Stat 2507 Assignment 1 Solution(Chapters 1-3) Winter 2019

Due: Monday, Feb 4, 2019 at the beginning of the class
Assignment 1 Solution has 11 questions, for a total of 100 marks

The marking scheme is as follows:

Question:	1	2	3	4	5	6	7	8	9	10	11	Total
Marks:	8	8	7	7	14	7	2	4	20	12	11	100
Score:												

Part I. Lab questions.

- Data used in this lab are in the Excel file on CuLearn of the course. You will need to copy the data from Excel and paste them into a Minitab worksheet (Open such a worksheet by double-clicking on Minitab).

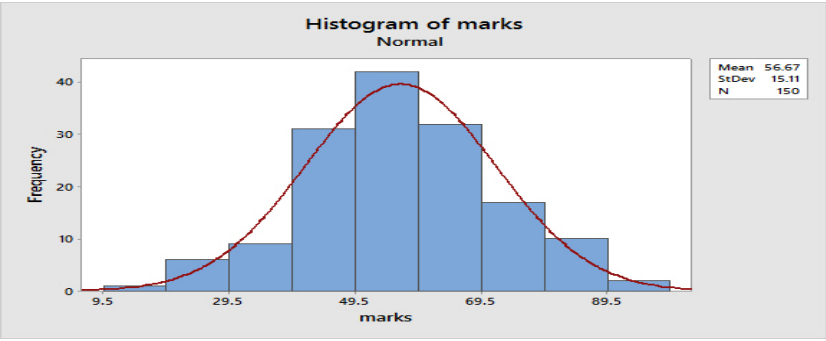
- Do not include ANY Minitab code to your assignment. Use spaces left to answer lab questions, and attach the printed graphs.

Q-1) ..... *8 marks*

1. The test marks of a sample of 150 students from a very large first year class are recorded in the column titled "marks" in the Excel file.

(**a**) (2 pts) Construct a frequency histogram for these data such that the first class interval is 9.5-19.5.

- *Enter the data in column C1*
- *Select Graph: Histogram. Enter C1 in the Graph variable window. Click OK to view the histogram*
- *Edit the horizontal axis scale. Double click on x-axis and under the Binning tab, select Interval Definition and Interval Type, choose Cutpoint and then enter the two endpoints 9.5 19.5 (with a blank space in between) for the first interval. This way Minitab will construct a histogram with the classes 9.5-19.5, 19.5-29.5, etc.*
- *Print your histogram and include it with your assignment*



(**b**) (2 pts) Describe the shape of distribution of this data set. **Bell and Symmetric**.

(**c**) (2 pts) What proportion of observations are higher than 49.5? $\frac{103}{150}$.

(**d**) (2 pts) The mean or the median, which one is greater?[1] **The same** why?[1]
Since the distribution is bell-shaped and symmetric

Q-2) ..... *8 marks*

(Refers to 'marks' data): Comparing Empirical rule to Tcbycheff theorem.

(**a**) (2 pts) Use *desc* command (Enable command editor "MTB >" by checking "Enable commands" from Editor in the bar menu)
to find the mean[1] ____**56.67**____ and the standard deviation[1] ____**15.11**____.

(**b**) (2 pts) Now, you will find out how many test marks fall between $\bar{x} \pm 2s$. You will use Minitab to construct a column C5 which will contain only values 1 or 0 according to whether the corresponding age (in column C1) falls in the interval $\bar{x} \pm 2s$, by typing in the following: *let c3=(c1>= $\bar{x} - 2 * s$ and c1<= $\bar{x} + 2 * s$).*
Note Before typing in you will replace $\bar{x}$ and s by their respective values found in part **a** above. Next you will check how many ages did fall in the interval $\bar{x} \pm 2s$ by typing in the following: *tally c3.*
The number ages did fall in the interval $\bar{x} \pm 2s$ is **143 students**

(**c**) (4 pts) What is the percentage of ages that fall between $\bar{x} \pm 2s$?[1] $\frac{143}{150} = 95.3\%$.
Is this value close to what the empirical rule suggests for the interval $\bar{x} \pm 2s$?[1]
The answer is Yes.

Remark 0.1:
The empirical rule says that ***if the distribution of a set of data is bell shaped and symmetrical*** *then approximately 95% of the measurements lie within $\bar{x} \pm 2s$.*

Does this value agree with Tchebysheff theorem?[1] ____**Yes**____ why?[1]
Tchebysheff theorem is applicable to all distributions.

Q-3) ..... *7 marks*

The data in the Excel file titled "beef" are the weights (in pounds) of ground beef packages in a supermarket.

(**a**) (2 pts) Construct a stem-and-leaf chart for this set of data (print your graph) to answer the following questions by:

- *Enter the data in column C2*
- *Select Graph: Stem-and-Leaf. Enter C2 in the Graph variable window. Click OK to view the graph*
- *Print your stem-and-leaf graph*

```
Stem-and-Leaf Display: beef
Stem-and-leaf of beef  N  = 27
Leaf Unit = 0.010
 1   7   5
 6   8   36999
 13  9   2366779
(3)  10  688
 11  11  2244788
 4   12  48
 2   13  8
 1   14
 1   15  2
```

(**b**) (2 pts) How many beef packages have the weight more than 1.17 lbs? **6 packages**

(**c**) (2 pts) How many beef packages have their weight exactly 1 lb or less? **13 packages**

(**d**) (1 pt) What is the median weight of the beef packages? ____**1.06 $**____

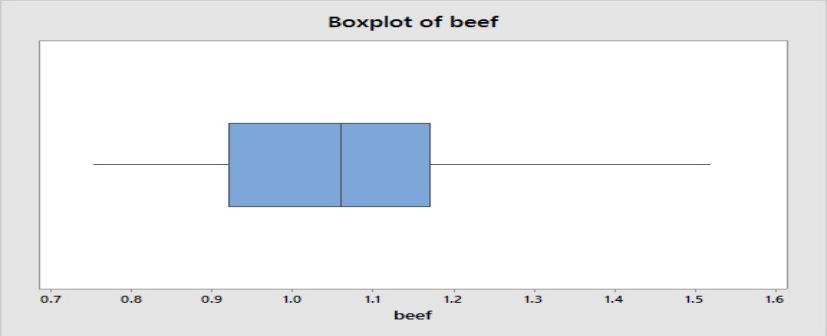
Q-4) ..... *7 marks*

Refers to "beef" data

(**a**) (4 pts) Construct a boxplot for this set of data(print it) to answer the following questions by:

- *Select Graph: Boxplots. In the Boxplots window Choose Simple Under "One Y". Click OK. Enter C2 in the variable window. Click OK to view the graph*
- *Print your boxplot*

.



(**b**) (1 pt) Based on the boxplot, how would you describe the shape of the distribution of this data set **fairly symmetrical**.

(**c**) (1 pt) The interquartile range (IQR) is approximately equal to

Approximately 1.17-0.92=0.25

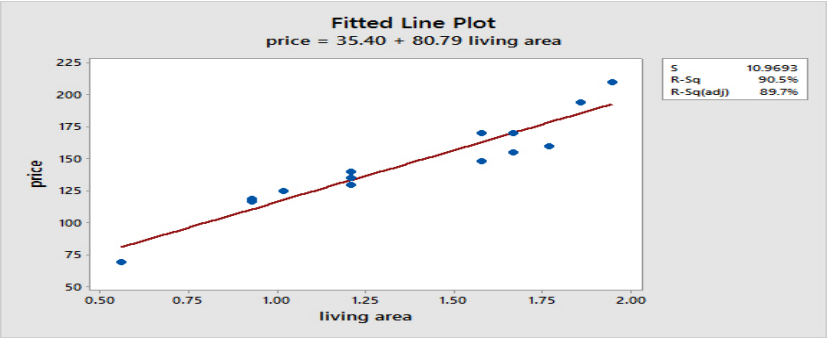
(**d**) (1 pt) Does the data set has any outliers. **No outliers**

Q-5) ..... *14 marks*

In the columns titled "price", "living area" and "bathrooms" in your excel file, are the data which relate condominium selling price (in thousands of $) to the living area (in hundreds of sq. meters) and number of bathrooms

(**a**) (3 pts) Construct a scatterplot with living area marked along horizontal axis and price along the vertical axis. Calculate the correlation coefficient: [1]

Correlation of Price and Living Area $= r = 0.951$.



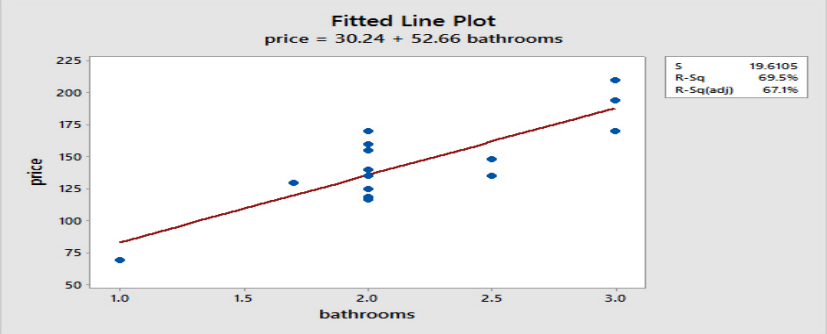
If appropriate, fit a least square regression line using living area to predict the response variable, price. What is the equation of regression line? [2]

• *Select Stat; Regression, then fitted line plot. Enter the response variable and the predictor variable. click OK.*

Regression Equation Price $= 35.40 + 80.79$**Living Area**

(**b**) (3 pts) b) Construct a scatterplot with bathrooms marked along horizontal axis and price along the vertical axis. Calculate the correlation coefficient: [1]

Correlation of Price and Bathrooms $= r = 0.834$.



If appropriate, fit a least square regression line using bathrooms to predict the response variable, price. What is the equation of regression line? [2]

Regression Equation Price $= 30.24 + 52.66 Batthroom$

(c) (8 pts) Predicted condominium price if the living area is 1.36? [2] **145.27**.
Predicted condominium price if the number of bathrooms is 1.5? [2] **109.23**.
Can you predict the number of Living Area from the price? [1] **No**. Why? [1]
Because the equation in (a) is meaningful only for predicting Price from Living area not the other way around.
Can you predict the number of bathrooms from the price? [1] **No**. Why? [1]
Because the equation in (b) is meaningful only for predicting Price from Bathrooms not the other way around.

Part II. 6 Long-answer questions; Give the solutions for the following questions in details.

Q-6) ..... *7 marks*

Identify each of the following variables as categorical (i.e. qualitative), or quantitative. If the variable is quantitative determine if it is discrete or continuous.

(**a**) (1 pt) amount of time it takes to assemble a puzzle **Continuous**

(**b**) (1 pt) province in which a person lives **Qualitative**

(**c**) (1 pt) number of students in STAT 2507 **Discrete**

(**d**) (1 pt) spring break destinations **Qualitative**

(**e**) (1 pt) mark out of 100 obtained on a statistic test **Continuous**

(**f**) (1 pt) letter grade obtained on a statistic test **Qualitative**

(**g**) (1 pt) number of snow days past winter **Discrete**

Q-7) ..... *2 marks*

Suppose a data set with 15 observations has an outlier present (i.e. the extreme observation very far away from the rest of the data).

(**a**) (1 pt) Would the mean of the data set be affected by this outlier? Explain:

Solution:
Yes, greatly affected

(**b**) (1 pt) Would the median of the data set be affected by this outlier? Explain:

Solution:
No, not affected

Q-8) ..... *4 marks*

A student took a Biology exam and scored 77. If the class exam scores were mound-shaped with a mean score of 70 and standard deviation of 16, use z-score to determine how the student placed comparing with the class average i.e. is the student's mark an outlier or not?

Solution:
The z-score for the measurement 77 is $z\text{-score} = \frac{77-70}{16} = 0.4375$. [2]
This value is not greater than 3 [1]. Meaning that 77 CAD is not unusual [1]

Q-9) ..... *20 marks*

The number of goals scored by Wayne Gretzky was recorded for seasons 1978-1999.
46, 51, 55, 92, 71, 87, 73, 52, 62, 40, 54, 40, 41, 31, 16, 38, 11, 23, 5, 23, 9

(**a**) (2 pts) What is the average number of goals scored by Wayne Gretzky?

Solution:

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^{21} X_i}{n} \quad [1] \\ &= 43.80952\, goals \quad [1] \end{aligned}$$

(**b**) (4 pts) What is the standard deviations of the number of goals?

Solution:

$$\text{Sample variance } = S^2 = \frac{\sum_{i=1}^{21} X_i^2 - \frac{\left(\sum_{i=1}^{21} X_i\right)^2}{n}}{n-1} \quad [1] \Rightarrow Std = \sqrt{606.5619} = 24.62848 \quad [1]$$
$$= \frac{52436 - \frac{920^2}{21}}{20} \quad [2]$$
$$= 606.5619$$

(**c**) (2 pts) Draw a stem-and-leaf plot (by hand).

Solution:

Unit of leaf= 1 [1]
0 | 59
1 | 16
2 | 33
3 | 18
4 | 0016
5 | 1245
6 | 2
7 | 13
8 | 7
9 | 2

The stem-and-leaf graph Worth's [1]

(**d**) (2 pts) Roughly, how would you describe the shape of the graph?

Solution:

Fairly symmetrical

(**e**) (6 pts) Calculate the 3 quartiles (Q_1, Q_2 = median, Q_3)?

Solution:

Location of Q_1: (21+1)*0.25=5.5 [1]
$Q_1 = X_{(5)} + 0.25(X_{(6)} - X_{(5)}) = 23 + 0.5(23 - 23) =$ **23 goals [1].**
Location of $Q_2 = median$: 0.5(21+1)=11 [1]
$median = X_{(11)} = 41$ **[1]**
Location of Q_3: (21+1)*0.75=16.5 [1]
$Q_3 = X_{(16)} + 0.5(X_{(17)} - X_{(16)}) = 55 + 0.5(62 - 55) =$ **58.5 goals [1].**

(**f**) (2 pts) What is the proportion of the measurements in $\bar{x} \pm 2s$?

Solution:

$\bar{x} \pm 2s = [43.81 - 2 * 24.63, 43.81 + 2 * 24.63] = [-5.447431, 93.066479]$ [1]
The proportion of the measurements in $[-5.45, 93.07]$ is $\frac{21}{21} = 1$ [1]

(**g**) (2 pts) (Refers to (**f**)) Is this proportion close to what the empirical rule suggests? Why?

Solution:

Yes, The empirical rule states that 95% of the population lies within two standard deviations of the mean. It was found that 100% of the population lies within $\mp 2s$ so the proportion is close to what the empirical rule suggests since the distribution is symmetric.

Q-10) ..... *12 marks*

The monthly utility bills for a household in Windsor, Ontario, were recorded for 12 consecutive months starting in January 2006:
204.94,180.00, 178.23, 176.43, 165.12, 236.72, 276.70, 309.70, 312.40, 238.66, 225.47, 222.23

(**a**) (2 pts) 25% of utility bills are smaller than what value?

Solution:

First, sort the data in an ascending order as follows
165.12,176.43,178.23,180.00,204.94,222.23,225.47,236.72,238.66,276.70,309.70,312.40
Location of Q_1 is $0.25(12+1) = 3.25$ [1]

$$Q_1 = X_{(3)} + 0.25(X_{(4)} - X_{(3)}) = 178.23 + 0.25(180.00 - 178.23) = 178.6725 \quad [1]$$

(**b**) (2 pts) 50% of utility bills are greater than what value?

Solution:

Location of Q_2, which is the same as the median m, is $0.5(12+1) = 6.5$ [1]

$$m = \frac{X_{(5)} + X_{(6)}}{2} = \frac{222.23 + 225.47}{2} = 223.85 \quad [1]$$

(**c**) (2 pts) 75% of utility bills are smaller than what value?

Solution:

Location of Q_3 is $0.75(12+1) = 9.75$ [1]

$$Q_3 = X_{(9)} + 0.75(X_{(10)} - X_{(9)}) = 238.66 + 0.75(276.70 - 238.66) = 267.19 \quad [1]$$

(**d**) (1 pt) Calculate the IQR (i.e. Inter-Quartile Range).

Solution:

$IQR = Q_3 - Q_1 = 267.19 - 178.6725 = 88.5175.$ [1]

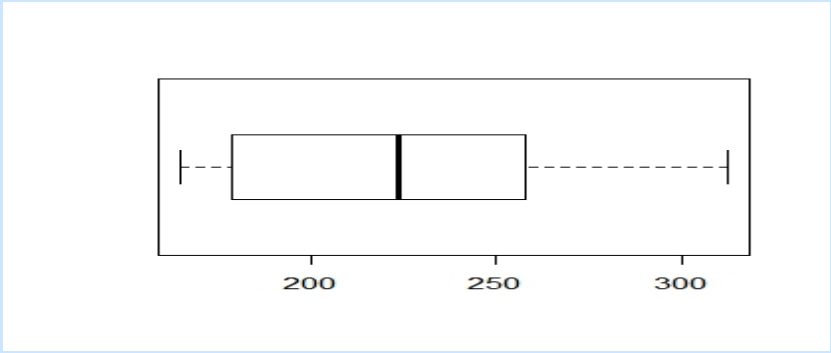
(**e**) (2 pts) Calculate the lower and upper fences.

Solution:

lower fence$= Q_1 - 1.5IQR = 178.6725 - 1.5(88.5175) = 45.89625$ [1]
upper fence$= Q_3 + 1.5IQR = 267.19 + 1.5(88.5175) = 399.9663$ [1]

(**f**) (3 pts) Construct a Boxplot (by hand). [2] Describe the distribution of the monthly utility bills and identify any outliers.

Solution:



There is no outlier. [1]

Q-11) ..... *11 marks*

A random sample of 10 students was taken from a population of first year students who have already entered the university or will do so in the immediate future. We wish to predict students' final grade in Calculus based on the mark obtained on a math test administered prior to the entrance to university

student	1	2	3	4	5	6	7	8	9	10
math score	39	43	21	64	57	47	28	75	34	52
Final grade	65	78	52	82	92	89	73	98	56	75

(**a**) (4 pts) Find the correlation coefficient between math test score and the final grade in Calculus.

Solution:

Std of math grades= $S_{math} = 16.58$ [1] and Std of final grades= $S_{final} = 15.115$ [1]

covariance of math and final grades = $S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1} = 210.4444$ [1]

$$r = \frac{S_{xy}}{S_{math}\ S_{stat}} = 0.84 \quad [1]$$

(**b**) (5 pts) Find the regression line that enables you to predict the final grades in Calculus based on the results of the math test.

Solution:

The equation of the required regression line is $final = a + b * math$, where
b is

$$b = r\frac{S_{final}}{S_{math}} = 0.84\frac{15.115}{16.58} = 0.7656 \quad [1]$$

a is the y intercept of the regression line and it is

$$a = 76 - 0.7656(46) = 40.7842 \quad [2].$$

The regression line is $final = 40.7842 + 0.7656 * math$ [2]

(**c**) (2 pts) What would be the predicted final grade in Calculus for a new student who scored 50 on a math test?

Solution:

predicted grade for stat $= 40.7842 + 0.7656 * (50) = 79.0642$ [2]