Contents

Chapter 1 Solutions 3
Chapter 2 Solutions 5
Chapter 3 Solutions 7
Chapter 4 Solutions 17
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

   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.

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.


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.
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. 
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 Graph]

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>750.00 ≤ x &lt; 800.00</td>
<td>1</td>
<td>1.67%</td>
<td>1</td>
<td>1.67%</td>
</tr>
<tr>
<td>800.00 ≤ x &lt; 850.00</td>
<td>3</td>
<td>5.00%</td>
<td>4</td>
<td>6.67%</td>
</tr>
<tr>
<td>850.00 ≤ x &lt; 900.00</td>
<td>13</td>
<td>21.67%</td>
<td>17</td>
<td>28.33%</td>
</tr>
<tr>
<td>900.00 ≤ x &lt; 950.00</td>
<td>14</td>
<td>23.33%</td>
<td>31</td>
<td>51.67%</td>
</tr>
<tr>
<td>950.00 ≤ x &lt; 1000.00</td>
<td>19</td>
<td>31.67%</td>
<td>50</td>
<td>83.33%</td>
</tr>
<tr>
<td>1000.00 ≤ x &lt; 1050.00</td>
<td>8</td>
<td>13.33%</td>
<td>58</td>
<td>90.67%</td>
</tr>
<tr>
<td>1050.00 ≤ x &lt; 1100.00</td>
<td>1</td>
<td>1.67%</td>
<td>59</td>
<td>98.33%</td>
</tr>
<tr>
<td>1100.00 ≤ x ≤ 1150.00</td>
<td>1</td>
<td>1.67%</td>
<td>60</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

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 5th observation. \( P_{23} = 15 \)
c) \( n = 20 \) so the location is \( .85(20) = 17 \) so we will sort the data and take the average of the 17th and 18th observations. \( P_{85} = \frac{66 + 84}{2} = 75 \)

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

<table>
<thead>
<tr>
<th>Campaign</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>40.000</td>
<td>41.000</td>
<td>42.000</td>
<td>44.000</td>
<td>46.000</td>
</tr>
<tr>
<td>#2</td>
<td>40.000</td>
<td>41.000</td>
<td>43.500</td>
<td>45.000</td>
<td>46.000</td>
</tr>
<tr>
<td>#3</td>
<td>44.000</td>
<td>46.000</td>
<td>48.000</td>
<td>51.000</td>
<td>52.000</td>
</tr>
</tbody>
</table>

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 50th percentile. By definition, 50% are above and 50% are below the 50th 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

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Population SD</th>
<th>Sample SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEN</td>
<td>54</td>
<td>671.000</td>
<td>12.426</td>
<td>4.969</td>
<td>5.015</td>
</tr>
<tr>
<td>3SYL</td>
<td>54</td>
<td>784.000</td>
<td>14.519</td>
<td>10.730</td>
<td>10.831</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEN</td>
<td>4.000</td>
<td>9.000</td>
<td>11.500</td>
<td>16.000</td>
<td>25.000</td>
</tr>
<tr>
<td>3SYL</td>
<td>0.000</td>
<td>6.000</td>
<td>11.500</td>
<td>22.000</td>
<td>43.000</td>
</tr>
</tbody>
</table>

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.
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.

![Graph of Jan Temp and July Temp]

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

29. a)

<table>
<thead>
<tr>
<th>Technology</th>
<th>Min</th>
<th>$Q_1$</th>
<th>Median</th>
<th>$Q_3$</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI-83/84</td>
<td>790.73</td>
<td>897.48</td>
<td>943.685</td>
<td>982.29</td>
<td>1113.16</td>
</tr>
<tr>
<td>TC-Stats</td>
<td>790.73</td>
<td>895.70</td>
<td>943.685</td>
<td>984.12</td>
<td>1113.16</td>
</tr>
</tbody>
</table>

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 $P_{20} = \frac{887.47 + 891.71}{2} = 889.59$.

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 $P_{80} = \frac{991.29 + 994.65}{2} = 992.97$. 
f) Both are the same position in from the ends. The are between the 12th and 13th 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 55th 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.