Data Analysis
an Applied Approach to Statistics
With Technology

Student Solution Manual

Brian Jean
David Meyers
Rene' Sporer

5th Edition

Solutions last updated 1 March 2018

Contents

Chapter 1 Solutions

1. a) Ratio b) Nominal c) Interval d) Ratio e) Ratio f) Nominal g) Nominal

   h) Ordinal i) Nominal

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.

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

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.

9. Answers will vary. Possible solutions include:

   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.

11. a) descriptive b) inferential c) inferential d) descriptive

13. a) Time interval between successive births. b) ratio c) continuous d) sample

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.

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. Continue with probing questions?Answers will vary.

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.

19. a) All college students.

    b) The sample is not well identified here, but clearly the sample consisted of college students who were asked about their use of these so called “focus or study” drugs.

    c) The true proportion of students who use these drugs.

    d) It is 31%.

    e) It would be inferential. The sample statistic was reported and then a statement was made about the population. This may be a little tricky in the case because it says “31% of college students report having participated ...” which is making reference to all college students Had it said “... 31% of those surveyed reported ...” then it would have been descriptive.

    f) Answers will vary here. One possible answer would be: The name of the drug/s used, the frequency used, how the drug was obtained and the gender of the user.

    g) For the above answers:

    Name: qualitative, nominal, discrete

    Frequency: quantitative, ratio, discrete

    How Obtained: Qualitative, nominal, discrete

    Gender: Qualitative, nominal, discrete.

21. Answers will vary.

23. Answers will vary.

Chapter 2 Solutions

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.

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

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.

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.

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.

11. a) Since the sample is random, the list of numbers will vary.

    b) The number of samples required is 25, so we have $500/25 = 20$. Next, we need a random start between 1 and 20. This means you need to use the same random number generator you used in the first part to generate a single random number from 1 to 20.

    c) Answers will vary.

    d) Answers will vary.

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.

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.

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.

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.

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.

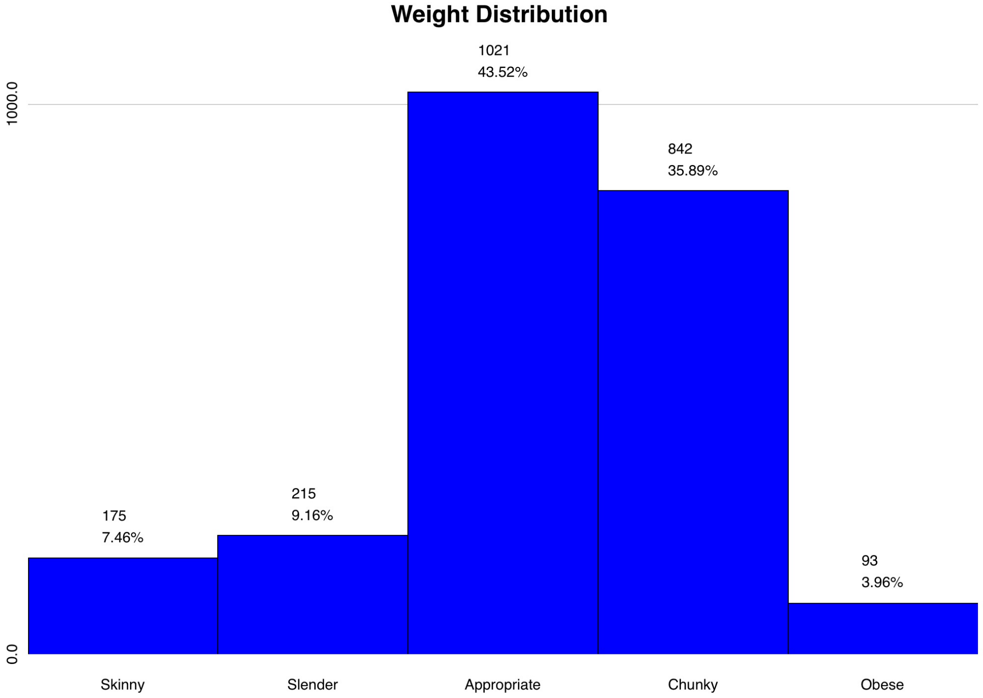
    b) Answers will vary

23. Collecting data is what we commonly do in a poll or retrospective study, such as obtaining data from medical records. Producing data involves an experiment where the situation is control and the data produced is a result of the experiment, such as the survival rate in a clinical trial testing a new cancer treatment.

25. This is an experiment. In an observational study, the data already exists and you simply go get it. In this example, the data is being produced and simultaneously collected.

Chapter 3 Solutions

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.



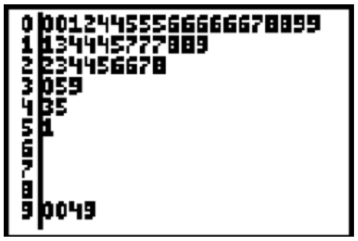
3. A histogram is used for quantitative data whereas a bar graph is used for qualitative data.

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.

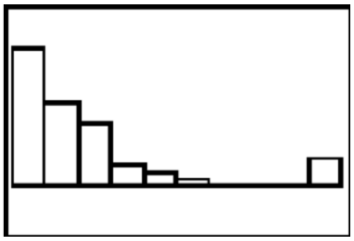
7. False. Stem-and-leaf displays have no meaning for qualitative data.

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
```

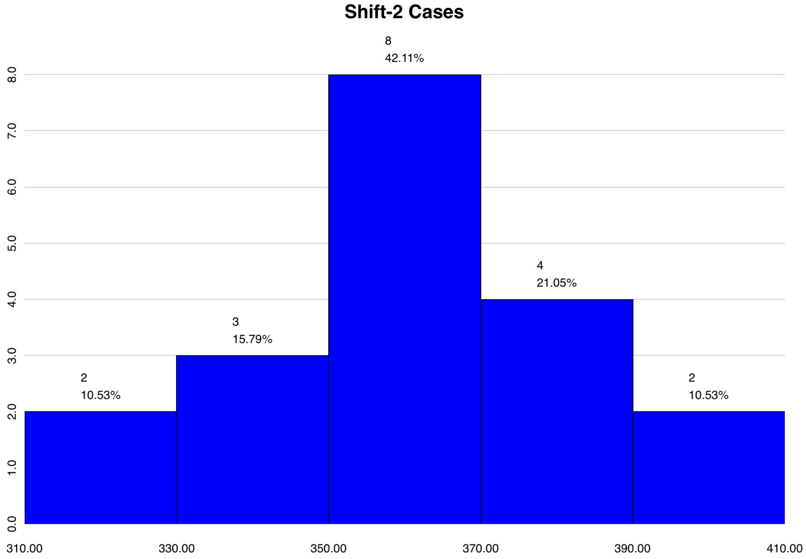


c) The distribution is skewed right.

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.

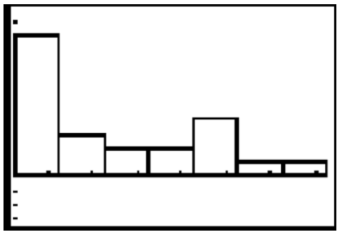
11. Variable: Number of cases produced daily. Measurement scale: ratio.



The distribution is approximately symmetric.

13. a) Ratio

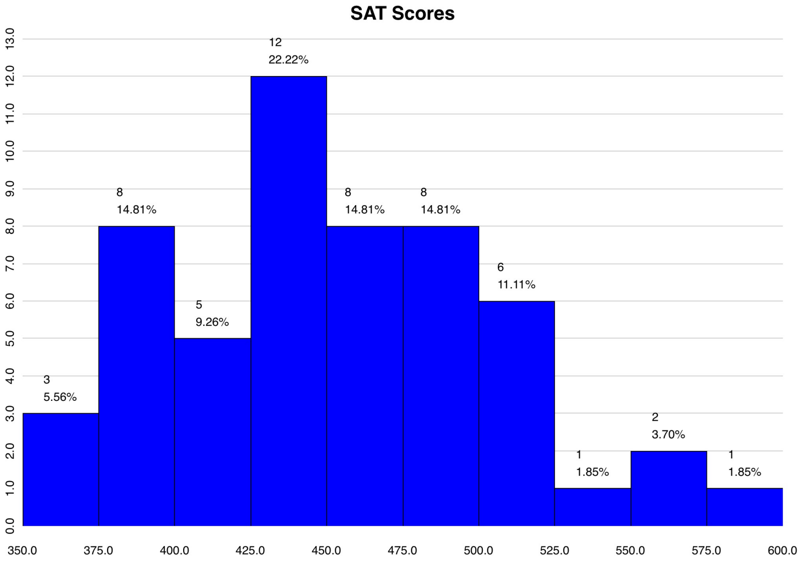
b) The data is skewed right. A student's response should have a properly labeled graphical display.



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.

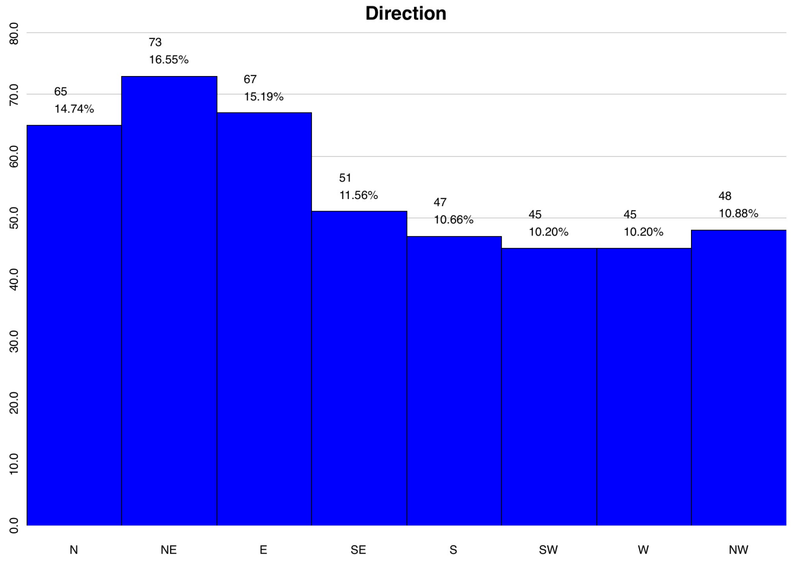
15. a) Ratio

b) The data is slightly skewed right.



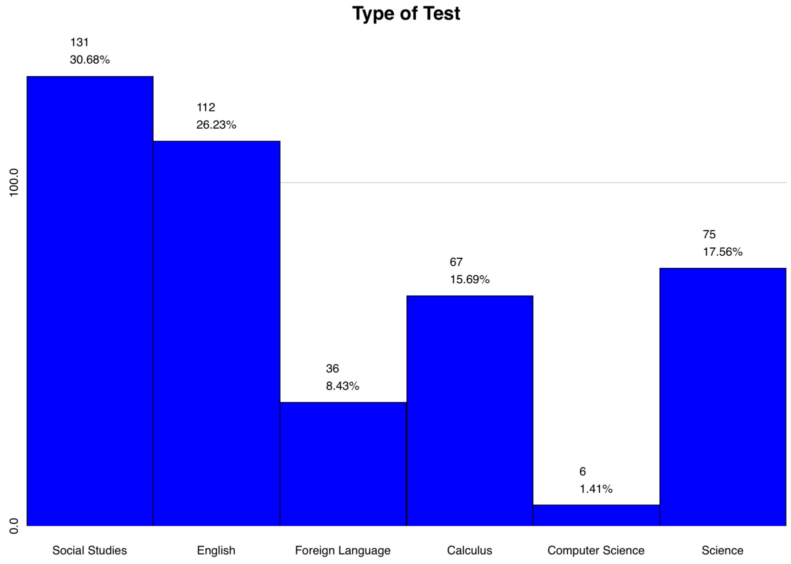
17. a) Ordinal. Regardless of which direction you start, the next category is predetermined due to the obvious order.

b)

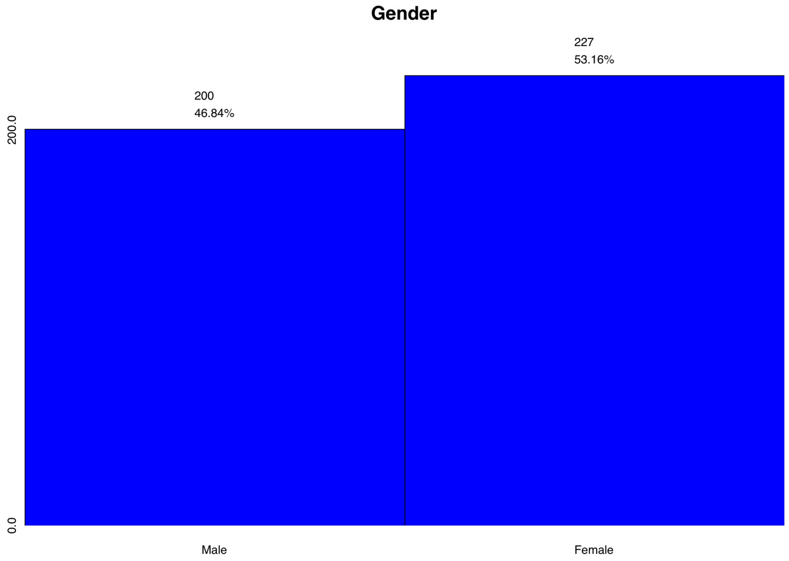


19. a) Both variables, gender and type of test, are nominal.

b)

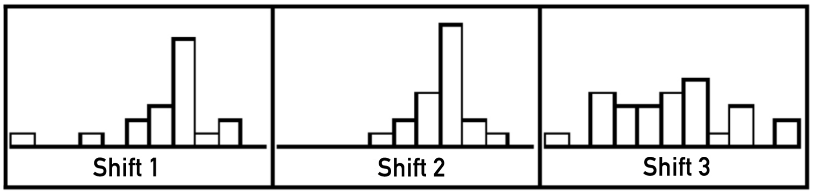


c)



21. a) Ratio.

b)



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 than shifts 1 and 2. Of the three shifts, shift 2 appears to be more consistent in their production.

23. a. WDS – Number of words in each advertisement. Discrete, Quantitative, Ratio

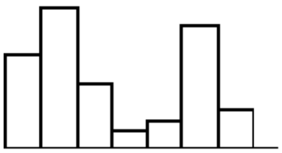
SEN – Number of sentences in each advertisement Discrete, Quantitative, Ratio

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

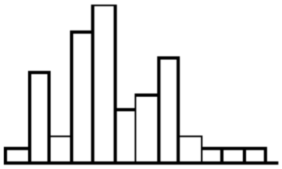
MAG – Which magazine in the sample. Discrete, Qualitative, Nominal

GROUP – Educational level of the magazine. Discrete, Qualitative, Ordinal

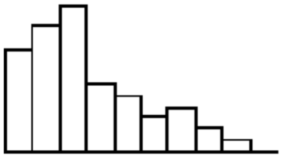
b.



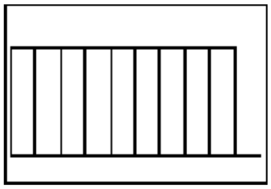
Number of Words in Ad
Bimodal



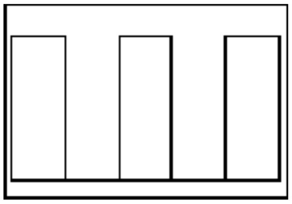
Number of Sentences in Ad
Approximately Bell Shaped



Number of 3 Syllable Words in Ad
Right Skewed



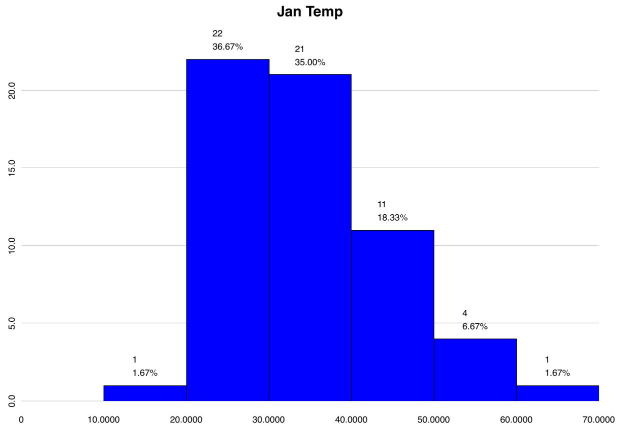
Magazines Used in Study
Uniform

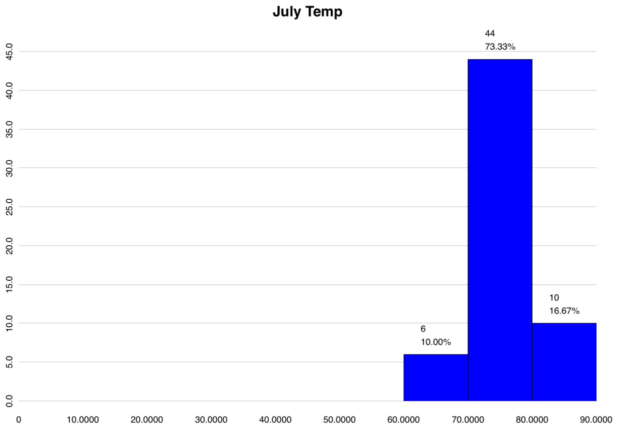


Education Level of Magazines Used in Study
Uniform

25. a) January Temperature and July Temperature are Continuous, Quantitative and Interval.

b)

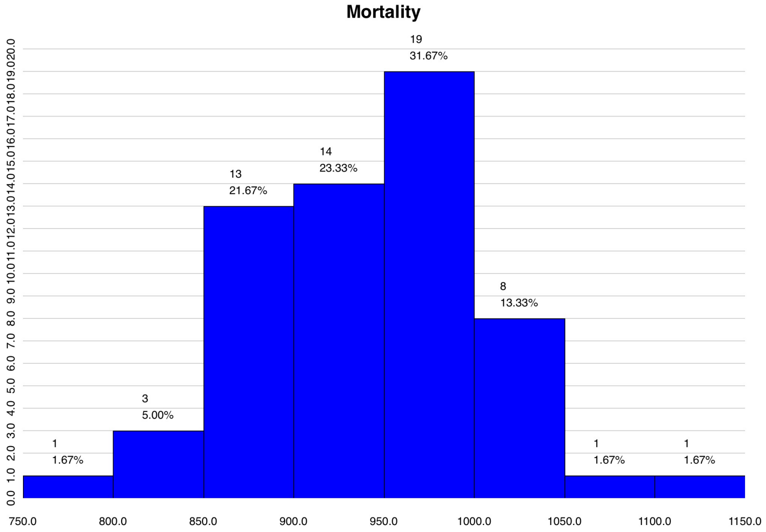


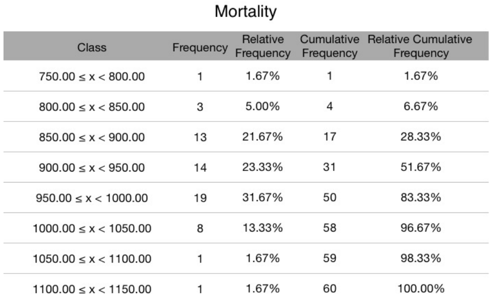


c) The temperatures in July are typically higher than the January temperatures.

27. a) Mortality Rate – Ratio.

b)





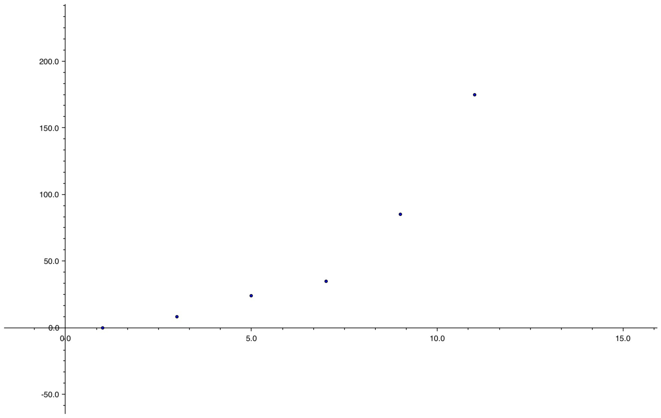
Mortality

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
750.00 ≤ x < 800.00	1	1.67%	1	1.67%
800.00 ≤ x < 850.00	3	5.00%	4	6.67%
850.00 ≤ x < 900.00	13	21.67%	17	28.33%
900.00 ≤ x < 950.00	14	23.33%	31	51.67%
950.00 ≤ x < 1000.00	19	31.67%	50	83.33%
1000.00 ≤ x < 1050.00	8	13.33%	58	96.67%
1050.00 ≤ x < 1100.00	1	1.67%	59	98.33%
1100.00 ≤ x < 1150.00	1	1.67%	60	100.00%

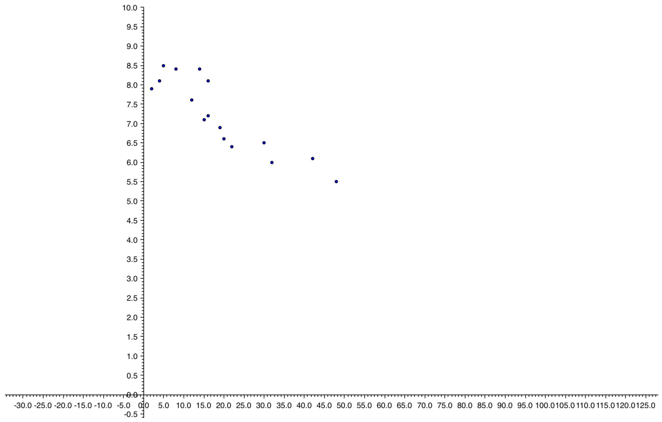
29. a) The level of measurement is nominal.

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

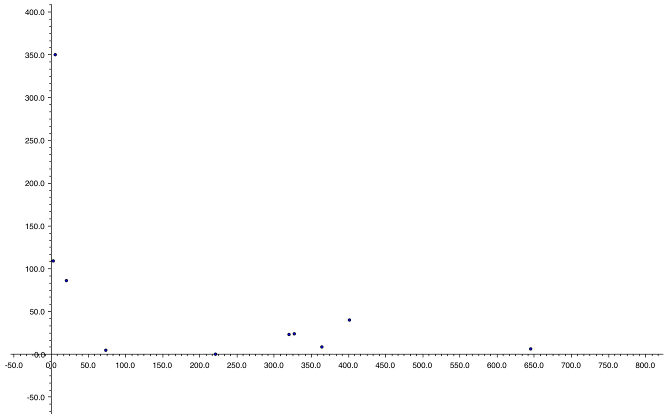
31. Scatter plot of the xy-data.



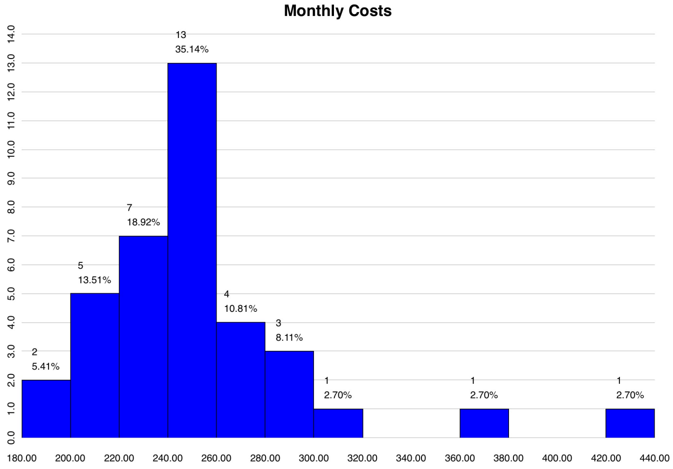
33. Cigarettes is on the horizontal axes. The shape is linear with a negative trend.



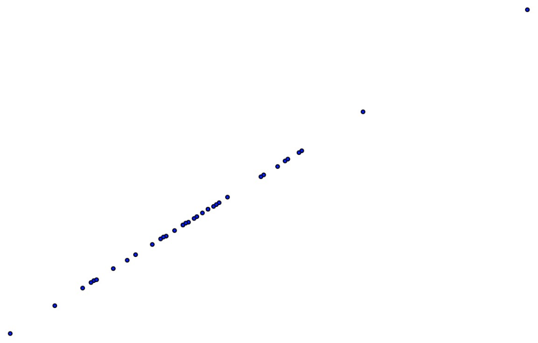
35. The shape is monotonic decreasing. There is clear curvature in the data.



37. The data is skewed right with two states clearly having more expensive health care costs that the others.



39. a) Scatter plot of the data.



b) There is a perfect, positive linear association. This means as the monthly costs increases the yearly costs increase in a perfectly predictable fashion.

c) No, it did not make any sense to graph this data. Clearly, annual costs are just a multiple of the monthly costs.

Chapter 4 Solutions

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. For non-symmetric data, the median is typically preferred.

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

3. Yes, the new average is 85.9. If the average of 37 exams was 86, then the total points was 86(37) = 3182. So the new total points for the class is $3182 + 76 + 81 + 97 = 3436$ so the new average is $\frac{3436}{40} = 85.9$.

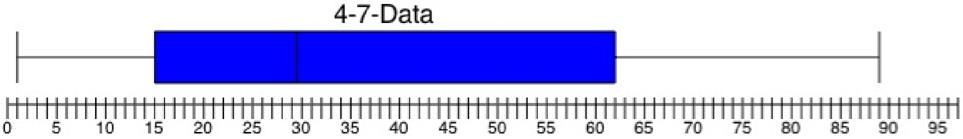
5. You would ask that your client receive the mean salary. The distribution of salaries is clearly skewed right so the mean salary will be higher than the median salary.

7. a) The TI-83/84 and TC-Stats calculates quartiles slightly differently so the solution will be slightly different depending on which technology you are using. Both answers are valid.

   TI-83/84: The values for the five-number summary are minimum = 1, Q1=16, median = 29.5, Q3 = 64, maximum = 89.

   TC-Stats: The values for the five-number summary are minimum = 1, Q1=15, median = 29.5, Q3 = 62, maximum = 89.

   b)



   c) $n = 20$ so the location is .15(20) = 3 so we will sort the data and take the average of the 3rd and 4th observations. $P_{15} = \frac{13 + 14}{2} = 13.5$

   d) $n = 20$ so the location is .23(20) = 4.6 so we will move up to the next observation, which is the 5^{th} observation. $P_{23} = 15$

e) $n = 20$ so the location is $.85(20) = 17$ so we will sort the data and take the average of the 17th and 18th observations. $P_{85} = \dfrac{66+84}{2} = 75$

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

   b) (TI-83/84)The values for the five-number summary are minimum = 16, Q1 = 28.5, median = 37, Q3 = 40.5, maximum = 47. (TC-Stats) The values for the five-number summary are minimum = 16, Q1 = 28, median = 37, Q3 = 40, maximum = 47.

   c) $0.8(28) = 22.4$ so we will go to the 23rd observation (after we sort the data) and report it as the 80^{th} percentile. $P_{80} = 42$. For the 90^{th} percentile we have $0.9(28) = 25.2$ so we will go to the 26th position and report it as the 90^{th} percentile. $P_{90} = 45$.

   d) Look at a box-plot of the data and you will quickly see the data is skewed left. Since it is not symmetric, the median should be used. For this data we have $\bar{x} = 34.5$ and $M = 37$, so 37 should be used.

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 so the median would be the single best descriptor of the middle' however, the mean can still be used to estimate the TOTAL number of squirrels. Using the mean 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 (69.933)(1478) = 103,360.974 or 103,361 squirrels.

13. The mean and median are the same if the distributional shape of the data is truly symmetric. In reality, we never have data that is truly symmetric, but often we see data that is approximately 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 reasonably symmetric.

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

    b) TI-83/84 and TC-Stats

Campaign	Min	Q_1	Median	Q_3	Max
# 1	40.0000	41.0000	42.0000	44.0000	46.0000
#2	40.0000	41.0000	43.5000	45.0000	46.0000
#3	44.0000	46.0000	48.0000	51.0000	52.0000

    c) The side-by-side box-plots are in the order Campaign #1 on the top, then Campaign #2 followed by Campaign #3.

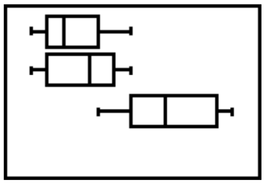


Figure 4.3: Stacked Box-and-Whisker from a TI-83/84.

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

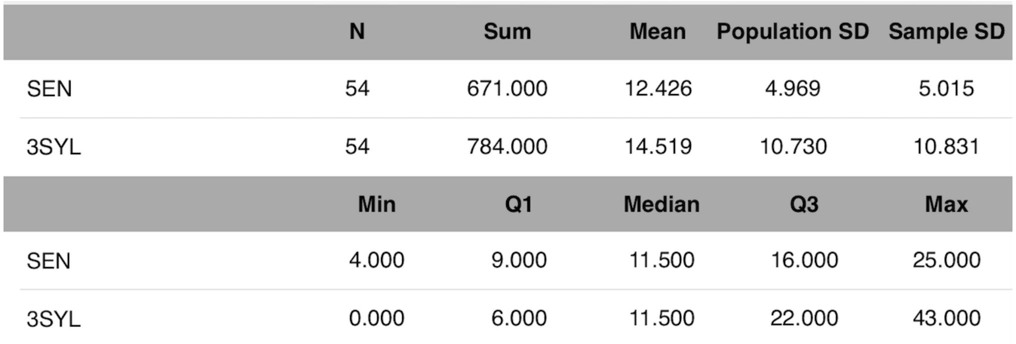
17. Answers will vary. The basic idea is that this was a silly statement, as worded. It is not possible for everyone to be above the 50^{th} percentile. By definition, 50% are above and 50% are below the 50^{th} percentile.

19. Yes. Consider the following data: 2, 10, 10, 10, 10, 10, 27, 27, 27, 27. This will result in the five-number-summary that was observed. You should create the box-and-whisker plot to verify.

21. Answers will vary.

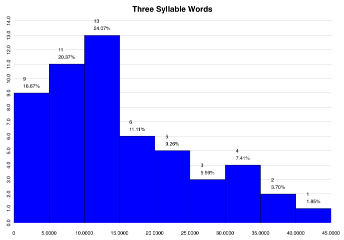
23. The data is skewed left. This can be easily observed once a box-plot is drawn.

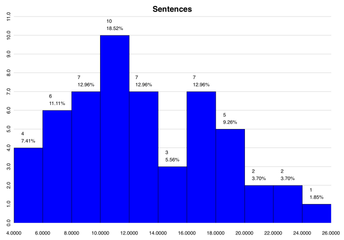
25. a) Summary statistics from TC-Stats

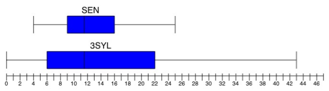


	N	Sum	Mean	Population SD	Sample SD
SEN	54	671.000	12.426	4.969	5.015
3SYL	54	784.000	14.519	10.730	10.831
	Min	**Q1**	**Median**	**Q3**	**Max**
SEN	4.000	9.000	11.500	16.000	25.000
3SYL	0.000	6.000	11.500	22.000	43.000

b) Both SEN and 2SYL are skewed right. SEN is not as heavily skewed, but definitely skewed as is even more evident in the box plots. As such, the median would be the most appropriate measure of the center.







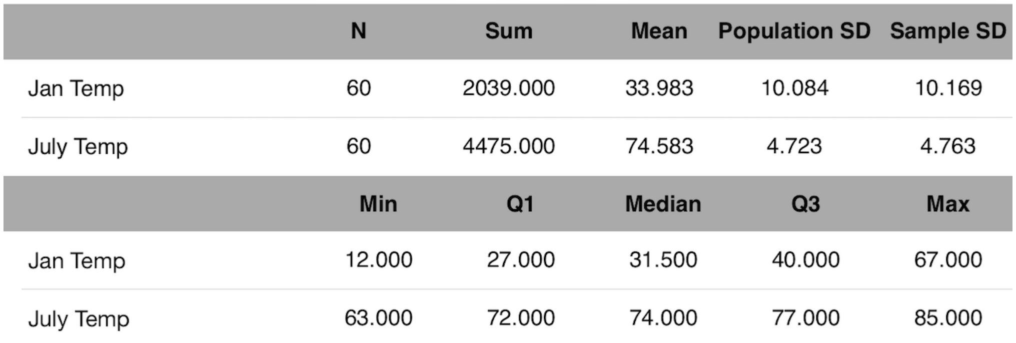
c) Since the data is not symmetric, the median is the most appropriate measure of center.

d) The median for the variable SEN is 11.5. That is telling us 50% of the advertisements have less than 11.5 sentences and 50% have more than 11.5 sentences.

e) The average for SEN is 12.426. That is telling us that "on average" the advertisements had 12.426 sentences.

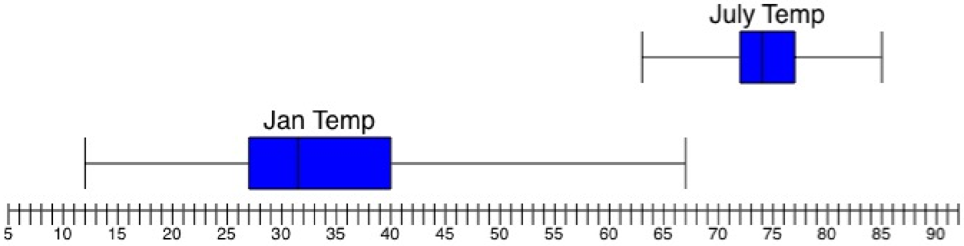
f) Both the 5-number summary and the box plots are shown above. In terms of which is more desirable, either could be justified, depending on the need/use so answers will vary.

27. a) January mean = 33.983, median = 31.5. July mean = 74.583, median = 74.000.



	N	Sum	Mean	Population SD	Sample SD
Jan Temp	60	2039.000	33.983	10.084	10.169
July Temp	60	4475.000	74.583	4.723	4.763
	Min	**Q1**	**Median**	**Q3**	**Max**
Jan Temp	12.000	27.000	31.500	40.000	67.000
July Temp	63.000	72.000	74.000	77.000	85.000

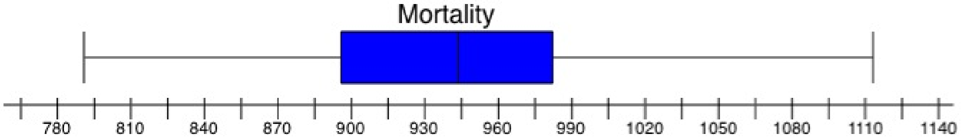
b) July appears to be reasonably symmetric; however, January is clearly skewed right. As such I would want to use the mean for July and the median for January. If I were comparing the two, then I would use the medians for both. It doesn't make any sense to compare the two using different measurements of the center.



c) Based on the graphs and the summary statistics, the mean and median temperatures in July are much high than that of January.

29. a)

Technology	Min	Q_1	Median	Q_3	Max
TI-83/84	790.73	897.48	943.685	984.12	1113.16
TC-Stats	790.73	895.70	943.685	982.29	1113.16



b) The distribution is approximately mound (bell) shaped.

c) Since the shape is reasonably mound (bell) shaped, I would use the mean.

d) In this case, $n = 60$ so the location for P_{20} will be $.20(60) = 12$. After the data is sorted, the percentile is identified as $P_{20} = \dfrac{887.47 + 891.71}{2} = 889.59$.

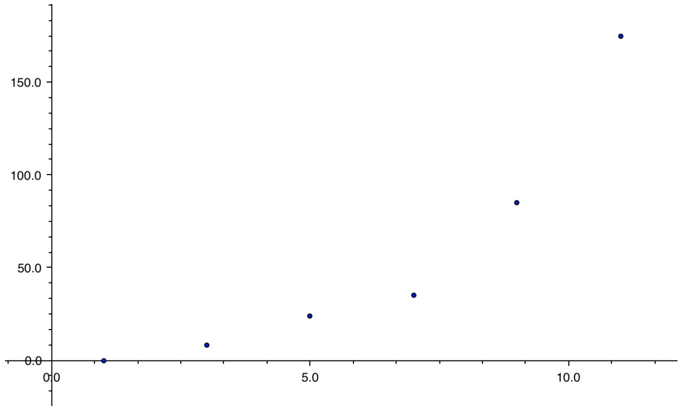
e) Once again, $n = 60$ so the location for P_{80} will be $.80(60) = 48$. After the data is sorted, the percentile is identified as $P_{80} = \dfrac{991.29 + 994.65}{2} = 992.97$.

f) Both are the same position in from the ends. The are between the 12^{th} and 13^{th} observation from each end. The definition of P_{20} is no more than 20% below leaving 80% above. The definition of P_{80} is no more than 80% below and no more than 20% above, so they are really mirror images of each other.

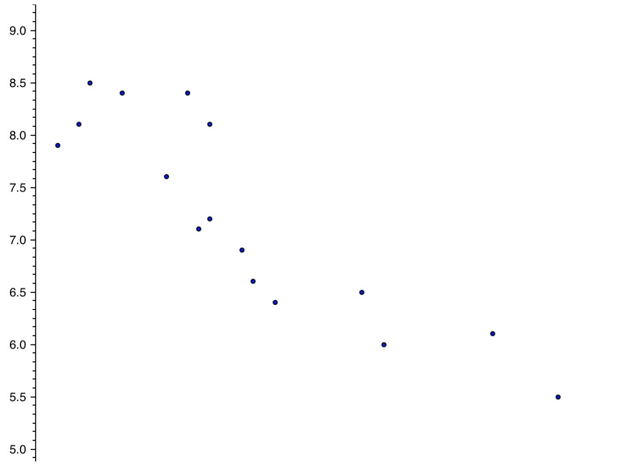
g) Once again we have $n = 60$ so the location for P_{91} will be $.91(60) = 54.6$ so we will move up to the 55^{th} ordered data value and identify $P_{91} = 1017.61$.

31. False. The measurement scale of both variables needs to be at least interval to calculate Pearson's Correlation Coefficient.

33. Pearson's Correlation coefficient is $r = 0.90$. It is not appropriate in this case because the data is obviously curved.

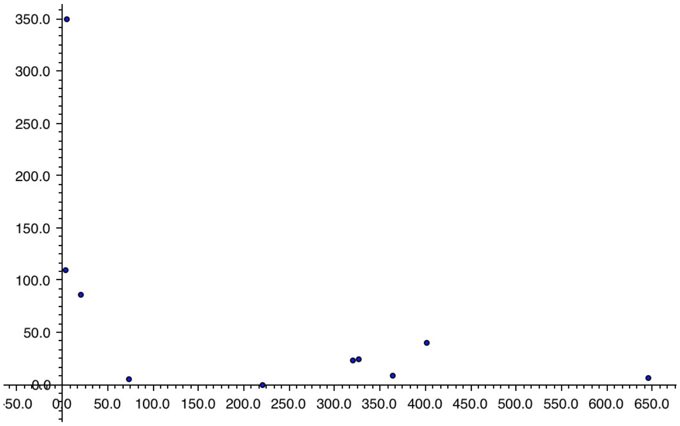


35. a) The scatter plot appears to be reasonably linear so Pearson's Correlation is appropriate.



b) $r = -0.90$ which means as the number of Cigarettes increases, the baby weigh decreases.

37. a) The data has clear curvature so Spearman's Correlation is the appropriate measure.



b) $r_S = -0.50$ which means as the number of one type of barnacles increases, the other decreases. This is suggesting they compete for space on the lobster.

Chapter 5 Solutions

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.

3. a) Sample mean, $(\bar{x}) = 10.7143$, sample median $(M) = 11.0000$. 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 $(s^2) = 2.87022^2 = 8.2380$, Sample standard deviation $(s) = 2.8702$.

5. All data must have the same value.

7. $\bar{x} = 5.7333$. $M = 5$. Range = 11 - 1 = 10. $s^2 = 3.2834^2 = 10.7807$. $s = 3.2834$.

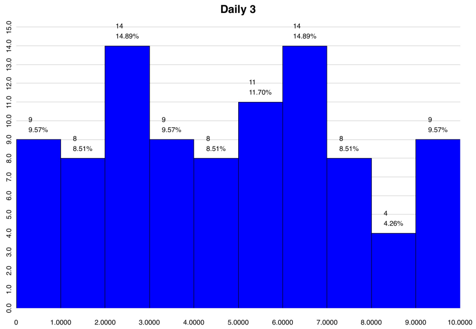
9. Answers will vary due to the randomness of the numbers generated.

11. Measurements of variability tell us how spread out our data is. Think of a simple scenario consisting of exam scores. A large variability indicates there were students on each end of the grade spectrum so. The student level of understanding that particular material is highly varied. If the measurements of variability are small, that is telling us the class as a whole has essentially the same level of understanding. That does not tell us how well they understand.

    As we saw in the Empirical Rule and Chebychev's Rule, the standard deviation (which is a measurement of variability) can be used effectively to identify unusual observations.

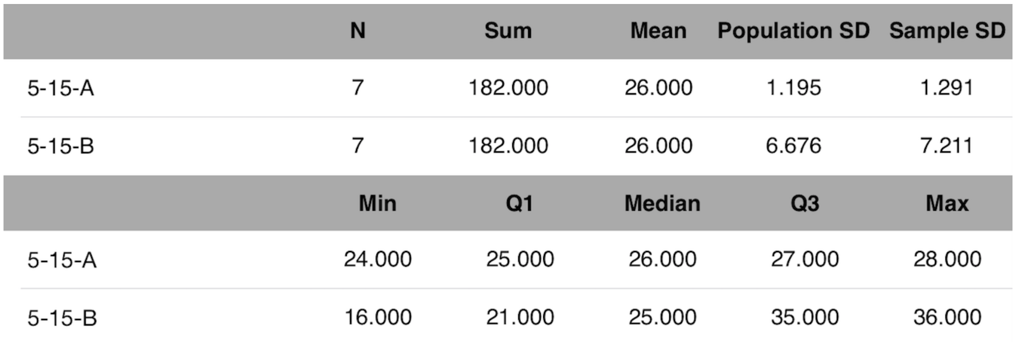
13. a) The variable of interest is the number drawn. The measurement scale is ordinal. Although numbers are used, mathematical operations with this data lacks meaning. Each number is simply a label having no more mathematical meaning than colors.

    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.



15. Answers will vary. This is one possible approach:

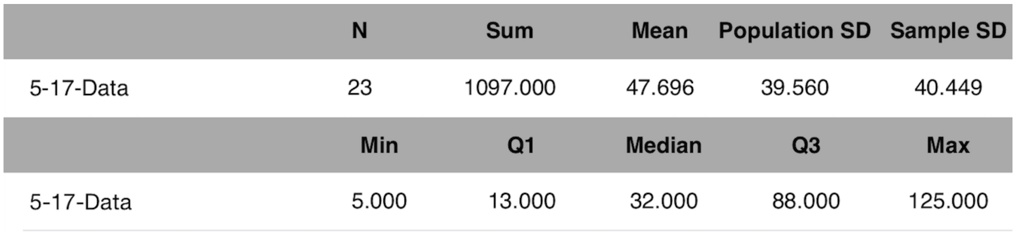
The summary statistics for these vehicles are very interesting. The average mileages are the same. Vehicle-A has a slight edge, by one mile, for the medians. Both the mean and the median describe what is "typical"; however, the standard deviation for Vehicle-B is over 5 times greater than Vehicle-A. This is a very curious result. The max is 8 miles higher and the min is 8 miles lower. This would concern me as to why there is such disparity. For that reason, I would probably go with Vehicle-A because it is much more consistent. The instances where I get the much greater mileage with Vehicle-B may be rare and would appear to be offset by the times when I would be getting the much lower mileage.



	N	Sum	Mean	Population SD	Sample SD
5-15-A	7	182.000	26.000	1.195	1.291
5-15-B	7	182.000	26.000	6.676	7.211

	Min	Q1	Median	Q3	Max
5-15-A	24.000	25.000	26.000	27.000	28.000
5-15-B	16.000	21.000	25.000	35.000	36.000

17. a) The mean, standard deviation and all elements for the 5 number summary are shown below. The proper symbols are: $\bar{x} = 47.696,\ s = 40.449,\ min = 5,\ Q_1 = 13,\ M = 32,\ Q_3 = 88,\ max = 125$.



	N	Sum	Mean	Population SD	Sample SD
5-17-Data	23	1097.000	47.696	39.560	40.449

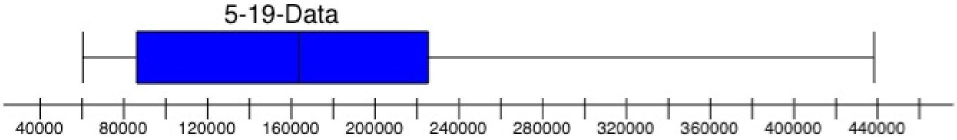
	Min	Q1	Median	Q3	Max
5-17-Data	5.000	13.000	32.000	88.000	125.000

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

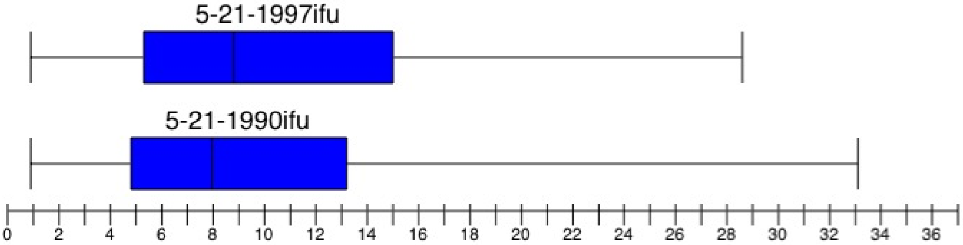
19. a) The random variable is the population of a county. The measurement scale is ratio.

b) $\bar{x} = 176024.1500, \quad M = 163586.0000$. Yes, there does appear to be a big difference between the sample mean and median, they are 12438.5 a part 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.

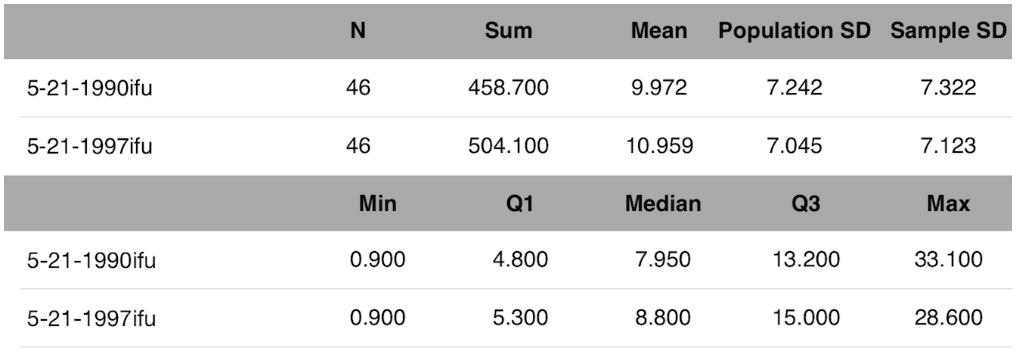


21. Variable: Level of unemployment. Scale: Ratio. b) Stacked box–plots.



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) The median unemployment for these reporting countries was 7.95% in 1990 and grew slightly to 8.80% in 1997. If this trend is accurate, there appears to be a growing level of unemployment among women in these countries.



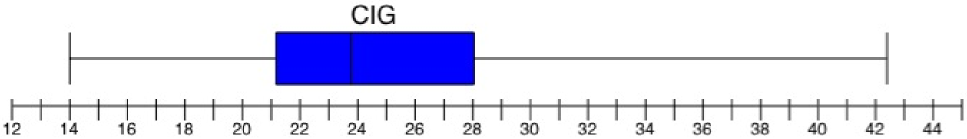
	N	Sum	Mean	Population SD	Sample SD
5-21-1990ifu	46	458.700	9.972	7.242	7.322
5-21-1997ifu	46	504.100	10.959	7.045	7.123
	Min	**Q1**	**Median**	**Q3**	**Max**
5-21-1990ifu	0.900	4.800	7.950	13.200	33.100
5-21-1997ifu	0.900	5.300	8.800	15.000	28.600

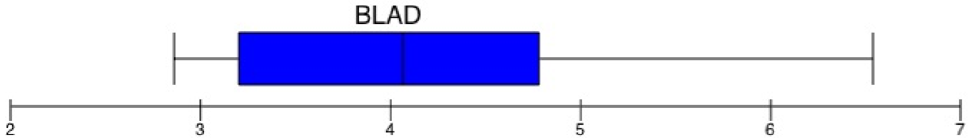
23. a) Hundreds of Cigarettes per Thousand – Continuous, Quantitative, Ratio

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

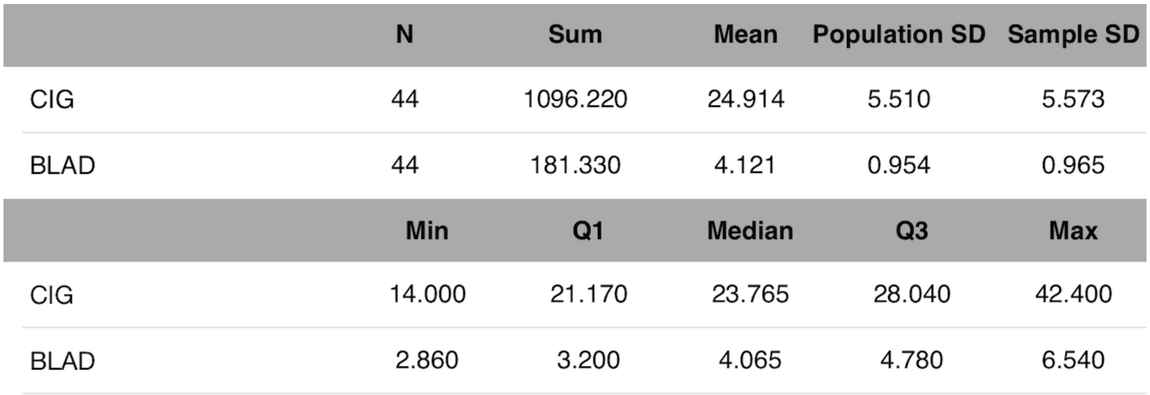
b) Cigarettes by State for the U.S. – Skewed right.

Bladder Cancer Deaths – Skewed Right



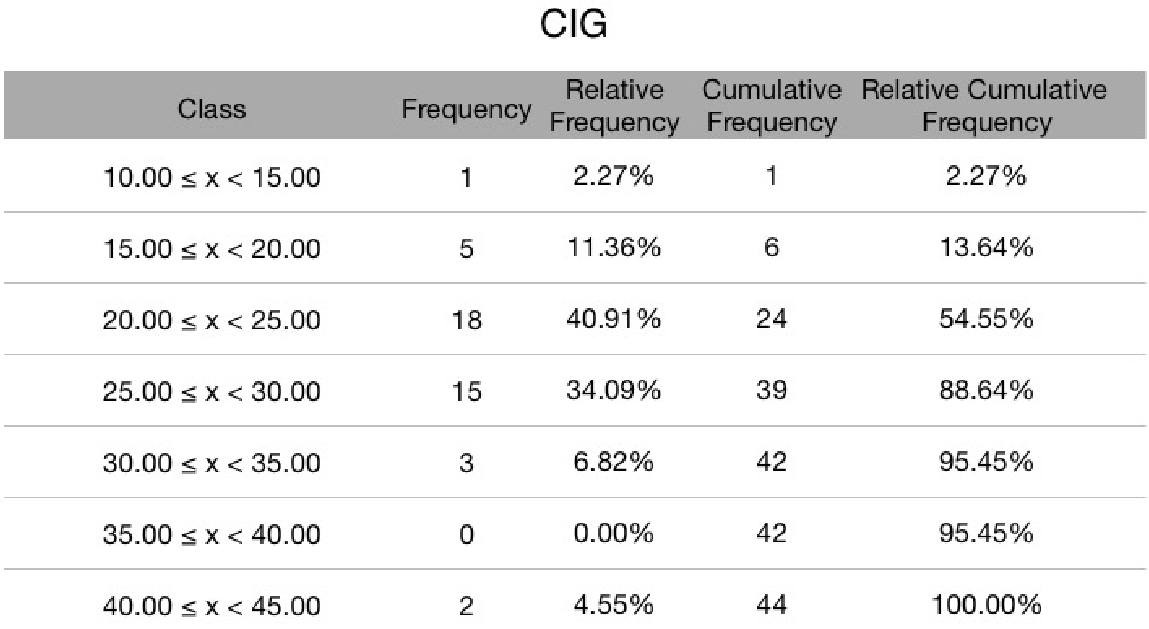


c. The mean and medians are shown below.



	N	Sum	Mean	Population SD	Sample SD
CIG	44	1096.220	24.914	5.510	5.573
BLAD	44	181.330	4.121	0.954	0.965
	Min	Q1	Median	Q3	Max
CIG	14.000	21.170	23.765	28.040	42.400
BLAD	2.860	3.200	4.065	4.780	6.540

We would not be reporting a mode with this data. Rather, we would be reporting a modal class and that will depend on the lower bounds you set and the class widths. As an example, the frequency distribution table for cigarettes is shown below. The left end point used was 10 and the class width was set to 5. The modal class is then 20–25 since that class has the highest frequency, which is 18..



CIG

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
10.00 ≤ x < 15.00	1	2.27%	1	2.27%
15.00 ≤ x < 20.00	5	11.36%	6	13.64%
20.00 ≤ x < 25.00	18	40.91%	24	54.55%
25.00 ≤ x < 30.00	15	34.09%	39	88.64%
30.00 ≤ x < 35.00	3	6.82%	42	95.45%
35.00 ≤ x < 40.00	0	0.00%	42	95.45%
40.00 ≤ x < 45.00	2	4.55%	44	100.00%

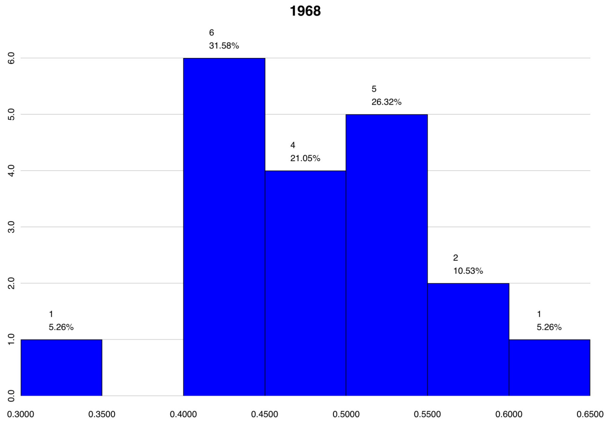
d. CIG: Range = 42.4 – 14 = 28.4, $s = 5.5733$, $s^2 = 5.5733^2 = 31.0617$.

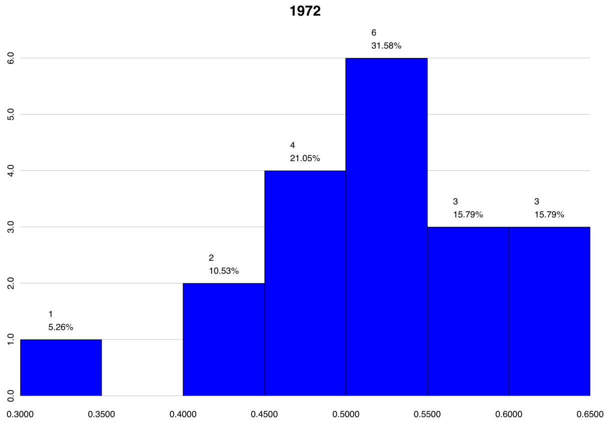
BLAD: Range = 6.54 - 2.86 = 3.68, $s = 0.9649$, $s^2 = 0.9649^2 = 0.9310$.

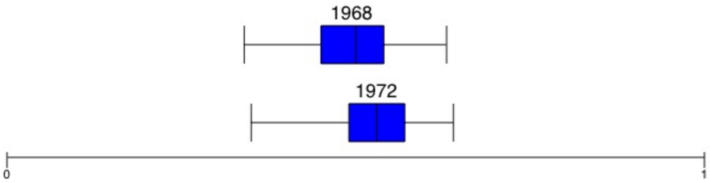
e. Summary statistics are shown for both above. Based on the fact that both distributions are skewed, the median would be the single best measure of the center. Likewise, the 5-number summary would be used to describe the the spread of the data; however, the 5-number summary is not a single measurement. The standard deviation would be the single best measurement because Chebychev's Rule still allows us to use the standard deviation to describe the spread of the data.

25. a) 1968 and 1972 are both Continuous, Quantitative and Ratio.

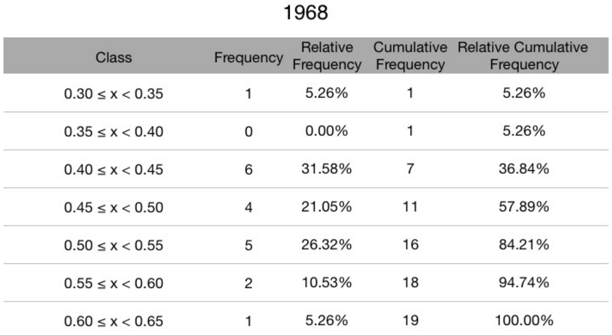
b)





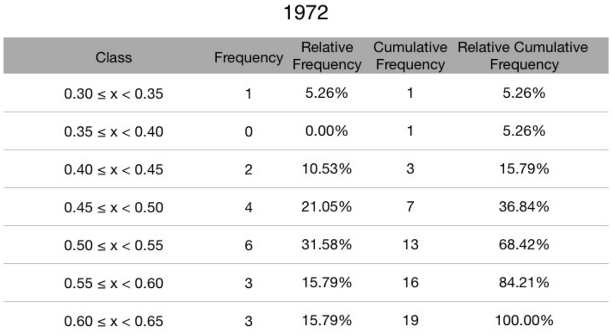


c) The summary statistics are shown below. I am not sure the "mode" makes as much sense here as does the "modal class." The modal class for 1968 is [0.40, 0.45]. The modal class for 1972 is [0.50, 0.55].



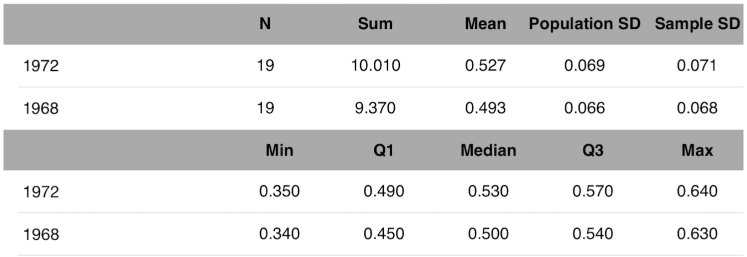
1968

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
0.30 ≤ x < 0.35	1	5.26%	1	5.26%
0.35 ≤ x < 0.40	0	0.00%	1	5.26%
0.40 ≤ x < 0.45	6	31.58%	7	36.84%
0.45 ≤ x < 0.50	4	21.05%	11	57.89%
0.50 ≤ x < 0.55	5	26.32%	16	84.21%
0.55 ≤ x < 0.60	2	10.53%	18	94.74%
0.60 ≤ x < 0.65	1	5.26%	19	100.00%



1972

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
0.30 ≤ x < 0.35	1	5.26%	1	5.26%
0.35 ≤ x < 0.40	0	0.00%	1	5.26%
0.40 ≤ x < 0.45	2	10.53%	3	15.79%
0.45 ≤ x < 0.50	4	21.05%	7	36.84%
0.50 ≤ x < 0.55	6	31.58%	13	68.42%
0.55 ≤ x < 0.60	3	15.79%	16	84.21%
0.60 ≤ x < 0.65	3	15.79%	19	100.00%



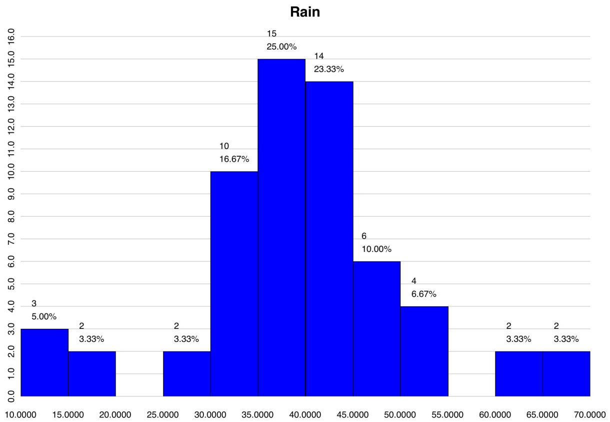
	N	Sum	Mean	Population SD	Sample SD
1972	19	10.010	0.527	0.069	0.071
1968	19	9.370	0.493	0.066	0.068
	Min	**Q1**	**Median**	**Q3**	**Max**
1972	0.350	0.490	0.530	0.570	0.640
1968	0.340	0.450	0.500	0.540	0.630

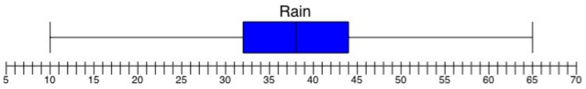
d) The range for 1968 is 0.63 - 0.34 = 0.29. The range for 1972 is 0.64 - 0.35 = 0.29. The standard deviations are displayed above. $s_{1972} = 0.071$ so the variance, $s^2_{1972} = 0.071^2 = 0.005041$. Likewise for 1968, $s_{1968} = 0.068$ so the variance, $s^2_{1968} = 0.068^2 = 0.0046$.

e) Both of these distribution appear to be relatively symmetric so I wold chose the standard deviations for each. The histograms, because of the starting points and class widths, may or may not appear to be approximately symmetric. This is a perfect example when we should consider both the histograms and the box-ploys to help us decide the overall general shape of the distribution.

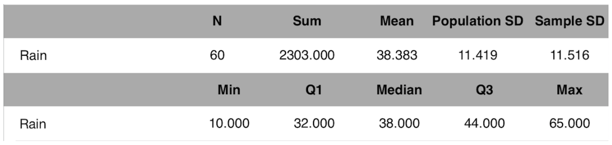
27. a) Rain – Continuous, Quantitative and Ratio.

b)

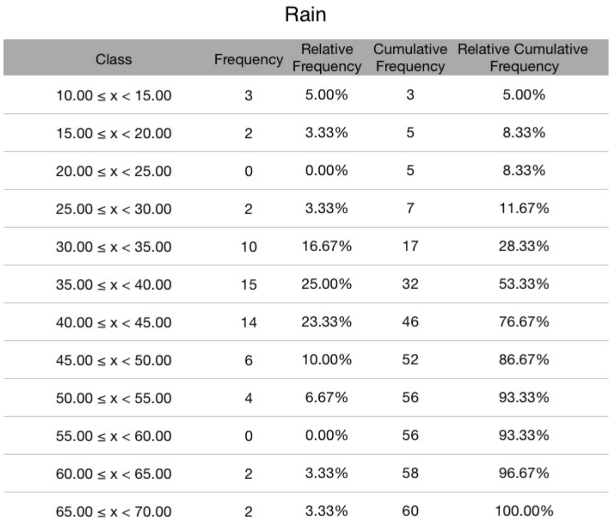




c) The modal class makes more sense than the mode here. The modal class is [35, 40].



	N	Sum	Mean	Population SD	Sample SD
Rain	60	2303.000	38.383	11.419	11.516
	Min	**Q1**	**Median**	**Q3**	**Max**
Rain	10.000	32.000	38.000	44.000	65.000



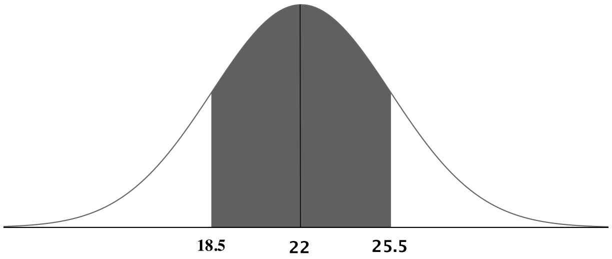
Rain

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
10.00 ≤ x < 15.00	3	5.00%	3	5.00%
15.00 ≤ x < 20.00	2	3.33%	5	8.33%
20.00 ≤ x < 25.00	0	0.00%	5	8.33%
25.00 ≤ x < 30.00	2	3.33%	7	11.67%
30.00 ≤ x < 35.00	10	16.67%	17	28.33%
35.00 ≤ x < 40.00	15	25.00%	32	53.33%
40.00 ≤ x < 45.00	14	23.33%	46	76.67%
45.00 ≤ x < 50.00	6	10.00%	52	86.67%
50.00 ≤ x < 55.00	4	6.67%	56	93.33%
55.00 ≤ x < 60.00	0	0.00%	56	93.33%
60.00 ≤ x < 65.00	2	3.33%	58	96.67%
65.00 ≤ x < 70.00	2	3.33%	60	100.00%

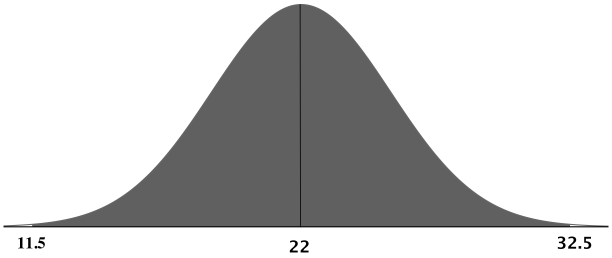
d) The sample standard deviation is given above as 11.516 so the sample variance, $s^2 = 11.516^2 = 132.8487$.

e) Since the data is reasonably symmetric, I would use the Empirical Rule.

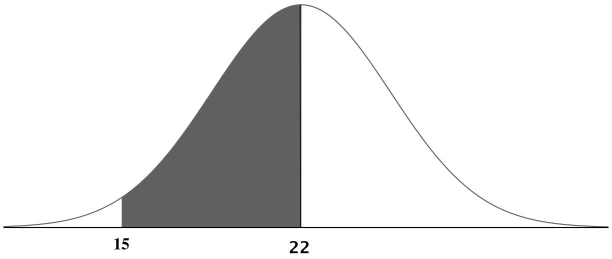
29. a) 18.5 to 25.5 is 22 ± 3.5 which is one standard deviation from the mean. As such, based on the Empirical rule, I would expect to find approximately 68% of the data between 18.5 and 25.5.



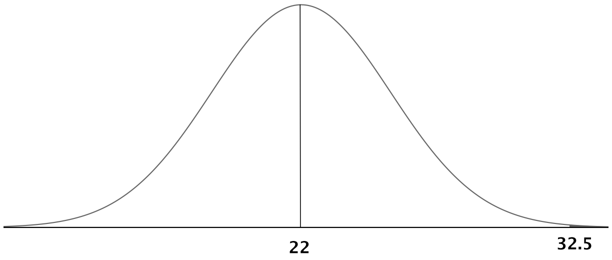
b) 11.5 to 32.5 is $22 \pm 3(3.5)$ which is three standard deviations from the mean. As such, based on the Empirical rule, I would expect to find approximately 99.7% (or essentially all) of the data between 11.5 and 32.5.



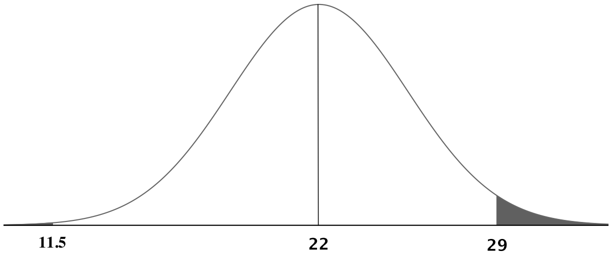
c) 15 to 22 is $22 - 2(3.5)$ which is two standard deviations below the mean. As such, based on the Empirical rule, I would expect to find approximately 95% of the data within two standard deviations of the mean. This is actually half of that so I would expect to see 95% $over 2 = 47.5\%$.



d) Greater than 32.5 is greater than three standard deviations above the mean.I really wouldn't expect much at all, if anything. We know there is 99.7% within three standard deviations which tells us there is 100% - 99.7% = 0.03% beyond three standard deviations. This means I would expect $\frac{0.3\%}{2} = 0.15\%$. Realistically, I wouldn't expect to see anything that far out.



e) we need to break these down into pieces. The first piece will e less than 11.5. 11.5 is three standard deviations from the mean so we know from part (d) to expect 0.15%. Greater than 29 is greater than two standard deviations above the mean. We know to expect 95% within two standard deviatins so beyond two standard deviations leaves us 5%. Half of 5% is 2.5% so we expect a total of 2.5% + 0.15% which is 2.65%.



31. The standard deviation is very important because it gives a method of describing unusual observations for any distribution. The standard deviation is "*standardized*" making it useful to compare unusual observations from different distributions. More than three standard deviations above the mean is very unusual, regardless of what the actual value of the standard deviation is.

Chapter 6 Solutions

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."

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) $P(\textit{exactly one head is observed}) = \frac{3}{8}$. $P(\textit{at least } 1 \textit{ head was observed}) = \frac{7}{8}$.

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

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.

7. Answers will vary. A common approach may be to show that a probability can be represented as the number of times an outcome is observed divided by the number of trials. As such, it cannot be observed less than zero times or greater than the total number of trials.

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

11. True. By definition, two events that are mutually exclusive are dependent because if one happens, it completely dictates the probability of the other (which is zero).

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 $\frac{0}{S(n)}$, which is 0.

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

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.

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 of Y indicates the trait was passed on by that parent.

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

    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:

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

    Since the two events are independent, the answer is simply $0.93^2 = 0.8649$.

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

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.

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

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

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

    b) $P(Belief = yes) = \dfrac{806}{1076} = 0.7491$

c) $P(Belief\, and\, female) = \dfrac{435}{1076} = 0.4043$

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

e) If independent then $P(Belief = yes)$ and $P(Belief = yes|Males)$ will be the same. We already found $P(Belief = yes)$ is $\dfrac{806}{1076}$. $P(Belief = yes|Males) = \dfrac{371}{488}$. These two probabilities are deferent so the two events are dependent, not independent.

f) $806 : 270 \approx 2.9852 : 1$

g) $117 : 371 \approx 0.3154 : 1$

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

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

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.

$\dfrac{37+30+5+1+41+34}{427} = \dfrac{148}{427} = 0.3466$

We could also have solved this by using a compliment. Find the probability they did take the social studies, English or foreign language tests (the part you do not want) and subtract that answer from 1.

d) $\dfrac{1}{427} = 0.0023$

e) $P(female \cup Science\, test) = P(female) + P(Science\, test) - P(female \cap Science\, test) =$

$\dfrac{227}{427} + \dfrac{75}{427} - \dfrac{34}{427} = \dfrac{268}{427} \approx 0.6276$

f) $5 : 427 - 5 = 5 : 422$

g) $\dfrac{227+42}{427} = 0.6300$

h) $P(Female\,|\,English) = \dfrac{n(Female \cap English)}{n(English)} = \dfrac{70}{112} = 0.6250$

i) $\dfrac{42}{200} = 0.2100$

Chapter 7 Solutions

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.

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

5. a) Discrete b) Continuous c) Continuous d) Discrete e) Continuous
f) Discrete g) Discrete h) Discrete

7. False. It is possible, just not probable. If the distribution is discrete, I would not expect the calculated mean to match the theoretical but it would not surprise me if it did. If the distribution is continuous then it is not really possible because we do not have the ability to measure anything with infinite precision. You may "see" it happen, but that would only be because we measure everything discretely even though it may be continuous in nature.

9. This is a proper probability distribution since all probabilities are on [0, 1] and they all add up to $\frac{45}{45} = 1$.

x	$P(X = x)$
1	$\frac{1^3+3}{45} = \frac{4}{45}$
2	$\frac{2^3+3}{45} = \frac{11}{45}$
3	$\frac{3^3+3}{45} = \frac{30}{45}$

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% because 0.461 + 0.31 + 0.045 + **.184**= 1.

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:

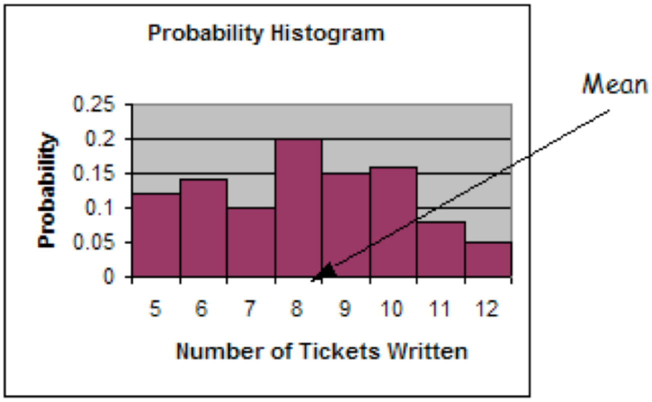
$$\mu = \sum x\, P(X = x) = 5(0.12) + 6(0.14) + 7(0.10) + 8(-.20) + 9(0.15) + 10(0.16) + 11(0.08) + 12(0.05) = 8.17$$

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

$$\sigma^2 = \sum (x-\mu)^2\, P(X = x) = (5-8.17)^2(0.12) + (6-8.17)^2(0.14) + (7-8.17)^2(0.10) + (8-8.17)^2(.20)$$
$$+(9-8.17)^2(0.15) + (10-8.17)^2(0.16) + (11-8.17)^2(0.08) + (12-8.17)^2(0.05) = 4.0211$$

This implies the standard deviation, σ is $\sqrt{4.0211} = 2.0053$.

i) A probability histogram is simply a histogram with the relative frequency (probability) on the y-axes rather than frequency. This is a bit awkward with a TI-83/84 or TC-Stats. Rather than spending a bunch of time trying to get these two pieces of technology to cooperate, it is easier to do by hand. Below is the desired histogram created with Excel.



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

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.

Chapter 8 Solutions

A Note To Students: Problem 1 asks you to outline the required assumptions to use a binomial distribution. The answer is given very generically. Each problem needs to have the assumptions outlined in such a way as to connect the assumption with the problem. As an example, if there are 25 trials in a particular scenario then you may say something like "n is fixed at 25." That shows you connected that assumption to the scenario you are working with. Each assumption needs to be reasonably connected to the scenario you are working with to ensure the binomial probability distribution can be properly applied to that scenario. The following solutions do not go into that much detail although your responses should.

It would be very wise on your part for you to ask your professor how much detail regarding the assumptions he/she wants you include in your solutions.

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 , 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.

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.

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.

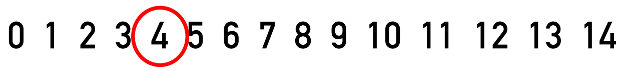
h) No. The random variable is the amount of time rather than a success or failure

7. False. Typically we see multiple trials, but the properties for a binomial still hold for n = 1. When $n = 1$, we actually have what is known as a Bernouli trial. It is a single trial.

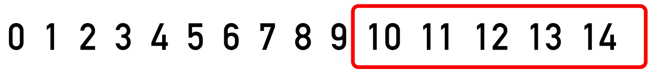
9. Suppose you were selling candy bars for \$1.00 each and were told you will keep 30% of the total amount of what you sell. If you sold 100 boxes then you would expect to profit \$30.00. You came to this conclusion by simple arithmetic: $0.30(100) = 30$. The expected value is another term for mean.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

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

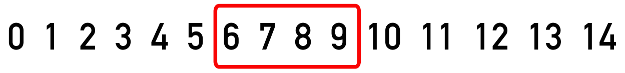
13. a) This scenario is describing a binomial pdf. It is asking $P(X = 4 \,|\, \pi = 0.763, n = 14) = 1.8968E{-4} \approx 0.0002$.



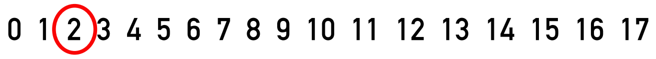
b) This scenario is describing a binomial cdf. It is asking $P(X \geq 10 \,|\, \pi = 0.763, n = 14) = 0.7786$.



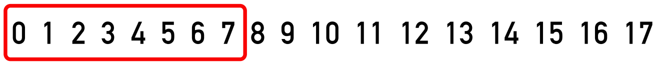
c) This scenario is describing a binomial cdf. It is asking $P(6 \leq X \leq 10 \,|\, \pi = 0.763, n = 14) = 0.2199$.



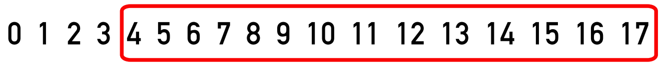
15. a) This scenario is describing a binomial pdf. It is asking $P(X = 2 \,|\, \pi = 0.52, n = 17) = 0.0006$



b) This scenario is describing a binomial cdf. It is asking $P(X \leq 7 \,|\, \pi = 0.52, n = 17) = 0.2577$



c) This scenario is describing a binomial cdf. It is asking $P(X \geq 4 \,|\, \pi = 0.52, n = 17) = 0.9960$



d) This scenario is describing a binomial pdf. It is asking $P(X = 17 \,|\, \pi = 0.52, n = 17) = 1.486E{-}05 = 0.000014$ or approximately 0.

$$0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16\ \ \textcircled{17}$$

e) This scenario is describing a binomial cdf. It is asking $P(4 \leq X \leq 7 \,|\, \pi = 0.52, n = 17) = 0.2538$.

$$0\ \ 1\ \ 2\ \ 3\ \ \boxed{4\ \ 5\ \ 6\ \ 7}\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16\ \ 17$$

f) This scenario is describing a binomial cdf. It is asking $P(X \leq 1 \,|\, \pi = 0.52, n = 17) = 7.401E - 05 = 0.000074$ which is approximately 0.

$$\boxed{0\ \ 1}\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16\ \ 17$$

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

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

$$0\ \ 1\ \ 2\ \ 3\ \ 4\ \boxed{5}$$

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

$$\boxed{0\ \ 1\ \ 2\ \ 3}\ 4\ \ 5$$

c) This is asking for $P(X \geq 4 \,|\pi = 0.6250,\ n = 5) = 0.3815$.

$$0\ \ 1\ \ 2\ \ 3\ \boxed{4\ \ 5}$$

19. Probability of a success (infant death) is $\dfrac{18}{97} = 0.1856$.

a) The expected value (mean of the distribution) is $\mu = n\pi = 40(0.1856) = 7.4240$. We would expect to see, on average, 7.4240 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. $P(X \leq 2|n = 40, \pi = 0.1856) = 0.0137$. Since the probability is small, we will conclude the efforts of NHTSA have been successful.

$$\boxed{0\ \ 1\ \ 2}\ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ \ldots\ 20\ \ 21\ \ 22\ \ 23\ \ldots\ 38\ \ 39\ \ 40$$

21. a) This is asking for a binomial cdf. $P(X \geq 10 \,|\, n = 18, \pi = 0.47) = 0.3110$.

$$0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \boxed{10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16\ \ 17\ \ 18}$$

b) This is asking for a binomial cdf. $P(X \leq 9 \,|\, n = 18, \pi = 0.47) = 0.6890$

$$\boxed{0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9}\ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16\ \ 17\ \ 18$$

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 $0.3110 + 0.6890 = 1.0000$, which can also be seen by combining the two diagrams.

$$\boxed{0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9}\boxed{10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16\ \ 17\ \ 18}$$

d) The expected value (mean of the distribution) is $\mu = n\pi = 18(0.47) = 8.46$. 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.

23. a) This is asking for a binomial cdf. $P(X \leq 5 \,|\, n = 25, \pi = 0.08) = 0.9877$.

$$\boxed{0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5}\ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ldots\ 22\ \ 23\ \ 24\ \ 25$$

b) This is asking for a binomial cdf. $P(4 \leq X \leq 7 \,|\, n = 25, \pi = 0.08) = 0.1346$.

$$0\ \ 1\ \ 2\ \ 3\ \boxed{4\ \ 5\ \ 6\ \ 7}\ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ldots\ 22\ \ 23\ \ 24\ \ 25$$

c) This is asking for a binomial cdf. $P(X \leq 5 \,|\, n = 25, \pi = 0.39) = 0.0367$

$$\boxed{0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5}\ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ldots\ 22\ \ 23\ \ 24\ \ 25$$

d) The expected value for the alcohol source is $\mu = n\pi = 25(0.08) = 2$. The expected value for the government source is $\mu = n\pi = 25(0.39) = 9.75$.

25. a) This is describing a binomial pdf scenario where, for red beans, the probability of success is 0.27. $P(X = 4 \,|\, n = 25, \pi = 0.27) = 0.0906$.

$$0\ \ 1\ \ 2\ \ 3\ \boxed{4}\ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ldots\ 22\ \ 23\ \ 24\ \ 25$$

b) This is describing a binomial cdf scenario where the probability of "not black beans" $= 1 - P(Black\,Beans) = 1 - 0.50 = 0.50$. $P(X \geq 13 \,|\, n = 25, \pi = 0.5000) = 0.5000$.

$$0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ \boxed{13\ 14\ \ldots\ 22\ 23\ 24\ 25}$$

c) This is describing a binomial cdf scenario where the probability of pinto beans $= 0.13$. $P(X \leq 4 \,|\, n = 25, \pi = 0.13) = 0.7817$.

$$\boxed{0\ 1\ 2\ 3\ 4}\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ \ldots\ 22\ 23\ 24\ 25$$

d) This is asking black or red beans, so the probability of success will be $P(Black)+P(Red) = 0.50+0.27 = 0.77$. This is describing a pdf scenario so we have $P(X = 20 \,|\, n = 25, \pi = 0.77) = 0.1836$.

$$0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ \ldots\ 19\ \boxed{20}\ 21\ 22\ 23\ 24\ 25$$

e) Black Beans: $\mu_{black} = 25(0.50) = 12.5$, Red Beans: $\mu_{red} = 25(0.27) = 6.75$, Pinto Beans: $\mu_{pinto} = 25(0.13) = 3.25$, Navy Beans: $\mu_{navy} = 25(0.10) = 2.5$. Each average represents the number of shipments, out of 25, that we expect to consist of that type bean.

Chapter 9 Solutions

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.

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

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.

7. a) $z = \dfrac{12.31 - 6.3}{3.17} = 1.8959$

   b) $z = \dfrac{8.2 - 6.3}{3.17} = 0.5994$

   c) $z = \dfrac{191 - 121}{14} = 5.0000$

   d) $z = \dfrac{0 - 121}{14} = -8.6429$

9. a $z - score$ is a value from the standard normal distribution. The value of the $z - score$ represents the number of standard deviations from the mean. As such, a $z - score$ of -3.4 is very unusual because it is 3.4 standard deviations below the mean. From the empirical rule we know any observation two or more standard deviations from the mean is unusual.

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

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.

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. The solution is 9.5855.

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. The solution is 4.5458.

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. The solution is 3.0145.

17. a) 0.1810

    b) 0.50

    c) 0.0929 + 0.1564 = 0.2493

    d) 0.5516

19. a) 0.3065

    b) 0.6776

    c) 0.1666

    d) 0

21. a) This is asking for $P(X > 55 \,|\, \mu = 50, \sigma = 10)$, which is the same as saying $P(X > 55)$ given that $X \sim N\left(50, 10^2\right)$. The solution is is 0.3085.

    b) This is asking for $P\left(47 < X < 59\right)$ which is the same as $P\left(47 \leq X \leq 59\right)$ because the normal distribution is continuous. The answer is 0.4339

23. a)The empirical rule states that approximately 95% of the data will fall within two standard deviations of the mean. The information provided suggests the mean is then mid point between 5.7 and 16, which is 10.85. $\dfrac{5.7 + 16}{2} = 10.85$. The mean of the distribution is 10.85 billion barrels.

    b) The standard deviation can also be estimated by $\dfrac{16 - 5.7}{4} = 2.5750$. The reason we divide by 4 is because there are 2 standard deviations each side of the mean.

    c) Since we now have the mean and standard deviation, we know $B \sim N\left(10.85, 2.5750^2\right)$ where these values are in billions. This problem is asking us to find $P(B \leq 14\,Billion)$ which is 0.8894.

    d) This problem is asking us to find $P(B \geq 14\,Billion)$ which is 0.1106

    e) The two events are compliments of each other, as such, the probabilities sum to 1.

    f) Based on the empirical rule, we expect to see 99.7% (or almost all) within 3 standard deviations. That gives us 5.7 - 2.575 = 3.1250 and 16 + 2.575 = 18.575. This calculation is reasonable because we already said that 5.7 was two standard deviations below the mean and 16 was two standard deviations above the mean, so we just moved out one more standard deviation. We would get the same value if we used $\mu \pm 3\sigma$ which would be $10.85 \pm 3(2.5750)$.

25. Based on the distribution, approximately 1.31% of the rats were able to run the maze in less than 28 minutes.

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 798.6445

29. a) If 70% deliver within 10 days, the assuming symmetry we know 35% deliver 10 days over and 35% deliver 10 days prior to the projected delivery date. This is very near the approximate 68% that we expect to find ± one standard deviation of the mean. That is telling us the standard deviation is approximately 10. This is not exact, just a real good estimate.

    We can get more exact by equating to the standard normal. If 35% deliver 10 days over then 35% + 50% = 85% deliver less than 290 days (280 + 10). That means 290 is P_{85}. We can quickly confirm P_{85} for the standard normal is 1.0364, so converting 290 to a z-score will result in 1.0364. This gives us:

    $$1.0364 = \frac{290 - 280}{\sigma}$$

    Solving for σ gives us 9.6487. So we see our estimate of 10 is actually pretty good.

    b) There is an argument for both ways we are about to present this answer. The first assumes ±37 days is approximately two standard deviations from the mean. Simple algebra tells us the standard deviation is then approximately 18.5 days. The second approach would to assume the ±37 days represents three standard deviations in each direction, in which case the estimate for the standard deviation would be 12.33 days.

    c) Use $Days \sim N\left(280,\, 10^2\right)$ and either $Days \sim N\left(280,\, 18.5^2\right)$ or $Days \sim N\left(280,\, 12.33^2\right)$ for the inverse normal to find the 95^{th} percentiles. If you used $\sigma = 18.5$ then you would find $P_{95} = 310.4298$. If you used $\sigma = 12.33$ then you would find $P_{95} = 300.2810$.

Chapter 10 Solutions

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

3. 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. Refer to question one regarding the sampling distribution.

5. 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.

7. a) 0.1056

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

   c) 6.452E-7 or approximately 0.0000.

   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. You can verify this by finding z-scores.

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

   b) The probability is 0.7335. This is based on a lower bound of 211, an upper bound of 215, $\mu = 213$. If you are using a TI-83/84 then you will enter $\frac{\sigma}{\sqrt{n}}$ as $\sqrt{\frac{81}{25}} = \frac{9}{5}$. If you are using TC-Stats you will enter $\sigma = 9$ and $n = 25$. If you didn't know $\sqrt{81} = 9$ then you could always enter sqrt(81) which TC-Stats recognizes as $\sqrt{81}$ and let TC-Stats calculate it for you.

   c) 0.0956

11. Based on what is given, we know: $Scores \sim N\left(28.6, 4.3^2\right)$

    a) $P(Score > 33) = 0.1531.$

    b) $\bar{Scores} \sim N\left(28.6, \frac{4.3^2}{75}\right) \quad P(\bar{x} > 30.2) = 0.0006$

    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.

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.

13. a) $\hat{P} \sim N\left(\frac{2}{3}, \frac{\frac{2}{3}\left(1-\frac{2}{3}\right)}{200}\right)$

b) 0.0107

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

15. For the CLT to apply we need $n\pi$ and $n(1-\pi)$ to be both at least 5. In this instance we have: $18(.568) = 10.2240$ and $18(1-0.568) = 7.7760$. The other assumption is that of $n \geq 20$. Here we have $n = 18$, so technically the assumptions are violated; however, the n of at least 20 is not nearly as important as the other two assumptions, which have been reasonably satisfied. So the CLT can be reasonably applied here.

b) $\hat{P} \sim N\left(0.568, \frac{0.568(1-0.568)}{18}\right)$.

c) 0.6595

17. 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.

19. a) $\overline{Length} \sim N\left(19, \frac{2.9^2}{43}\right)$

b) (17.6733, 20.3267). This was calculated based on the Empirical Rule telling us we expect approximately 99.7% or "all" of the data to fall within 3 standard deviations of the mean where $\sigma_{\bar{x}} = \frac{2.9}{\sqrt{43}}$.

c) (10.30, 27.70). This was calculated based on the Empirical Rule telling us we expect approximately 99.7% or "all" of the data to fall within 3 standard deviations of the mean where $\sigma = 2.9$.

d) $P_{95} = 19.7274$

e) Yes, this is unusual. The probability is only 0.0105.

f) The top 10% would have average body lengths of 19.5668 or more.

g) (17.8050, 20.1950). The reason these changed is because the sampling distribution has also changed. It is now $\overline{Length} \sim N\left(19, \frac{2.9^2}{53}\right)$. The standard deviation of the sampling distribution is now much smaller.

Chapter 11 Solutions

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.

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.

5. (a) (-8.754, 29.374)

   (b) (-5.256, 25.876)

   (c) (-0.6968, 21.317)

   (d) (1.7842, 18.836)

   (e) (4.2813, 16.339)

   (f) (6.0471, 14.573)

   (g) Each interval gets increasingly smaller as the sample size is increased. This makes sense the standard deviation used to calculate the confidence interval is $\frac{\sigma}{\sqrt{n}}$ so the ratio gets smaller and smaller as the sample size gets larger making the confidence interval get smaller and smaller.

7. (a) I am 90% confident the true mean time to relieve a minor or moderate headache with this new pain medication is between 12.5525 and 13.0875 minutes.

   I am 95% confident the true mean time to relieve a minor or moderate headache with this new pain medication is between 12.4899 and 13.1501 minutes.

   (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 for the 90% confidence interval than the 95% confidence interval.

9. The assumptions **MUST** be addressed. The assumption is that of the sample means being distributed normally: $\bar{X} \sim N$. This is reasonable because we are told the parent distribution is normal.

   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. 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, but this is not an issue because the confidence interval is designed to capture the true mean somewhere in the interval.

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. (Note: A student should provide a rough sketch of the normal plot they produced.)

    (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.

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” (the word proof is being used lightly here – evidence is probably a much better word) 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.

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.

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 under the central limit theorem for proportions. The assumption of independence is met by reasonable sampling.

    There are actually two ways to go about the confidence interval, depending on the technology you have. If you are using TC-Stats, then you can simply enter the value 0.58 for $\hat{p}$. If you do, then you will have a CI of $(.4902, 0.6698)$.However, if we look closely at this we see the reported proportion is 58%. It is doubtful those conducting the study actually ended up with 58%. If we do a little algebra we see (.58)(116) = 67.28. 67.28 people did not report a specific outcome. Rather, the observed proportion was undoubtedly $\frac{67}{116} = 0.5776$, which was clearly rounded to 58% in the report. If you use $\hat{p} = \frac{67}{116}$ then 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% (or 49.02% and 66.98% – depending on which method you used).

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

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.

(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.

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.

    (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.

23. (a) To choose the correct point estimate we will use a normal plot to determine if the data is approximately normal. The normal plot is questionable. Because of this I will also look at a box-plot to help me make my decision. The box-plot is clearly skewed left so I will use the sample median for the best point estimate for the center.

    (b) $M = 362.5$.

    (c) Since the data is not normal and the CLT does not apply, we will build a confidence interval for θ, the population median. We were not able to get a 98% confidence interval, rather it turned out to be a 98.3% CI: (335, 367).

    (d) A sample size of at least 50 is needed if you used the sample standard deviation as the estimate for σ, 48 if you used the range/4 as an estimate for σ.

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 use the large sample case. 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. The location of the end points are the 7th and the 16th position. We are approximately 95% confident the true median birth mortality rate is between 15 and 66.

    If you used the small sample case, which TC-Stats does, then you would have found a 96.5% CI of (13,66).

    (d) Since this is an interval built around the median we are not able to get exactly a 95% confidence interval. That is due to the discrete nature of the method used for a confidence interval for θ. The confidence interval is only approximately 95%.

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.

29. 95% confidence interval for the mean based on the t-distribution: (23.220, 26.609).

    An approximate 95% confidence interval for the median: (22.060, 26.180) based on positions 15 and 28.

    The CI for the mean is most appropriate because the CLT applies (n = 44).

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

    (b) Based on the normal plot (Note: A student should produce a rough sketch of the normal plot they produced.) and small sample size, a CI for the median (θ) is most appropriate.

    We will report an approximate 95% confidence interval for the median as (6, 41) eights of an inch. In reality, the confidence level is 96.48%.

    (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.

Chapter 12 Solutions

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 (α) 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.

3. (a) The parameter of interest is the mean, μ, so we want to do a hypothesis test for μ.

   (b) 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.

   (c) We will be concerned with the following hypotheses.

   $$H_0 : \mu = 10 \quad H_A : \mu < 10$$

   A reasonable level of risk for this problem is 0.05 or 5%. This level is often left to the researcher.

   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 calculated value 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.

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 it was given to us. Since we don't know the population standard deviation, we will use a t-statistic. The assumption of normality was satisfied because we were told the distribution was normal in the problem. The hypotheses to be tested is as follows.

   $$H_0 : \mu = 36.2 \quad H_A : \mu > 36.2 \quad \alpha = 0.01$$

The test statistic is 4.8937 and the p-value is approximately zero. There is sufficient evidence to show jogger's maximal oxygen intake is greater than that of an average adult.

7. (a) The hypotheses we are concerned with is about a population proportion.

$$H_0 : \pi = 0.50 \quad H_A : \pi < 0.50 \quad \alpha = 0.05$$

We will use a one-sample proportion z-test statistic with z = -3.4731 and the p-value is 0.0003 so we will reject the null hypothesis. There is strong evidence to support the citizen's group claim.

(b) To allow for the use of the z-test, we must check the assumption of normality. The sample size is 642, which is greater than or equal to 20. There are 277 successes ($n\hat{p}$) and 365 failures ($n(1-\hat{p})$), which satisfy the normality assumption through the central limit theorem for proportions.

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:

$$H_0 : \mu = 0 \quad H_A : \mu > 0 \quad \alpha = 0.05$$

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 (a value for α was not given so 0.05 was assumed) 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.

11. (a) Assumption: the population is distributed normally ($X \sim N$). The normality assumption is not an issue because the central limit theorem applies (n=45).

(b) The hypotheses we are interested in concern the mean.

$$H_0 : \mu = 2 \quad H_A : \mu > 2 \quad \alpha = 0.05$$

The test statistic is a t which has a value of 3.577. The p-value is 0.0004. This is strong evidence to show the wells have a barium level higher than the EPA maximum contaminant level goal.

13. (a) The hypotheses we are interested in concern the mean.

$$H_0 : \mu = 505 \quad H_A : \mu > 505 \qquad \alpha = 0.05$$

The test statistic is a t with a value of -7.764. The p-value comes out to be 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.

In many ways, this was a silly test and should have never been actually run. The sample statistic is $\bar{x} = 451.3518$ which is less than the hypothesized mean of 505. It is simply not possible to show the mean is larger than 505 when the the value of the sample mean is less than 505 to begin with. This would immediately guarantee a p-value of at least 0.50. If anything, it appears they may be admitting with lower than the national average for verbal scores.

(b) The assumption is reasonably satisfied by the Central Limit Theorem since $n = 54$ which is clearly much greater than 30.

15. 12.15 (a) The hypotheses will involve the population proportion.

$$H_0 : \pi = 0.50 \quad H_A : \pi > 0.50 \qquad \alpha = 0.05$$

With $\alpha = 0.05$ 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.

(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) where the number of success is $n\hat{p}$ and the number of failures is $n(1 - \hat{p})$. These values are both greater than 5 and so satisfying the central limit theorem.

17. (a) The hypotheses we are interested in concern the mean.

$$H_0 : \mu = 138 \quad H_A : \mu \neq 138 \qquad \alpha = 0.02$$

We will do a t-test since the population standard deviation is unknown. The normality assumption is reasonably satisfied because n = 31 which is greater than 30.

The test statistic is 3.9003 and the p-value is approximately 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.

(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 reasonable 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.

19. (a) The hypotheses we are concerned with involve the population proportion.

$$H_0 : \pi = 0.48 \quad H_A : \pi \neq 0.48 \qquad \alpha = 0.05$$

With $alpha = 0.05$ we will fail to reject the null hypothesis because he p-value is 0.1092. There is no evidence that the true proportion of voters that consider themselves middle class has changed.

(b) The assumption needed for this test is that of the sample 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 ($n\hat{p}$)) is 728 and the number of failures ($n(1 - \hat{p})$ is 855.

21. If the after group has a higher level of aggression behavior, then 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:

$$H_O : \mu = 0 \quad H_A : \mu > 0 \qquad \alpha = 0.05$$

where the mean makes reference to the mean of the differences. We will then work only with the differences, which would now be in a new list.

The normal plot does not look normal, but could be considered to be in a gray area. Because of this, I will also use a box-plot. The box-plot clearly indicates the data is not reasonably normal. As such, the appropriate test will be a sign test for the median.

$$H_O : \theta = 0 \quad H_A : \theta > 0 \qquad \alpha = 0.05$$

The p-value from the sign test os 0.50 so we will clearly fail to reject the null hypothesis and conclude there is not enough evidence to suggest a higher level of aggressiveness after being exposed to higher levels of cadmium.

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.

$$H_O : \theta = 0 \quad H_A : \theta > 0 \qquad \alpha = 0.05$$

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

Chapter 13 Solutions

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 of the 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.

3. The assumptions are approximate normality in both populations 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.

5. Practical significance (also commonly referred to as Clinical 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 analysis 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? Statistical significance is really just a ”small” p-value.

7. We must assume that both populations are normally distributed since we are told so in the problem. The hypothesis we are interested in involve two independent means. The hypothesis statement is:

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

   We need to determine if we will pool the variances or not so we will use an F-test.

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

   resulted in a p-value of 0.0653. This suggests the variances for the two populations are similar which indicates **we should pool** the standard deviations when testing the means.

   Now that we know to pool the variances, we can complete the 2-sample t-test and when we do so we find a p-value of approximately zero. This indicates there is a very strong evidence to suggest the homes in the southern part of town have a higher value than the homes in the northern part of town.

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. Once we have the differences calculated (the order is important) we need to address the assumption of normality with a normal plot of the differences. The order that we subtract is important because will ultimately tell us what the alternative hypothesis will look like.In this case I used Before - After.

   The normal plot shows gross departures from normality, so we will complete a sign test. The hypothesis to be tested is:

$$H_0 : \theta_d = 0 \quad H_A : \theta_d > 0$$

   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.

11. Since we are interested in the variability, we will be concerned with the following hypotheses.

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

   The assumptions to be satisfied are independence between data sets and normality in each data set. The independence assumption is reasonable because the data came from two different shifts. There is no reason to suspect dependence between shifts. Since we have data, we will check the normality assumption with normal plots. The normal plots for Shift 2 and Shift 3 show no obvious deviations from normality, we will continue with the F-test. (Note: A student should draw normal plots to support their conclusion.)

   The F-test resulted in a p-value of 0.0004 therefor we conclude there is strong evidence suggesting the variance for shift 3 is greater than the variance for shift 2.

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

15. This is a test of variation so the hypotheses we are interested in will involve the population variance.

$$H_0 : \frac{\sigma_{80}^2}{\sigma_{70}^2} = 1 \quad H_A : \frac{\sigma_{80}^2}{\sigma_{70}^2} > 1$$

   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 resulted in a p-value of 0.2750 so we conclude there is not enough evidence to reject the null hypothesis. There is not enough evidence to suggest that the variation in the annual releases for the 1970?s is less than that of the 1980?s.

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%.

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

Assumptions:
Both sample proportions are distributed normally. This can be shown by: $n_F\hat{p}_F \geq 5\,,\;\; n_F(1-\hat{p}_F) \geq 5\,,\; n_F \geq 20$. Which are all true (students should do the math to show this is true). Likewise we need to show the same for males, $n_M\hat{p}_M \geq 5\,,\;\; n_M(1-\hat{p}_M) \geq 5\,,\; n_M \geq 20$ which is also true.

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

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

TI-83/84 users will note that 1 is in the interval showing lack of significance. TC-Stats users will note the p-value is 0.0709 and come to the same conclusion based on the following hypothesis statement:

$$H_0 : \lambda = 1 \quad H_A : \lambda > 1 \quad \alpha = 0.05$$

(b) Yes, the conclusion using both techniques agrees.

19. (a) The parameters we are interested in comparing are the means of each of the groups. The hypothesis statement is:

$$H_o : \mu_{R1} - \mu_{R2} = 0 \qquad H_A : \mu_{R1} - \mu_{R2} > 0 \qquad \alpha = 0.02$$

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. Before 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.

$$H_0 : \frac{\sigma^2_{R1}}{\sigma^2_{R2}} = 1 \quad H_A : \frac{\sigma^2_{R1}}{\sigma^2_{R2}} \neq 1$$

The p-value from the F-test is 0.1073 so we will conclude the variances are similar therefore pool them during 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. (b) The 98% confidence interval is (-4.721, 3.9490). We are 98% confident that the true difference in the means is between -4.721 and 3.9490. Zero is in the interval, so there is no real difference in the two means. This agrees with the hypothesis test.

21. (a) Since we are dealing with percentages, we will form hypotheses concerning the difference between the two population proportions. It is reasonable to assume the data is independent. Both sample sizes are greater than or equal to 20 and the number of successes and failures from each sample is greater than or equal to 5 (students need to show this is true). The assumptions for a two-sample z-test for proportions are reasonably satisified. The hypotheses to be tested are as follows.

$$H_o : \pi_G - \pi_B = 0 \qquad H_A : \pi_G - \pi_B > 0 \qquad \alpha = 0.05$$

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.

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

Sample Size ($n_1 = n_2$)	p-value
1,000	0.3273
10,000	0.0786
20,000	0.0227

(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.

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 find the differences by subtracting the 1968 data from the 1972 data.

$$H_0 : \mu_D = 0 \quad H_A : \mu_D > 0 \qquad \alpha = 0.05$$

Assumptions: The differences are distributed normally.

The normal plot of the differences indicates a gross violation of the assumptions so we will use a nonparametric approach (the sign test). The requires we write our hypothesis statement as:

$$H_0 : \theta_D = 0 \quad H_A : \theta_D > 0 \qquad \alpha = 0.05$$

The results are: n = 19, Above = 13, Equal = 4, Below = 2.

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.

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. We will still use the Wilcoxon Rank-Sum, but we can make our hypothesis statement about the means rather than being forced to switch to the medians.

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

    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.

27. (a) Since the children have the same mothers, we are approaching this as dependent data. The question being asked is if there is a difference, which indicates a two-tailed test. As such, it doesn't matter which direction we subtract.

    $$H_0 : \mu_D = 0 \quad H_A : \mu_D \neq 0 \qquad \alpha = 0.05$$

    The CLT applies since the sample size is 284. The one sample t-test performed on the differences resulted in a p-value of 1.447E05, clearly indicating a difference.

    (b) There are many possible confounding factors. Possibilities will not be listed here. Rather, that may be an excellent topic for a class discussion.

Chapter 14 Solutions

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 two-sample t-test.

3. The assumptions for ANOVA consist of the following.

   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 decide if a gross violation of the normality assumption has occurred.

   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 a visual inspection.

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 statistically different. If zero is not in the interval, then the means are said to be statistically different.

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 one or more of the assumptions for ANOVA are violated, then Kruskal-Wallis should be used.

9. Note: Students should show all graphs along with all hypothesis statements.

   (a) 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. If the two procedures result in a different conclusion then you MUST decide if the data is more reasonably normal, or not. Box-plots maybe useful for this.

The p-value for Kruskal-Wallis is 0.0148. Since the significance level is 1%, we would fail to reject the null hypothesis if Kruskal-Wallis is the procedure that is used.

If the data were normal we would need to address similar variance assumption using an F-test. Promotions 1 and 3 are the groups to be tested because they have the largest and smallest sample standard deviations. Based on a p-value of 0.4027 we will fail to reject the null hypothesis and conclude that all population variances are homogeneous (similar).

With the equal variances assumption satisfied, we will continue with the ANOVA. Based on the p-value of 0.0030 we would reject the null hypothesis and conclude there is a difference in the distribution centers. Since ANOVA and Kruskal-Wallis brought us to different conclusions, we must now use the procedure we had decided upon earlier. If Kruskal-Wallis were the procedure used, then you would fail to conclude that a difference exists between the three promotion types. It would then not be possible to determine which of the promotions are different. If ANOVA were the procedure used, then you would conclude that a difference does exist between the three promotion types. It would then be possible to determine which of the three promotion types are different based on Fisher's LSD which tells us Promotions 1 and 2 are not different from each other; however, Promotions 1 and 2 are both different from Promotion-3. Promotions 1 and 2 are both smaller than Promotion 3.

11. 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 normal plots of the March and April data suggest normality; however, the May data is questionable. A quick look at a box plot confirms the data is not normal so we will use the Kruskal-Wallis test.

$$H_0 : \theta_{March} = \theta_{April} = \theta_{May} \; H_A : \qquad \text{At least one } \neq$$

Based on a p-value of 0.0079 we conclude there is sufficient evidence to suggest the staffing needs are not the same during the first week of the three months examined.

13. 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.

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 \quad H_A : \text{ At least one } \neq$$

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 types is different from the others.

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.

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.

$$H_0 : \mu_{BK} = \mu_M = \mu_{CJ} = \mu_{JB} \qquad \text{At least one } \neq$$

The normal plots for the data from each restaurant appears reasonably normal with the exception of the Carls Jr. data; however, with only three data point that is not much of a surprise. This is an example of how tough it can be at times to complete a meaningful data analysis with small sample sizes. Three data points is really not much to work with.

Based on summary statistics, we will address the similar variance assumption for ANOVA based on the standard deviations of Carl's Jr. and McDonalds.

$$H_0 : \frac{\sigma^2_M}{\sigma^2_{CJ}} = 1 \quad H_0 : \frac{\sigma^2_M}{\sigma^2_{CJ}} \neq 1 \qquad \alpha = 0.05$$

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.

19. Clearly, we want to address the means if possible. All of the normal plots look reasonably normal. Students should produce a copy of the normal plot they produced.

$$H_0 : \mu_{22mm} = \mu_{19mm} = \mu_{16mm} \qquad \text{At least one } \neq$$

To address the similar variance we will use an F-test using the data for 19mm and 16mm since those represent the largest and smallest variances (based on the sample standard deviations).

$$H_0 : \frac{\sigma^2_{16mm}}{\sigma^2_{19mm}} = 1 \quad H_0 : \frac{\sigma^2_{16mm}}{\sigma^2_{19mm}} \neq 1 \qquad \alpha = 0.05$$

Based on a p-value of 0.4220 for the F-test, we will fail to reject and conclude the variances are similar. This means we can continue on with the ANOVA.

ANOVA Results: The p-value for the ANOVA is 0.0002 so we will reject the null hypothesis and conclude at least one of the means is different.

Based on Fisher's LSD, we can see 16mm and 19mm are not different from each other; however, both the 16mm and the 19mm are different from the 22m. Both the 16mm and the 19mm have a higher HIC meaning a higher likelihood of a head injury as compared to the 22mm.

Chapter 15 Solutions

1. This is a goodness of fit problem. 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. The smallest expected value is 49, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$\begin{array}{l} H_O : \pi_1 = 0.47,\, \pi_2 = 0.46, \pi_3 = 0.07 \\ H_A : \text{At least one } \neq \end{array}$$

3. This is a goodness of fit problem. 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. The expected values are all 1107.6, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$\begin{array}{l} H_O : \pi_1 = \pi_2 = \pi_3 = \pi_3 = \pi_5 = {}^1\!/_5 \\ H_A : \text{At least one} \neq \end{array}$$

5. This is a goodness of fit problem. 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. All of the expected values are 55.125, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$\begin{array}{l} H_O : \pi_1 = \pi_2 = \pi_3 = \pi_3 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = {}^1\!/_8 \\ H_A : \text{At least one } \neq \end{array}$$

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.

9. This is a goodness of fit problem. Based on the p-value of 0.5885 we will fail to reject the null hypothesis. There is not enough evidence to suggest the game is not fair. All of the expected values are 55.5, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$\begin{array}{l} H_O : \pi_1 = \pi_2 = \pi_3 = \pi_3 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_9 = \pi_{10} = 0.10 \\ H_A : \text{At least one } \neq \end{array}$$

11. This is a test of independence. Based on the p-value of 0.0004, we will reject the null hypothesis. There is sufficient evidence to suggest the hours worked and class status are dependent. The smallest expected value is 10, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$\begin{aligned} H_0 &: \text{ Hours worked and Class Status are independent.} \\ H_A &: \text{ Hours worked and Class Status are dependent} \end{aligned}$$

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.

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. All of the expected values are 100, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

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

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. The smallest expected value is 56.25, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

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

19. This is a test of independence. Based on the p-value of 6.483E-10 (approximately 0.0000), we will reject the null hypothesis. There is sufficient evidence to suggest that Ethnicity and Ideology are dependent. The smallest expected value is 7.2868, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$H_0 : \text{ Ethnicity and Ideology are Independent}$$

$$H_A : \text{ Ethnicity and Ideology are Dependent}$$

21. This is a test of independence. Based on the p-value of 0.5977, we will fail to reject the null hypothesis. There is not enough evidence to suggest the belief it is possible to someday eliminate racism and race are dependent. The smallest expected value is 10.2902, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$H_0 : \text{ Ethnicity and Belief are Independent}$$

$$H_A : \text{ Ethnicity and Belief are Dependent}$$

23. (a) This is a test of independence. Based on the p-value of 1.712E-10 (approximately 0.0000), we will reject the null hypothesis. There is enough evidence to suggest Ethnicity and belief that law enforcement tend to display racist tendencies are dependent. The smallest expected value is 17.2934, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$H_0: \text{ Ethnicity and belief that law enforcement tend to display racist tendencies are Independent}$$
$$H_A: \text{ Ethnicity and belief that law enforcement tend to display racist tendencies are Dependent}$$

    (b) Answers will vary.

25. (a) This is a test of independence. Based on the p-value of 2.309E-13 (approximately 0.0000), we will reject the null hypothesis. There is enough evidence to suggest Ideology and being in favor of Homeland Security using racial profiling is are dependent. The smallest expected value is 54.4474, which is clearly greater than 5, so the assumptions regarding the expected values are satisfied.

$$H_0: \text{ Ideology and being in favor of Homeland Security using racial profiling is are Independent}$$
$$H_A: \text{ Ideology and being in favor of Homeland Security using racial profiling is are Dependent}$$

    (b) Answers will vary.

27. This is clearly a goodness-of-fit test. We want to determine of the observed counts match what should have occurred if the sample did reasonably represent the population from which it was sampled.

$$H_0: \ \pi_A = 0.016 \ \ \pi_B = 0.066 \ \ \pi_H = 0.57 \ \ \pi_W = 0.292 \ \ \pi_{Mx} = 0.024 \ \ \pi_O = 0.032$$
$$H_A: \text{ At least one is } \neq$$

    There are a total of $n = 367$ observations. The smallest expected value is 5.8720 so all expected values are at least 5. The p-value is 0.7985 so we will fail to reject the null hypothesis meaning there is not enough evidence to suggest the observed counts differ statistically from the published distribution.

29. This is clearly a goodness-of-fit test. We want to determine of the observed counts match the old market share or if the market share has changed.

$$H_0: \ \pi_A = 0.42 \ \ \pi_B = 0.36 \ \ \pi_C = 0.22 \qquad H_A: \text{ At least one is } \neq$$

    There are a total of $n = 250$ observations. The smallest expected value is 55 so all expected values are at least 5. The p-value is 0.0171 so we will reject the null hypothesis meaning there is enough evidence to suggest the market share has changed. Further examination shows Company C has increased their market share whereas the others have both decreased their market share.

Chapter 16 Solutions

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.

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

5. The interval is impossible since correlation cannot take on a value less than -1.

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.

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

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 monotonic increasing (curved) behavior in the data. As such, Spearman's correlation is the more appropriate measure of association. The value of Spearman's correlation is 1.

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.

    c) The 95% confidence interval for ρ is (-0.9594, -0.6916). I am 95% confident that the true population correlation between the number of cigarettes smoked per day and the birth weight of children is between -0.9594 and -0.6916.

15. a) A scatter plot, with O. tridens on the x-axis and O. lowei on the y-axis, is clear that the trend is monotonic decreasing.

    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 $r_s = -0.4788$.

    c) If the appropriate correlation coefficient is significant (small enough p-value to reject a null hypothesis which states the correlation is zero) and the correlation is negative, then it is telling us as one population of one barnacle increases the other decreases which indicates they are competing for space.

d) The value for r_s is negative, which suggests they may be competing for space; however, with a p-value of 0.081, there is not enough evidence to suggest the observed value for r_s is statistically significant meaning we can not claim the barnacles compete for space.

$$H_0 : \rho_s = 0$$
$$H_A : \rho_s < 0$$

The 95% confidence interval for the value for ρ_s is $(-0.8517, 0.2159)$.

17. a) The scatter plot appears to be linear so Pearson's Correlation Coefficient would be the most appropriate measurement of association. The value is $r = 0.7896$.

    b) The regression equation is: $V\hat{elocity} = 0.3991 + 0.0014(Distance)$.

    c) The meaning of b_0 is simply that of the y-intercept since it does not make any sense to say a nebula is 0 distance from earth.

    d) b_0 is not statistically significant based on a p-value of 0.6296 or approximately 0.0000. The p-value for b_1 is 3.719E-06 or approximately 0.0000 indicating it is statistically significant.

    e) Yes. The residual plot looks basically like a blob indicating a reasonably constant variance, and not autocorrelated.

    f) It suggests the universe is expanding because the further away, the faster the velocity away from us.

19. a) $\hat{y} = 22.1627 + 0.3632(SPEND)$

    b) The value of b_0 is the y-intercept. It is simply the y-intercept in this case since $0.00 spent is not a data point. The value of b_1 is the slope of the regression line. It is saying for every additional million dollars in advertising, they can expect 0.3632(1,000) or 3632 additional impressions per week.

21. a) We will start the analysis with a look at the scatter plot between CIG and BLAD. The scatter plot appears to have an positive linear trend suggesting it is reasonable to continue the analysis by calculating Pearson's correlation coefficient.

    The value of r is 0.0.7036 and the p-value for the hypothesis test is approximately 0.0000 (reported as 4.982E-08). 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 Bladder Cancer.