Stat 2507 **Assignment 1 Solution** **Summer 2016**

Due: Wednesday, Jun 1, 2016 in the class @ 6:05pm-7:30pm

You should pick up your marked assignment from the TA during your lab time
Assignment 1 Solution has 9 questions, for a total of 70 marks

The marking scheme is as follows:

Question:	1	2	3	4	5	6	7	8	9	Total
Marks	5	5	10	5	5	5	15	10	10	70
Score:										

Activate "Enable command" from Editor on the toolbars menu when the session window is active.

Question 1) ..... *5 marks*

Generate data of 200 numbers and store them in C2. Think about this data as the prices of a sample of 200 houses, in 1000 dollars.

```
MTB > set C1
DATA> 1:200
DATA> end
MTB > Base 2.
MTB > random 200 C11;
SUBC> normal 50 10.
MTB > let C2=ROUND(5*log(C1)+C11,0)
MTB > Stem-and-Leaf C2
MTB > Sort C2 C3;
SUBC> By C2.
MTB > describe C3
```

Draw a stem-and-leaf plot of these 200 prices and then use your plot to answer the following questions:

(a) (1 mark) The maximum price is **$ 99**.

(b) (1 mark) The minimum price is **$ 37**.

(c) (1 mark) The median price is **$ 71**.

(d) (1 mark) 36% of the prices are less than or equal to **$ 67**.

(e) (1 mark) What is the shape of the distribution of the prices? **bell-shape**.

Question 2) ..... *5 marks*

Use the "describe" command to answer parts (a), (b), and (c) of Question 1.

(a) (1 mark) (a) **99**

(b) (1 mark) (b) **37**

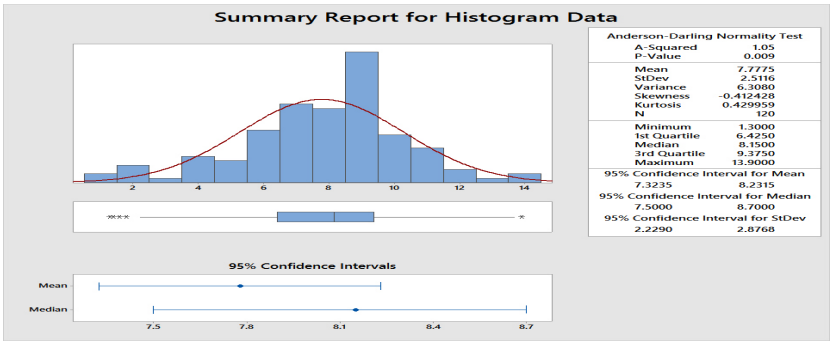
(c) (1 mark) (c) **71**

(d) (2 marks) The average price and the Std of the price of a house are **$71.145** and **$11.508**, respectively.

Question 3) ..... *10 marks*

Use the Excel Data posted on the Culearn.

(a) (1 mark) Construct a frequency histogram for "Histogram Data" in the data file by clicking on **Graph → histogram**



(b) (1 mark) Check the shape of this frequency histogram.

◯ Skewed to the left ◯ Skewed to the right √ **Symmetrical**

(c) (2 marks) What relation do you see between the mean and the median of this data set?

Solution:
Sine the shape of the data is symmetric then the mean and median are close to each other.

(d) (2 marks) Find the mean and median using the command *desc* for the column Histogram data, and interpret the result.

Solution:
The mean **(7.77)** is smaller than the median **(8.15)**. The mean is pulled towards the left while the median is not; hence, the shape is skewed to the left but according to the histogram it looks symmetrical because the mean and median are very close to each other.

(e) (2 marks) Would you use Chebychev's theorem or the empirical rule for this data?

Solution:
Empirical rule because the shape of the distribution is bell-shaped.

(f) (2 marks) For the class with a midpoint of 6, specify:

i. the class width?

Solution:
the class width is the difference between any two midpoints. Therefore the class width $= (5 - 4) = 1$

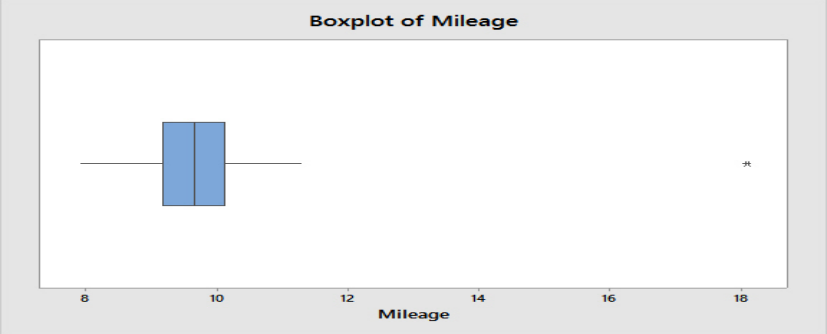
ii. the lower and upper class boundaries?

Solution:
The lower and upper class boundaries are $6 - \frac{1}{2} = 5.50$ and $6 + \frac{1}{2} = 6.50$ respectively.

Question 4) ..... *5 marks*

The data in the column "Mileage" are the Mileage in kilometers of 20 cars on one liter of gasoline.

(a) (1 mark) Construct a boxplot for the data in column Mileage.



(b) (1 mark) Do you see any outliers? **Yes** How many? **1 Outlier**

(c) (1 mark) Look at the boxplot and at the measurements. List the value(s) you think can be outliers. **The largest value 18.1**

(d) (1 mark) Compute the lower fence $= 9.175 - 1.5 * (10.125 - 9.175) = 7.75$
and the upper fence $= 10.125 + 1.5 * (10.125 - 9.175) = 11.55$

(e) (1 mark) Compare your candidates for outliers in part (c) to the lower and the upper fence.
$18.1 > 11.175$

Question 5) ..... *5 marks*

The data in columns "Brain" and "Body" are averages of brain weights and body weights of a number of mammal species.

(a) (1 mark) Compute the correlation coefficient between the Brain and the Body weights. Click on **Stat** $\rightarrow$ **Basic Statistics** $\rightarrow$ **Correlation** $\rightarrow$ **put both "Brain" and "Body"**

Pearson correlation of Body and Brain $= r = 0.934$

(b) (1 mark) What is the equation of the regression line when the Body weight is used to predict the Brain weight? **Stat** $\rightarrow$ **Regression** $\rightarrow$ **Fitted line plot** $\rightarrow$ **select "Brain" and "Body"** **Regression Equation Brain** $= -56.9 + 0.9029 Body$

(c) (1 mark) Use the above regression line to predict the Brain weight for Body weight=110.
Predicted Brain weight $= 0.9029 * 110 - 56.855 =$**42.464**

(d) (1 mark) What is the equation of regression line when the Brain weight is used to predict the Body weight.

Regression Equation Body $= 91.00 + 0.9665 Brain$

(e) (1 mark) Predict the Body weight of a mammal species with Brain weight=5.
Predicted Body weight $= 0.9665 * 5 + 91.0044 =$ **95.8369**

Question 6) ..... *5 marks*

Identify each of the following variables as categorical, discrete, or continuous. Use space provided.

(a) (1 mark) Blood type for a randomly selected person: **Categorical or Qualitative**

(b) (1 mark) Amount of snow (in inches) of the next snow storm in Ottawa: **Quantitative and Continuous**

(c) (1 mark) Daily exchange rate of Canadian dollar versus US dollar: **Quantitative and Continuous**

(d) (1 mark) The number of car accidents in the Ottawa area tomorrow: **Quantitative and Discrete**

(e) (1 mark) The brands of ice cream that you purchase: **Qualitative or Categorical**

Question 7) ..... *15 marks*

The numbers of rooms for n homes recently sold were;

$$8, 8, 8, 5, 9, 8, 7, 6, 6, 7, 7, 7, 7, 9, 35$$

(a) (3 marks) Compute p_i percentile where $i = 10, 65, 90$.

Solution:

First sort the data $5, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 35$.
The position of the percentile p is $\frac{p}{100}(n+1)$

$$\text{Position of } 10^{th} = 0.10(n+1) = 0.10(16) = 1.60 \Rightarrow p_{10} = 5 + 0.6(6-5) = 5.6$$
$$\text{Position of } 65^{th} = 0.65(n+1) = 0.65(16) = 10.4 \Rightarrow p_{65} = 8 + 0.4(8-8) = 8$$
$$\text{Position of } 90^{th} = 0.90(n+1) = 0.90(16) = 14.4 \Rightarrow p_{90} = 9 + 0.4(35-9) = 19.4$$

(b) (4 marks) Compute Q_1, Q_2, Q_3 and IQR.

Solution:

First sort the data $5, 6, 6, \textcircled{7}, 7, 7, 7, \textcircled{7}, 8, 8, 8, \textcircled{8}, 9, 9, 35$.
The position of the quartiles Q_1, Q_2, Q_3 are $\frac{25}{100}(n+1)$, $\frac{50}{100}(n+1)$, $\frac{75}{100}(n+1)$ respectively

$$\text{Position of } Q_1 = 0.25(n+1) = 0.25(16) = 4 \Rightarrow Q_1 = 7$$
$$\text{Position of } Q_2 = 0.50(n+1) = 0.50(16) = 8 \Rightarrow Q_2 = 7$$
$$\text{Position of } Q_3 = 0.75(n+1) = 0.75(16) = 12 \Rightarrow Q_3 = 8$$

$IQR = Q_3 - Q_1 = 8 - 7 = 1$

(c) (2 marks) Compute the mean, and the mode.

Solution:

$\bar{x} = \frac{\sum_{i=1}^{15} X_i}{n} = \frac{5+6+6+7+7+7+7+7+8+8+8+8+9+9+35}{15} = 9.13$ and $mode = 7$

(d) (3 marks) Compute the range, variance and standard deviation.

Solution:

$Range = max - min = 35 - 5 = 30$

$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{15} X_i^2 - n\bar{x}^2\right) = \frac{1}{14}\left(1985 - 15 \times 83.41778\right) = 52.41 \Rightarrow s = \sqrt{52.41} = 7.24$

(e) (2 marks) Is the largest number unusually big? why?

Solution:

Yes since $z - score = \frac{x-\bar{x}}{s} = \frac{35-9.13}{7.24} = 3.57 > 3$

(f) (1 mark) Plot the boxplot(by hand) and label all five summary statistics with the outliers.

Question 8) ..... *10 marks*

For n young patients, catheters were fed from a principal vein into the heart. The necessary catheter length and the patients, height are measured with the following results:

Patient	1	2	3	4	5	6	7	8	9	10
Height (in inches)	42.8	63.5	37.5	39.5	45.5	38.5	43	22.5	37	23.5
Catheter Length (centimeters)	37	50	34	36	43	28	37	20	34	30

(a) (5 marks) Calculate $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} y_i^2$, $\sum_{i=1}^{n} x_i y_i$, S_x, S_y, S_{xy} where x is the height and y is the length

Solution:

$\sum_{i=1}^{n} x_i = 42.8 + \cdots + 23.5 = 393.3 \Rightarrow \bar{x} = 39.33$ and $\sum_{i=1}^{n} x_i^2 = 16659.59$

$\sum_{i=1}^{n} y_i = 37 + \cdots + 30 = 349 \Rightarrow \bar{y} = 34.9$ and $\sum_{i=1}^{n} y_i^2 = 12779$

$\sum_{i=1}^{n} x_i y_i = 1583.6 + \cdots + 705.0 = 14494.1$

$S_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{9}(16659.59 - 10 \times 39.33^2) = 132.3446 \Rightarrow S_x = 11.50$

$S_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{9}(12779 - 10 \times 34.9^2) = 66.54444 \Rightarrow S_y = 8.16$

$S_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) = \frac{1}{9}(14494.1 - 0.1 \times 393.3 \times 349) = 85.32556$

(b) (2 marks) Use part (a) to obtain the linear regression equation of Catheter Length on Height.

Solution:

The regression line is $y = a + bx$

where

correlation coefficient $r = \frac{S_{xy}}{S_x S_y} = \frac{85.32556}{11.50 \times 8.16} = 0.9092$

$b = r\frac{S_y}{S_x} = 0.9092\frac{8.16}{11.5} = 0.6447$ and $a = \bar{y} - b\bar{x} = 34.9 - 0.6447 \times 39.33 = 9.5431$

Catheter Length = 9.5431 + 0.6447 × Height

(c) (2 marks) Use part (a) to obtain the linear regression equation of Height on Catheter Length.

Solution:

The regression line is $x = c + dy$

where

correlation coefficient $r = \frac{S_{xy}}{S_x S_y} = \frac{85.32556}{11.50 \times 8.16} = 0.9092$

$d = r\frac{S_x}{S_y} = 0.9092\frac{11.5}{8.16} = 1.282$ and $c = \bar{x} - d\bar{y} = 39.33 - 1.282 \times 34.9 = -5.420$

Height = -5.420 + 1.282×Catheter Length

(d) (½ mark) Estimate the required Catheter Length for a patient whose Height is 36 inches.
Catheter Length = 9.5431 + 0.6447 × 36=32.752

(e) (½ mark) Estimate the height of a patient who requires a Catheter with length 22 centimeters.
Height = -5.420 + 1.282×22=22.784

Question 9) ..... *10 marks*

Show that $S_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]$ is the same as $S_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right]$

Solution:

$$\begin{aligned}
S_{xy} &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i - \bar{x}\bar{y})\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n} \bar{x}y_i + \sum_{i=1}^{n} \bar{x}\bar{y}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y}\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\frac{\sum_{i=1}^{n} x_i}{n}\frac{\sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right]
\end{aligned}$$