STAT 2507 Assignment # 1 (Chapters 1-3) Fall 2016

Last Name ————————————-, First Name ————————-

Student # ————————————
Lab Section (IMPORTANT) ————————————

Due Wednesday October 12 IN CLASS.

Total mark 100

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.

1. The age (in years) of 100 randomly selected tourists in a resort recorded in the column titles "Age" in the Excel file.
   a. Construct a frequency histogram for these data such that the first class interval is 0.83 - 10.93. [2]

   - *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, choose Cutpoint and then enter the two endpoints 0.83 10.93 (with a blank space in between) for the first interval. This way Minitab will construct a histogram with the classes 0.83 - 10.93, 10.93-21.03, etc.*
   - *Print your histogram and include it with your assignment*

   b. Describe the shape of distribution of this data set.[2] **Skewed to the right**
   c. What proportion of observations are older than 10.93?[2] **90/100**
   d. The mean or the median, which one is greater? [1] **The mean** why? [1] **Since the distribution is skewed to the right**

2. (Refer to "Age" data): Comparing Empirical rule to Tcbycheff theorem.
a. Use *desc* command (Enable command editor "MTB >" by checking "Enable commands" from Editor in the bar menu)
to find the mean [1] **27.67** and the standard deviation [1] **15.12**

b. [2] The following is meant to check how many student heights 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 c5=(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 **b.** above.
Next you will check how many ages did fall in the interval $\bar{X} \pm 2S$ by typing in the following: *tally c5*
c. What is the percentage of ages that fall between $\bar{X} \pm 2S$? [1] 94/100.
Is this value close to what the empirical rule suggests for the interval $\bar{X} \pm 2S$. **The answer is Yes but some may answer No, as the distribution is not symmetrical so both Yes and No should be accepted** [1]
Remark: 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 $\bar{X} \pm 2S$. This does NOT mean that for skewed distributions we cannot have 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**

3. The average sales (in Canadian dollars) per customer transaction of 45 randomly selected convenience stores were recorded. The data are listed in the column titled "Sales in CAD per customer transaction-2014" in the excel file.
Construct a stem-and-leaf chart for this set of data (print your graph)[2] 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 and include it with your assignment*

a. How many stores had average sales per transaction less than 13

CAD? [2] **9 stores**
b. How many stores had average sales per transaction of 16 CAD or more? [2] **15 stores**
c. what is the median of the average sales per transaction? [1] **14.9 $**

4. Refer to the "Sales in CAD per customer transaction-2014" data set above.
Construct a boxplot for this set of data(print it) [4] 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 and include it with your assignment*

   a. Based on the boxplot, how would you describe the shape of the distribution of this data set [1] **fairly symmetrical**
   b. The interquartile range (IQR) is approximately equal to [1] **Approximately 16.5-13.5=3, Or any number between 2 and 4 is acceptable**
   c. Does the data set has any outliers. [1] **No outliers**

5. Columns E, F and G are Height, Weight and (body mass index) BMI for 20 patients.
   a. Construct a scatterplot with height marked along the horizontal axis and weight marked along the vertical axis. Calculate the correlation coefficient: [1] **0.926**. If appropriate, fit a least square regression line using height to predict weight (Response variable). What is the equation of regression line? [2] **Weight=-322.96+7.195*Height**
   • *Select Stat; Regression. Enter the response variable and the predictor variable. click OK.*
   b. Construct a scatterplot with height marked along the horizontal axis and BMI marked along the vertical axis. Calculate the correlation coefficient: [1] **0.8135**. If appropriate, fit a least square regression line using height to predict BMI (Response variable). What is the equation of regression line? [2] **BMI=-21.23+0.6951*Height**
   c. Construct a scatterplot with weight marked along the horizontal axis and BMI marked along the vertical axis. Calculate the correlation coefficient: [1] **0.936**. If appropriate, fit a least square regression line using weight to predict BMI. What is the equation of regression line? [2] **BMI=8.92+0.103*Weight**
   d. What is the predicted BMI if weight is 134? [1] **22.722** . If weight is 200. [1] **29.518**. Can you predict the weight if BMI is 25? [1] **NO**.

Why? [1] **Because the equation in (c) is meaningful only for predicting BMI from weight not the other way around**.

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

1. Identify each of the following variables as categorical (i.e. qualitative), discrete or continuous.

   a. [1] Number of times per year a person catches a cold **Discrete**
   b. [1] Wind speed (Km/hour) in Chicago **Continuous**
   c. [1] The color of a ball drawn from a box containing two red, and 3 white balls **Qualitative**
   d. [1] Monthly unemployment rate in Canada **Continuous**
   e. [1] The month in which Ottawa's first major Winter storm will happen in 2017 **Qualitative**

2. A data set consists of 10 values that are fairly close together. The largest value is replaced by another value but the new value is an outlier (very far away from the other ones).

   a. [2] How is the mean affected? **Greatly affected**
   b. [2] How is the median affected? **Not affected**

3. [4] The average monthly salary of full professors in universities in Canada is 10,800 CAD with the standard deviation 1500 CAD. If a full professor's monthly salary is 12,000 CAD, would consider this salary unusual? (Hint: Use the z-score to justify your answer)
   Solution
   The z-score for the measurement 12,000 is z-score=$\frac{12000-10800}{1500}$=0.8. [2]
   This value is not greater than 3 [1]. Meaning that 12,000 CAD is not unusual [1]

4. The waiting time (in minutes) to speak to an agent of an insurance company of 20 randomly selected callers are recorded as below.

   3.06, 2.98, 2.86, 2.81, 2.99, 2.71, 3.11, 2.98, 3.03, 2.95, 2.81, 2.94, 3.09, 2.85, 3.05, 2.96, 3.05, 2.84, 2.84, 2.97
   a. [2] What is the average waiting time according to this data
   Solution

   $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ [1]
   = 2.944 minutes. [1]
   b. [4] What is the standard deviation of the waiting time data
   Solution

$$Sample\ variance = S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} \quad [1]$$

$$= \frac{173.5688 - \frac{(58.88)^2}{20}}{19} \quad [2]$$
$$= 0.01189 \quad [1]$$

c. [4] Construct a stem-and-leaf graph for this data (by hand)

Unit of leaf= 0.01 [1]
27 | 1
28 | 114456
29 | 4567889
30 | 35569
31 | 1

The stem-and-leaf graph worths [3]

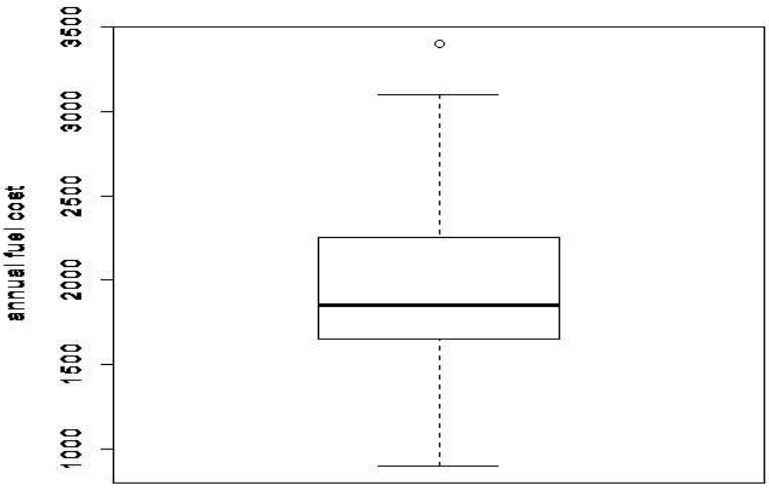
d. [2] Roughly, how would you describe the shape of the graph
Solution
Fairly symmetrical
e. [6] What are the values of the three quartiles (Q_1, median, Q_3)
Location of Q_1: (20+1)*0.25=5.25 [1]
$Q_1 = X_{(5)} + 0.25 * (X_{(6)} - x_{(5)}) = 2.84 + 0.25 * (2.85 - 2.84)$ **= 2.8425 minutes [1].**
Location of $Q_2 = median$: (20+1)*0.5=10.5 [1]
$median = (X_{(10)} + X_{(11)})/2 = (2.96 + 2.97)/2 = 2.965$ **[1]**
Location of Q_3: (20+1)*0.75=15.75 [1]
$Q_3 = X_{(15)} + 0.75 * (X_{(16)} - x_{(15)}) = 3.03 + 0.75 * (3.05 - 3.03)$ **= 3.045 minutes [1].**
f. [2] What is the proportion of the measurements in $\bar{X} \pm 2S$
Solution:
$\bar{X} \pm 2S$=[2.725,3.162] [1] which contains 17/20 =0.85 [1]
g. [2] (Refers to (**f**)) Is this proportion close to what the empirical rule suggests? Why?
Solution: Considering the relatively small size of the sample, 0.85 can be viewed as approximately close to 0.95 as suggested by the empirical rule [2]

5. The annual fuel cost in CAD for 24 popular vehicles in Canada are as listed below. (Source Statistic Canada Website).

3100, 3100, 900, 1000, 1750, 1750
2250, 1850, 1850, 2250, 3400, 2550
1600, 1700, 2250, 1750, 2000, 2150
2400, 1550, 1550, 1100, 1700, 1850

Solution: First, sort the data in an ascending order as follows
900, 1000, 1100, 1550, 1550, 1600, 1700, 1700, 1750, 1750, 1750, 1850, 1850, 1850, 2000, 2150, 2250, 2250, 2250, 2400, 2550, 3100, 3100, 3400
a. [2] 75% of the annual fuel costs are greater than what value.
Location of Q_1 is 0.25*(24+1)=6.25 [1]
So, $Q_1 = 1600 + 0.25(1700 - 1600) = 1625$. [1]
b. [2] 50% of the annual fuel costs are greater than what value.
Location of Q_2, which is the same as the median m, is 0.5*(24+1)=12.5 [1]
So, $m = 1850 + 0.5(1850 - 1850) = 1850$. [1]
c. [2] 75% of the annual fuel costs are smaller than what value.
Location of Q_3 is 0.75*(24+1)=18.75 [1]
So, $Q_3 = 2250 + 0.75(2250 - 2250) = 2250$. [1]
c. What is the range that contains approximately 50% of the middle values of the annual fuel costs. (Hint: Think of the interpretation of the IQR)
IQR contains approximately 50% of the middle values of the measurements and in this case it is $IQR = Q_3 - Q_1 = 2250 - 1625 = 625$. [1]
e. Compute the lower and upper fences [2] and construct a box-plot (by hand) for this data set [1], and identify any potential outliers [1].
lower fence= $Q_1 - 1.5 * IQR = 1625 - 1.5(625) = 687.5$ [1]
upper fence= $Q_3 + 1.5 * IQR = 2250 + 1.5(625) = 3187.5$ [1]
There is one outlier which is 3400. [1]

[1] for the graph



6. The math and stats final grades (out of 100) for 6 second year students are recorded in the following table.

student	1	2	3	4	5	6
math grade	15	88	91	72	64	58
stats grade	5	92	94	72	62	54

a. [4] Find the correlation coefficient between math and stats grads
Solution:
sd of math grades=$S_{math} = 27.58$ [1]
sd of stat grades= $S_{stat} = 32.646$ [1]
covariance of math and stat grades $= S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{j=1}^{n} x_i)(\sum_{k=1}^{n} y_k)}{n}}{n-1} = 900.0667$ [1]

$$r = \frac{S_{xy}}{S_{math}\ S_{stat}} = 0.9996 \text{ [1]}$$

b. [5] Find the required regression line that enables you to predict the stats grades of the students based on their math grads.
Solution:

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

b is $b = r\dfrac{S_{stat}}{S_{math}} = 0.9996 * \dfrac{32.646}{27.58} = 1.1832$ [1]

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

$a = 63.1666 - 1.1832 * 64.6666 = -13.3469$ [2].

The regression line is $stat = -13.3469 + 0.9996 * math$ [2]

c. [2] What is your prediction for the stats grade of a student whose math grade is 75.

Solution:

predicted grade for stat= -13.3469 +0.9996*(75)= 61.623 [2]