I.A. Levels of Data, Types of Samples

I.A.1. Categorize the following data according to level: nominal, ordinal, interval, or ratio.

(a) Time of first class.
(b) Length of time to complete an exam.
(c) Course evaluation scale: poor, acceptable, good.
(d) The length of time it takes for someone to run a marathon.
(e) The time of day the marathon starts.
(f) The name of the city in which the marathon is run.

**Answer.**  
(a) Interval—differences in time are meaningful, but ratios are not. For example, a first class at 3:00pm is not 1.5 times later than a first class at 2:00pm.
(b) Ratio—differences in time make sense as do ratios. For example, if Student A takes 50 minutes and Student B takes 100 minutes, it makes sense to say that Student B took twice as long as Student A.
(c) Ordinal—the categories can be ranked, but differences between ranks do not make sense.
(d) Ratio—similar to (b).
(e) Interval—similar to (a)
(f) Nominal—categories cannot even be ranked.

I.A.2. To estimate the average GPA of all La Sierra Students, President Wisbey computed the average GPA obtained in his Advanced Christian Theology class.

(a) Identify the variable?
(b) What is the implied population?
(c) What is the sample?
(d) What type of sample was this?

**Answer.**  
(a) GPA
(b) The GPA’s of all La Sierra University Students.
(c) The GPA’s of all students in Dr. Geraty’s Advanced Hebrew Grammar class.
(d) This is a sample of convenience.

I.A.3. (a) If your instructor were to compute the class mean of this test when it is graded, and use it to estimate the average for all tests taken by this class this quarter, would this be an example of descriptive or inferential statistics? Explain.

(b) A study on attitudes about smoking is conducted at a college. The students are divided by class, and then a random sample is selected from each class. What type of sampling technique is this (e.g,
simple random, convenient, stratified, systematic, cluster)? Explain why this type of sample is not a simple random sample.

(c) A politician wishes to determine the reading level of 5th graders in her State. She does not have funding to test all 5th graders in her state, so she randomly selects some of the schools in her state and tests every 5th grader in those schools. What type of a sample is this?

**Answer.** (a) Inferential: using a sample mean (one test) to estimate the population (all tests, quizzes and assignments) mean.

(b) Stratified: a random sample is taken from each class (strata). This cannot be a simple random sample because it requires elements from each class. A simple random sample need not have representation from each class.

(c) Cluster: the population is divided into groups, some of the groups are selected randomly and everyone in those groups is tested.

I.A.4. To compute the average amount of medical insurance its patients had in 2006, a hospital considered taking the following types of samples. For each case classify the sample as simple random, stratified, systematic, cluster, convenient.

(a) The hospital numbered its complete list of patients from 2006 and used a random number generator to select 100 patients from the list from which to collect the data.

(b) The hospital decided to collect the data from the first 50 patients admitted on July 4, 2006.

(c) The hospital randomly chose a patient, and collected data from that patient and every 200th patient admitted thereafter.

**Answer.** (a) Simple random; (b) convenient; (c) systematic.

I.A.5. Explain how you could use the table of random numbers in your text to help design a true false test of 10 questions so that the pattern of answers is random.

**Answer.** Randomly select a starting spot in the table. If the digit is odd, make a question with a false answer, if the digit is even make a question with a true answer. Proceed along the row for 10 such digits. For example, if the starting point had been the beginning of the 3rd row, the digits are: 59654 71966 which leads to answers of F F T F T F F F T T.

I.A.6. (a) Describe the difference between *population data* and *sample data*.

(b) Describe the difference between *inferential* and *descriptive* statistics.

(c) Describe the difference between *quantitative* and *qualitative* variables.

**Answer.** See text.

I.A.7. A medical school is investigating new eye drops as a treatment for glaucoma. Out of 63 volunteers, 35 will get the new eye drops. The others will get the currently used (not new) eye drops. The new eye drops come in a grey plastic bottle and the old eye drops come in a red plastic bottle. Neither the patient nor the doctor knows which color contains which eye drops. After six months, eye pressure on each patient is measured and then a sealed report revealing medication is opened.
(a) Explain how you would make this a randomized two treatment experiment.

(b) Is this a double-blind experiment? Explain.

**Answer.** (a) Assign numbers to each of the 63 volunteers, use a random number generator to randomly choose 35 distinct numbers from 1 to 63. Volunteers with those 35 numbers will get the new eye drops, the other volunteers will get the old eye drops.

(b) This is a double blind experiment, because neither the doctor nor the patient knows which are the new eye drops.

I.A.8. (Focus Problem: Fireflies) Suppose you are conducting a study to compare firefly populations exposed to normal daylight/darkness conditions with firefly populations exposed to continuous light (24 hours a day). You set up two firefly colonies in a laboratory environment. The two colonies are identical except that one colony is exposed to normal daylight/darkness conditions and the other is exposed to continuous light.

Each colony is populated with the same number of mature fireflies. After 72 hours, you count the number of living fireflies in each colony.

(a) Is this an experiment or observation study? Explain.

(b) Is there a control group? Is there a treatment group? If so which is which?

(c) What is the variable in this study?

(d) What is the level of measurement of the variable?

**Answer.** (a) It is an experiment because a treatment (continuous light) is deliberately being imposed on one of the groups.

(b) The control group is the group one that is exposed to normal daylight/darkness conditions and the treatment group is the one that receives continuous light.

(c) The variable is the number of living fireflies at the end of 72 hours.

(d) The level of measurement is ratio, because both differences and ratios make sense for population numbers, e.g., one population is twice as large as the other, and e.g. one population is 2000 more than the other.

I.B. Organizing and Presenting Data

I.B.1. Consider the data (which are systolic blood pressures of 50 subjects):

126,126,128,128,128,130,130,130,130,130,132,132,134,134,136,136,138,140,140,146,  
148,152,152,152,156,160,190,200,208,208

(a) What class width should be chosen if you would like to have 6 classes.

(b) Suppose you dont want a class width of 19, but would like a class width of 15 irrespective of how many classes that would give you. Complete the following table for this data given the first class has limits 100 – 114.

(c) Draw a frequency histogram using the table in (b).
(d) Draw an Ogive using the table in (b).

(e) Draw a relative frequency histogram for the table in (b).

(f) Describe in words how to construct each of the above mentioned types of graphs.

**Answer.** (a) \((208 - 100)/6 = 18\). Go to the next higher whole number to ensure that all of the data is covered. Thus a class width of 19 would be suitable.

(b)

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Lower Boundary</th>
<th>Upper Boundary</th>
<th>Midpoint</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>114</td>
<td>99.5</td>
<td>114.5</td>
<td>107</td>
<td>10</td>
<td>10</td>
<td>.20</td>
</tr>
<tr>
<td>115</td>
<td>129</td>
<td>114.5</td>
<td>129.5</td>
<td>122</td>
<td>15</td>
<td>25</td>
<td>.30</td>
</tr>
<tr>
<td>130</td>
<td>144</td>
<td>129.5</td>
<td>144.5</td>
<td>137</td>
<td>14</td>
<td>39</td>
<td>.28</td>
</tr>
<tr>
<td>145</td>
<td>159</td>
<td>144.5</td>
<td>159.5</td>
<td>152</td>
<td>6</td>
<td>45</td>
<td>.12</td>
</tr>
<tr>
<td>160</td>
<td>174</td>
<td>159.5</td>
<td>174.5</td>
<td>167</td>
<td>1</td>
<td>46</td>
<td>.02</td>
</tr>
<tr>
<td>175</td>
<td>189</td>
<td>174.5</td>
<td>189.5</td>
<td>182</td>
<td>0</td>
<td>46</td>
<td>.00</td>
</tr>
<tr>
<td>190</td>
<td>204</td>
<td>189.5</td>
<td>204.5</td>
<td>197</td>
<td>2</td>
<td>48</td>
<td>.04</td>
</tr>
<tr>
<td>205</td>
<td>219</td>
<td>204.5</td>
<td>219.5</td>
<td>212</td>
<td>2</td>
<td>50</td>
<td>.04</td>
</tr>
</tbody>
</table>

(c) — (f): For (c) — (e), answers can be found in the following [worksheet](#). For (f), see your text for the precise descriptions.

**I.B.2.** Consider the collection of 30 data

| 2 3 4 4 5 6 8 9 11 12 14 15 18 19 21 |
| 23 23 25 27 34 40 45 51 56 58 62 73 83 91 98 |

Complete the following table given that the first class has limits 1 – 20.

**Answer.**

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Lower Boundary</th>
<th>Upper Boundary</th>
<th>Midpoint</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>.5</td>
<td>20.5</td>
<td>10.5</td>
<td>14</td>
<td>14</td>
<td>14/30 ≈ .467</td>
</tr>
<tr>
<td>21</td>
<td>40</td>
<td>20.5</td>
<td>40.5</td>
<td>30.5</td>
<td>7</td>
<td>21</td>
<td>7/30 ≈ .233</td>
</tr>
<tr>
<td>41</td>
<td>60</td>
<td>40.5</td>
<td>60.5</td>
<td>50.5</td>
<td>4</td>
<td>25</td>
<td>4/30 ≈ .133</td>
</tr>
<tr>
<td>61</td>
<td>80</td>
<td>60.5</td>
<td>80.5</td>
<td>70.5</td>
<td>2</td>
<td>27</td>
<td>2/30 ≈ .067</td>
</tr>
<tr>
<td>81</td>
<td>100</td>
<td>80.5</td>
<td>100.5</td>
<td>90.5</td>
<td>3</td>
<td>30</td>
<td>3/30 = .100</td>
</tr>
</tbody>
</table>
I.B.3. Consider the following table

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Lower Boundary</th>
<th>Upper Boundary</th>
<th>Midpoint</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>.5</td>
<td>15.5</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>.04</td>
</tr>
<tr>
<td>16</td>
<td>30</td>
<td>15.5</td>
<td>30.5</td>
<td>23</td>
<td>5</td>
<td>7</td>
<td>.10</td>
</tr>
<tr>
<td>31</td>
<td>45</td>
<td>30.5</td>
<td>45.5</td>
<td>38</td>
<td>24</td>
<td>31</td>
<td>.48</td>
</tr>
<tr>
<td>46</td>
<td>60</td>
<td>45.5</td>
<td>60.5</td>
<td>53</td>
<td>13</td>
<td>44</td>
<td>.26</td>
</tr>
<tr>
<td>61</td>
<td>75</td>
<td>60.5</td>
<td>75.5</td>
<td>68</td>
<td>6</td>
<td>50</td>
<td>.12</td>
</tr>
</tbody>
</table>

(a) Construct a relative frequency histogram using the information given in the table above.
(b) Construct an ogive using the information given in the table above.
(c) Construct a frequency histogram using the information given in the table above.

Answer. For (a) and (b) see answers to Question 4, Test I, Winter 2007. For (c), the difference between a relative frequency histogram and a frequency histogram is that for the latter the heights of the bars are the relative frequencies of the class.

I.B.4. Make a stem and leaf display for the following data.

58  52  68  86  72  66  97  89  84  91  91  92  66  68  87  86
73  61  70  75  72  73  85  84  90  57  77  76  84  93  58  47

Answer.

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

I.C. Numerical Representations of Data

I.C.1. Consider the following sample consisting of 20 numbers.

20  23  24  24  25  26  28  29  31  32
33  34  35  37  41  46  52  53  58  98

(a) Given that $\sum x = 749$ and $\sum x^2 = 34109$ find the mean, variance and standard deviation for this sample.
(b) Find the $Q_1$, $Q_2$, $Q_3$ and the IQR for the data and then construct a box-and-whisker plot for the data.

Answer. (a) The mean is $\bar{x} = \frac{749}{20} = 37.45$. Because this is a sample, the variances is

$$s^2 = \frac{34109 - \frac{749^2}{20}}{19} \approx 318.892$$
and the standard deviation is \( s = \sqrt{s^2} \approx 17.858 \).

(b) The quartiles are \( Q_1 = 25.5 \), \( Q_2 = 32.5 \), and \( Q_3 = 43.5 \) and the inter quartile range is \( IQR = Q_3 - Q_1 = 18 \). For a sketch of the box plot see answers to Test 1, Winter 2007, Question 5.

**I.C.2.** Given the data 9,12,15,17,18,19,23,45,52,61,63,63. One has \( \sum x = 397 \) and \( \sum (x - \mu)^2 \approx 5206.917 \). Find:

(a) the mean; (b) the sample variance (c) the population variance
(d) the sample standard deviation (e) the median (f) the 50th percentile (g) the mode

**Answer.** (a) the mean is \( \frac{397}{12} = 33.08 \).

(b) the sample variance is \( s^2 = \frac{5205.917}{11} = 473.36 \).

(c) the population variance is \( \sigma^2 = \frac{5205.917}{12} = 433.90915 \).

(d) the sample standard deviation is \( s = \sqrt{473.36} = 21.76 \).

(e) the median is \( \frac{19 + 23}{2} = 21 \).

(f) the 50th percentile of the data is 21 (the same as the median).

(g) the mode is 63 the most common data.

**I.C.3.** The depth of ground water is given in the following grouped data table.

<table>
<thead>
<tr>
<th>Distance from ground to water level (ft), ( x )</th>
<th>15 – 19</th>
<th>20 – 24</th>
<th>25 – 29</th>
<th>30 – 34</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of wells, ( f )</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

(a) Estimate the mean depth of the ground water.

(b) Estimate the sample standard deviation for the depth of the ground water.

(c) Estimate the coefficient of variation for this data.

**Answer.** (a) \( \bar{x} \approx \frac{\sum xf}{n} \), where \( n = \sum f \) and so

\[
\bar{x} \approx \frac{(17)(3) + (22)(5) + (27)(8) + (32)(4)}{20} = \frac{505}{20} = 25.25.
\]

(b) \( s \approx \sqrt{\frac{\sum x^2 f - (\sum xf)^2/n}{n - 1}} \), and \( \sum x^2 f = (17^2)(3) + (22^2)(5) + (27^2)(8) + (32^2)(4) = 13215 \), and so

\[
s \approx \sqrt{\frac{13215 - \frac{505^2}{20}}{19}} \approx \sqrt{24.4079} \approx 4.940
\]

(c) \( C.V. = \frac{s}{\bar{x}} \cdot 100% \approx \frac{4.940}{25.25} \cdot 100% \approx 19.57% \)

**I.C.4.** (Short answer general) (a) Given a collection of data with lowest number 4 and highest number 100. What class width should be chosen if 6 classes are desired?
(b) Given a collection of ordered data with 65 numbers, in what position is the median?
(c) What percentile is the third quartile \( Q_3 \)?
(d) In a set of data, approximately what percentage of the data lie at or above the 63rd percentile?
(e) If you are among 8000 people that took a test and you scored at the 79th percentile, approximately how many people scored at your score or lower? Approximately how many people scored at your score or higher?
(f) If you are among 4000 students taking the MCAT, and you wish to score at least the 95th percentile, what is the maximum number of students that can score at least as well or better than you?

**Answer.**

(a) \((100 - 4)/6 = 16\), and we use the next higher whole number, so the class width should be 17.
(b) The median is in the 33rd position.
(c) The 75th percentile.
(d) 37% of the data.
(e) Approximately \(.79)(8000) = 6320\) scores were less than or equal to your score, and approximately 1680 were greater than or equal to your score.
(f) Not more than 5% of 4000, or 200 can score as well or better than you if you are to achieve the 95th percentile, or better.

I.C.5. A population is known to have a mean of 50 and a standard deviation of 6.
(a) Use Chebyshev’s theorem to find an interval that contains at least 75% of the data.
(b) Use Chebyshev’s theorem to find an interval that contains at least 24/25 of the data.
(c) At least what portion of data is contained in the interval from 26 to 74?

**Answer.**

(a) This is the data within two standard deviations of the mean, hence the interval \( 50 \pm 2(6), \) i.e. from 38 to 62.
(b) This is the data within 5 standard deviations of the mean, hence the interval from 20 to 80.
(c) The interval from 26 to 74 includes data that is within 4 standard deviations of the mean. According to Chebyshev’s theorem, this must include at least \( 1 - \frac{1}{16} = 15/16 \) of the data, that is, at least 93.75% of the data.

I.C.6. Consider the following data of 26 numbers.
8, 35, 47, 48, 51, 57, 60, 64, 64, 65, 66, 70, 72 76, 78, 80, 82, 84, 85, 89, 90, 90, 93, 94, 96, 111
(a) Find the median of the data.
(b) Find Q1, Q3 and the IQR. Construct a box and whisker plot for the data.
(c) Compute the interval \((Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR)\). Data outside of this interval are identified as suspected outliers. Are there any suspected outliers in the above data?

**Answer.**

(a) Because 26 is even, the median is the average of the 26/2 = 13th place and the 14th
place, therefore the median is \((72+76)/2 = 74\)

(b) The first quartile is the median of the 13 numbers below the median 74 of the entire set. Hence the first quartile is \(Q_1 = 60\) and the third quartile is the median of the 13 numbers above 74. Therefore the third quartile is \(Q_3 = 89\).

The IQR is the inter quartile range, \(IQR = Q_3 - Q_1 = 89 - 60 = 29\).

See section 3.4 in the text for construction of the box and whisker plot. Note the lower whisker will go down to 33, the bottom of the box will start at 60, the line in the box will be at 74, the top of the box will be at 89, the upper whisker will go to 97.

(c) The interval is \((16.6, 132.5)\), so 8 is a suspected outlier.

I.C.7. (General Questions on Means) (a) A student receives grades of A,A,B,C,C and is surprised to receive of GPA of 3.5 because the average grade of A,A,B,C,C is a B since

\[
\frac{4+4+3+2+2}{5} = \frac{15}{5} = 3.
\]

Explain how the GPA could be 3.5.

(b) Bob is pleased to learn from his boss that his annual salary is in the top 10 percent of all salaries in the company. A month later Bob learns that his salary is less than half of the mean salary in the company and suspects his boss lied to him. Explain how Bob’s boss could have told the truth.

(c) An army crosses a river whose average depth is 1 foot. Explain how several soldiers could have drowned crossing the river because they cannot swim.

(d) Ken had an average of 50% on exams and 90% on assignments in his class, so he computed that his average should be 70% (a C). Why was he shocked when he saw that his grade was a D?

Answer. (a) GPA’s are weighted averages. For example if the A’s were received in 5 unit classes and the B was received in a 4 unit class and the C’s were received in 1 unit classes, then the GPA is computed as follows.

\[
GPA = \frac{\sum xw}{\sum w} = \frac{4(5) + 4(5) + 3(4) + 2(1) + 2(1)}{5 + 5 + 4 + 1 + 1} = \frac{56}{16} = 3.5.
\]

(b) For example consider a company with 20 employees whose annual salaries in thousands are

\[
20 \ 20 \ 25 \ 25 \ 30 \ 30 \ 30 \ 30 \ 30 \ 40
40 \ 40 \ 40 \ 40 \ 45 \ 45 \ 45 \ 50 \ 60 \ 2000
\]

where Bob’s salary is 60K per year. The average annual salary is

\[
\frac{\sum x}{20} = \frac{2680}{20} = 134K
\]

which is more than double Bob’s salary. Notice the difference if you computed a 5% trimmed mean.

(c) (Hint) Think of a wide river that is 6 inches deep for the majority of the width and then has a deep but rather narrow trench. (An unlikely possibility is that the river is 1 foot deep all the way across and the soldiers that could not swim were not in good shape and ended up falling face first in the water because of tiredness.)
(d) (Hint) The exams must have a higher weighting. For example, what is his average if the exams are worth 80% of his overall grade, and the assignments are worth 20%? (An unlikely possibility is that Ken wasn’t paying attention, and in that class 70% is a D.)

I.C.8. (General Questions on Measures of Data) Consider a data set of 25 distinct measurements with mean $A$, median $B$, and range $R$.

(a) If the largest number is increased by 250, what is the new mean?
(b) If the largest number is increased by 250, what is the new median?
(c) If the largest number is increased by 250, what is the new range?
(d) If every number is increased by 10, what is the effect on the mean, median, range and standard deviation?
(e) In what position of the ordered data is the median?
(f) Explain why the data does not have a mode.
(g) Is it possible for the mean and median to be equal?

Answer. (a) the mean is increased by 10.
(b) The median remains the same.
(c) The range increases by 250.
(d) The mean and median increase by 10, then range and standard deviation remain the same.
(e) The median is in the 13th place.
(f) All numbers are distinct, so there is no most frequently occurring number.
(g) Yes, for example let the numbers be 1, 2, 3, . . . , 24, 25. Then the median and mean are both 13.

I.C.9. (Weighed Average) Suppose the weightings in a class are such that the tests are worth 60% of the grade, the assignments are worth 10% of the grade, the quizzes are worth 5% of the grade, and the final is worth 25% of the grade.

(a) Calculate the overall grade of a student that has 90% on assignments, 95% on quizzes, 75% on tests and 80% on the final exam.

(b) Is the answer in (a) higher or lower than if you had taken the average of 90%, 95%, 75% and 80%? Explain why in this example you would expect the weighted average to be lower.

(c) Suppose a student has 90% on assignments, 95% on quizzes, 75% on tests going into the final exam. What percentage is needed on the final exam for the student to finish with an overall grade of 82% in the class?

Answer. (a) $0.1(90) + 0.05(95) + 0.60(75) + 0.25(80) = 78.75\%$ (Note in this case the weights were used as 0.1, 0.05, 0.60 and 0.25 so they sum to 1).

(b) The weighted average is lower because the lower scores were in the categories that were weighted more heavily.

(c) Let $x$ be the needed percentage; then solve

$0.1(90) + 0.05(95) + 0.6(75) + 0.25x = 82$
to get $58.75 + .25x = 82$, and so $0.25x = 82 - 58.75 = 23.25$, or $x = 4(23.25) = 93$. Thus, 93% is needed on the final exam.

I.C.10. On-time percentages are given for two airlines in Phoenix, Los Angeles and Seattle for 2006.

<table>
<thead>
<tr>
<th>Crashcade Airlines</th>
<th>Los Angeles</th>
<th>Phoenix</th>
<th>Seattle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Fights</td>
<td>1000</td>
<td>500</td>
<td>3500</td>
</tr>
<tr>
<td>On time %</td>
<td>90</td>
<td>95</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pacific Worst Airlines</th>
<th>Los Angeles</th>
<th>Phoenix</th>
<th>Seattle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Fights</td>
<td>250</td>
<td>4500</td>
<td>250</td>
</tr>
<tr>
<td>On time %</td>
<td>85</td>
<td>90</td>
<td>80</td>
</tr>
</tbody>
</table>

(a) Calculate the on-time percentage average for these three cities for each airline. Do this as a weighted average where the weight for each airline and city is the number of flights.

(b) Given that the on-time percentage for Crashcade Airlines is 5% higher in each city, does the answer in (a) surprise you? Why or why not?

Answer. (a) For Crashcade we compute

$$\frac{\sum xw}{\sum w} = \frac{(1000)(90\%) + (500)(95\%) + (3500)(85\%)}{1000 + 500 + 3500} = \frac{435,000}{5000} = 87\%$$

For Pacific Worst we compute

$$\frac{\sum xw}{\sum w} = \frac{(250)(85\%) + (4500)(90\%) + (250)(80\%)}{250 + 4500 + 250} = \frac{446,250}{5000} = 89.25\%$$

You should get the same answer if you computed this as follows (we illustrate with Crashcade): Out of Los Angeles, 90% of 1000 flights = 900 flights were on time, out of Phoenix, 95% of 500 = 475 flights were on-time, out of Seattle 85% of 3500 = 2975 flights were on time. Therefore, Crashcade had a total of 900 + 475 + 2975 = 4350 out of 5000 flights on time, which is 87%.

(b) On the surface it is very surprising that Pacific Worst has a better overall on-time percentage. However, this happens because Pacific Worst’s schedule is heavily weighted to flights in Phoenix where they have there best on-time percentage, whereas Crashcade’s flights heavily weighted in Seattle where they have their worst on-time percentage.

I.C.11. Consider the following ordered data of 21 numbers

$$34 \ 36 \ 40 \ 46 \ 46 \ 48 \ 51 \ 54 \ 57 \ 58 \ 59 \ 60 \ 61 \ 62 \ 63 \ 64 \ 66 \ 70 \ 78 \ 85 \ 101$$

Find $Q_1$, $Q_2$, $Q_3$, the IQR and then construct a box and whisker plot for the data.

Answer. $Q_1$ is the median of the first 10 data (all data below the position of the median of the entire set, which is 11th position), so $Q_1 = 47$; $Q_2$ is the median, so $Q_2 = 59$ and $Q_3$ is the median of the last 10 data (all data above the position of the median of the entire set), so $Q_3 = 65$.

$$\text{IQR} = Q_3 - Q_1 = 65 - 47 = 18$$

The box plot then has the lower whisker starting at 34 and ending at the bottom edge of the box which is at 47, the line in the box is at the median, i.e. 59 and the upper edge of the box is at 65. The upper whisker starts there and ends at 101. See text for further examples.