Unit 1 Assignment Solutions

1. (a) categorical and nominal
   (b) categorical and ordinal
   (c) quantitative
   (d) categorical and ordinal
   (e) categorical and nominal
   (f) quantitative
   (g) categorical and ordinal

2. (a) The minimum is 0.178 and the maximum is 0.340.

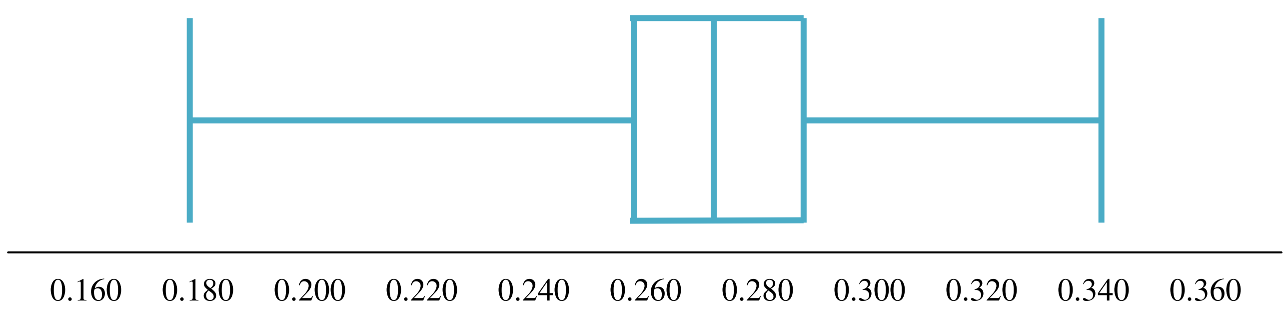
   The median is in position (n + 1)/2 = (34 + 1)/2 = 17.5, so the median is the average of the 17^{th} and 18^{th} ordered observations. Therefore, the median is equal to (0.271 + 0.275)/2 = 0.273.

   The first quartile is the median of all data values less than the median, and so Q1 is in position (17 + 1)/2 = 9. Therefore, Q1 = 0.256.

   The third quartile is the median of all data values greater than the median, and so Q3 is in position (17 + 1)/2 = 9 above the median (or equivalently, the ninth position counting down from the maximum). Therefore, Q3 = 0.288.

   The five number summary is therefore 0.178 0.256 0.273 0.288 0.340.

   (b) The quantile boxplot is shown below:



   The distribution appears to be skewed to the left.

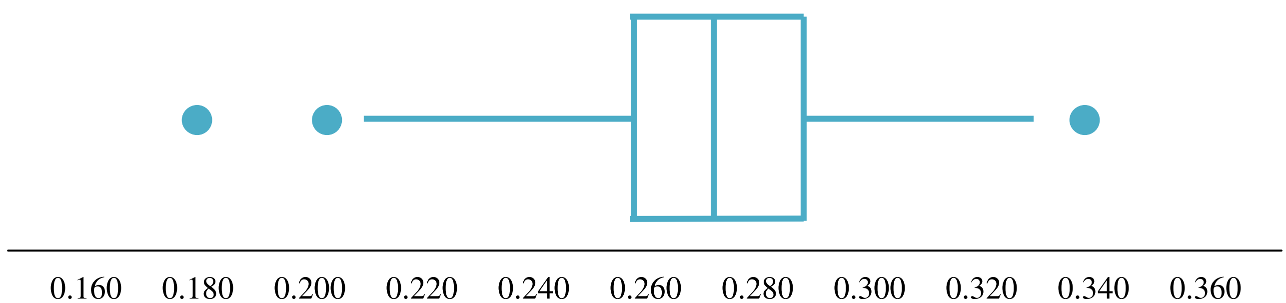
(c) We calculate the values of the lower and upper fences:

LF = Q1 – 1.5(IQR) = Q1 – 1.5(Q3 – Q1) = 0.256 – 1.5(0.288 – 0.256) = 0.256 – 0.048 = 0.208.
UF = Q3 + 1.5(IQR) = Q3 + 1.5(Q3 – Q1) = 0.288 + 1.5(0.288 – 0.256) = 0.288 + 0.048 = 0.336.

There are two data values less than LF (0.178 and 0.202) and so they are labelled as outliers. The whisker on the left now extends to the next lowest value, which is 0.210 (the "new" minimum).

There is one data value greater than UF (0.340) and so it is labelled as an outlier. The whisker on the right now extends to the next highest value, which is 0.333 (the "new" minimum).

The outlier boxplot is shown below:



When we remove the outliers from consideration, we see that the distribution is actually Approximately symmetric.

3. (a) The mean is equal to

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{9.7+3.1+10.5+5.6+14.9+6.7+8.3}{7} = \frac{58.8}{7} = 8.4 \text{ km.}$$

The variance is equal to

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$= \frac{(9.7-8.4)^2 + (3.1-8.4)^2 + (10.5-8.4)^2 + (5.6-8.4)^2 + (14.9-8.4)^2 + (6.7-8.4)^2 + (8.3-8.4)^2}{6}$$

$$= \frac{1.69+28.09+4.41+7.84+42.25+2.89+0.01}{6} = \frac{87.18}{6} = 14.53 \text{ km}^2$$

The standard deviation is therefore equal to $s = \sqrt{s^2} = \sqrt{14.53} = 3.812$ km.

(b) We know that y = a + bx, and so

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\left(\sum_{i=1}^{n} a + b\sum_{i=1}^{n} x_i\right) = \frac{na}{n} + b\frac{\sum_{i=1}^{n} x_i}{n} = a + b\overline{x}.$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(a + bx_i - (a + b\overline{x}))^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(bx_i - b\overline{x})^2}{n-1}}$$

$$= \sqrt{\frac{b^2\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = |b|\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = |b|\, s_x$$

(c) If we let X be the distance of the trip and we let Y be the fare, then

$$y = 3.30 + 1.25x.$$

So we use the results in (b), where a = 3.30 and b = 1.25 to find the mean and standard deviation of fares.

$$\bar{y} = a + b\bar{x} = 3.30 + 1.25(8.4) = \$13.80 \quad \text{and} \quad s_y = |b| s_x = 1.25(3.812) = \$4.77.$$

(d) If the mean remains unchanged, then the distance of the eighth trip must be equal to the mean, which is 8.4 km. This is shown below:

$$\bar{x}_{new} = 8.4 \Rightarrow \frac{\sum_{i=1}^{8} x_i}{8} = \frac{\sum_{i=1}^{7} x_i + x_8}{8} = \frac{58.8 + x_8}{8} = 8.4 \Rightarrow x_8 = 8(8.4) - 58.8 = 8.4 \text{ km}.$$

(e) The numerator in the standard deviation will not change because, for this eighth observation, $x_i - \bar{x} = 0$. However, the denominator will increase from 6 to 7, because the sample size has increased by one. Since the numerator stays the same and the denominator increases, the value of the standard deviation will decrease.

Another way to look at it: We can think of the sample variance (loosely speaking) as a sort of average squared deviation from the mean. If we add an observation which is exactly equal to the mean, the average squared deviation will decrease. Since the variance decreases, the standard deviation also decreases.

4. We know that $\overline{x}_f = 3.22$, $s_f = 0.62$, $\overline{x}_m = 3.22$, $s_m = 0.78$. The standard deviation for the whole class would be calculated as

$$s_c = \sqrt{\frac{\sum_{i=1}^{27}(x_i - \overline{x}_c)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{27}(x_i - 3.22)^2}{26}}$$

We know that the mean GPA of the whole class is $\overline{x}_c = 3.22$, since this is the mean for both the males and females. Therefore,

$$\sum_{i=1}^{27}(x_i - 3.22)^2 = \sum_{\substack{i=1\\ \text{female}}}^{12}(x_i - 3.22)^2 + \sum_{\substack{i=1\\ \text{male}}}^{15}(x_i - 3.22)^2$$

We find these two sums as follows:

$$s_f = \sqrt{\frac{\sum_{i=1}^{12}(x_i - 3.22)^2}{11}} \quad \Rightarrow \quad \sum_{\substack{i=1\\ \text{female}}}^{12}(x_i - 3.22)^2 = 11 s_f^2 = 11(0.62)^2 = 4.2284$$

$$s_m = \sqrt{\frac{\sum_{i=1}^{15}(x_i - 3.22)^2}{14}} \quad \Rightarrow \quad \sum_{\substack{i=1\\ \text{male}}}^{15}(x_i - 3.22)^2 = 14 s_m^2 = 14(0.78)^2 = 8.5176$$

And so we have

$$s_c = \sqrt{\frac{\sum_{i=1}^{27}(x_i - 3.07)^2}{26}} = \sqrt{\frac{4.2284 + 8.5176}{26}} = \sqrt{\frac{12.7460}{26}} = \sqrt{0.4902} = 0.7001$$

Note that we were only able to do this because the means for the males and females were equal. If the sample means had been different, we could not have done this calculation without the original data values. The overall standard deviation is close to the average of the standard deviations of the males and females, but this is a coincidence in this example.

5. We know that $\bar{x} = 78$ and $s^2 = 47.2$. The student's grade in biology was $x_1 = 87$, her grade in history was $x_2 = 74$, and her grades in geography and chemistry were the same ($x_3 = x_4$) and her calculus grade x_5 was the lowest.

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{87+74+x_3+x_4+x_5}{5} = \frac{161+x_3+x_4+x_5}{5} = \frac{161+2x_3+x_5}{5} = 78$$

$$\Rightarrow 2x_3 + x_5 = 5(78) - 161 = 229 \;\Rightarrow\; x_5 = 229 - 2x_3.$$

Now we have

$$s^2 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{n-1} = \frac{(87-78)^2 + (74-78)^2 + (x_3-78)^2 + (x_4-78)^2 + (x_5-78)^2}{4}$$

$$= \frac{81+16+(x_3-78)^2 + (x_4-78)^2 + (x_5-78)^2}{4} = \frac{97+2(x_3-78)^2 + (x_5-78)^2}{4}$$

$$= \frac{97+2(x_3-78)^2 + (229-2x_3-78)^2}{4} = \frac{97+2(x_3-78)^2 + (151-2x_3)^2}{4}$$

$$= \frac{97+2x_3^2 - 312x_3 + 12168 + 22801 - 604x_3 + 4x_3^2}{4} = \frac{6x_3^2 - 916x_3 + 35066}{4} = 47.2$$

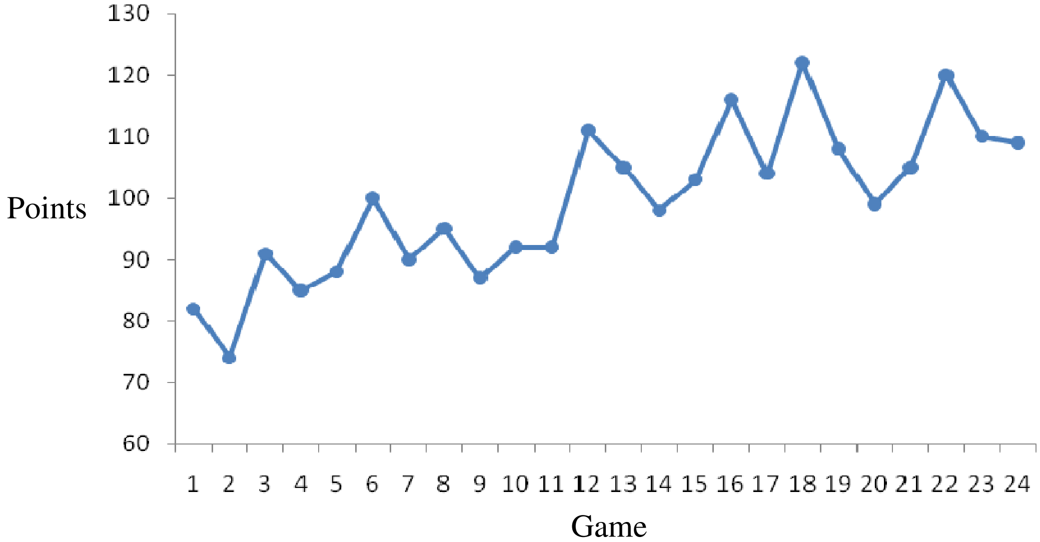
$$\Rightarrow 6x_3^2 - 916x_3 + 35066 = 4(47.2) = 188.8 \;\Rightarrow\; 6x_3^2 - 916x_3 + 34877.2 = 0$$

We use the quadratic formula with $a = 6$, $b = -916$ and $c = 34877.2$ to solve for x_3:

$$x_3 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-(-916) \pm \sqrt{(-916)^2 - 4(6)(34877.2)}}{2(6)} = \frac{916 \pm \sqrt{2003.2}}{12} = 72.6 \text{ or } 80$$

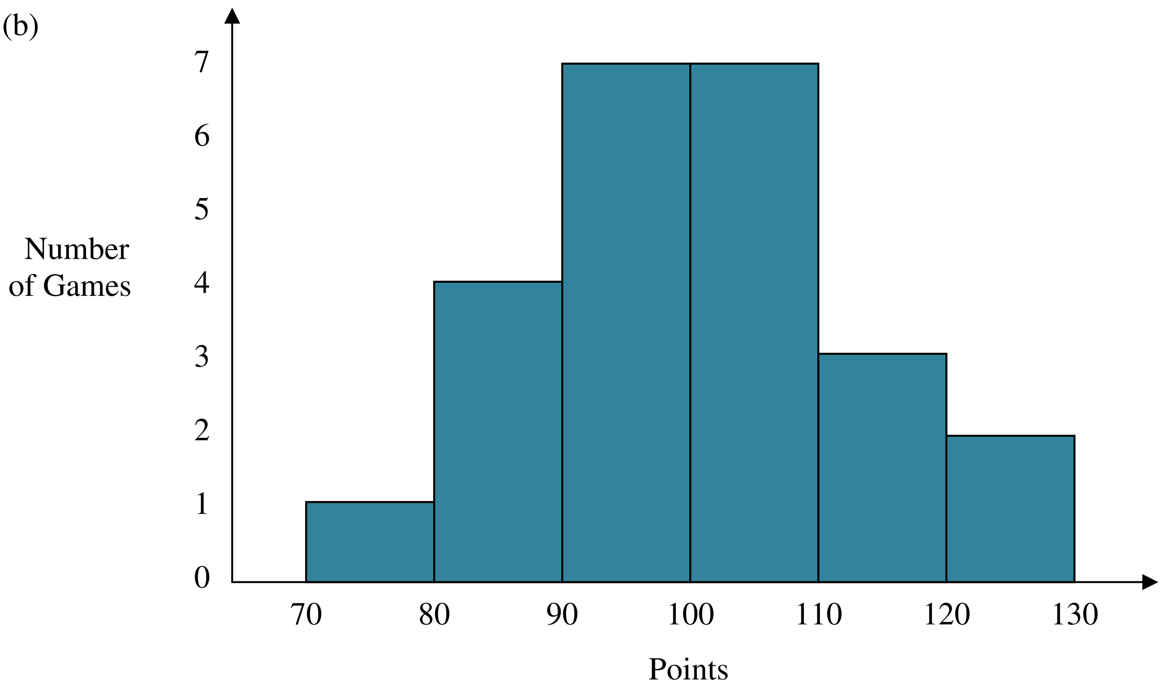
If x_3 was 72.6, then we would have $x_5 = 229 - 2(72.6) = 83.8$, but this can't be the student's calculus grade, because we know this grade is the lowest. Therefore, we must have that $x_3 = x_4 = 80$ and $x_5 = 229 - 2(80) = 69$.

6. (a)



Points
130
120
110
100
90
80
70
60
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Game

(b)



Number of Games
7
6
5
4
3
2
1
0
70 80 90 100 110 120 130
Points

(c)

7	4
8	2 5 7 8
9	0 1 2 2 5 8 9
10	0 3 4 5 5 8 9
11	0 1 6
12	0 2

(d) The timeplot displays the most information, as it displays both the data values and the order in which they were observed. The stemplot displays less information than a timeplot, as the data values are still given, but we don't know the time order. The histogram displays the least information, as we can see the number of observations in each interval, but we don't know the specific data values.

Another way to look at it – If we had a timeplot, we would be able to construct a stemplot, but the converse is not true. If we had a stemplot, we would be able to construct a histogram, but the converse is not true.