 a. **January Temperature:** Mean = 33.9833, Median = 31.50, Mode = 24

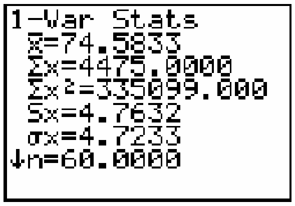


```
1-Var Stats
 x̄=33.9833
 Σx=2039.0000
 Σx²=75393.0000
 Sx=10.1689
 σx=10.0838
↓n=60.0000
```

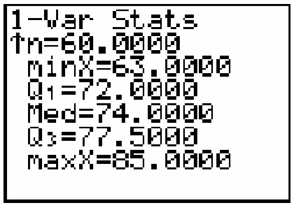


```
1-Var Stats
↑n=60.0000
 minX=12.0000
 Q1=27.0000
 Med=31.5000
 Q3=40.0000
 maxX=67.0000
█
```

July Temperature: Mean = 74.5833, Median = 74, Mode = 72

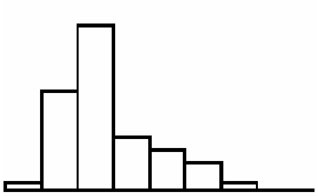


```
1-Var Stats
 x̄=74.5833
 Σx=4475.0000
 Σx²=335099.000
 Sx=4.7632
 σx=4.7233
↓n=60.0000
```

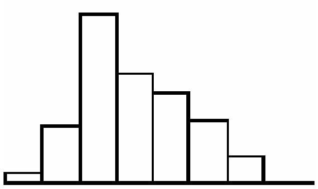


```
1-Var Stats
↑n=60.0000
 minX=63.0000
 Q1=72.0000
 Med=74.0000
 Q3=77.5000
 maxX=85.0000
```

b. **January Temperatures:** The proper measure of center would be the median of 31.50 degrees since the data is right skewed.

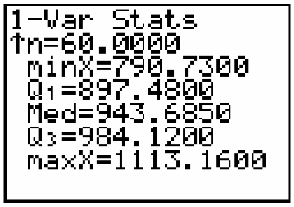


July Temperatures: The proper measure of center would be the median of 74 degrees since the data is right skewed.



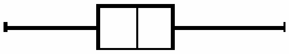
c. The mean January temperature of 33.9833 degrees is lower than the mean July temperature of 74.5833 degrees. The medians also show the difference in temperature.

4.29 a. Minimum = 790.73, 1st Quartile = 897.48, Median = 943.685, 3rd Quartile = 984.12, Maximum = 1113.16



```
1-Var Stats
↑n=60.0000
 minX=790.7300
 Q1=897.4800
 Med=943.6850
 Q3=984.1200
 maxX=1113.1600
```

b. Mortality Rate - The distribution is approximately bell shaped.



c. The proper measure of center would be the mean since the data set is bell shaped symmetric.

d. Location = 60(0.20) = 12 so we will average the 12th and the 13th position
P20 = 889.59

$$\frac{(887.47 + 891.71)}{2} = 889.59$$

e. Location = 60(0.80) = 48 so we will average the 48th and the 49th position
P80 = 992.97

$$\frac{(991.29 + 994.65)}{2} = 992.97$$

f. The twentieth percentile separates the bottom twenty percent of the data from the top eighty and the eightieth percentile separates the bottom eighty percent from the top twenty percent.

CHAPTER 4

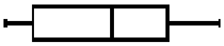
CHAPTER 4

5 Measurements of Variability

Review Exercise Solutions

5.1 The sample standard deviation is the square root of the sample variance. The sample variance is an average of the squared distances between the observed data values and the sample mean. Thus, the variance is a measurement of data dispersion based on the sample mean. If data is skewed, it is generally agreed that the better measurement of the center of the data is the median rather than the mean. It would seem intuitive to then base a measurement of the data dispersion based on the appropriate measure of the center. As such, the 5-number summary may be a more appropriate overall measure of data dispersion for skewed data.

5.3 a) Sample mean = 10.7143, sample median = 11.0000. The mean is the more appropriate measure of center. This is based on the fact that the box plot appears relatively symmetric.



b) Range = 8, Sample variance = $2.8702^2 = 8.2380$, Sample standard deviation = 2.8702.

5.5 All data must have the same value.

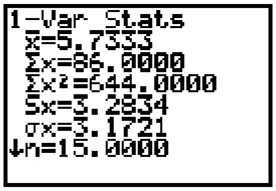
5.7 Sample mean = 5.7333

Sample median = 5.0000

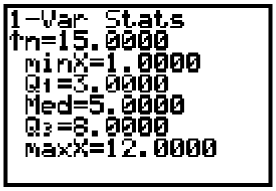
Range = 12 - 1 = 11

Sample variance = $3.2834^2 = 10.7807$

Sample standard deviation = 3.2834



```
1-Var Stats
 x̄=5.7333
 Σx=86.0000
 Σx²=644.0000
 Sx=3.2834
 σx=3.1721
↓n=15.0000
```



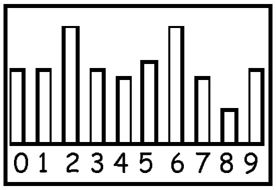
```
1-Var Stats
↑n=15.0000
 minX=1.0000
 Q1=3.0000
 Med=5.0000
 Q3=8.0000
 maxX=12.0000
```

5.9 Answers will vary due to the randomness of the numbers generated.

5.11 Answers will vary.

5.13 a) The random variable is the number drawn. The measurement scale is nominal. Although numbers are used, mathematical operations with this data lacks meaning. Each number is simply a label having no more mathematical meaning than colors.

b) A bar graph is the appropriate graphical representation because the data is nominal. If the game is "fair" then we would expect to see the frequencies for each group to be approximately uniformly distributed. Although not perfect, the distribution does appear to be approximately uniform so the game does

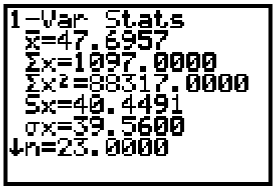


appear to be a “fair” game. The low observed frequency of 8’s is of some concern, but when the small sample size is taken into consideration the amount of concern diminishes.

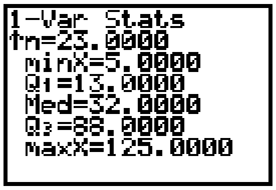
5.15 Answers will vary.

5.17 a) The mean, standard deviation and all elements for the 5 number summary are shown in the accompanying screen shots.

b) The mean is substantially greater then the median which suggests the data is skewed right. A box plot quickly confirms this observation. Due to the skewness of the data, the median would be the better measurement of the distribution center.



```
1-Var Stats
 x̄=47.6957
 Σx=1097.0000
 Σx²=88317.0000
 Sx=40.4491
 σx=39.5600
↓n=23.0000
```

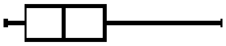


```
1-Var Stats
↑n=23.0000
 minX=5.0000
 Q1=13.0000
 Med=32.0000
 Q3=88.0000
 maxX=125.0000
```

5.19 a) The random variable is the population of a county. The measurement scale is ratio.

b) Sample mean = 176024.1500. Sample median = 163586.0000. Yes, there does appear to be a big difference between the sample mean and median, they are 12438.5 apart in a distribution that has a range of 378039. As such, we would expect the distribution to be skewed right (the sample mean is bigger than the sample median).

c) Yes. It demonstrates that the sample data is in fact skewed right.

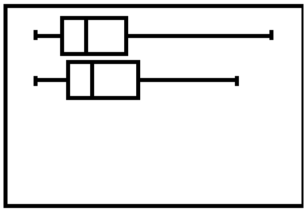


5.21 a) Variable: Level of unemployment. Scale: Ratio.

b) Females are on top.

c) Both data sets are skewed right so the medians are the appropriate measure of the center. The five-number summary would be the appropriate measure of spread.

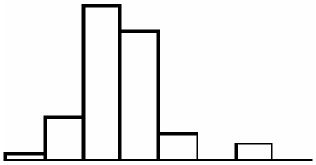
d) Approximately 50% of the countries surveyed reported an unemployment rate for women of approximately 7.95% or less.

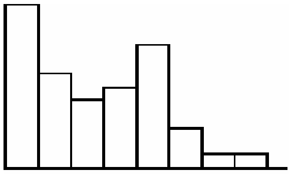


5.23 a) Hundreds of Cigarettes per Thousand - Continuous, Quantitative, Ratio

Bladder Cancer Deaths per 100K - Continuous, Quantitative, Ratio

b) Cigarettes by State for the U.S. - Approximately Bell Shaped / Bladder Cancer Deaths - Skewed Right

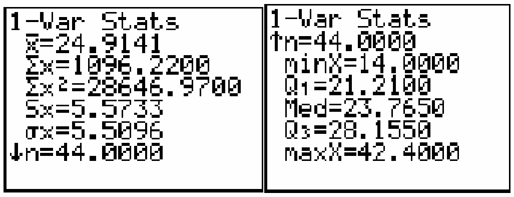




Cigarettes:

c. Mean = 24.9141, Median = 23.7650, Mode = 23.44

d. Range = 42.4 - 14 = 28.4, Standard Deviation = 5.5733, Variance = 5.57332 = 31.0617

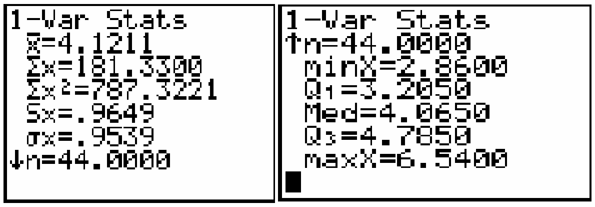


```
1-Var Stats            1-Var Stats
 x̄=24.9141             ↑n=44.0000
 Σx=1096.2200            minX=14.0000
 Σx²=28646.9700          Q1=21.2100
 Sx=5.5733               Med=23.7650
 σx=5.5096               Q3=28.1550
↓n=44.0000               maxX=42.4000
```

Bladder Cancer:

c. Mean = 4.1211, Median = 4.0605, Modes = 2.90, 3.72, 4.04, 4.46, 4.78

d. Range = 6.54 - 2.86 = 3.68, Standard Deviation = 0.9649, Variance = 0.96492 = 0.9310



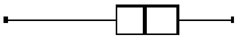
```
1-Var Stats            1-Var Stats
 x̄=4.1211              ↑n=44.0000
 Σx=181.3300             minX=2.8600
 Σx²=787.3221            Q1=3.2050
 Sx=.9649                Med=4.0650
 σx=.9539                Q3=4.7850
↓n=44.0000               maxX=6.5400
```

e. **Cigarettes:** The data is approximately symmetric so the proper numerical summary is the mean, 24.9141, and the standard deviation, 5.5733.

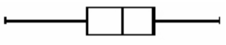
Bladder Cancer: The data is right skewed so the proper numerical summary is the median, 4.0650 and the 5 number summary, Minimum = 2.86, First Quartile = 3.2050, Median = 4.0650, Third Quartile = 4.7850, Maximum = 6.54

5.25 a. 1968 and 1972 are both Continuous, Quantitative and Ratio.

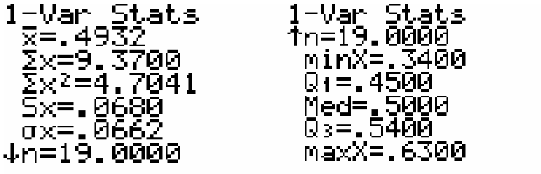
b. Proportion of Women in the Work Force in 1972 in Major Cities, approximately symmetric.



Proportion of Women in the Work Force in 1968 in Major Cities, approximately symmetric.

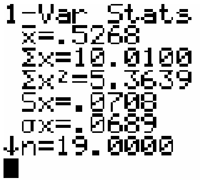


c. **1968:** Mean = 0.4932, Median = 0.5, Mode = 0.45

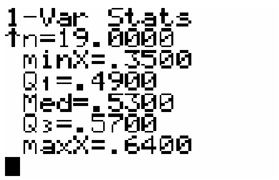


```
1-Var Stats            1-Var Stats
 x̄=.4932               ↑n=19.0000
 Σx=9.3700               minX=.3400
 Σx²=4.7041              Q1=.4500
 Sx=.0680                Med=.5000
 σx=.0662                Q3=.5400
↓n=19.0000               maxX=.6300
```

1972: Mean = 0.5268, Median = 0.53, Mode = 0.45, 0.50, 0.52, 0.53, 0.55, 0.57, 0.64



```
1-Var Stats
 x̄=.5268
 Σx=10.0100
 Σx²=5.3639
 Sx=.0708
 σx=.0689
↓n=19.0000
```



```
1-Var Stats
↑n=19.0000
 minX=.3500
 Q1=.4900
 Med=.5300
 Q3=.5700
 maxX=.6400
```

d. **1968:** Range = 0.63 - 0.34 = 0.29, Standard Deviation = 0.68, Variance = 0.682 = 0.4624

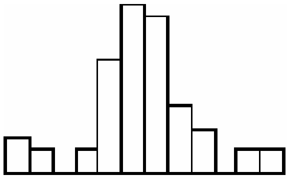
1972: Range = 0.64 - 0.35 = 0.29, Standard Deviation = 0.0708, Variance = 0.07082 = 0.0050

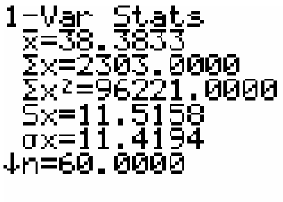
e. **1968:** The data is approximately symmetric so the proper numerical summary is the mean, 0.4932, and the standard deviation, 0.68.

1972: The data is approximately symmetric so the proper numerical summary is the mean, 0.5268, and the standard deviation, 0.0708.

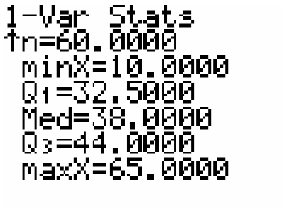
5.27 a. Rain - Continuous, Quantitative and Ratio.

b. Annual Rainfall in U.S. Cities





```
1-Var Stats
 x̄=38.3833
 Σx=2303.0000
 Σx²=96221.0000
 Sx=11.5158
 σx=11.4194
↓n=60.0000
```



```
1-Var Stats
↑n=60.0000
 minX=10.0000
 Q1=32.5000
 Med=38.0000
 Q3=44.0000
 maxX=65.0000
```

c. Rainfall: Mean = 38.3833, Median = 38, Mode = 35, 36, 42

d. Rainfall: Range = 65 - 10 = 55, Standard Deviation = 11.5158, Variance = 11.51582 = 132.6136

e. For the variable Rainfall, the Empirical Rule would be the most appropriate since the data is approximately bell shaped symmetric.

5.29. a. 68% b. 99.7% c. 47.5% d. 0.15% e. 2.65%

6 Probability

Review Exercise Solutions - Sections 1 - 4

6.1 Empirical probability is a probability that is calculated based on something that is actually observed, such as the number of heads observed when flipping a coin 1000 times. Theoretical probabilities are the true probabilities that can be calculated based on an understanding of the process under various assumptions. The theoretical probability of observing a head is 0.50 under the assumption that the coin is "fair."

6.3 a) The sample space can be enumerated as S = {(H,H,H), (T,H,H), (H,T,H), (H,H,T), (T,T,H), (H,T,T), (T,H,T), (T,T,T)} where the ordered triple (nickel, dime, quarter) is used to represent the possible outcomes.

b) P(exactly one head is observed) = 3/8. P(at least 1 head was observed) = 7/8.

c) Answers will vary because it is based on actually flipping coins.

6.5 The information is incorrect. It is not mathematically possible for a probability to be greater than 1. The person may be reporting the odds of rain, but not the probability of rain.

Review Exercise Solutions - Sections 5, 6 and 7

6.9 For every 1.03 days of rain we observed 1 day of no rain.

6.11 True. By definition, two events that are mutually exclusive are dependent.

Review Exercise Solutions - Section 8

6.13 Answers will vary. The main idea is that simple probabilities are represented by the number of ways the event can occur divided by size of the sample space. Since the event never occurs, the probability would then be $\frac{0}{\text{S(n)}}$, which is 0.

6.15

$$a)\ P(B^c) = 1 - P(B) = 1 - 0.70 = 0.30$$

$$b)\ P(A \cup B) = P(A) + P(B) - P(A \cap B) \Rightarrow\ P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

$$= 0.40 + 0.70 - 0.80 = 0.30$$

$$c)\ P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.30}{0.70} = 0.4286$$

d) If events A and B were independent then P(A|B) would be the same as P(A), but as we can see from the work above, they are not the same.

CHAPTER 6

6.17 This problem is easily addressed if we first recognize that there are several independent events given in the setup. First, the mother passing on the trait and the father passing on the trait (or not) are independent. Likewise, the first child having the trait (or not) and the second child having the trait (or not) are also independent.

The next thing we should do is list the sample space. A subscript if Y indicates the trait was passed on by that parent.

$$S = \left\{ \left(M_Y, F_Y \right), \left(M_N, F_N \right), \left(M_N, F_Y \right), \left(M_Y, F_N \right) \right\}$$

With this information we can address the question.

a) P(first born has the trait) = $P(M_Y, F_Y) = 0.20(0.35) = 0.0700$

b) P(second born has the trait) = $P(M_Y, F_Y) = 0.20(0.35) = 0.0700$

c) P(both 1st and 2nd born have the trait) = 0.70(0.70)=0.0049 due to independence.

d) P(both 1st and 2nd born do not have the trait). To answer this, we must find the probability that the first born does not have the trait. This will be given as:

$$P\left(M_N, F_N \right) + P\left(M_N, F_Y \right) + P\left(M_Y, F_N \right) = 0.80(0.65) + 0.80(0.35) + 0.20(0.65) = 0.93$$

Since the two events are independent, the answer is simply .

e) Again, independence plays an important roll. The answer is 0.0700.

6.19 a) Answers will vary. The preferred solution would say that the color of car a person is driving has nothing to do with the speed that person drives; however, a student could successfully show that brighter colored cars, such as red, draw the attention of law enforcement which results in a higher incidence of traffic tickets.

b) Dependent. The more hours you work on your job, the less time you have to study for your class.

c) Independent. There is no reasonable connection between the two events that can be made.

d) Answers will vary. An argument for dependence can be made in that larger people have both larger hands and larger feet, although exceptions can always be found.

e) Answers will vary. The preferred solution is dependence in that larger families are seldom seen with single parents although exceptions can easily be found.

6.21 a) 403-403(0.461+0.31+0.045)=74.152. 74 people answered "Don't know."

b) Odds = $\dfrac{P(success)}{P(failure)} = \dfrac{0.461}{1 - 0.461} = 0.8553$.

CHAPTER 6

6.23 a) The random variables are (1) Gender, which is nominal and (2) Belief in the Afterlife, which is also nominal.

b) $$P(Belief\ = yes) = \frac{806}{1076} = 0.7491 \quad \cdot$$

c) $$P(Belief\ and\ \ female) = \frac{435}{1076} = 0.4043 \quad .$$

d) $$P(Belief\ = no\,|\ female) = \frac{153}{588} = 0.2602 \quad .$$

e) If independentt then P(Belief = yes) and P(Belief = yes | Males) will be the same. We already found P(Belief = yes) is 806/1076. P(Belief = yes | Males) = 371/488. These two probabilities are deferent hence the two events are dependent, not independent.

f) $$\frac{806}{270} = 2.9852$$

g) $$\frac{92}{219} \approx 0.4200$$

6.25 a) (1) Gender, nominal, (2) Type of exam taken, nominal.

b) $\frac{37 + 30}{427} = 0.1569$

c) To do this problem we must first assume that the data consists of those persons who took only one test. If a person in this study took more than one test then there would be no way we could answer this question. $\frac{148}{427} = 0.3466$

d) $\frac{1}{427} = 0.0023$ e) $\frac{23 + 34}{427} = 0.1335$ f) $\frac{5}{427} = 0.0117$

g) $\frac{227 + 42}{427} = 0.6300$ h) $\frac{70}{112} = 0.6250$ i) $\frac{62 + 42}{200} = 0.5200$

CHAPTER 6

7 Random Variables and Probability Distributions

Review Exercise Solutions

7.1 Answers will vary. The key idea is that a random variable records the outcomes of an experiment or process in which data is generated randomly.

7.3 False. A continuous random variable has an uncountable (infinite) number of possible values.

7.5 a) Discrete b) Continuous c) Continuous d) Discrete

e) Continuous f) Discrete g) Discrete h) Discrete

7.7 False. It is very possible to collect sample data that has an average equal to the theoretical mean if the distribution is discrete. It is not possible for continuous distributions; however, you may see many instances where a calculated mean from data collected from a continuous distribution is equal to the population mean. This is only because all empirical data is measured discretely even though the random variable is actually continuous. Humans do not have the ability to measure anything with infinite precision.

7.9 Yes.

1) Each probability is greater than or equal to 0 and less than or equal to 1.

2) The sum of the probabilities is 45/45 = 1.

x	P(x)
1	4/45
2	11/45
3	30/45

7.11 a) The random variable is how often batteries in a smoke alarm should be changed, possibly represented by *T*.

b) Discrete.

c) Yes. All probabilities are greater than or equal to 0 and less than or equal to 1. In addition, the sum of the probabilities is equal to 1. Four possible responses were offered however the fourth response, “don’t know” is not listed. Based on the information given it must be equal to 18.4%.

7.13 a) Yes. All probabilities are greater than or equal to 0 and less than or equal to 1. In addition, the sum of the probabilities is equal to 1

b) More than 9 tickets means 10, 11 or 12 tickets. That probability is 0.16 + 0.08 + 0.05 = 0.29.

c) The probability of less than 5 tickets, based on the data collected, is 0 because 5 was the minimum number of tickets recorded.

d) At least 9 tickets is 9, 10, 11 or 12 tickets. That probability is 0.15 + 0.16 + 0.08 + 0.05 = 0.44.

e) No more than 6 tickets means 5 or 6 tickets. That probability is 0.12 + 0.14 = 0.26.

f) From 6 to 10 tickets means 6 tickets were written, or 7 tickets, or ... , or 10 tickets. That probability is 0.14 + 0.10 + 0.20 + 0.15 + 0.16 = 0.75.

g) The average number of tickets can be found by:

$$\mu = \sum xP(X = x) = 5(0.12) + 6(0.14) + 7(0.10) + 8(0.20) + 9(0.15) + 10(0.16) + 11(0.08 + 12(0.05)$$

$$= 8.17$$

h) The standard deviation is the square root of the variance. The variance is:

$$\sigma^2 = \sum (x - \mu)^2 P(X = x)$$

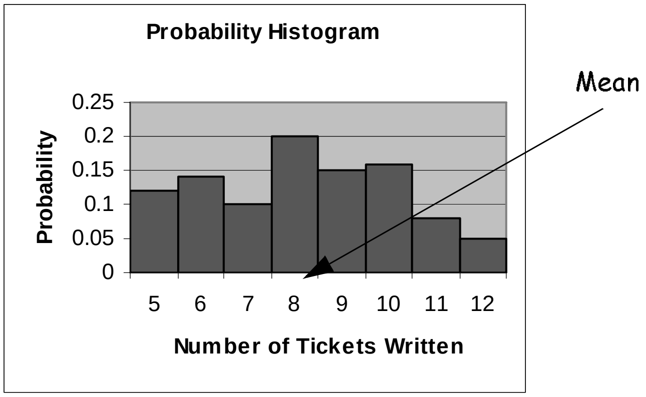
$$= (5 - 8.17)^2 (0.12) + (6 - 8.17)^2 (0.14) + (7 - 8.17)^2 (0.10) + (8 - 8.17)^2 (0.20) +$$

$$(9 - 8.17)^2 (0.15) + (10 - 8.17)^2 (0.16) + (11 - 8.17)^2 (0.08) + (12 - 8.17)^2 (0.05)$$

$$= 4.0211$$

This implies the standard deviation is 2.0053.

i) A probability histogram on a TI-83 is difficult. As such, this probability histogram was constructed using Excel.



j) An arrow has been added to the histogram approximating the location of the mean, which is near the center of the distribution.

7.15 The value 2.3 is simply an average. It does not mean any given household actually has 2.3 children. Suppose we went to 3 households and found they had 1, 4, 2 and 3 children. The average is 10/4 = 2.5. None of the households had 2.5 children, but I can recover the total number of children surveyed by multiplying the average, 2.5, by the number of households surveyed. 2.5(4) = 10.

8 The Binomial Probability Distribution

Review Exercise Solutions

8.1 1) Each trial is random and independent of the others.

2) The number of trials is fixed.

3) There are two possible outcomes, which we label as a success or failure.

4) The probability of a success, denoted by π, remains constant for each trial. The probability of success, plus the probability of a failure, is equal to one.

5) The random variable for a binomial experiment records the number of success in n trials.

8.3 This is a binary response experiment because we are looking for a 3. The value 3 is considered a success whereas all other values are considered a failure.

8.5 a) No. The question is asking for the number of traffic accidents in a 30 day period so the number of trials cannot be fixed. If the question had been "the number of days out of 30 that have a traffic accident" then it could be considered as a binomial random variable.

b) Yes. All of the criteria outlined in problem 8.1 are satisfied.

c) Yes. All of the criteria outlined in problem 8.1 are satisfied.

d) No. The number of trials is not constant.

e) No. The probability of success is changing for each trial because you are eating the cookie drawn rather than replacing it.

f) No. The number of trials is not constant.

g) Yes. All of the criteria outlined in problem 8.1 are satisfied.

h) No. The random variable is the amount of time rather than a success or failure.

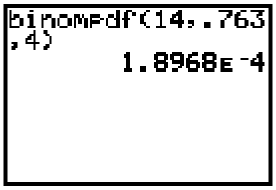
8.7 False. Typically we see multiple trials, but the properties for a binomial still hold for n = 1.

8.9 Suppose you were selling candy bars for $1.00 each and were told you will keep 30% of the total amount of what you sell. If you sold 100 boxes then you would **expect** to profit $30.00. You came to this conclusion by simple arithmetic: 0.30(100) = 30. The **expected value** is another term for *mean*. The mean for the binomial distribution is the number of success you would **expect** to see given a probability of success and the number of trials.

8.11 What is the probability of observing 7 success in 25 trials where the probability of success is 0.33?

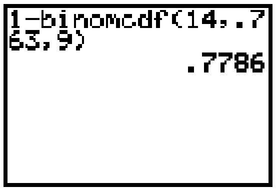
CHAPTER 8

8.13 a) $P(X = 4|n = 14, \pi = 0.763) = 0$ or approximately 0.



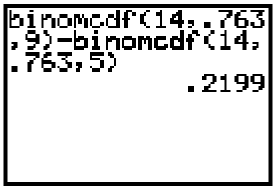
```
binompdf(14,.763
,4)
         1.8968E-4
```

b) $P(X \geq 10|n = 14, \pi = 0.763) = 1 - P(X \leq 9|n = 14, \pi = 0.763) = 0.7786$



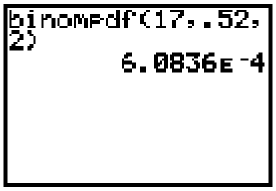
```
1-binomcdf(14,.7
63,9)
            .7786
```

c) $P(6 \leq X \leq 9|n = 14, \pi = 0.763) = P(X \leq 9|n = 14, \pi = 0.763) - P(X \leq 5|n = 14, \pi = 0.763) = 0.2199$



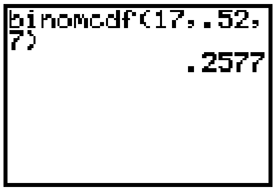
```
binomcdf(14,.763
,9)-binomcdf(14,
.763,5)
            .2199
```

8.15 a) 0.0006



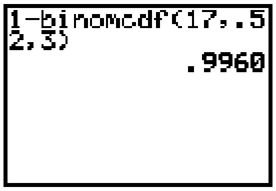
```
binompdf(17,.52,
2)
         6.0836E-4
```

b) 0.2577



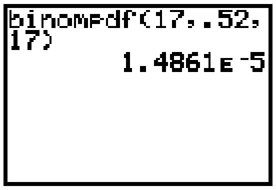
```
binomcdf(17,.52,
7)
            .2577
```

c) 0.9960



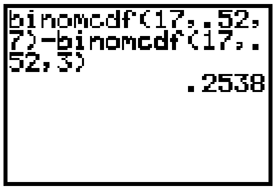
```
1-binomcdf(17,.5
2,3)
           .9960
```

d) 0.000014 or approximately 0



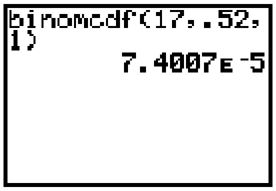
```
binompdf(17,.52,
17)
      1.4861E-5
```

e) 0.2538



```
binomcdf(17,.52,
7)-binomcdf(17,.
52,3)
           .2538
```

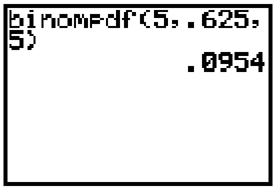
f) 0.000074 or approximately 0



```
binomcdf(17,.52,
1)
      7.4007E-5
```

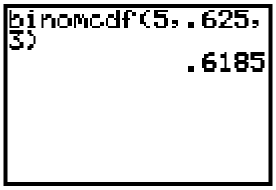
8.17 If the odds of winning is 5:3 then the probability of a win is $\frac{5}{5+3} = 0.6250$.

a) Probability they are undefeated is the same as “have no losses” or $P(X = 5|n = 5, \pi = 0.6250) = 0.0954$.



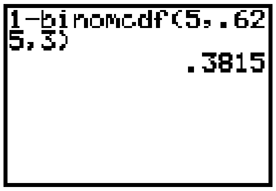
```
binompdf(5,.625,
5)
           .0954
```

b) No more than 3 games means 3 or fewer which is represented as $P(X \le 3 | n = 5, \pi = 0.6250) = 0.6185$.



```
binomcdf(5,.625,
3)
                .6185
```

c) $P(X \ge 4 | n = 5, \pi = 0.6250) = 0.3815$

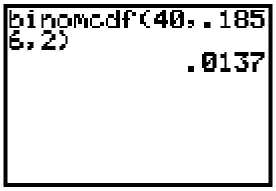


```
1-binomcdf(5,.62
5,3)
                .3815
```

8.19 Probability of a success (infant death) is 18/97 = 0.1856.

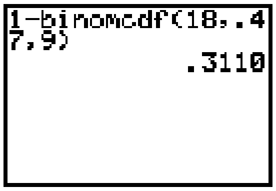
a) The expected value (mean of the distribution) is $40(0.1856) = 7.4240$. We would expect to see, on average, 7.424 infant deaths in every 40 child deaths.

b) The way we will approach this is to look at the probability of observing 2 or fewer infant deaths out of 40 child deaths. If this probability is small, then it will suggest the efforts of NHTSA to get the word out on the dangers of infants in the front seats of automobiles has been effective. If the probability is large, then it suggests their efforts have not been successful. $P(X \le 2 | n = 40, \pi = 0.1856) = 0.0137$. Since the probability is small, we will conclude the efforts of NHTSA have been successful.



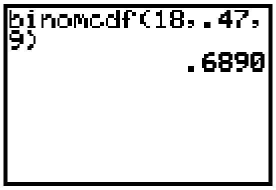
```
binomcdf(40,.185
6,2)
                .0137
```

8.21 a) $P(X \ge 10 | n = 18, \pi = 0.47) = 1 - P(X \le 9 | n = 18, \pi = 0.47) = 0.3110$



```
1-binomcdf(18,.4
7,9)
                .3110
```

b) $P(X \le 9|n = 18, \pi = 0.47) = 0.6890$

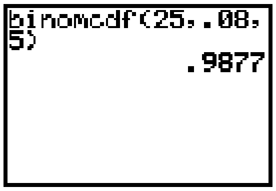


```
binomcdf(18,.47,
9)
               .6890
```

c) They are compliments of each other. The probability that at least 10 exercise three or more times per week plus the probability that no more than 9 exercise three or more times per week is 0.3110 + 0.6890 = 1.0000.

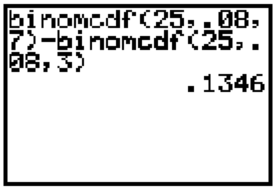
d) The expected value (mean of the distribution) is $18(0.47) = 8.46$. We would expect to see, on average, 8.46 elderly married males that exercise three or more times per week out of 18 elderly males surveyed.

8.23 a) $P(X \le 5|n = 25, \pi = 0.08) = 0.9877$



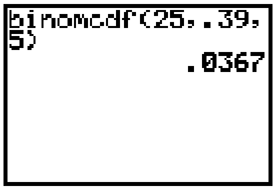
```
binomcdf(25,.08,
5)
               .9877
```

b) $P(3 < X < 8|n = 25, \pi = 0.08) = P(X \le 7|n = 25, \pi = 0.08) - P(X \le 3|n = 25, \pi = 0.08) = 0.1346$



```
binomcdf(25,.08,
7)-binomcdf(25,.
08,3)
               .1346
```

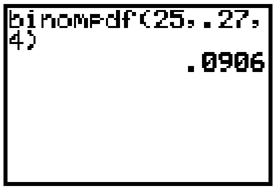
c) $P(X \le 5|n = 25, \pi = 0.39) = 0.0367$



```
binomcdf(25,.39,
5)
               .0367
```

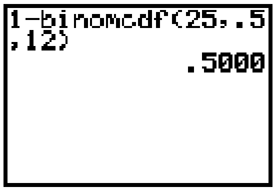
d) The expected value for the alcohol source is 25(0.08)=2. The expected value for the government source is 25(0.39)=9.75.

8.25 a) For red beans, the probability of success is 0.27. $P(X=5|n = 25, \pi = 0.27) = 0.0906$



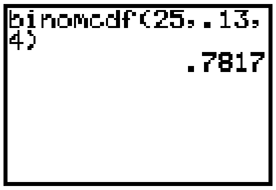
```
binompdf(25,.27,
4)
                 .0906
```

b) $P(X \geq 13|n = 25, \pi = 0.50) = 0.5000$



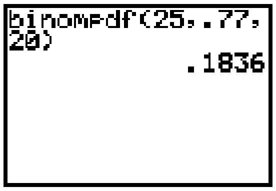
```
1-binomcdf(25,.5
,12)
                 .5000
```

c) $P(X \leq 4|n = 25, \pi = 0.13) = 0.7817$



```
binomcdf(25,.13,
4)
                 .7817
```

d) $P(X=20|n = 25, \pi = 0.77) = 0.1836$



```
binompdf(25,.77,
20)
                 .1836
```

e) Black Beans: 25(0.50) = 12.5, Red Beans: 25(0.27)=6.75, Pinto Beans: 25(0.13)=3.25, Navy Beans: 25(0.10)=2.5.

9 The Normal Distribution

Review Exercise Solutions

9.1 The standard normal distribution is a normal distribution that is centered at zero and has a standard deviation of one. Any normal distribution can be transformed to a standard normal distribution by converting its values to z-scores. Once this is done, z-scores have intuitive meaning to us because of the empirical rule. This provides us immediately with an intuitive measure for how “unusual” a particular observation may, or may not, be.

9.3 True. A normal distribution is symmetric. The mean is equal to the median in all symmetric distributions.

9.5 a) The distribution with a mean of 2 and standard deviation of 6 will have the highest values.

b) The distribution with a mean of 2 and standard deviation of 6 will have the lowest values.

c) The above are true because +/- 3 standard deviations for the distribution with a mean of 2 and standard deviation of 6 is -16 and 20 whereas +/- 3 standard deviations for the distribution with a mean of 4 and standard deviation of 3 is -5 and 13.

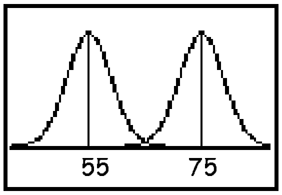
9.7 a) $z = \frac{12.31 - 6.3}{3.17} = 1.8959$ b) $z = \frac{8.2 - 6.3}{3.17} = 0.5994$ c) $z = \frac{2.1 - 6.3}{3.17} = -1.3249$

d) $z = \frac{15.8 - 6.3}{3.17} = 2.9968$

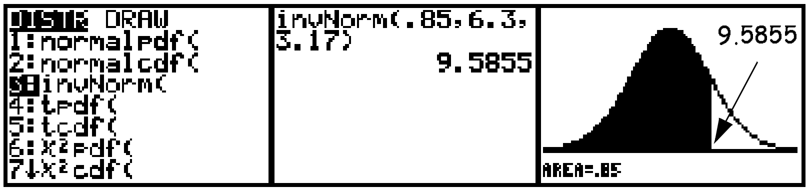
9.9 Answers will vary, but the basic idea is that a z-score is a standardized score. It represents the number of standard deviations a particular value is away from the mean where a positive value indicates it is above the mean and a negative value indicates it is below the mean. Based on the empirical rule, we expect almost all of our data to fall within 3 standard deviations of the mean, so if the data is approximately normal or at least reasonably bell shaped, then an observation 3.4 standard deviations below the mean would be very unusual. According the Chebychev’s rule, we expect *at least* 91.34% of our data to fall within 3.4 standard deviations of the mean, so once again, regardless of the shape of the distribution, an observation that is 3.4 standard deviations from the mean is considered to be unusual.

9.11 A z-score of -1.96 indicates the value of interest is 1.96 standard deviations *below* the mean.

9.13 The two distributions are identical in shape differing only in location. Both normal distributions have the same standard deviation which tells us the spread of the curves will be identical. Since the means are different the second curve is nothing more than the first curve with a horizontal translation of 20 units.



9.15 a) The 85th percentile is that value that will give you 85% of the data values below and 15% above. We can obtain that value from the inverse normal function on the TI-83. The solution is 9.5855.

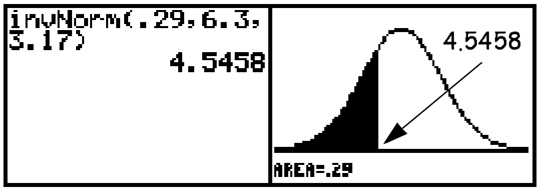


```
DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:tpdf(
5:tcdf(
6:χ²pdf(
7↓χ²cdf(
```

```
invNorm(.85,6.3,
3.17)
            9.5855
```

```
AREA=.85
```

b) The 29th percentile is that value that will give you 29% of the data values below and 71% above. We can obtain that value from the inverse normal function on the TI-83. The solution is 4.5458.

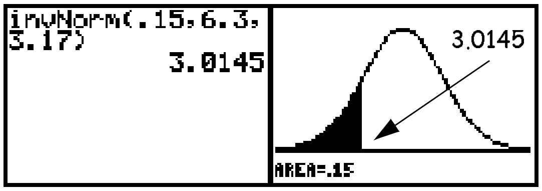


```
invNorm(.29,6.3,
3.17)
            4.5458
```

```
AREA=.29
```

c) There is no calculation to complete here. The 50th percentile is the same as the median which is the same as the mean for a normal distribution, which is 6.3.

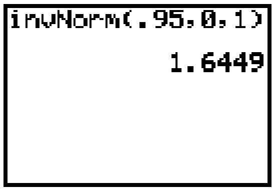
d) The 15th percentile is that value that will give you 15% of the data values below and 85% above. We can obtain that value from the inverse normal function on the TI-83. The solution is 3.0145.



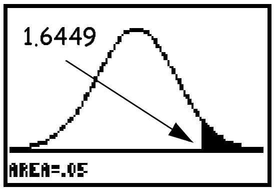
```
invNorm(.15,6.3,
3.17)
            3.0145
```

```
AREA=.15
```

9.17 a) The value that results in 5% above in a standard normal distribution is the same value that results in 95% below. This can be obtained from the inverse normal command on the TI-83. That value is 1.6449. If we attempt to convert the observed value, which is 45, to a z-score we will see the only missing value is the standard deviation. Since the problem told us 5% were above, we were able to find the actual z-score. With just a little algebra we are able to solve for the missing value, σ. The standard deviation is 4.8635.



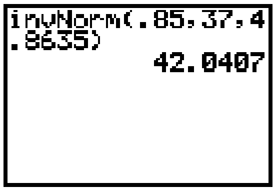
```
invNorm(.95,0,1)
          1.6449
```



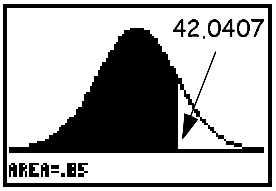
$$z=\frac{obs-\mu}{\sigma}\Rightarrow 1.6449=\frac{45-37}{\sigma}$$

$$\Rightarrow \sigma=\frac{45-37}{1.6449}=4.8635$$

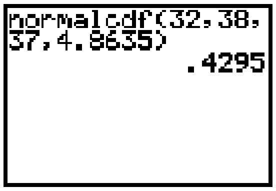
b) The 85th percentile can be obtained from the inverse normal command which is 42.0407.



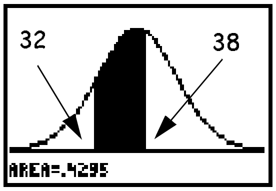
```
invNorm(.85,37,4
.8635)
         42.0407
```



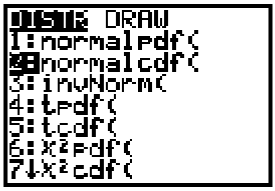
c) The percentage of cars that will drive within 3 mph of the speed limit is the same as saying P(32 < S < 38) for S a normally distributed random variable with a mean of 37 and a standard deviation of 4.8635. The answer is 42.95%.

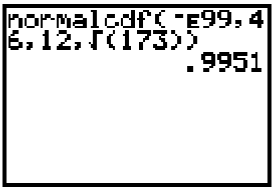


```
normalcdf(32,38,
37,4.8635)
           .4295
```

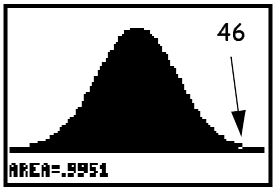


9.19 a) $P(R < 46)$

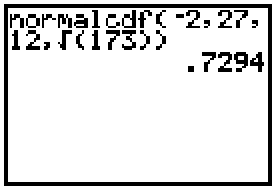




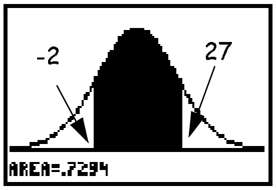
```
normalcdf(-E99,4
6,12,√(173))
           .9951
```



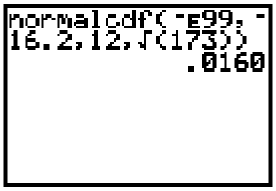
b) $P(-2 < R < 27)$



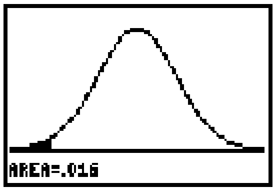
```
normalcdf(-2,27,
12,√(173))
           .7294
```



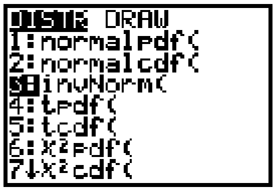
c) $P(R < -16.2)$

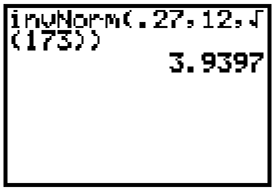


```
normalcdf(-E99,-
16.2,12,√(173))
            .0160
```

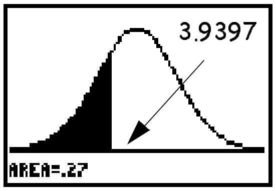


d) The 27th percentile is that value such that 27% of the data observations fall below and 73% fall above. This can be obtained from the inverse normal command on the TI-83. The answer is 3.9397.

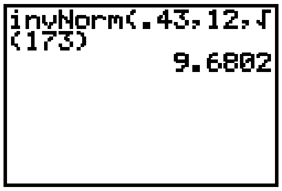




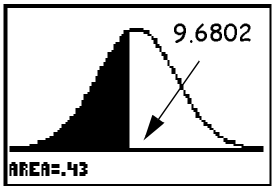
```
invNorm(.27,12,√
(173))
          3.9397
```



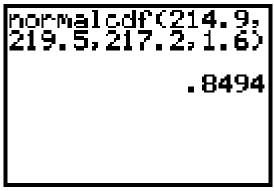
e) The problem is describing the 43th percentile. The answer is 9.6802.



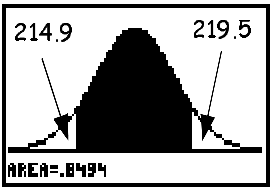
```
invNorm(.43,12,√
(173))
          9.6802
```



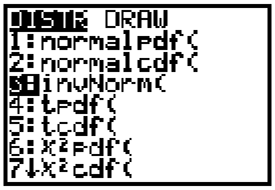
9.21 a) $P(214.9 < R < 219.5)$

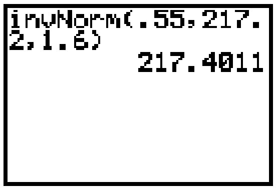


```
normalcdf(214.9,
219.5,217.2,1.6)
            .8494
```

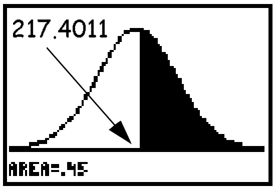


b) 1-0.45 = 0.55. This question is actually asking for the 55th percentile. The answer is 217.4011.

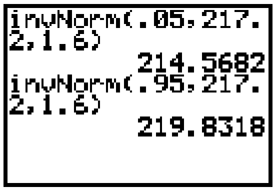




```
invNorm(.55,217.
2,1.6)
         217.4011
```



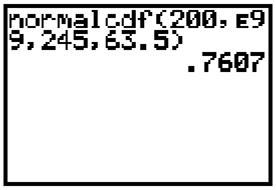
c) 5th percentile = 214.5682. The 95th percentile = 219.8318. The minimum and maximum amount of drink that is found to be acceptable is 214.5682 grams and 219.8318 grams.



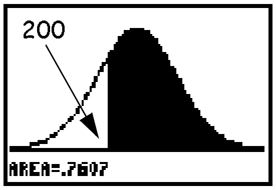
```
invNorm(.05,217.
2,1.6)
         214.5682
invNorm(.95,217.
2,1.6)
         219.8318
```

CHAPTER 9

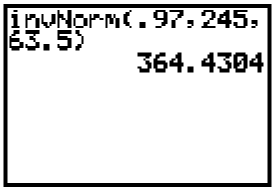
9.23 a) $P(D > 200)$



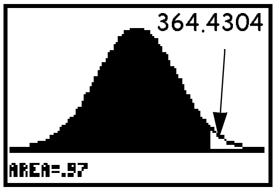
```
normalcdf(200,E9
9,245,63.5)
           .7607
```



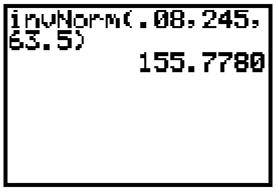
b) This is asking for the 97th percentile. The answer is 364.4304.



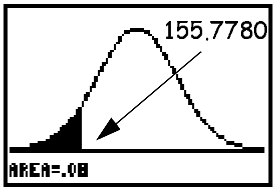
```
invNorm(.97,245,
63.5)
        364.4304
```



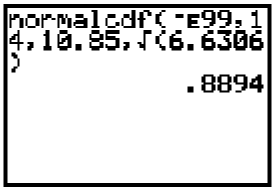
c) This is asking for the 8th percentile. The answer is 155.7780.



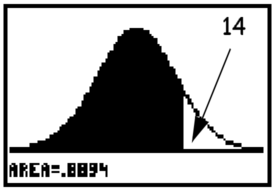
```
invNorm(.08,245,
63.5)
        155.7780
```



9.25 a) $P(B < 14)$



```
normalcdf(-E99,1
4,10.85,√(6.6306
)
           .8894
```



b) Approximately 8.9837 billion barrels.

9.27 The problem is asking for various percentiles based on a normal distribution with a mean of 973 and a standard deviation of 106.

Grade	Minimum Score
A	1082.8619
B	1013.8440
C	883.7881
D	798.6455
F	Below 798.6455

CHAPTER 9

10 Sampling Distributions

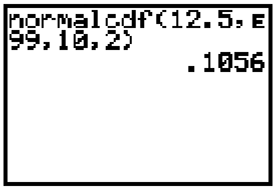
Review Exercise Solutions

10.1 Answers will vary. A key concept that should be presented is the idea of repeated sampling.

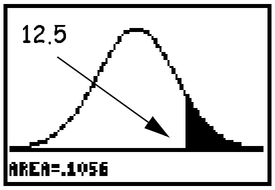
10.3 Answers will vary. The key concept is that the central limit theorem tells us that the sampling distribution of the sample mean will become more normal as the sample size increases.

10.5 Answers will vary. They key concept is the fact that the variance for the sampling distribution of the sample mean is divided by the square root of the sample size. This means as the sample size becomes larger, the variance, and hence standard deviation, becomes smaller.

10.7 a) $P(D > 12.5) = 0.1056$

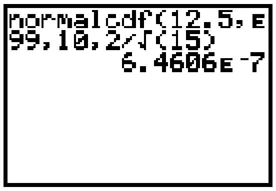


```
normalcdf(12.5,E
99,10,2)
           .1056
```

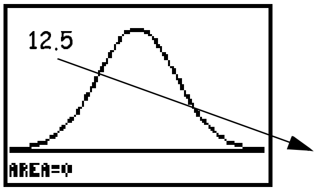


b) $\bar{X} \sim N\left(10, \frac{4}{15}\right)$

c)



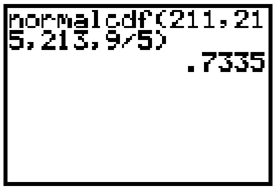
```
normalcdf(12.5,E
99,10,2/√(15)
       6.4606E-7
```



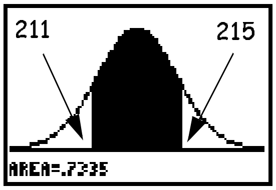
d) The difference between parts (a) and (c) is the standard deviation. In part (a), the observed value of 12.5 is only 1.25 standard deviations above the mean. In part (c), the observed value of 12.5 is 4.8412 standard deviations above the mean.

10.9 a) $\bar{X} \sim N\left(213, \frac{81}{25}\right)$

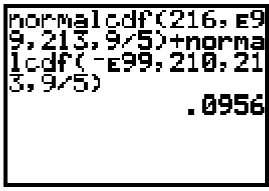
b)



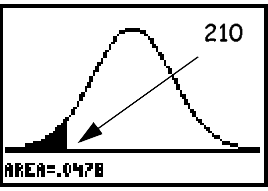
```
normalcdf(211,21
5,213,9/5)
           .7335
```

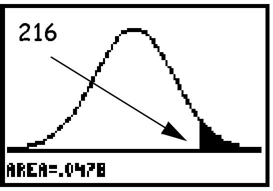


c) $P(\bar{X} > 216) + P(\bar{X} < 210)$



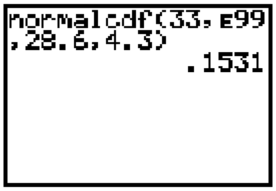
```
normalcdf(216,E9
9,213,9/5)+norma
lcdf(-E99,210,21
3,9/5)
           .0956
```



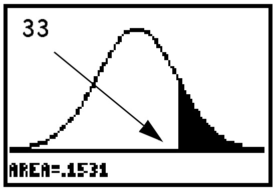


0.0478 + 0.0478 = 0.956

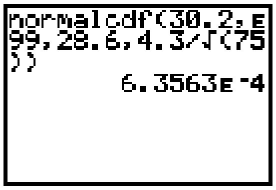
10.11 a) Scores ~ $N(28.6, 4.3^2)$



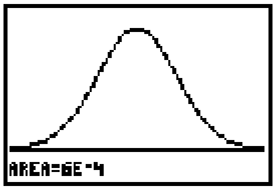
```
normalcdf(33,E99
,28.6,4.3)
           .1531
```



b) AverageScores ~ $N\left(28.6, \frac{4.3^2}{75}\right)$. No, an average of 30.2 from a sample of 75 that has a mean of 28.6 and a standard deviation of 4.3 is not at all expected. The probability of this happening, assuming the reported mean and standard deviation is correct, is approximately 0.0006, which is very unlikely.

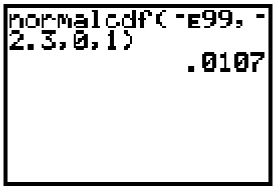


```
normalcdf(30.2,E
99,28.6,4.3/√(75
))
      6.3563E-4
```



c) We would expect to find the average score for a sample of 75 to be within two standard deviations of the mean, which is $28.6 \pm 2\left(\frac{4.3}{\sqrt{75}}\right)$ or approximately 27.6070 to 29.5930.

10.13 a) $\hat{P} \sim N\left(\frac{2}{3}, \frac{2/3\left(1 - 2/3\right)}{n}\right)$ b) $P\left(\hat{P} < \frac{118}{200}\right) = P\left(\frac{\hat{P} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} < \frac{\frac{118}{200} - \frac{2}{3}}{\sqrt{\frac{2/3\left(1/3\right)}{200}}}\right) = P(z < -2.3)$



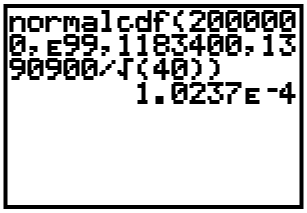
```
normalcdf(-E99,-
2.3,0,1)
           .0107
```

c) The program does seem to be working because the probability of observing only 118 out of 200 is very small.

10.15 a) Yes. A minimum of 20 samples is recommended before you apply the central limit theorem to the sampling distribution of a sample proportion; however, we have seen the distribution converge to normality very quickly when the population proportion is near 0.50. Since the proportion has remained constant at 56.8% for several years, it is reasonable to assume the population proportion is near 0.568 which is very close to 0.50.

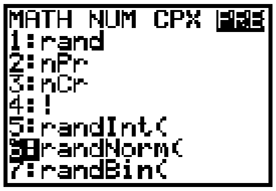
b) $\hat{P} \sim \left(0.568, \frac{0.568\left(1-0.568\right)}{18}\right)$ c) $P(\hat{P} > 0.52) = P\left(z > \frac{0.52-0.568}{\sqrt{\frac{0.568(1-0.568)}{18}}}\right) = P(z > -0.4111) = 0.6595$

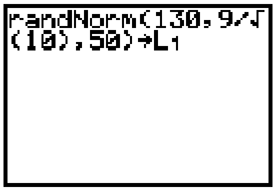
10.17 $P(\overline{X} > 2000000) = 0.0000$. We can use the central limit here to use the normal distribution because the sample size is 40. The probability of observing an average of 40 players with a salary of $2,000,000 or more is approximately 0.0001.



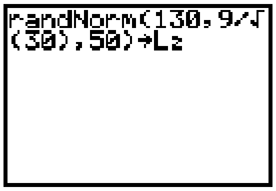
```
normalcdf(200000
0,E99,1183400,13
90900/√(40))
         1.0237E-4
```

10.19 Answers will vary. The TI-83 commands to generate the random samples and store them in lists are shown here.

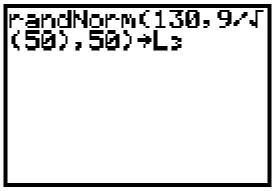




```
randNorm(130,9/√
(10),50)→L1
```



```
randNorm(130,9/√
(30),50)→L2
```



```
randNorm(130,9/√
(50),50)→L3
```

CHAPTER 10

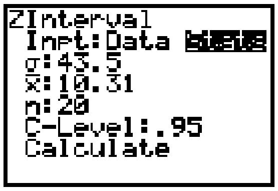
11 Confidence Intervals for Univariate Data

Review Exercise Solutions

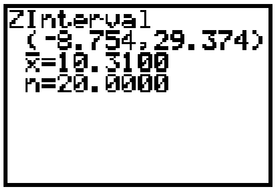
11.1 A point estimate is the sample statistic used to estimate a parameter. An interval estimate is an interval, calculated with sample data that will estimate the population parameter with a certain amount of confidence. The point estimate is expected to be close to the population parameter, but will almost never hit the parameter value exactly (in continuous distributions the probability is zero). However, the interval estimate will contain the population parameter with a certain probability and is usually a more reliable estimate.

11.3 True. In calculating sample size for either a proportion or a mean, the z-distribution is used to set the level of confidence. You cannot use the t-distribution since the value will depend on the sample size.

11.5 a) $n = 20, \bar{x} = 10.31$

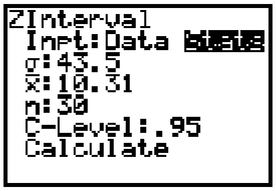


```
ZInterval
 Inpt:Data Stats
 σ:43.5
 x̄:10.31
 n:20
 C-Level:.95
 Calculate
```

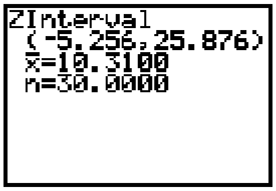


```
ZInterval
 (-8.754,29.374)
 x̄=10.3100
 n=20.0000
```

b) $n = 30, \bar{x} = 10.31$

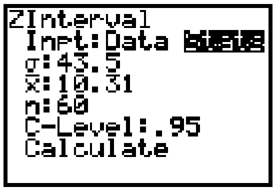


```
ZInterval
 Inpt:Data Stats
 σ:43.5
 x̄:10.31
 n:30
 C-Level:.95
 Calculate
```

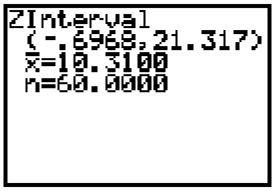


```
ZInterval
 (-5.256,25.876)
 x̄=10.3100
 n=30.0000
```

c) $n = 60, \bar{x} = 10.31$

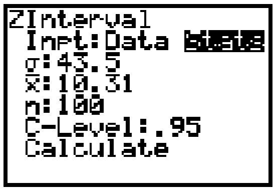


```
ZInterval
 Inpt:Data Stats
 σ:43.5
 x̄:10.31
 n:60
 C-Level:.95
 Calculate
```

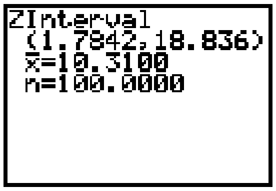


```
ZInterval
 (-.6968,21.317)
 x̄=10.3100
 n=60.0000
```

d) $n = 100, \bar{x} = 10.31$

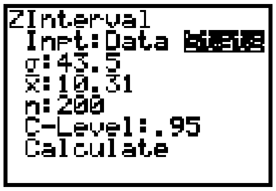


```
ZInterval
 Inpt:Data Stats
 σ:43.5
 x̄:10.31
 n:100
 C-Level:.95
 Calculate
```

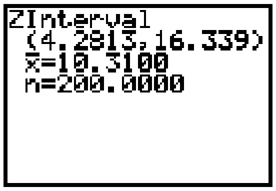


```
ZInterval
 (1.7842,18.836)
 x̄=10.3100
 n=100.0000
```

e) n = 200, $\bar{x}$ = 10.31

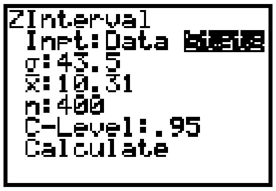


```
ZInterval
 Inpt:Data Stats
 σ:43.5
 x̄:10.31
 n:200
 C-Level:.95
 Calculate
```

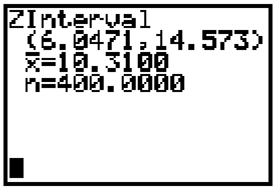


```
ZInterval
 (4.2813,16.339)
 x̄=10.3100
 n=200.0000
```

f) n = 400, $\bar{x}$ = 10.31



```
ZInterval
 Inpt:Data Stats
 σ:43.5
 x̄:10.31
 n:400
 C-Level:.95
 Calculate
```

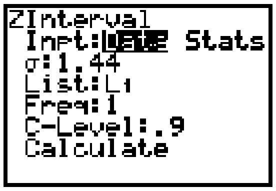


```
ZInterval
 (6.0471,14.573)
 x̄=10.3100
 n=400.0000
```

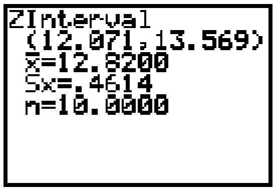
g) Each interval gets increasingly smaller as the sample size is increased. This makes sense, as the estimate for the mean would be more accurate for larger and larger sample sizes.

11.7 Assuming the data is in L1:

a) I am 90% confident the true mean time to relieve a minor or moderate headache with this new pain medication is between 12.071 and 13.569 minutes.

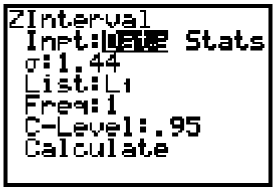


```
ZInterval
 Inpt:Data Stats
 σ:1.44
 List:L1
 Freq:1
 C-Level:.9
 Calculate
```

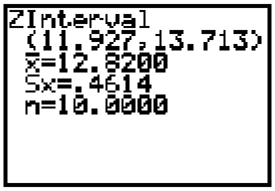


```
ZInterval
 (12.071,13.569)
 x̄=12.8200
 Sx=.4614
 n=10.0000
```

I am 95% confident the true mean time to relieve a minor or moderate headache with this new pain medication is between 11.927 and 13.7113 minutes.

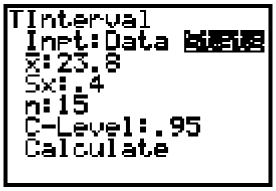


```
ZInterval
 Inpt:Data Stats
 σ:1.44
 List:L1
 Freq:1
 C-Level:.95
 Calculate
```

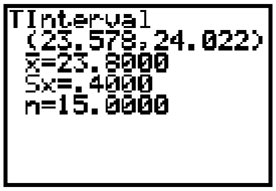


```
ZInterval
 (11.927,13.713)
 x̄=12.8200
 Sx=.4614
 n=10.0000
```

b) The 90% confidence interval is smaller than the 95% confidence interval because there is a greater probability of the interval *not containing* the true population parameter. This means there is more data outside the interval 90% confidence interval than the 95% confidence interval.



```
TInterval
 Inpt:Data Stats
 x̄:23.8
 Sx:.4
 n:15
 C-Level:.95
 Calculate
```

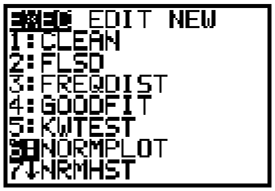


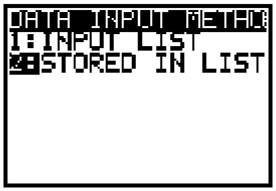
```
TInterval
 (23.578,24.022)
 x̄=23.8000
 Sx=.4000
 n=15.0000
```

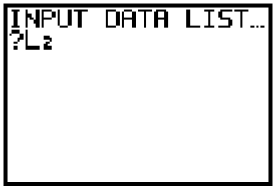
11.9 We are constructing a t-interval. This is due to the fact that the true standard deviation is unknown, all we have to work with is the sample standard deviation. The assumption of normality has been met by the statement in the problem "Assuming the distribution of the weight of the bags is approximately normal." I am 95% confident the true mean weight of the bags is between (23.578 and 24.022 ounces). The believed true mean is 24 ounces. The confidence interval has defined an interval in which we are 95% certain that the true mean resides. Since 24 ounces is within this interval there is no evidence to support the consumer group concern. The consumer group might be concerned that the confidence interval is not "centered" on

24 ounces. This is not an issue because the confidence interval is designed to capture the true mean somewhere in the interval.

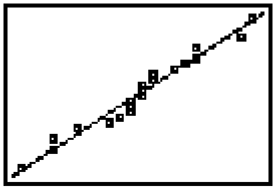
11.11 a) To determine the proper point estimate for this data, we will check a normal plot. Since this data appears to be approximately normal, we will calculate a sample mean for the point estimate. The sample mean is 80.85.





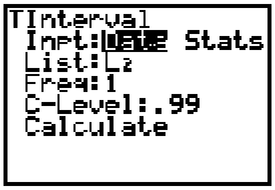


```
INPUT DATA LIST…
?L₂
```

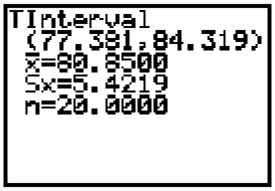


b) We will use a t-interval since the data is approximately normal and the population standard deviation is unknown. We are 99% confident the true mean score for this class is between 77.381 and 84.319.

c) This class seems to have lower scores than past classes since the “known” mean is not in the interval.



```
TInterval
 Inpt:Data Stats
 List:L₂
 Freq:1
 C-Level:.99
 Calculate
```



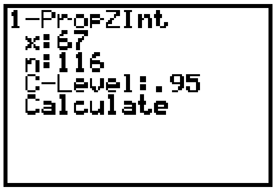
```
TInterval
 (77.381,84.319)
 x̄=80.8500
 Sx=5.4219
 n=20.0000
```

11.13 a) We are 95% confident that the true mean time for an account within the top 25% of accounts receivable to pay their bill will be between 34 days and 107 days.

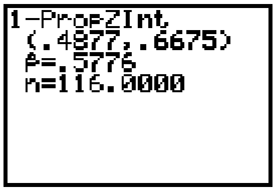
b) This is “proof” that these accounts will most likely take longer than one month to pay and in some cases, 3 months. This information verifies your beliefs and you can now aggressively pursue collection for these accounts. It will also allow you better information to plan cash flow when you know not to expect their payments right on time. Overall, it is a very usable confidence interval.

11.15 The statement “at least” means greater than or equal to. So, stating there will be “at least 5.7 billion barrels with 95% probability” means that the probability there are 5.7 billion barrels or more is 0.95. Similarly, “at least 16 billion barrels with 5% probability” means that the probability there are 16 billion barrels or more is 0.05. This means that there is 5% below 5.7 billion and 5% above 16 billion. This would be 10% overall error corresponding to a 90% confidence interval.

11.17 a) Checking assumptions for a confidence interval for proportions shows there would be (0.58)(116) = 67 successes and 116 - 67 = 49 failures. This qualifies the normality assumption



```
1-PropZInt
 x:67
 n:116
 C-Level:.95
 Calculate
```



```
1-PropZInt
 (.4877,.6675)
 p̂=.5776
 n=116.0000
```

under the central limit theorem for proportions. The assumption of independence is met by reasonable sampling. The confidence interval is (0.4877, 0.6675).

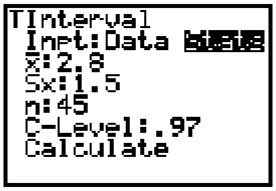
b) We are 95% confident the true increase in risk is between 48.77% and 66.75%.

c) A sample size of at least 375 is needed.

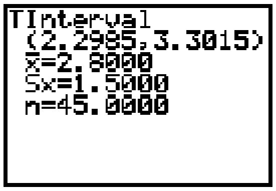
$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{ME}\right)^2 \Rightarrow n = \left(\frac{1.96\sqrt{0.58(1-0.58)}}{0.05}\right)^2$$

$$n = \left(\frac{1.96(0.4936)}{0.05}\right)^2 \Rightarrow n = 374.3255 \Rightarrow n = 375$$

11.19 a) The assumption of normality has been met by the central limit theorem since the sample size is 45. We are 97% confident the true mean level of Barium is between 2.2985 and 3.3015 parts per million.



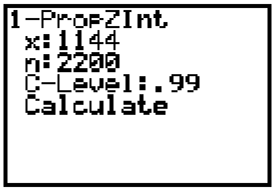
```
TInterval
 Inpt:Data Stats
 x̄:2.8
 Sx:1.5
 n:45
 C-Level:.97
 Calculate
```



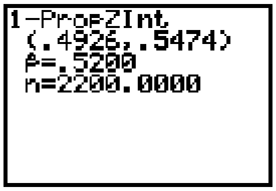
```
TInterval
 (2.2985,3.3015)
 x̄=2.8000
 Sx=1.5000
 n=45.0000
```

b) Since the target value for Barium is 2 ppm, we are reasonably sure these wells are exceeding the maximum contaminant level set by the EPA.

11.21 a) There are (2,200)(0.52) = 1144 successes (females) and 2,200-1144=1056 failures (males) in the sample. This meets the normality assumptions since there are more than 5 each of successes and failures and it is reasonable to assume that the voters are independent of each other. We are 99% confident the true proportion of women voters in the 2000 Presidential election is between 0.4926 and 0.5474.



```
1-PropZInt
 x:1144
 n:2200
 C-Level:.99
 Calculate
```



```
1-PropZInt
 (.4926,.5474)
 p̂=.5200
 n=2200.0000
```

b) We must assume that the ABC News poll was conducted fairly and randomly. We must, in essence, assume it is a fair representation of all voters participating in the 2000 Presidential election.

c) A sample size of at least 1066 is needed.

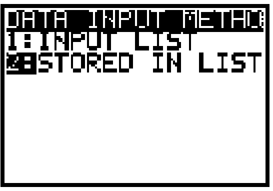
$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{ME}\right)^2 \Rightarrow n = \left(\frac{1.96\sqrt{0.52(1-0.52)}}{0.05}\right)^2$$

$$n = \left(\frac{1.96(0.4996)}{0.03}\right)^2 \Rightarrow n = 1065.40 \Rightarrow n = 1066$$

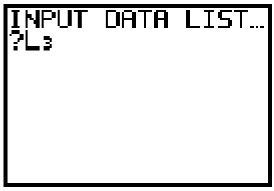
CHAPTER 11

11.23 a) To choose the correct point estimate we will use a normal plot to determine if the data is approximately normal. Our data is in L3. The normal plot shows no obvious problems so we will assume the population is approximately normal and use the mean as the point estimate for the center of this data.

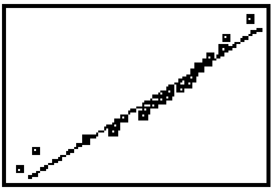
EXEC EDIT NEW
1:CLEAN
2:FLSD
3:FREQDIST
4:GOODFIT
5:KWTEST
6:NORMPLOT
7↓NRMHST



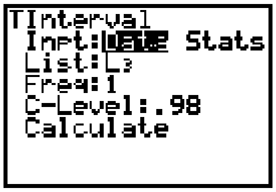
b) The sample mean is 357.5455.



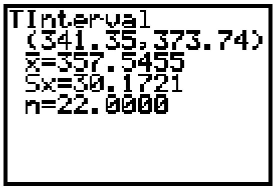
```
INPUT DATA LIST...
?L3
```



c) We will use a t-interval since the population standard deviation is unknown. The normality assumption was met by observing the normal plot. We are 98% confident the true mean production for shift 2 is between 341.35 and 373.74 cases.

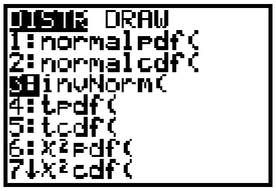


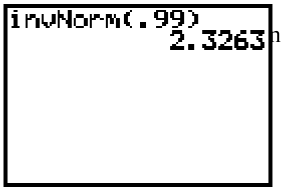
```
TInterval
 Inpt:Data Stats
 List:L3
 Freq:1
 C-Level:.98
 Calculate
```



```
TInterval
 (341.35,373.74)
 x̄=357.5455
 Sx=30.1721
 n=22.0000
```

d) A sample size of at least 50 is needed.



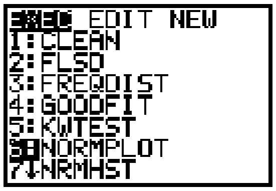


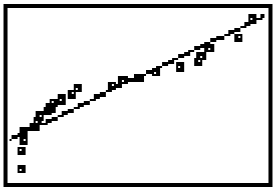
```
invNorm(.99)
            2.3263
```

$$n = \left(\frac{z_{\alpha/2}\sigma}{\text{ME}}\right)^2 \Rightarrow n = \left(\frac{2.3263(30.17)}{10}\right)^2$$

$$n = 49.2586 \Rightarrow n = 50$$

11.25 a) To choose the correct point estimate, we use the normal plot. This data shows obvious deviations from normality. This is the reason to choose the median as the proper measure of center.





b) The median birth rate is 32.

c) Since the sample size is greater than 20, we will construct a confidence interval for a proportion of 0.50 and use this to show the location of the end points of an approximate 95% confidence interval for the median. The location of the end points comes to the 7th and the 16th position. We are approximately 95% confident the true median birth mortality rate is between 15 and 66.

$$0.50 \pm 1.96\sqrt{\frac{0.50(0.50)}{23}}$$

$$0.50 \pm 1.96\sqrt{0.0109}$$

$$0.50 \pm 0.2043$$

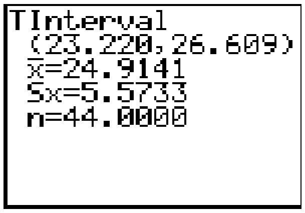
$$(0.2957, 0.7043)$$

$$23(0.2957, 0.7043)$$

$$(6.8011, 16.1989)$$

d) Since this is an interval built around the median and looking only at the location of the data, the confidence interval is only approximately 95%.

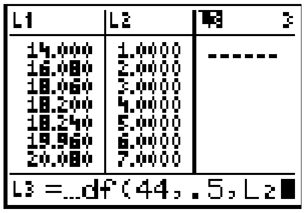
11.27 If we hold the sample size constant, the margin of error will increase as the level of confidence increases. The probability we *will not* include the true parameter in the interval gets smaller as we increase the level of confidence. This means the confidence interval actually gets wider, which in turn means the margin of error is increasing.



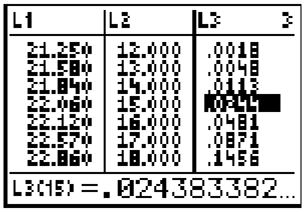
```
TInterval
 (23.220,26.609)
 x̄=24.9141
 Sx=5.5733
 n=44.0000
```

11.29 First, let's take a look at the parametric technique. The parametric technique will be a confidence interval based on the t-distribution because we do not know the population standard deviation but we can find the sample standard deviation. A 95% confidence interval for the mean based on the t-distribution: (23.220, 26.609). The Central Limit Theorem applies (n=44) so we really don't have to worry about checking a normal plot.

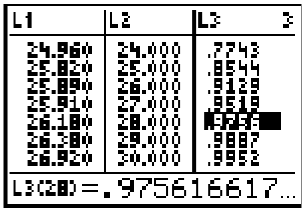
The nonparametric technique will be a confidence interval for the median. There were two methods presented in the text for finding a CI for the median. Here, we will use the technique based on the binomial distribution although the "large sample" case could easily be applied. The CI we will report is (22.060, 26.180). The actual level is 1 - ((1-0.9756) + 0.0244) = 0.9512. So we actually have a 95.12% CI for the median.



```
L1       L2       L3       3
14.000   1.0000   ------
16.080   2.0000
18.060   3.0000
18.200   4.0000
18.240   5.0000
19.960   6.0000
20.080   7.0000
L3 =…df(44,.5,L2
```



```
L1       L2       L3       3
21.250   12.000   .0018
21.580   13.000   .0048
21.840   14.000   .0113
22.060   15.000   [illegible]
22.120   16.000   .0481
22.570   17.000   .0871
22.860   18.000   .1456
L3(15)=.024383382…
```

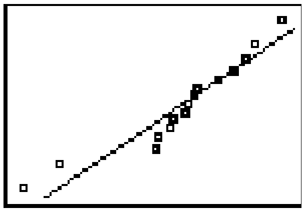


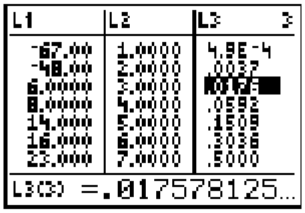
```
L1       L2       L3       3
24.960   24.000   .7743
25.820   25.000   .8544
25.890   26.000   .9129
25.910   27.000   .9519
26.180   28.000   [illegible]
26.380   29.000   .9887
26.920   30.000   .9952
L3(28)=.975616617…
```

The most approprtae CI to report the the CI for the mean because the CLT applies (as earlier stated).

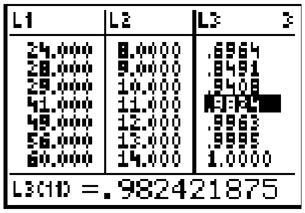
11.31 a) The random variable is the difference in heights of the plants, the measurement scale is ratio.

b) Based on the normal plot and small sample size, a CI for the median is most appropriate. We will report an approximate 95% confidence interval for the median as (6, 41) eights of an inch. In reality, the CI is 96.48% (the sum of the area in both tails).





```
L1       L2       L3       3
-67.00   1.0000   4.9E-4
-48.00   2.0000   .0037
6.0000   3.0000   [illegible]
8.0000   4.0000   .0592
14.000   5.0000   .1509
16.000   6.0000   .3036
23.000   7.0000   .5000
L3(3) =.017578125…
```



```
L1       L2       L3       3
24.000   8.0000   .6964
28.000   9.0000   .8491
29.000   10.000   .9408
41.000   11.000   [illegible]
49.000   12.000   .9963
56.000   13.000   .9995
60.000   14.000   1.0000
L3(11) =.982421875
```

c) Yes, it does look like one of the fertilizers is superior to the others. The data consists of the difference between the two fertilizers. If they were both the same then we would expect the differences to be, on average, zero. The CI does not include zero as a possible answer so it would be reasonable to conclude one fertilizer is resulting in greater growth than the other.

12 Univariate Hypothesis Testing

Review Exercise Solutions

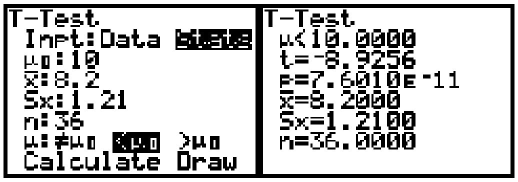
12.1 A p-value is the calculated probability that a value as least as extreme as your sample statistic will occur in the hypothesized distribution. In simpler terms, it is the probability as extreme as your sample statistic will occur randomly, given what you believed about your distribution (the null hypothesis) is true. This leads to the interpretation that the p-value is the calculated probability you will make a type I error if you reject the null hypothesis. Alpha is the reasonable risk you are willing to accept in making your decision. It is the level of type I error (rejecting the null hypothesis when it is actually true) you find reasonable. If the p-value, the calculated probability of making a type I error, is smaller than the amount of risk you are willing to take then you will reject the null hypothesis. If the probability of making a type I error (the p-value) is not smaller than the reasonable risk (alpha) you established, then you would have a greater risk of making a type I error than you feel is reasonable so you will fail to reject the null hypothesis

12.3 a) We will be concerned with the following hypotheses. A reasonable level of risk for this problem is 0.05 or 5%. This level is often left to the researcher.

$$H_o : \mu = 10$$

$$H_a : \mu < 10$$

The population standard deviation is unknown so, we will use a t-distribution, that is, our test statistic will be a “t”. The value of our test statistic is -8.9256 and the p-value is approximately zero. The value shown in the window is 7.6010 E-11 is scientific notation for 0.000000000076010. As you can see, that is very small. In fact, any value that is zero in the first four decimal places will be considered zero. Since the probability of this test statistic occurring, given the null hypothesis is true, is zero and as such, much smaller than our reasonable level of risk. We can be quite sure that our hypothesized mean is not what we believe, but something smaller. We will reject the null hypothesis and say there is strong evidence to show the students are not studying enough.



```
T-Test
 Inpt:Data Stats
 μ0:10
 x̄:8.2
 Sx:1.21
 n:36
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```

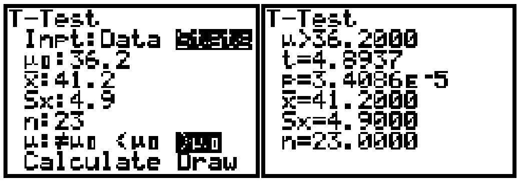
```
T-Test
 μ<10.0000
 t=-8.9256
 p=7.6010E-11
 x̄=8.2000
 Sx=1.2100
 n=36.0000
```

b) The parameter of interest is the mean, so we will need to verify the assumption of normality. Since the sample size is greater than 30 (n=36) this is reasonably satisfied by the central limit theorem.

12.5 We are testing the mean to see if joggers have a higher intake of oxygen than the average adult. A reasonable level of risk for this problem is 0.01. Since we don’t know the population standard deviation, we will use a t-statistic. The assumption of normality was stated in the problem. The test statistic is 4.8937 and the p-value is approximately zero. There is sufficient evidence to show joggers maximal oxygen intake is greater than that of an average adult.

$$H_o : \mu = 36.2$$

$$H_a : \mu > 36.2$$



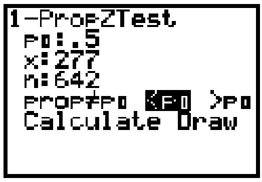
```
T-Test
 Inpt:Data Stats
 μ0:36.2
 x̄:41.2
 Sx:4.9
 n:23
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```

```
T-Test
 μ>36.2000
 t=4.8937
 p=3.4086E-5
 x̄=41.2000
 Sx=4.9000
 n=23.0000
```

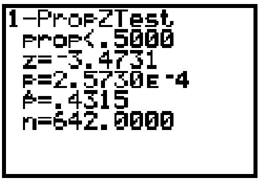
12.7 a) The hypotheses we are concerned with involve population proportions. We will use a one sample proportion z-test statistic with z = -3.4731 and the p-value is zero so we will reject the null hypothesis. There is strong evidence to support the citizens group claim.

$H_o : \pi = 0.50$

$H_a : \pi < 0.50$



```
1-PropZTest
 p0:.5
 x:277
 n:642
 prop≠p0 <p0 >p0
 Calculate Draw
```



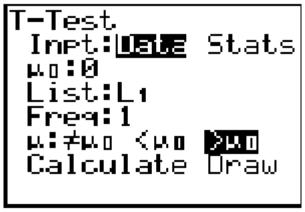
```
1-PropZTest
 prop<.5000
 z=-3.4731
 p=2.5730E-4
 p̂=.4315
 n=642.0000
```

b) To allow for the use of the z-test, we must check the assumption of normality. There are 277 successes and 365 failures which satisfy the normality assumption through the central limit theorem for proportions.

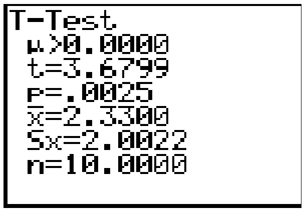
12.9 If the drug did not increase the number of sleep hours then we would expect the average to be zero. This leads us to the null and alternative hypothesis statements as:

$$H_O : \mu = 0 \qquad H_A : \mu > 0$$

The normal plot suggests the sample data is reasonably normal so we will complete a hypothesis test based on the t-distribution. Based on the p-value of 0.0025, we will reject the null hypothesis in favor of the alternative hypothesis and conclude that there is sufficient evidence to suggest the use of laevohysocyamine hydrobromide results in an increase number of sleep hours.



```
T-Test
 Inpt:Data Stats
 μ0:0
 List:L1
 Freq:1
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```

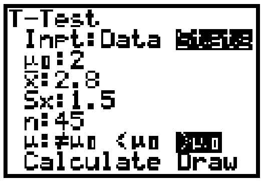


```
T-Test
 μ>0.0000
 t=3.6799
 p=.0025
 x̄=2.3300
 Sx=2.0022
 n=10.0000
```

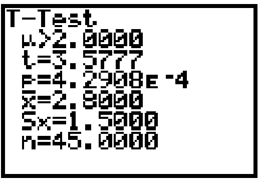
12.11 a) The hypotheses we are interested in concern the mean. The test statistic is shown along with a p-value of approximately zero (0.0004). This is strong evidence to show the wells have a barium level higher than the EPA maximum contaminant level goal.

$H_o : \mu = 2$

$H_a : \mu > 2$



```
T-Test
 Inpt:Data Stats
 μ0:2
 x̄:2.8
 Sx:1.5
 n:45
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```



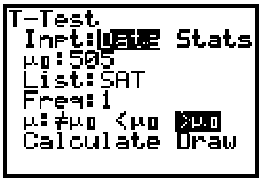
```
T-Test
 μ>2.0000
 t=3.5777
 p=4.2908E-4
 x̄=2.8000
 Sx=1.5000
 n=45.0000
```

b) The assumption for this test is approximate normality. This is reasonably satisfied through the central limit theorem since the sample size is 45.

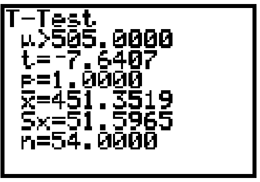
12.13 a) The hypotheses we are interested in concerns the mean. The test statistic is shown along with a p-value of 1. There is no evidence to support the claim that California colleges and universities admit students with higher than average verbal scores.

$H_o : \mu = 505$

$H_a : \mu > 505$

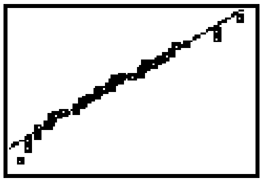


```
T-Test
 Inpt:Data Stats
 μ0:505
 List:SAT
 Freq:1
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```



```
T-Test
 μ>505.0000
 t=-7.6407
 p=1.0000
 x̄=451.3519
 Sx=51.5965
 n=54.0000
```

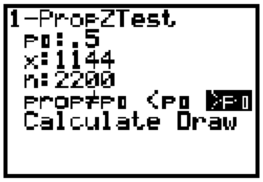
b) The assumption needed is that the data is normal. We can satisfy this one of two ways. First with a normal plot and second with the Central Limit Theorem. In general, if we have data, we will always look at a normal plot. This plot shows no great deviations from normality.



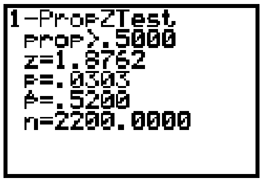
12.15 The hypotheses will involve the population proportion. With a practical level of significance of 5% we will reject the null hypothesis. The p-value is 0.0303 and thus is evidence that the true proportion of women voters has increased.

$H_o : \pi = 0.50$

$H_a : \pi > 0.50$



```
1-PropZTest
 p0:.5
 x:1144
 n:2200
 prop≠p0 <p0 >p0
 Calculate Draw
```



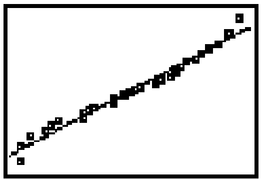
```
1-PropZTest
 prop>.5000
 z=1.8762
 p=.0303
 p̂=.5200
 n=2200.0000
```

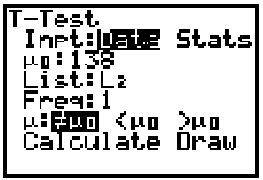
b) Our test statistic will be a z-score and normality is the assumption to use this. We have 1144 successes (women voters) and 1056 failures (male voters). These values are both greater than 5 and so satisfying the central limit theorem.

12.17 a) The hypotheses we are interested in concern the mean. We will do a t-test since the population standard deviation is unknown. The data is checked for normality by using a normal plot. The test statistic is 3.9003 and the p-value is zero. This is strong evidence against the null hypothesis. There is a significant difference between the budgeted census and the actual census for the month of August.

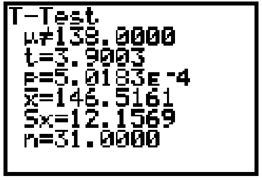
$H_o : \mu = 138$

$H_a : \mu \neq 138$

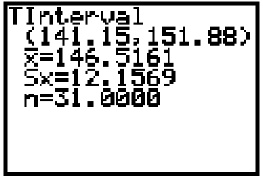




```
T-Test
 Inpt:Data Stats
 μ0:138
 List:L2
 Freq:1
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```



```
T-Test
 μ≠138.0000
 t=3.9003
 p=5.0183E-4
 x̄=146.5161
 Sx=12.1569
 n=31.0000
```



```
TInterval
 (141.15,151.88)
 x̄=146.5161
 Sx=12.1569
 n=31.0000
```

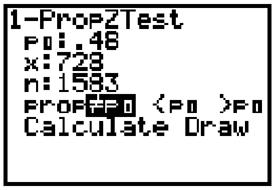
b) The 98% confidence interval for the true mean number of beds occupied is (141.15, 151.88). With 98% confidence the true mean number of beds occupied is between 141.15 and 151.88.

c) The confidence interval does not contain the budgeted census of 138. This means that 138 is not a possibility for the true mean number of beds occupied with 98% certainty. The hypothesis test rejected the idea that the true mean number of beds occupied was equal to 138. The two methods do agree.

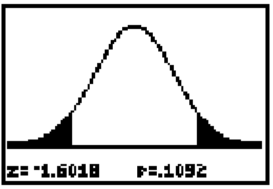
12.19 a) The hypotheses we are concerned with involve the population proportion. With a practical level of significance of 5% we will fail to reject the null hypothesis. The p-value is 0.1092. There is no evidence that the true proportion of voters that consider themselves middle class has changed.

$H_o : \pi = 0.48$

$H_a : \pi \neq 0.48$



```
1-PropZTest
 p0:.48
 x:728
 n:1583
 prop≠p0 <p0 >p0
 Calculate Draw
```



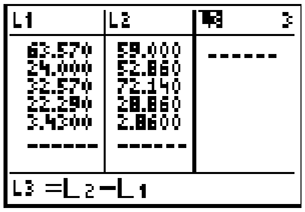
```
z=-1.6018     p=.1092
```

b) The assumption needed for this test is that of the sampling distribution of the sample proportions is normal. This is satisfied by the central limit theorem for proportions since the sample size is 1583, the number of successes is 728 and the number of failures is 855.

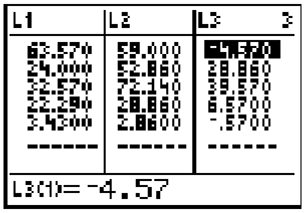
12.21 If the after group has a higher level of aggression behavior, the subtracting the before from the after data should result in predominately positive numbers. If there is no difference between the two groups then we would expect the average to be zero. We can subtract the data values and formulate a null and alternative hypothesis as:

$$H_O : \mu = 0 \quad H_A : \mu > 0$$

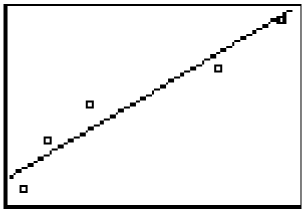
where the mean makes reference to the mean of the differences. We will then work only with the differences, which are shown in L3 below.



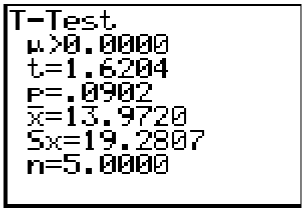
```
L1       L2       L3      3
63.570   59.000   ------
24.000   52.860
32.570   72.140
22.290   28.860
3.4300   2.8600
------   ------
L3 =L2-L1
```



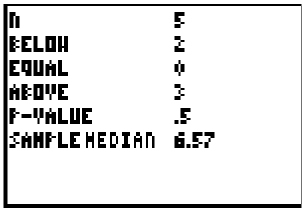
```
L1       L2       L3      3
63.570   59.000   -4.570
24.000   52.860   28.860
32.570   72.140   39.570
22.290   28.860   6.5700
3.4300   2.8600   -.5700
------   ------   ------
L3(1)=-4.57
```



Based on the normal plot, you could mount an argument either way. Since it is so close to call, I will perform both the t-test and the sign test. If the results agree then the normality decision is not relevant. If the results do not agree, then I will go with the sign test because I believe the normal plot is leaning toward not being normal.



```
T-Test
 μ>0.0000
 t=1.6204
 p=.0902
 x̄=13.9720
 Sx=19.2807
 n=5.0000
```

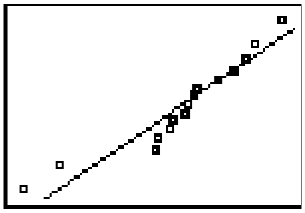


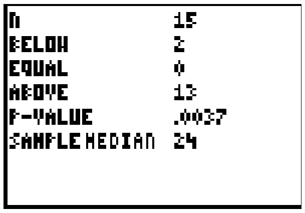
```
n              5
BELOW          2
EQUAL          0
ABOVE          3
P-VALUE        .5
SAMPLE MEDIAN  6.57
```

The p-values are different, as would be expected, but the conclusions are the same. There is not enough evidence to suggest a higher level of agrees in the fish after being exposed to Cadmium.

12.23 This is similar to the problem in the previous chapter. In the previous chapter we looked at the differences and calculated a confidence interval. Here, we will also look at the differences but rather than a confidence interval, we will do a hypothesis test. Based on the normal plot and small sample size, a hypothesis test for the median is most appropriate.

$$H_O : \theta = 0 \quad H_A : \theta > 0$$





```
n              15
BELOW          2
EQUAL          0
ABOVE          13
P-VALUE        .0037
SAMPLE MEDIAN  24
```

The p-value is 0.0037 so we will reject the null hypothesis and conclude that cross-fertilization is more effect than self-fertilization.

13 Comparing Two Population Parameters

Review Exercise Solutions

13.1 When testing one mean, you have an idea or a value that is known and you test to see if your sample has this same mean. When testing two means, you don’t have to have a preconceived notion of the value of the true parameters. You have the ability of comparing two independent groups without prior knowledge of the true mean for either populations. Other differences are noticed in the mechanics and assumptions. In testing two means, you must decide if the variances are the same before testing the means. The assumption that the data is from independent data sets is also unique to the two mean test.

13.3 The assumptions are approximate normality in both samples and independence between samples. If you have not been told the data was from a normal distribution, you would check the normality by using either the central limit theorem or a normal plot (if you have data). The assumption of independence refers not only to the lack of “pairing” but also that the data sets are truly independent of each other in that the gathering of one data set in no way influenced the second data set.

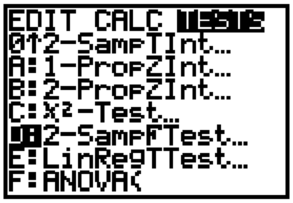
13.5 Practical significance is the ability to use the information that has been derived from statistics. It is highly dependent upon reason and common sense. Statistical significance can be found in almost any question if you gather enough data. It is possible to show a statistically significant difference between two parameters at a very high level of accuracy, but, the question is, is it practically significant? That is, does this difference you have shown, have any real use?

13.7 The hypothesis we are interested in involve two independent means. The hypothesis statement is:

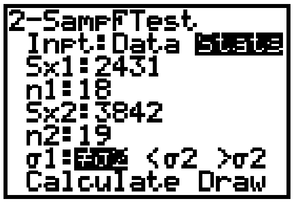
$$H_0: \mu_s - \mu_n = 0 \qquad H_A: \mu_s - \mu_n > 0$$

The assumption of normality is not a problem as we are told to make the necessary assumption within the problem. Using an F-test to check for similar variances and entering the southern group as group 1 yields a p-value of 0.0653. This suggests the variances for the two populations are similar which implies we should pool the standard deviations when testing the means.

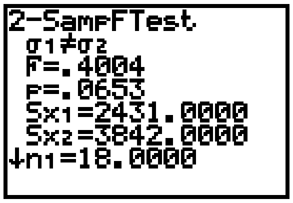
$$H_0: \frac{\sigma_s^2}{\sigma_n^2} = 1 \qquad H_A: \frac{\sigma_s^2}{\sigma_n^2} \neq 1$$



```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```

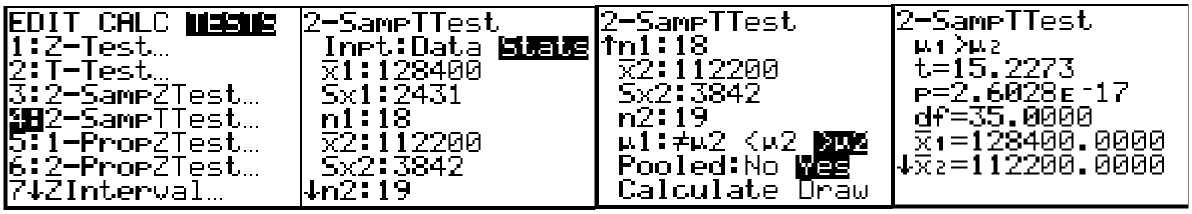


```
2-SampFTest
 Inpt:Data Stats
 Sx1:2431
 n1:18
 Sx2:3842
 n2:19
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



```
2-SampFTest
 σ1≠σ2
 F=.4004
 p=.0653
 Sx1=2431.0000
 Sx2=3842.0000
↓n1=18.0000
```

The screen shots for the two independent samples t-test are given below. With a p-value of approximately zero there is a very strong evidence to suggest that the homes in the southern part of town have a higher value than the homes in the northern part of town.



```
EDIT CALC TESTS
1:Z-Test…
2:T-Test…
3:2-SampZTest…
4:2-SampTTest…
5:1-PropZTest…
6:2-PropZTest…
7↓ZInterval…
```

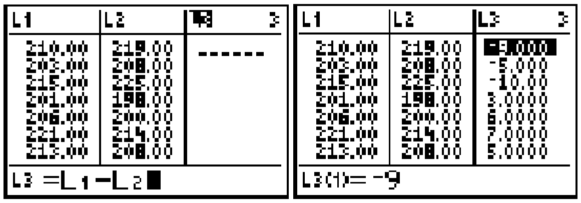
```
2-SampTTest
 Inpt:Data Stats
 x̄1:128400
 Sx1:2431
 n1:18
 x̄2:112200
 Sx2:3842
↓n2:19
```

```
2-SampTTest
↑n1:18
 x̄2:112200
 Sx2:3842
 n2:19
 μ1:≠μ2 <μ2 >μ2
 Pooled:No Yes
 Calculate Draw
```

```
2-SampTTest
 μ1>μ2
 t=15.2273
 p=2.6028E-17
 df=35.0000
 x̄1=128400.0000
↓x̄2=112200.0000
```

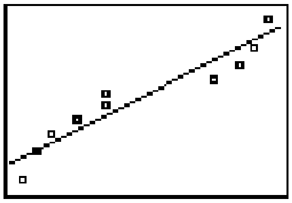
13.9 This data is obviously dependent or paired. Therefore the hypotheses we are interested in will involve the mean of the differences. We will need to calculate the differences and then do a one sample t-test if all assumptions have been met. To calculate the differences we will use the Edit screen of the TI-83. After entering the "Before" data in L1 and the "After" data in L2, we will subtract them and put the differences in L3 as shown. Next is verifying the assumption of normality with a normal plot on the differences.

$$H_o : \mu_d = 0$$
$$H_a : \mu_d > 0$$

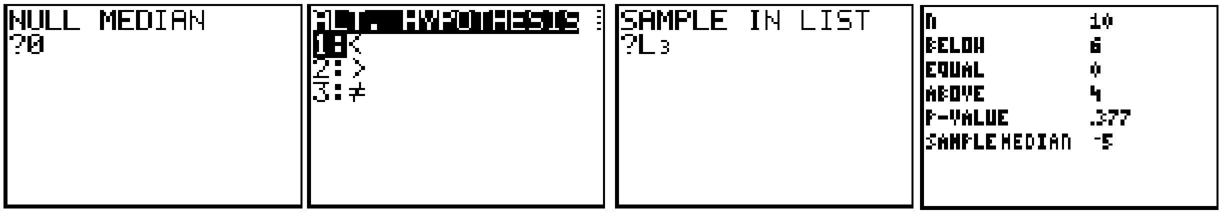


```
L1       L2       L3     3
210.00   219.00   ------
203.00   208.00
215.00   225.00
201.00   198.00
206.00   200.00
221.00   214.00
213.00   208.00
L3 =L1-L2
```

```
L1       L2       L3     3
210.00   219.00   -9.000
203.00   208.00   -5.000
215.00   225.00   -10.00
201.00   198.00   3.0000
206.00   200.00   6.0000
221.00   214.00   7.0000
213.00   208.00   5.0000
L3(1)=-9
```



The normal plot shows a gross departure from normality, so we will continue with the sign test, restating the null and alternative hypothesis referencing the median, rather than the mean. With a p-value of 0.377, there is no evidence to suggest the special blend of herbs increase the strength of men between 35 and 65 years of age.



```
NULL MEDIAN
?0
```

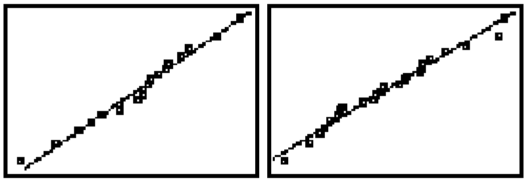
```
ALT. HYPOTHESIS
1:<
2:>
3:≠
```

```
SAMPLE IN LIST
?L3
```

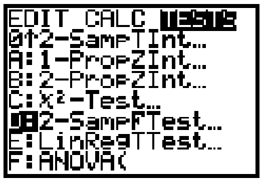
```
n              10
BELOW          6
EQUAL          0
ABOVE          4
P-VALUE        .377
SAMPLE MEDIAN  -5
```

13.11 Since we are interested in the variability, we will be concerned with the following hypotheses. The assumptions to be satisfied are independence between data sets and normality in each data set. The independence assumption is reasonable because the data cae from two different shifts. There is no reason to suspect a dependence between shifts. Since we have data, we will check this assumption with normal plots. The normal plots for Shift 2 and Shift 3 are shown.

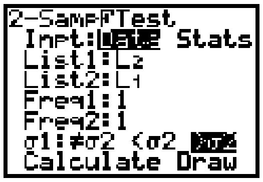
$$H_o : \frac{\sigma^2_{S3}}{\sigma^2_{S2}} = 1$$
$$H_a : \frac{\sigma^2_{S3}}{\sigma^2_{S2\,1}} > 1$$



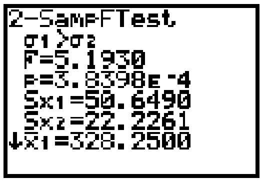
Since there are no obvious deviations from normality, we will continue with the F-test of two variances.



```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```



```
2-SampFTest
 Inpt:Data Stats
 List1:L2
 List2:L1
 Freq1:1
 Freq2:1
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



```
2-SampFTest
 σ1>σ2
 F=5.1930
 p=3.8398E-4
 Sx1=50.6490
 Sx2=22.2261
↓x̄1=328.2500
```

Shift 3 data is in list L_2 and Shift 2 in list L_1. With a p-value of approximately zero there is strong evidence to suggest that the variance for shift 3 is greater than the variance for shift 2.

13.13 a) This is independent data. There are two separate groups with no connection other than they are all 2nd graders.

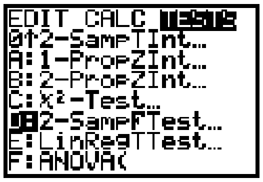
b) This is dependent data. There are two measurements on the same person, making the data dependent upon each experimental object. In this case, an experimental object is the person in the weight loss program.

c) This is independent data. There are two separate groups of babies. There is no connection between the two groups other than they are all babies.

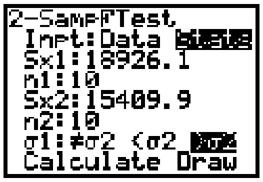
13.15 This is a test of variation so the hypotheses we are interested in will involve the population variance. Since the data is known to be approximately normal, we will move on to the assumption of independence between the data sets. This is reasonably satisfied by the simple fact they are from different decades and different facilities. The two sample F-test follows below. With a p-value of 0.2750, there is not enough evidence to reject the null hypothesis. There is no statistical evidence to show that the variation in the annual releases for the 1970's is less than that of the 1980's.

$$H_0 : \frac{\sigma^2_{80}}{\sigma^2_{70}} = 1$$

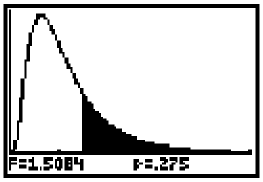
$$H_A : \frac{\sigma^2_{80}}{\sigma^2_{70}} > 1$$



```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```



```
2-SampFTest
 Inpt:Data Stats
 Sx1:18926.1
 n1:10
 Sx2:15409.9
 n2:10
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



```
F=1.5084     p=.275
```

13.17 a) The proportion of females in favor of going to war reported in this survey was 199/217 or approximately 91.71%.The proportion of males was 201/211 or approximately 95.26%

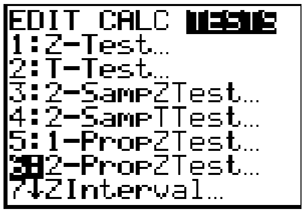
$$H_O : \pi_M = \pi_F \quad H_A : \pi_M > \pi_F \qquad \alpha = 0.05$$

Assumptions: Both sample proportions are distributed normally. This can be shown by:

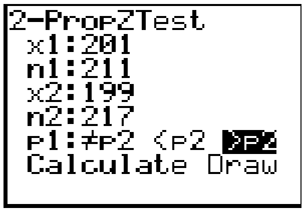
$$n_F\,\hat{p}_F \geq 5,\quad n_F(1-\hat{p}_F)\geq 5,\; n_F \geq 20$$

$$n_M\,\hat{p}_M \geq 5,\quad n_M(1-\hat{p}_M)\geq 5,\; n_M \geq 20$$

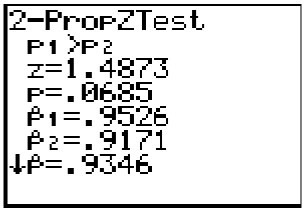
Which are all true (students should do the math to show this is true).



```
EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7↓ZInterval...
```



```
2-PropZTest
 x1:201
 n1:211
 x2:199
 n2:217
 p1:≠p2 <p2 >p2
 Calculate Draw
```



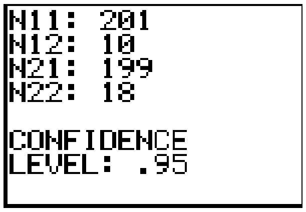
```
2-PropZTest
 p1>p2
 z=1.4873
 p=.0685
 p̂1=.9526
 p̂2=.9171
↓p̂=.9346
```

Based on the p-value of 0.0685, we will fail to reject the null hypothesis and conclude that there is not enough evidence to suggest the proportion of males that were in favor of going to war was greater than females.

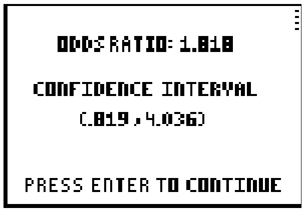
The same analysis can be done with an odds ratio by setting up the following table:

In favor:	Yes	No
Males	201	10
Females	199	18

$$\hat{\lambda}=\frac{201\cdot 18}{199\cdot 10}=1.8181 \qquad 95\%\;\; CI\;\;(0.819, 4.036)$$



```
N11: 201
N12: 10
N21: 199
N22: 18

CONFIDENCE
LEVEL: .95
```



```
ODDS RATIO: 1.818

CONFIDENCE INTERVAL
(.819 , 4.036)

PRESS ENTER TO CONTINUE
```

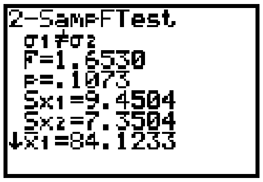
Since 1 is in interval we conclude there is insufficient evidence to suggest the odds of a male in favor of going to war is greater than the odds of a female.

b) Yes, the conclusion using both techniques do agree.

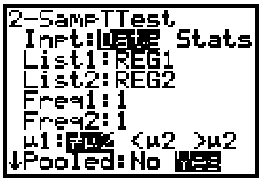
CHAPTER 13

$H_o : \mu_{R1} - \mu_{R2} = 0$
$H_a : \mu_{R1} - \mu_{R2} \neq 0$

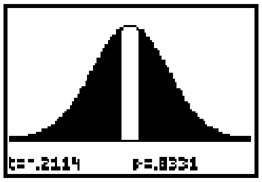
13.19 a) The parameters we are interested in are the means of each of the groups. The hypotheses will be as shown. Since we are concerned with the mean, we will have to verify the data is approximately normal. Since there are more than 30 data points for each data set, the Central Limit Theorem will apply. We will proceed with a t-test for two independent samples. We need to know if the variances are similar so we will know if we should pool the variances or not during the t-test procedure. We will use an F-test to make a determination regarding the variances. The results are shown below. There is not enough evidence to show a difference in the variances, so we will pool the standard deviations for the t-test. With a p-value of 0.8331, there is not enough evidence to show a difference in the customer service scores that were reported by the two regions.



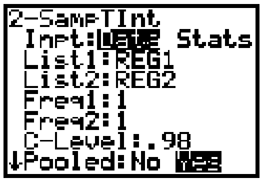
```
2-SampFTest
 σ1≠σ2
 F=1.6530
 p=.1073
 Sx1=9.4504
 Sx2=7.3504
↓x̄1=84.1233
```



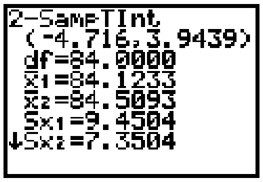
```
2-SampTTest
 Inpt:Data Stats
 List1:REG1
 List2:REG2
 Freq1:1
 Freq2:1
 μ1:≠μ2 <μ2 >μ2
↓Pooled:No Yes
```



```
t=-.2114        p=.8331
```



```
2-SampTInt
 Inpt:Data Stats
 List1:REG1
 List2:REG2
 Freq1:1
 Freq2:1
 C-Level:.98
↓Pooled:No Yes
```



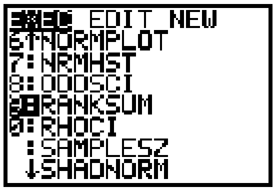
```
2-SampTInt
 (-4.716,3.9439)
 df=84.0000
 x̄1=84.1233
 x̄2=84.5093
 Sx1=9.4504
↓Sx2=7.3504
```

b) The 98% confidence interval is shown above. We are 98% confident that the true difference in the means is between -4.716 and 3.9439. Zero is in the interval, so there is no real difference in the two means. This agrees with the hypothesis test.

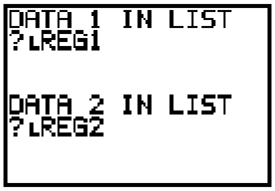
However, if you looked at the normal plots and observed the violation of the normality assumption, then we would change our hypotheses to concern the medians, not the means. The proper test would be the Wilcoxon Rank-Sum Test.

$$H_o : \theta_{R1} - \theta_{R2} = 0$$
$$H_a : \theta_{R1} - \theta_{R2} \neq 0$$

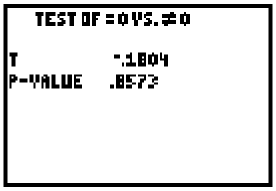
With a p-value of 0.8573 there is not enough evidence to show the medians customer service scores are different.



```
[illegible] EDIT NEW
6↑NORMPLOT
7:NRMHST
8:ODDSCI
9:RANKSUM
0:RHOCI
 :SAMPLESZ
 ↓SHADNORM
```



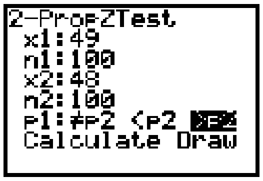
```
DATA 1 IN LIST
?LREG1

DATA 2 IN LIST
?LREG2
```



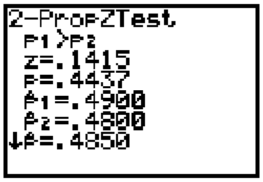
```
  TEST OF =0 VS.≠0

T            -.1804
P-VALUE      .8573
```

$H_o : \pi_G - \pi_B = 0$
$H_a : \pi_G - \pi_B > 0$

13.21 a) Since we are dealing with percentages, we will form hypotheses with the population parameter for the true proportion. Since we have 49 successes in the Gore group and 48 successes in the Bush group, we meet the normality assumption. It is reasonable to assume the data is independent. The z-score is 0.1415 and the p-value is 0.4437. There is not enough evidence to show the proportion of people that voted for Gore that worked full time for pay is greater than that of the voters that voted for Bush.

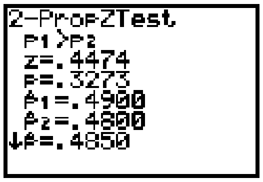


```
2-PropZTest
 x1:49
 n1:100
 x2:48
 n2:100
 p1:≠p2 <p2 >p2
 Calculate Draw
```

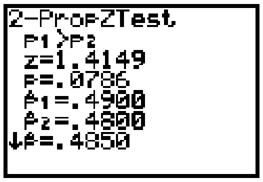


```
2-PropZTest
 p1>p2
 z=.1415
 p=.4437
 p̂1=.4900
 p̂2=.4800
↓p̂=.4850
```

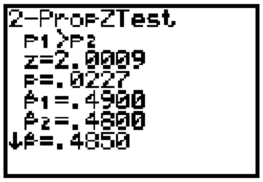
b) Repeating the same test for samples of size 1,000, 10,000 and 20,000, the results are shown below.



```
2-PropZTest
 p1>p2
 z=.4474
 p=.3273
 p̂1=.4900
 p̂2=.4800
↓p̂=.4850
```



```
2-PropZTest
 p1>p2
 z=1.4149
 p=.0786
 p̂1=.4900
 p̂2=.4800
↓p̂=.4850
```



```
2-PropZTest
 p1>p2
 z=2.0009
 p=.0227
 p̂1=.4900
 p̂2=.4800
↓p̂=.4850
```

c) The difference in the proportions becomes significant between the sample sizes of 10,000 and 20,000.

d) This is statistical significance. Practically speaking, there is no real difference between the proportions 0.49 and 0.48 in this problem.

e) There is no real usefulness of the significance found with the sample of 20,000 concerning voter descriptive statistics.

13.23 a) The random variables are the percent of women participation in the workforce for 1972 and 1968.

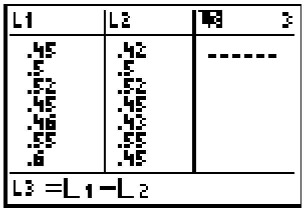
b) The population of interest is entire United States work force.

c) The data was collected in an effort to determine if the amount women participate in the workforce has increased from 1968 to 1972.

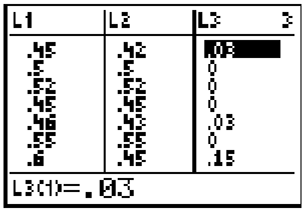
d) We will use subscripts of 68 and 72 to represent 1968 and 1972.

$$H_0 : \mu_{72-68} = 0 \qquad H_A : \mu_{72-68} > 0 \qquad \alpha = 0.05$$

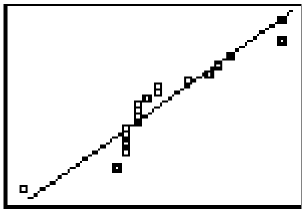
Assumptions: The differences are distributed normally.



```
L1     L2     L3     3
.45    .42    ------
.5     .5
.52    .52
.45    .45
.46    .43
.55    .55
.6     .45
L3 =L1-L2
```



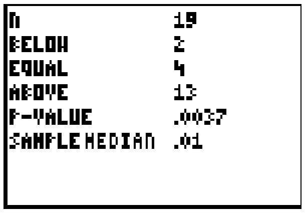
```
L1     L2     L3     3
.45    .42    .03
.5     .5     0
.52    .52    0
.45    .45    0
.46    .43    .03
.55    .55    0
.6     .45    .15
L3(1)=.03
```



The normal plot of the differences indicates a gross violation of the assumptions so we will use a nonparametric approach (the sign test).

Based on the p-value (0.0037) we will reject the null hypothesis and conclude that there is sufficient evidence to suggest the percent of women participating in the workforce had increased from 1968 to 1972.

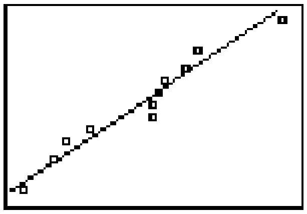
$$H_0 : \theta_{72-68} = 0 \quad H_A : \theta_{72-68} > 0 \quad \alpha = 0.05$$

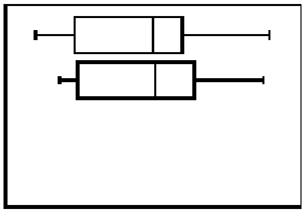


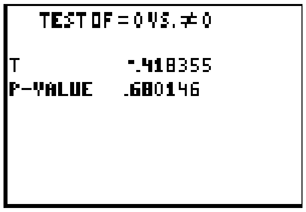
```
n              19
BELOW          2
EQUAL          4
ABOVE          13
P-VALUE        .0037
SAMPLE MEDIAN  .01
```

13.25 This problem asks if there is evidence to suggest a difference between the two groups, not suggesting which group was suspected to be bigger than the other. This means we will be doing a two tailed test.

Assumptions: (a) The distribution of the sample data from both the REG and KILN groups are each normal. (b) The two groups are independent of each other. The independence assumption is clear (the student should specify why). The normality assumption will be addressed with normal plots. The normal plot for REG indicates a gross violation so there is no need to even look at the normal plot for KILN. We will use a nonparametric approach, the Wilcoxon Rank-Sum. Based on the box-plots, the general shape of both distributions are the same which means we can construct a test regarding the means, rather than the medians.







```
   TEST OF = 0 VS. ≠ 0

T           ⁻.418355
P-VALUE     .680146
```

$$H_0 : \mu_R - \mu_K = 0 \quad H_A : \mu_R - \mu_K \neq 0 \quad \alpha = 0.05$$

Based on the p-value of 0.68, we will fail to reject the null hypothesis and conclude there is not sufficient evidence to suggest a difference between the two processes.

CHAPTER 13

14 Correlation and Simple Linear Regression

Review Exercise Solutions

14.1 There are two major problems. First, a correlation of 1.37 is mathematically impossible. The friend needs to recalculate the correlation coefficient, clearly a mistake was made. The second problem is with the friend's reference to the "*causation factor.*" Pearson's correlation is a measurement of linear association. A conclusion regarding causation is not appropriate.

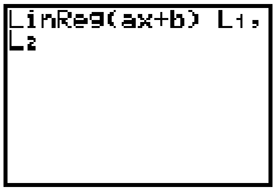
14.3 False. Pearson's correlation is a measurement of linear association. Calculating Pearson's correlation is not meaningful for qualitative data.

14.5 The interval is not possible because correlation cannot take on a value less than -1.

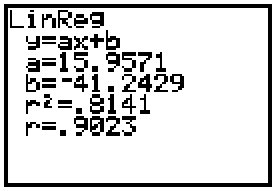
14.7 The primary use for regression analysis is to fit a mathematical model to data for prediction purposes. In this text, we limited our study to linear models consisting of one predictor variable and one response variable.

14.9 A residual is the difference between the predicted value of the response variable and the actual observed value of the response variable.

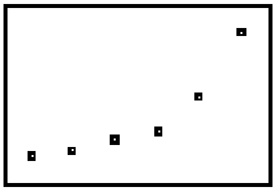
14.11 The value for Pearson's correlation is 0.9023; however, based on the scatter plot it is clear Pearson's correlation is not the appropriate measure of association. The scatter plot clearly shows curvature in the data. As such, Spearman's correlation is the more appropriate measure of association. The value of Spearman's correlation is 1.

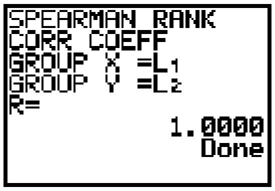


LinReg(ax+b) L_1,
L_2



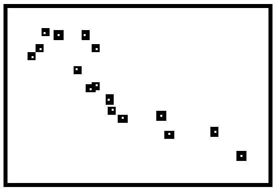
LinReg
y=ax+b
a=15.9571
b=⁻41.2429
r^2=.8141
r=.9023



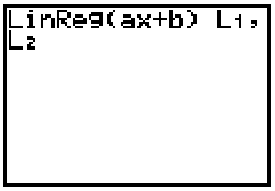


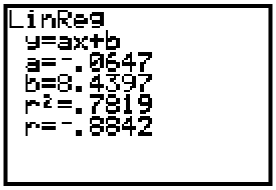
SPEARMAN RANK
CORR COEFF
GROUP X =L_1
GROUP Y =L_2
R=
1.0000
Done

14.13 a) The scatter plot suggests a negative linear trend. As such, Pearson's correlation would be the appropriate measure of the strength of that association.



b) The value of Pearson's correlation is -0.8842. This is indicating a strong negative association. In terms of this data, Pearson's correlation is suggesting that as the number of cigarettes the mother smokes each day increases, the birth weight of their children decreases.


```
LinReg(ax+b) L1,
L2
```


```
LinReg
 y=ax+b
 a=-.0647
 b=8.4397
 r²=.7819
 r=-.8842
```

c) The 95% confidence interval for ρ is (-0.9594, -0.6916). Both the hand calculations and the output from the TI-83 program RHOCI are shown below.

$$r' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$\Rightarrow r' = \frac{1}{2}\ln\left(\frac{1+(-0.8842)}{1-(-0.8842)}\right) = \frac{1}{2}\ln\left(\frac{0.1158}{1.8842}\right) = -1.3947$$

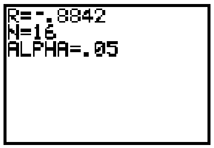
$$\sigma_{r'} = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{13}}$$

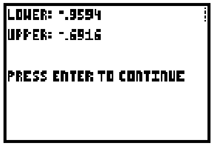
$$r' \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}} \Rightarrow -1.3947 \pm 1.96\sqrt{\frac{1}{13}} \Rightarrow (-1.9383,\ -0.8511)$$

$$\left(\frac{e^{2L}-1}{e^{2L}+1}, \frac{e^{2U}-1}{e^{2U}+1}\right)$$

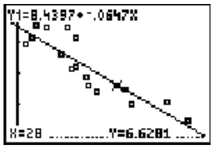
$$\left(\frac{e^{2(-1.9383)}-1}{e^{2(-1.9383)}+1}, \frac{e^{2(-0.8511)}-1}{e^{2(-0.8511)}+1}\right)$$

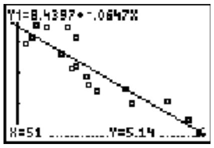
$$(-0.9594, -0.6916)$$

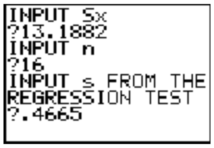

```
R=-.8842
N=16
ALPHA=.05
```

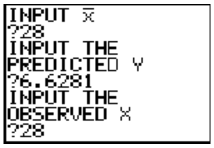

```
LOWER: -.9594
UPPER: -.6916

PRESS ENTER TO CONTINUE
```

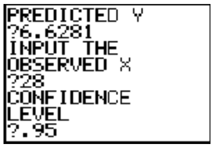
d) The prediction interval is calculated by using the PREDCI program.

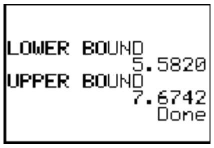


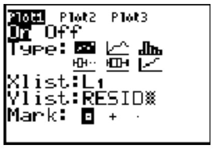


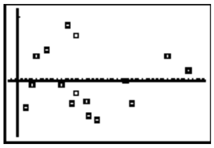

```
INPUT Sx
?13.1882
INPUT n
?16
INPUT s FROM THE
REGRESSION TEST
?.4665
```

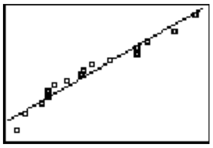

```
INPUT x̄
?28
INPUT THE
PREDICTED Y
?6.6281
INPUT THE
OBSERVED X
?28
```


```
PREDICTED Y
?6.6281
INPUT THE
OBSERVED X
?28
CONFIDENCE
LEVEL
?.95
```


```
LOWER BOUND
          5.5820
UPPER BOUND
          7.6742
            Done
```


```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:RESID
Mark: ▪ + ·
```



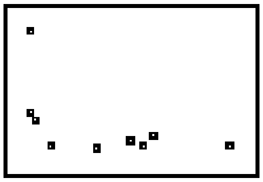


The process is the same for the other prediction interval.

The validity of the prediction interval is associated with the assumptions for regression analysis. Based on the residual plot, the residuals appear to have a reasonably constant variance. The normal plot also looks pretty good.

CHAPTER 14

14.15 a)



O. tridens is on the x-axis.

O. lowei is on the y-axis.

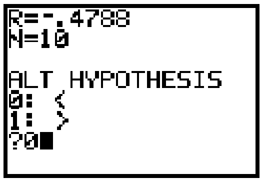
b) Due to the curvature of the data, Spearman’s correlation is the appropriate measurement of association. The value for Spearman’s correlation coefficient is -0.4788.

c) With a p-value of 0.081, there is not enough evidence to suggest the barnacles compete for space. The 95% confidence interval for the value of Spearmans Correlation Coefficient is

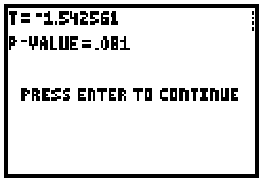
(-0.8517, 0.2159).

$H_0 : \rho_s = 0$

$H_A : \rho_s < 0$

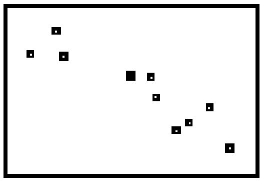


```
R=-.4788
N=10

ALT HYPOTHESIS
0: <
1: >
?0
```



```
T=-1.542561
P-VALUE=.081

 PRESS ENTER TO CONTINUE
```

14.17 a) There does appear to be a negative linear association. The value of Pearson’s correlation coefficient is -0.9088. The interval estimate, specifically a 95% confidence interval, is (-0.9764, -0.6792).



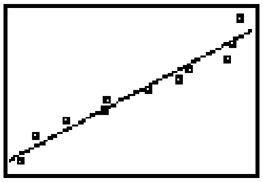
BTU is on the x-axis and Temp is on the y-axis.

b) The logical choice would be to predict the amount of energy used based on outside temperature. As such, Temp would be the predictor variable and BTU the response variable.

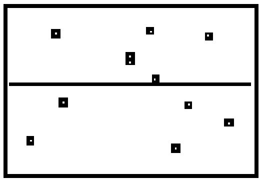
c) Yes, it appears appropriate based on the scatter plot. The regression equation is: BTU = 4.8407 - 0.0557(Temp)

d) The assumptions are based on the residuals: The residuals are random, independent and distributed normally with a mean of 0 and a constant variance. A normal plot of the residual suggests the normality assumption is reasonably satisfied. The independent, randomness and constant variance can be seen from the scatter plot of the predictor variable and the residuals.

Normal Plot of Residual



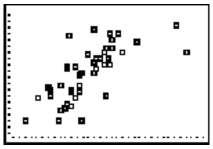
Predictor variable by Residuals

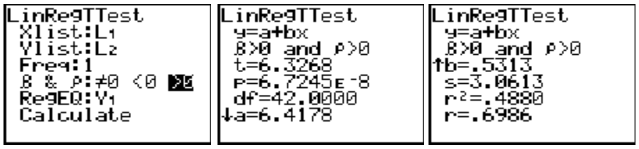


e) The predicted value of BTU for a temperature of 25 is: BTU = 4.8407 - 0.0557(25). BTU = 3.4482. Note that the minimum temperature in our sample data is 15 degrees and the maximum value is 45 degrees. The temperature of 25 degrees used as a predictor is well within the data range of Temp.

f) The prediction interval, calculated using the PREDCI program, is (2.8144, 4.0820).

4.19 a) We will start the analysis with a look at the scatter plot. The scatter plot appears to have an positive linear trend suggesting it is reasonable to continue the analysis by calculating Pearson's correlation coefficient. I will use the LinRegTTest command in the TI-83 to calculate the value of r because it will also calculate a bunch of other stuff I may later need.



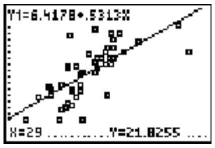


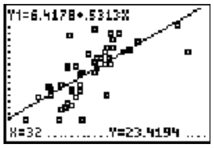
```
LinRegTTest        LinRegTTest        LinRegTTest
 Xlist:L1           y=a+bx             y=a+bx
 Ylist:L2           β>0 and ρ>0        β>0 and ρ>0
 Freq:1             t=6.3268          ↑b=.5313
 β & ρ:≠0 <0 >0     p=6.7245E-8        s=3.0613
 RegEQ:Y1           df=42.0000         r²=.4880
 Calculate         ↓a=6.4178           r=.6986
```

$H_0 : \rho = 0 \qquad H_A : \rho > 0$

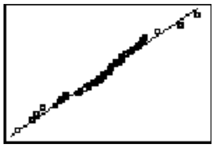
The value of r is 0.6986 and the p-value for the hypothesis test is approximately 0.0000 (reported as 0.000000067245). This indicates that there is sufficient evidence to suggest a positive linear association exists between the number of cigarettes smoked and the number of deaths from lunch cancer.

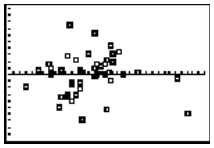
b) Answers will vary. The two values chosen here, for illustration, are cigarettes = 29 and 32. The predicted values are 21.8255 and 23.4195 respectively. This is saying that when the number of cigarettes smoked is 29 (hundred per capita) the expect number of deaths from lung cancer is 21.8255 per 10000 thousand persons. The statement for cigarettes = 32 is similar.





Note that the assumptions of normality and constant variance must also be addressed. A normality plot of the residuals and a scatter plot of cigarettes versus the residuals suggests these needed assumptions are reasonably satisfied.





c) The prediction intervals will vary, depending on the "new" values chosen. The prediction interval is appropriate based on the assumptions which have been addressed above.

14.21 The constant variance assumption is essential for the prediction interval because the value of the sample standard deviation of the residuals is used in the construction of the confidence interval. If the variance is non-constant then we would have to account for that in our prediction interval calculations. When making simple predictions (simply the predicted value without the prediction interval) a violation of the constant variance assumption is not an issue.

15 Analysis of Variance

Review Exercise Solutions

15.1 Analysis of Variance (ANOVA) is equivalent to a two sample t-test for independent data if conducted with only two samples. The strength of ANOVA is that it allows us to compare means of multiple samples simultaneously. It is a natural extension of the 2-sample t-test.

15.3 The assumptions for ANOVA consist of:

1) *Independent samples from more than two groups.* The independence assumption is reasoned out based on your knowledge of the data.

2) *Each sample is randomly obtained from a population that is normally distributed.* The randomness assumption is again reasoned out. Normal plots are produced for each sample which allow us to make a decision regarding the normality of each sample individually.

3) *All of the samples have similar variances.* This assumption can be satisfied in several ways. First, a formal F-test can be used on the largest and smallest sample variances. Second, the ratio of the largest and smallest sample variances can be examined. If the ratio is greater than 4 or less than 1/4, then we conclude the variances are not similar. Third, you can exam box plots and make a decision based on visual inspection.

15.5 Fisher's LSD produces a confidence interval for the difference of the two means. If zero is in the interval then the means are said to not be different. If zero is not in the interval, then the means are said to be different.

15.7 a) Kruskal-Wallis is essentially an ANOVA computed using the ranks of the data rather than the raw data values. By using the ranks, the magnitude of outliers that may result in a violation of the normality assumption or similar variances assumption is lessened. Kruskal-Wallis addresses the medians of the distributions whereas ANOVA addresses the means. Under additional assumptions, specifically the similarity of distribution shapes, Kruskal-Wallis can also be used to address the means.

b) The two tests are similar in that they both allow us to address the equality of the centers of multiple distributions. If it is safe to make the additional assumption that the shapes of all of the populations being sampled are similar, then Kruskal-Wallis allows you to continue discussing the mean of the distributions rather than the medians.

c) If the assumptions for the ANOVA are satisfied, then you should use ANOVA. If the similar variances and/or normality assumption is violated, then Kruskal-Wallis is the proper tool to use.

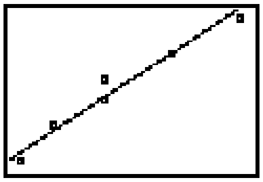
15.9 a) Before we can answer the question regarding a difference in the mean sales between the three policies, we must address the necessary assumptions for ANOVA. Since we have multiple groups and are interested in the means, we want to use ANOVA if possible.

The independence and random assumptions appear reasonable based on the information provided in the problem. The normality assumption can be addressed using normal plots. The sample data from Promotion #1 and Promotion #3 are normal; however, promotion #2 is questionable. The logical approach at this point is to complete both an ANOVA and Kruskal-Wallis. The reason is due to the fact that Promotion #2 is questionable. If it were a gross violation then ANOVA would no longer be considered. The idea is to see if what appears to be a violation in the normality assumption is sufficient to have an effect on our conclusion. If you felt the violation was not severe enough to be concerned with, then you would continue to check the required assumptions for ANOVA and skip Kruskal-Wallis.

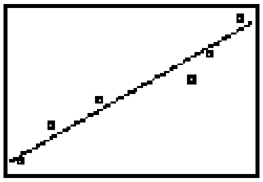
$H_0: \mu_1 = \mu_2 = \mu_3$

H_A: At least one ≠

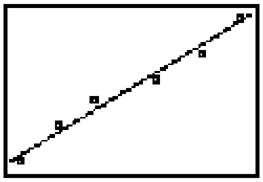
Promotion #1

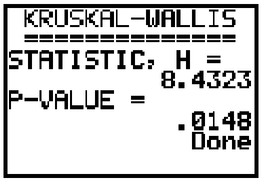


Promotion #2



Promotion #3





```
KRUSKAL-WALLIS
==============
STATISTIC, H =
        8.4323
P-VALUE =
         .0148
          Done
```

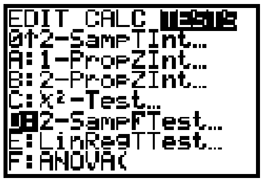
The Kruskal-Wallis resulted in a small p-value so we would reject the null hypothesis and conclude there is a difference in distribution centers. We will now continue with an ANOVA.

Similar variances assumption:

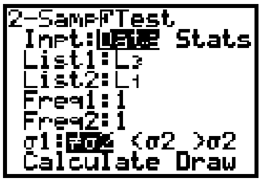
Sample	Sample Mean	Standard Deviation
Promotion #1	5.1667	0.4719
Promotion #2	5.3333	0.5785
Promotion #3	6.4833	0.7026

$$H_0: \frac{\sigma_3^2}{\sigma_1^2} = 1$$

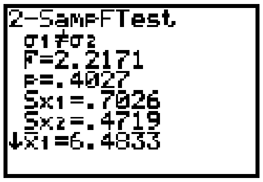
$$H_A: \frac{\sigma_3^2}{\sigma_1^2} \neq 1$$



```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```



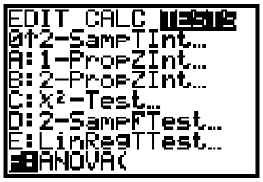
```
2-SampFTest
 Inpt:Data Stats
 List1:L3
 List2:L1
 Freq1:1
 Freq2:1
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



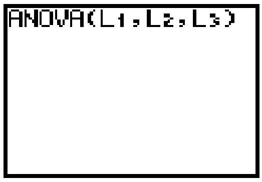
```
2-SampFTest
 σ1≠σ2
 F=2.2171
 p=.4027
 Sx1=.7026
 Sx2=.4719
↓x̄1=6.4833
```

Fail to reject the null hypothesis so conclude the equal variance assumption is reasonably satisfied (p-value = 0.4027).

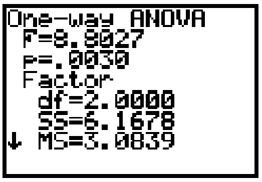
With the equal variances assumption satisfied, we will continue with the ANOVA. Based on the p-value of 0.0030 we would also reject the null hypothesis and conclude there is a difference in the distribution centers. SInce the ANOVA and Kruskal -Wallis brought us to the same conclusion, it seems reasonable to conclude what we believed may have been a violation of the normality assumption for Promotion #2 was not severe enough to have an effect on our conclusion. We may now continue with Fisher's LSD in an attempt to identify which proportions are different from the others.



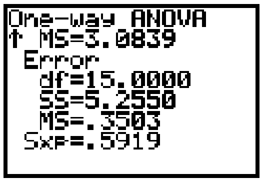
```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```



```
ANOVA(L1,L2,L3)
```



```
One-way ANOVA
 F=8.8027
 p=.0030
 Factor
  df=2.0000
  SS=6.1678
↓ MS=3.0839
```



```
One-way ANOVA
↑ MS=3.0839
 Error
  df=15.0000
  SS=5.2550
  MS=.3503
 Sxp=.5919
```

b)

Results from Fisher's LSD

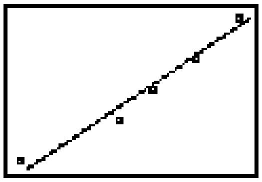
Comparison	Lower	Upper	0 in the interval?
#1 vs. #2	-0.8719	0.5387	yes
#1 vs. #3	-2.0219	-0.6113	no
#2 vs. #3	-1.8553	-0.447	no

The mean from promotion #3 is different from the other two promotions, but the other two promotions are not different from each other.

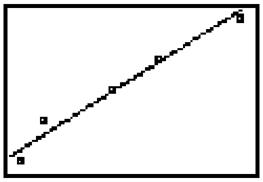
15.11 Before we can answer the question regarding whether or not there is a difference between the first week of each month, we must address the necessary assumptions for ANOVA. Since we have multiple groups and are interested in the means, we want to use ANOVA if possible.

The independence and random assumptions appear reasonable based on the information provided in the problem. The normality assumption can be addressed using normal plots. All three normal plots appear normal so the assumption is reasonably satisfied so we will check the similar variances assumption.

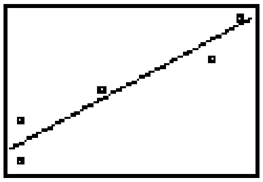
March



April



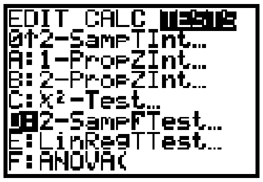
May



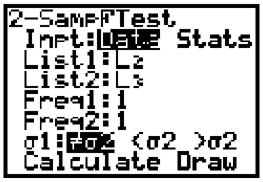
Sample	Mean	Standard Devation
March	181.4	7.6026
April	143.2	7.6942
May	146.6	3.7815

$$H_0 : \frac{\sigma_2^2}{\sigma_3^2} = 1$$

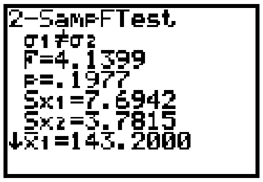
$$H_A : \frac{\sigma_2^2}{\sigma_3^2} \neq 1$$



```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```

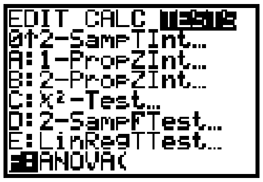


```
2-SampFTest
 Inpt:Data Stats
 List1:L2
 List2:L3
 Freq1:1
 Freq2:1
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```

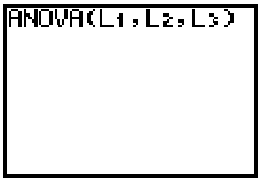


```
2-SampFTest
 σ1≠σ2
 F=4.1399
 p=.1977
 Sx1=7.6942
 Sx2=3.7815
↓x̄1=143.2000
```

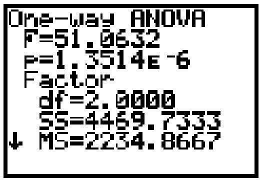
Based on the p-value of 0.1977, we will fail to reject the null hypothesis and conclude the similar variance assumption is reasonably satisfied.



```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```



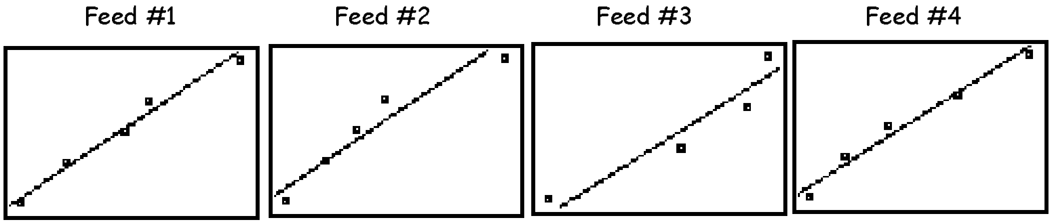
```
ANOVA(L1,L2,L3)
```

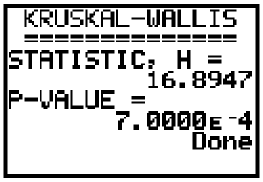


```
One-way ANOVA
 F=51.0632
 p=1.3514E-6
 Factor
  df=2.0000
  SS=4469.7333
↓ MS=2234.8667
```

Based on the p-value of 0.0000013, we will reject the null hypothesis. There is sufficient evidence to suggest the staffing needs are not the same during the first week of the three months examined. There was no desire here to identify which months were different from the others so there is no need to use Fisher's LSD. The question that was asked "Is there a difference" has been answered.

15.13 Before we can answer the question regarding a difference in the mean weight of the piglets, we must address the necessary assumptions for ANOVA. Since we have multiple groups and are interested in the means, we want to use ANOVA if possible. The independence and random assumptions appear reasonable based on the information provided in the problem. The normality assumption can be addressed using normal plots. The sample data from Feed #3 demonstrates a gross violation, as such, we will concentrate our efforts on the distribution medians rather than the means and use the Kruskal-Wallis test.





```
KRUSKAL-WALLIS
==============
STATISTIC: H =
       16.8947
P-VALUE =
     7.0000E-4
          Done
```

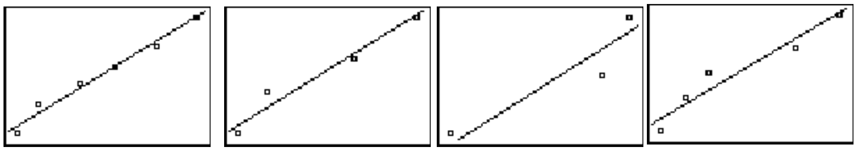
Based on a p-value of approximately 0.0000, we will reject the null hypothesis. There is sufficient evidence to suggest at least one of the four feed type medians is different from the others.

15.15 The process is called an analysis of variance because the decision is based on an F-test that is constructed from the variance measured within each group and the variance measured between each group.

15.17 a) The random variable is calories. The measurement scale is ratio.

b) This question will be addressed with a hypothesis test regarding the means.

The normal plots for the data from each restaurant appears reasonably normal with the exception of the Carl's Jr. data; however, with only three data point that is not much of a surprise.

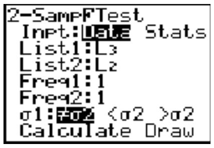


Burger King McDonalds Carl’s Jr. Jack In The Box

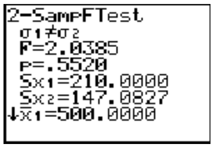
Restaurant	Sample Mean	Sample S.D.	n
Burger King	533.3333	178.1759	6
McDonalds	425.0000	147.0827	4
Carl's Jr.	500.0000	210.0000	3
Jack In The Box	474.0000	163.1870	5

Based on the above chart, we will address the similar variance assumption for ANOVA based on the variance (standard deviation) of Carl's Jr. and McDonalds.

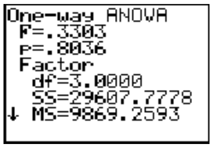
$$H_0 : \frac{\sigma^2_{Carl}}{\sigma^2_{Mac}} = 1 \qquad H_A : \frac{\sigma^2_{Carl}}{\sigma^2_{Mac}} \neq 1$$



```
2-SampFTest
 Inpt:Data Stats
 List1:L3
 List2:L2
 Freq1:1
 Freq2:1
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



```
2-SampFTest
 σ1≠σ2
 F=2.0385
 p=.5520
 Sx1=210.0000
 Sx2=147.0827
↓x̄1=500.0000
```

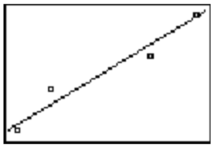


```
One-way ANOVA
 F=.3303
 p=.8036
 Factor
  df=3.0000
  SS=29607.7778
↓ MS=9869.2593
```

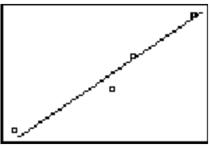
Based on the p-value of 0.5520 we will fail to reject the null hypothesis and conclude the variances are reasonably similar. This will now allow us to complete the ANOVA. The ANOVA p-value is 0.8036, which is HUGE, so we will fail to reject and come to the conclusion that there is no statistical difference between the four restaurants. If we chose to complete a KW test, rather than an ANOVA (due to the questionable normal plot) we would have calculated a p-value of 0.8715. This suggests our decision regarding the normality for this scenario did not matter. Either procedure brings us to the same conclusion.

15.19 The normal plots look reasonably normal. Variety-3 looks a bit questionable, but not too bad.

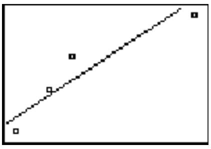
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad H_A : \text{ At least one } \neq$$



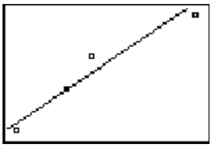
Varieyt-1



Variety-2



Variety-3

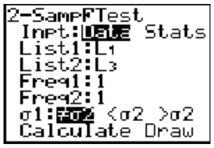


Variety-4

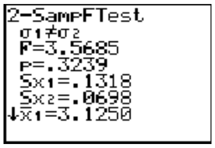
Variety	**Mean**	**Standard Deviation**
1	3.1250	0.1318
2	3.0625	0.0998
3	2.650	0.0698
4	3.2325	0.1223

Equal Variance Assumption

$$H_0 : \frac{\sigma^2_1}{\sigma^2_3} = 1 \quad H_A : \frac{\sigma^2_1}{\sigma^2_3} \neq 1$$



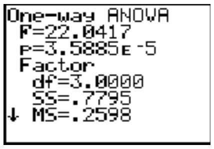
```
2-SampFTest
 Inpt:Data Stats
 List1:L1
 List2:L3
 Freq1:1
 Freq2:1
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



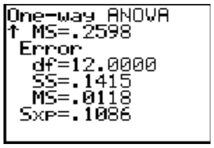
```
2-SampFTest
 σ1≠σ2
 F=3.5685
 p=.3239
 Sx1=.1318
 Sx2=.0698
↓x̄1=3.1250
```

Based on the p-value of 0.3239 we will fail to reject the null hypothesis and conclude the variances are similar. We can now continue with the ANOVA

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad H_A :$ At least one $\neq$

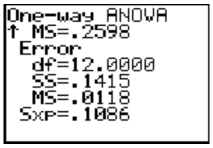


```
One-way ANOVA
 F=22.0417
 p=3.5885E-5
 Factor
  df=3.0000
  SS=.7795
↓ MS=.2598
```



```
One-way ANOVA
↑ MS=.2598
 Error
  df=12.0000
  SS=.1415
  MS=.0118
 Sxp=.1086
```

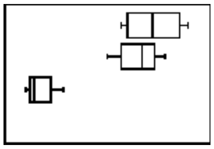
Based on the p-value of 0.0000 (0.000035885) we will conclude there is a difference between the varieties of alfalfa.

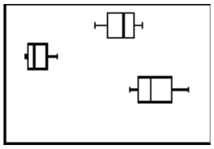


```
One-way ANOVA
↑ MS=.2598
 Error
  df=12.0000
  SS=.1415
  MS=.0118
 Sxp=.1086
```

Comparison	**Lower**	**Upper**	**Conclusion**
Variety-1 vs. Variety-2	-0.0983	0.2233	Not Different
Variety-1 vs. Variety-3	0.3142	0.6358	Different
Variety-1 vs. Variety-4	-0.2683	0.0533	Not Different
Variety-2 vs. Variety-3	0.2517	0.5733	Different
Variety-2 vs. Variety-4	-0.3308	-0.0092	Different
Variety-3 vs. Variety-4	0-.7433	-0.4217	Different

Based on Fisher's LSD test, it appears Variety-3 differs from all other varieties In addition, Varity-2 differs from Variety-4. No other differences are detected. Unfortunately, the TI-83 can only display three box-plots at the same time. The first graphic shows the box plots for varieties 1, 2, and 3. The second shows the box-plots for varieties 2, 3, and 4. From the box-plots it is no surprise that Varity-3 differs from all others; however, Variety-2 differing from Variety-4 is not at all obvious.





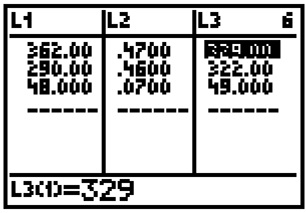
16 Categorical Data Analysis

Review Exercise Solutions

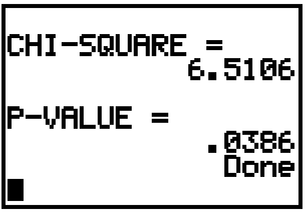
16.1 This is a goodness of fit problem.The observed frequencies were stored in L1, the theoretical probabilities in L2 and the expected frequencies (700*L2) in L3. The good fit program was then used to calculate the value of the test statistic and the p-value. Based on the expected values, the assumptions for the goodness of fit test have been satisfied. Based on the p-value of 0.0386 we will reject the null hypothesis. There is sufficient evidence to suggest there was a migration away from the democratic party in this district after 2000 presidential election.

$H_O : \pi_1 = 0.47, \pi_2 = 0.46, \pi_3 = 0.07$

H_A : At least one $\neq$



```
L1      L2      L3     6
362.00  .4700   329.00
290.00  .4600   322.00
48.000  .0700   49.000
------  ------  ------

L3(1)=329
```

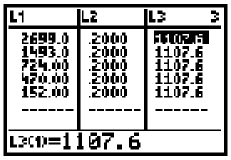


```
CHI-SQUARE =
         6.5106
P-VALUE =
          .0386
           Done
█
```

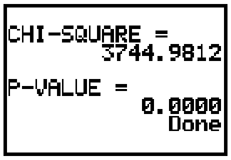
16.3 This is a goodness of fit problem. If there is not a preference in age group to ask for cigarettes, then the theoretical probability for each group will be 1/5. The observed frequencies were stored in L1, the theoretical probabilities in L2 and the expected frequencies (5538*L2) in L3. The good fit program was then used to calculate the value of the test statistic and the p-value. Based on the expected values, the assumptions for the goodness of fit test have been satisfied. Based on the p-value of 0.0000 we will reject the null hypothesis. There is sufficient evidence to suggest the proportion of adults approached for tobacco is not evenly distributed among the age groups listed.

$H_O : \pi_1 = \pi_2 = \pi_3 = \pi_3 = \pi_5 = {}^{1}\!/_{5}$

H_A : At least one $\neq$



```
L1      L2      L3     3
2699.0  .2000   1107.6
1493.0  .2000   1107.6
724.00  .2000   1107.6
470.00  .2000   1107.6
152.00  .2000   1107.6
------  ------  ------

L3(1)=1107.6
```

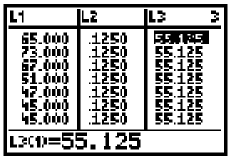


```
CHI-SQUARE =
      3744.9812
P-VALUE =
         0.0000
           Done
```

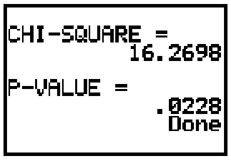
16.5 This is a goodness of fit problem. If there is not a preference in direction, then the theoretical probability for each group will be 1/8. The observed frequencies were stored in L1, the theoretical probabilities in L2 and the expected frequencies (441*L2) in L3. The good fit program was then used to calculate the value of the test statistic and the p-value. Based on the expected values, the assumptions for the goodness of fit test have been satisfied. Based on the p-value of 0.0228 we will reject the null hypothesis. There is sufficient evidence to suggest the birds do have a directional preference for their nests.

$H_O : \pi_1 = \pi_2 = \pi_3 = \pi_3 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = {}^{1}\!/_{8}$

H_A : At least one $\neq$



```
L1      L2      L3     3
65.000  .1250   55.125
73.000  .1250   55.125
67.000  .1250   55.125
51.000  .1250   55.125
47.000  .1250   55.125
45.000  .1250   55.125
45.000  .1250   55.125

L3(1)=55.125
```



```
CHI-SQUARE =
        16.2698
P-VALUE =
          .0228
           Done
```

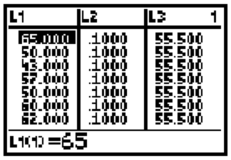
CHAPTER 16

16.7 Answers will vary. The basic idea is that the one sample proportion test is very much like a Chi-square goodness of fit where there are only two categories.

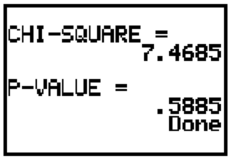
16.9 This is a goodness of fit problem. If the game is fair then the theoretical probability for each number will be the same, or 1/10. The observed frequencies were stored in L1, the theoretical probabilities in L2 and the expected frequencies (555*L2) in L3. The good fit program was then used to calculate the value of the test statistic and the p-value. Based on the expected values, the assumptions for the goodness of fit test have been satisfied. Based on the p-value of 0.5885 we will fail to reject. There is not enough evidence to suggest the game is not fair.

$H_O : \pi_1 = \pi_2 = \pi_3 = \pi_3 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_9 = \pi_{10} = 0.10$

H_A : At least one $\neq$



```
L1        L2       L3       1
65.000    .1000    55.500
50.000    .1000    55.500
43.000    .1000    55.500
57.000    .1000    55.500
50.000    .1000    55.500
60.000    .1000    55.500
62.000    .1000    55.500
L1(1) =65
```

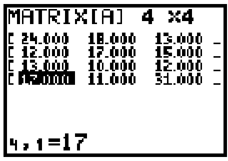


```
CHI-SQUARE =
           7.4685
P-VALUE =
            .5885
             Done
```

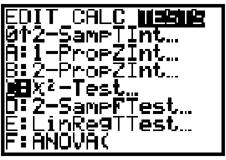
16.11 This is a Chi-squared test of independence. We will place the observed values in matrix-A. The TI-83 will calculate the expected values and place them in Matrix-B. Based on the expected values in matrix-B, the assumptions for Chi-squared are satisfied. Based on the p-value, we will reject the null hypothesis. There is sufficient evidence to suggest the hours worked and class status are dependent.

H_o : Hours worked and Class Status are independent.

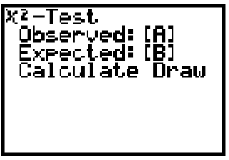
H_A : Hours worked and Class Status are dependent



```
MATRIX[A]  4 ×4
[ 24.000   18.000   13.000  _
[ 12.000   17.000   15.000  _
[ 13.000   10.000   12.000  _
[ 17.000   11.000   31.000  _

4,1=17
```

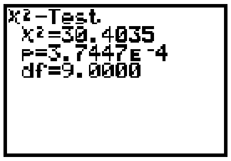


```
EDIT CALC TESTS
0↑2-SampTInt…
A:1-PropZInt…
B:2-PropZInt…
C:χ²-Test…
D:2-SampFTest…
E:LinRegTTest…
F:ANOVA(
```

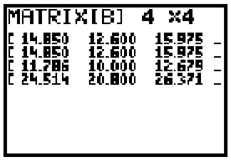


```
χ²-Test
 Observed: [A]
 Expected: [B]
 Calculate Draw
```

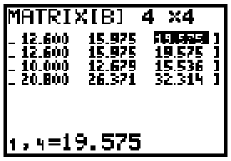
Matrix [B]



```
χ²-Test
 χ²=30.4035
 p=3.7447E-4
 df=9.0000
```



```
MATRIX[B]  4 ×4
[ 14.850   12.600   15.975  _
[ 14.850   12.600   15.975  _
[ 11.786   10.000   12.679  _
[ 24.514   20.800   26.371  _
```



```
MATRIX[B]  4 ×4
_ 12.600   15.975   19.575 ]
_ 12.600   15.975   19.575 ]
_ 10.000   12.679   15.536 ]
_ 20.800   26.371   32.314 ]

1,4=19.575
```

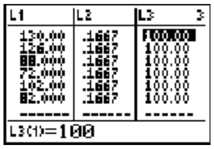
Only three columns of the matrix are shown at the same time when the calculator is set to four decimal places.

16.13 This is not a goodness of fit test. The problem is that the data consists of averages, not count data. The original data should be obtained then an ANOVA or Kruskal-Wallis test be performed, which ever is most appropriate.

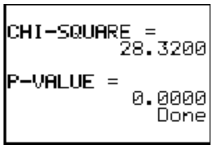
CHAPTER 16

16.15 This is a goodness of fit test. Based on the p-value of approximately zero, we will reject the null hypothesis and conclude there is evidence to suggest the die is not fair. List L3 contains the expected values which are needed to address the assumptions for this test.

$$H_0: \pi_1 = {}^1\!/_6, \pi_2 = {}^1\!/_6, \pi_3 = {}^1\!/_6, \pi_4 = {}^1\!/_6, \pi_5 = {}^1\!/_6, \pi_6 = {}^1\!/_6$$
$$H_A: \text{At least one } \neq$$



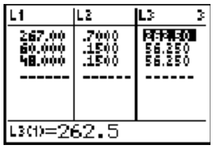
```
L1      L2      L3     3
130.00  .1667   100.00
126.00  .1667   100.00
88.000  .1667   100.00
72.000  .1667   100.00
102.00  .1667   100.00
82.000  .1667   100.00
------  ------  ------
L3(1)=100
```



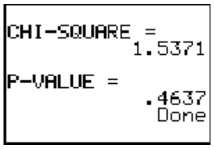
```
CHI-SQUARE =
       28.3200
P-VALUE =
        0.0000
          Done
```

16.17 This is a goodness of fit test. Based on the p-value of 0.4637, we will fail to reject the null hypothesis and conclude there is not enough evidence to suggest the professor is incorrect in his beliefs regarding where students are purchasing their books. List L3 contains the expected values which are needed to address the assumptions for this test.

$$H_0: \pi_{BS} = 0.70, \pi_{OL} = 0.15, \pi_{MO} = 0.1$$
$$H_A: \text{At least one } \neq$$



```
L1      L2      L3     3
267.00  .7000   262.50
60.000  .1500   56.250
48.000  .1500   56.250
------  ------  ------
L3(1)=262.5
```



```
CHI-SQUARE =
        1.5371
P-VALUE =
         .4637
          Done
```

CHAPTER 16