Assignment 1 Solutions

1. Gender is not a quantitative variable, so it must be categorical. Since there is *not* a natural ordering to the values of this variable, it is **categorical and nominal**.

2. Number of hours of sleep is a **quantitative** variable since the data is recorded as numbers and these numbers do not serve purely as labels.

3. Bedtime range is not a quantitative variable since the values are not obtained by a measurement, so it must be categorical. Since there is a natural ordering to the values of this variable, it is **categorical and ordinal**.

4. The population being studied here is all STAT 2000 students who had STAT 1000 final exam scores over 75%. From this population, a sample is selected, and the variable of interest is STAT 2000 midterm score, so this variable is displayed on the horizontal axis. The vertical axis of a histogram is always frequency.

5. There are six fares between \$5.00 and \$10.00 and two fares between \$0.00 and \$5.00. So there are eight fares less than \$10.00. There are 50 taxi rides in total, so the percentage of fares less than \$10.00 is $\frac{8}{50} \times 100\% = 16\%$.

6. When we look for a trend, we are look for a long-term pattern. Even though the team scored fewer points in game 6 than in game 5, there is still an upward trend overall, so the statement is TRUE.

7. The teams would have tied one game if the value on the vertical axis were the same for any given game. This does not happen, so the statement is FALSE.

8. For games 1-5, the team (represented by the red line) has scored more points than the opponent. For games 6-8, the team has scored fewer points than the opponent, so the statement is TRUE.

9. The total salary of the original seven employees is $7 * \$53,000 = \$371,000$. After removing the salary of the fired employee and adding the salaries of the newly hired employees, the total salary is $\$371,000 - \$71,000 + 3 * \$59,000 = \$477,000$. There are now nine employees in total, so the new mean annual salary is $\frac{\$477,000}{9} = \$53,000$.

10. The student's final percentage grade is calculated as a weighted average, using the weights given to different assessment items. We are given the desired weighted average (75%), we have all of the weights needed in the calculation, and we need to solve for the score on the missing assessment item (the final exam):

$$\overline{x}_W = \frac{x_1w_1 + x_2w_2 + x_3w_3 + x_4w_4}{w_1 + w_2 + w_3 + w_4}$$

$$75 = \frac{(72)(25) + (80)(15) + (94)(10) + (x_4)(50)}{25 + 15 + 10 + 50}$$

$$7500 = 1800 + 1200 + 940 + 50x_4$$

$$x_4 = 71.2$$

11. The distribution of the weights of the pumpkins is skewed to the left since there is a peak in the $700 - 800$ range, and there is a long tail extending toward low data values. The mean will get drawn in the direction of the long tail, but the median will not be affected, so the mean will be smaller than the median (i.e., the median will be larger than the mean).

12. We first need to find the median of this data set to help us find Q_1 and Q_3. The median is in position $(n + 1)/2 = 31/2 = 15.5$, i.e., the average of the 15^{th} and 16^{th} ordered values. Therefore, the median is $(7.0 + 7.2)/2 = 7.1$.

    The first quartile is the median of the first 15 ordered data values, so Q_1 is in position $(15 + 1)/2 = 8$. Therefore, $Q_1 = 6.0$.

    The third quartile is the median of the last 15 ordered data values, so $Q3$ is in position $(15 + 1)/2 = 8$ above the median, or, equivalently, the median is the 8^{th} largest data value. Therefore, $Q3 = 8.0$.

    The interquartile range is $Q_3 - Q_1 = 8.0 - 6.0 = 2.0$.

13. We first find the position of the median. The median is in position $(n + 1)/2 = (99 + 1)/2 = 50$ of the ordered data values.

    There are 49 data values before the median and 49 data values after the median. The third quartile is the median of the last 49 data values, so Q_3 is in position $(49+1)/2 = 25$ among the last 49 data values – that is, the 25^{th}. There is 1 data value ≥ 250, $1 + 1 = 2$ data values ≥ 225, $1+1+3 = 5$ data values ≥ 200, $1+1+3+5 = 10$ data values ≥ 175, $1 + 1 + 3 + 5 + 6 = 16$ data values ≥ 150, and $1 + 1 + 3 + 5 + 6 + 10 = 26$ data values ≥ 125. Therefore, the 25^{th} largest data value must lie in the $125 - 150$ interval.

14. The third quartile of fuel economies for Asian cars is approximately in line with the maximum fuel economy of European cars. So the statement is TRUE.

15. Variability of the distributions can be judged by using range as a measure of spread. The range of fuel economies for American cars is not the largest range; instead, the range of fuel economies for European cars is the largest. The statement is FALSE.

16. The distribution of fuel economies for European cars is skewed to the left since 50% of the data values between the median and the minimum cover a much larger range than the 50% of data values between the median and the maximum. The statement is FALSE.

17. We first need to find the median of this data set to help us find Q_1 and Q_3. The median is in position $(n+1)/2 = 55/2 = 27.5$, i.e., the average of the 27^{th} and 28^{th} ordered values. Therefore, the median is $(162+164)/2 = 163$.

    The first quartile is the median of the first 27 ordered data values, so Q_1 is in position $(27+1)/2 = 14$. Therefore, $Q_1 = 157$.

    The third quartile is the median of the last 27 ordered data values, so $Q3$ is in position $(27+1)/2 = 14$ above the median. Equivalently, we could count 14 positions down from the maximum. Therefore, $Q3 = 168$.

    Next, we calculate the lower and upper fences:

    $LF = Q_1 - 1.5IQR = Q_1 - 1.5(Q_3 - Q_1) = 157 - 1.5(168 - 157) = 157 - 16.5 = 140.5$

    $UF = Q_3 + 1.5IQR = Q_3 + 1.5(Q_3 - Q_1) = 168 + 1.5(168 - 157) = 168 + 16.5 = 184.5$

    Any data value less than 140.5 or greater than 184.5 will be labeled as an outlier. We see that there are 2 outliers on the left (137 and 140) and one outlier on the right (193).

18. The left whisker extends to the lowest data value which is not an outlier (i.e., 142). The right whisker extends to the highest data value which is not an outlier (i.e., 184).

19. The mean is $\dfrac{6+10+3+7+4}{5} = 6$. The deviations and the squared deviations are shown below:

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
6	6 - 6 = 0	0
10	10 - 6 = 4	16
3	3 - 6 = -3	9
7	7 - 6 = 1	1
4	4 - 6 = -2	4

    The sum of the last column is $\sum(x_i - \overline{x})^2 = 30$. So the standard deviation is $\sqrt{\dfrac{30}{n-1}} = \sqrt{\dfrac{30}{4}} = 2.74$.

20. Remember that standard deviation is a measure of spread around the mean. In Data Set A, there is one observation close to the mean and four observations far from the mean. That is, four out of the five observations are far from the mean. In Data Set B, two additional observations are added (in the 4-6 interval and the 12-14 interval) which are relatively closer to the mean than the original average distance from the mean. In addition, there is an additional observation close to the mean (in the 8-10 interval), further decreasing the average distance from the mean. So Data Set B has the smaller standard deviation.

21. The corrected data values involve subtracting 10 from one value and adding 10 to another value. Hence, the sum of the corrected data values will be the same as the sum of the original data values. Also, the number of observations is unchanged. So the mean will not change.

    The corrected data values are the smallest and largest data values. The median is the "middle" data value, so it will not change.

    Standard deviation is a measure of spread around the mean. The corrected data values will now be further away from the mean than they were originally, which will increase the standard deviation.