Z. Montazeri
M. Nasari
A. Sivathayalan

STAT 2507 Solution-Assignment # 1 Winter 2016

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 individual radiation dose per year is measured by U.S Regulatory Commission at nuclear power reactors. A sample of 50 individual radiation measurements is given in column A (Source: U.S Nuclear Regulatory Commission).
   a. Construct a frequency distribution and histogram for these data such that the first class interval is 0-0.1.[4]

   - *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 type.*
   - *Print your histogram*

   b. Describe the distribution of this data set.[2] **Skewed to the right**
   c. What proportion of observations are less that 0.5?[2] **46/50=0.92**
   d. Multiply each observation in the table by 100. Construct a histogram for these new transformed data, (don't need to print it). Compare the two histograms, are the shapes similar? Describe any differences.[2] **The shape of two histograms are same but the scale in x-axis is different.**

2. The weekly weights (in pounds) of plastic discarded by 62 households are listed in column D. Construct a stem-and-leaf chart for this set of data (print your graph)[4] to answer the following questions:
   a. How many families have plastic waste weighing below 1 pound?[2] **14**
   b. How many families have plastic waste weighing 3 pounds or more?[2] **9**
   c. Roughly what is the weekly median plastic waste? [1] **about 1.6 or 1.7**

3. Refers to the same data set above.
   Construct a boxplot for this set of data(print it) [4] to answer the following questions:
   a. The median is approximately equal to **1.6 or 1.7** [1]
   b. The interquartile range is approximately equal to **about 2.8-1.2=1.6** [1]
   c. The data set has any outlier value [1]**yes**

4. Refers to the same data set above. Construct a frequency histogram for the data above and find the shape of this data set? [2] **skewed to the right**.

5. Refers to the same data set above.
   a. Use *desc* command in Minitab (Enable command editor "MTB >" by checking "Enable commands" from Editor in the bar menu) to find the mean **1.9**, median **1.6**, 1st quartile **1.17**, 3rd quartile **2.7** of the weekly plastic waste for these 62 households.[4]
   b. 75% of weekly plastic waste is less than what value?[1] **2.7**.

6. **Illustration of the empirical rule**: Data set in column F corresponds to heights (in inches) of 100 students . Copy them in Minitab worksheet, for example in column C3.
   a. Construct a frequency histogram for this data, can be considered as fairly symmetric? [1] **Yes**
   b. Use *desc* command to find the mean $\bar{x}$=**67.88** and the standard deviation s=**4.47**[2]
   c. [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 C4 which will contain only values 1 or 0 according to whether the corresponding height (in column C3) falls in the interval $\bar{x} \pm 2s$, by typing in the following: *let c4=(c3>= $\bar{x} - 2 * s$ and c3<= $\bar{x} + 2 * s$).*
   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 heights did fall in the interval $\bar{x} \pm 2s$ by typing in the following: *tally c4*
   What is the percentage of heights that fall between $\bar{x} \pm 2s$. Answer **96%**. According to the empirical rule what percentage of heights that should be between $\bar{x} \pm 2s$? **95%**

7. Columns H, I and J are Height, Weight and 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? [0.5] **22.71-2** . If weight is 200. [0.5] **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**.

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. [2] Number of weekly accidents on Montreal-Ottawa portion of High Way 417 West.**discrete**
   b. [2] Birth country of the next United Nations secretary general. **categorical**
   c. [2] Weight of the next baby born at Ottawa General Hospital. **continuous**

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 effected? **A large effect**
   b. [2] How is the median effected? **No effect**

3. According to Statistics Canada, the monthly earnings of workers in the mining industry was \$3840, with a standard deviation of \$240. A mine worker claims to earn \$4325. each month. Find the z-score corresponding to this worker's wage. Is the amount unusual? [3]
   Solution:
   z-score of \$4325 is given by
   $$\frac{4325 - 3840}{240} = 2.02$$
   which is (barely) above 2. So we would rather consider this monthly earning as usual.

4. Given below are times (in seconds) between an order being placed and the food being received at a fast food restaurant's drive-through window. Compute the mean, median, range, variance, and standard deviation of these times and specify the unit for each of these statistics. 135 90 121 159 177 135 227.[10]
Solution:
The mean time is given by

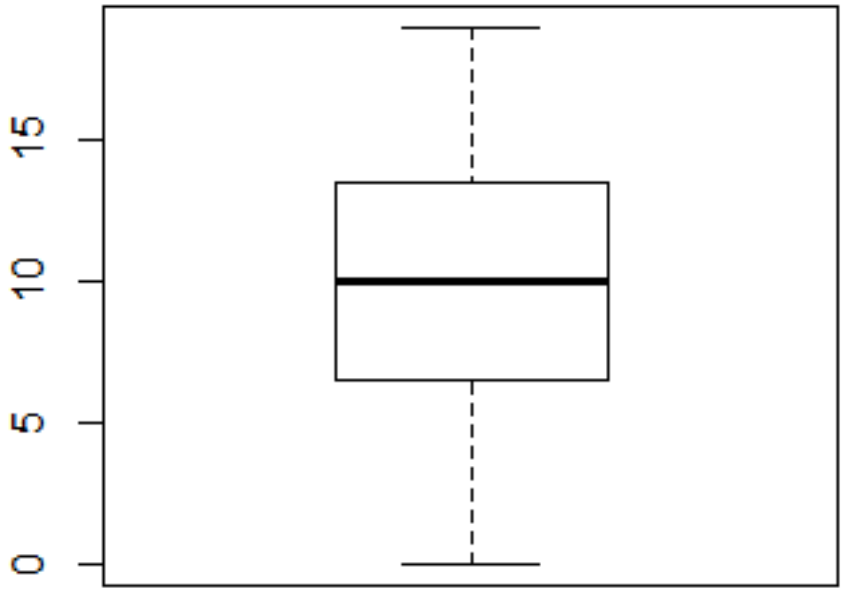
$$\bar{x} = \frac{1}{7}(135 + 90 + 121 + 159 + 177 + 135 + 227) = \frac{1044}{7} = 149.14 \text{ seconds.}$$

The median is obtained as follows: first we reorder the observations:
90 121 135 135 159 177 227. As the sample size $n = 7$ we get the median position as (.5)(7+1)=4 and hence the median time is is 135 seconds.
The range is maximum-minimum= 227-90=137 sec.
The variance is given by

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j^2 - \frac{\left(\sum_{j=1}^{n} x_j\right)^2}{n}\right) \\ &= \frac{1}{6}\left(135^2 + 90^2 + 121^2 + 159^2 + 177^2 + 135^2 + 227^2 - \left(\frac{(1044)^2}{7}\right)\right) \\ &= \frac{1}{6}(167330 - 155705.1) = \frac{11624.86}{6} = 1937.47 \text{sec}^2. \end{aligned}$$

and the standard deviation is $S = \sqrt{S^2} = \sqrt{1937.47} = 44.01$ sec.

5. Construct a box plot (by hand) for the data and identify any outliers:[10]
19, 12, 16, 0, 14, 9, 6, 1, 12, 13, 10, 19 ,7, 5, 8
Solution:
Order data from minimum to maximum 0, 1, 5, 6, 7, 8, 9, 10, 12, 12, 13, 14, 16, 19, 19
$Q_1 = 6,\ Q_2 = median = 10,\ Q_3 = 14,\ IQR = 14 - 6 = 8$
Lower fence=$Q_1 - 1.5IQR = 6 - 1.5(8) = -6$,
Upper fence=$Q_3 + 1.5IQR = 14 + 1.5(8) = 26$



No outlier.

6. The number of household members, x, and the amount spent on groceries per week, y, rounded to the nearest dollar are measured for 8 households in a suburb of Ottawa. The data are shown below.

x	5	2	2	1	4	3	5	3
y	140	50	55	35	95	70	130	65

a. Compute $\bar{y}, S_x, S_y$, and S_{xy}[8]
b. Compute the correlation coefficient r between x and y [6]. What would you estimate a household of 6 to spend on groceries per week [3].

Solution:

a.

$$\bar{y} = \frac{\sum y_i}{n} = \frac{640}{8} = 80,\ S_x = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} = \sqrt{\frac{93 - (25)^2/8}{7}} = 1.458$$

$$S_y = \sqrt{\frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1}} = \sqrt{\frac{61400 - (640)^2/8}{7}} = 38.173,$$

$$S_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i \sum y_i)}{n}}{n-1} = \frac{2380 - (25 * 640)/8}{7} = 54.286$$

b.

$$r = \frac{\sum x_i y_i - \frac{(\sum x_i \sum y_i)}{n}}{\sqrt{(\sum x_i^2 - (\sum x_i)^2/n)(\sum y_i^2 - (\sum y_i)^2/n)}} = \frac{2380 - (25)(640)/8}{\sqrt{(93 - (25)^2/8)(61400 - (640)^2/8)}} = 0.9756$$

We need to find the regression line

$$b = r(\frac{S_y}{S_x}) = 0.9756(38.173/1.458) = 25.546,\ a = \bar{y} - b\bar{x} = 80 - 25.546(25/8) = 0.168$$

The regression line is $y = a + bx = 0.168 + 25.546(x)$, so for $x = 6$, the estimated value for y is $y = 0.168 + 25.546(6) = 153.44$