STAT 2507 Solution - Assignment # 1 (Chapters 1&2) Fall 2015

Due: Thursday October 1 in class; NO LATE ASSIGNMENT ACCEPTED

Total of marks=100.

Part I. Lab Questions. Use only the blanks left to answer lab questions. Print all histograms, boxplots, stem-and-leaf plots, etc, you are asked to generate.

Click on 'Editor' and then on 'Enable Commands' to enable the 'MTB' prompt. Type the following Minitab commands to generate 200 numbers to be saved in a vector called c2:

MTB > set c1
DATA> 1:200
DATA> end
MTB > random 200 c11;
SUBC> normal 1400 100.
MTB > let c2=1000*log(c1)+c11 (c2 contains the data)

Think of these 200 numbers in c2 as the prices of a sample of 200 used cars, in dollars.

1. Draw a stem-and-leaf plot of these 200 prices (you can do this by typing 'stem c2') and then use your plot to answer the following questions:

   (a) [2] The maximum price is $6800 (±200).

   (b) [2] The minimum price is $ 1200 (±200).

   (c) [2] The median price is $ 5900 (±100).

   (d) [2] 36% of the prices are less than or equal to $ 5600 (±100).

   (e) [2] What is the shape of the distribution of the prices? Answer: Skewed to the left.

2. Use the 'describe' command to answer parts (a), (b), and (c) of Question 1, (type describe c2).
   Answer: (a) 6829.2(±150), (b) 1282.7(±150), (c) 5995.2(±50). [6]
   (d) [4] The average price and the standard deviation of the price of a used car are, respectively, $ 5717.2(±150) and $ 952.6(±50).

3. (a) [2] What proportion of the prices are within 2 standard deviations of the mean price (i.e., what proportion of the prices fall in the interval $\bar{x}\pm 2s$)? Answer: 189/200 ($(189 \pm 3)/200$).

(b) [2] Answer part (a) using Tchebysheff's Theorem. At least $150/200 = 0.75$.

(c) [2] Answer part (a) using the Empirical Rule. Approximately $190/200 = 0.95$.

4. Now make the following transformation of the used car prices (that are in c2): (new price) = 1.6× (price) + 200. This can be done in Minitab by using the 'let' command:

   MTB > let c10=1.6*c2 + 200

   (a) [8] The z-scores corresponding to the two smallest values in c10 are $-4.65(\pm 0.8)$ and $-3.80(\pm 0.5)$ respectively. Are these two values outliers? Yes.
   Why ? $|Z - score| > 3$.

   (b) Use a Boxplot of the new prices in the vector c10 to answer the following questions (click on 'graph' and then on 'Boxplot' and click 'ok'):

   (i) [2] The median is approximately equal to \$ 9800($\pm$200).

   (ii) [2] The interquartile range is approximately equal to \$ 1800($\pm$200).

   (iii)[2] The range of the data (new prices) is approximately equal to \$ $11200 - 2200 = 9000(\pm 300)$.

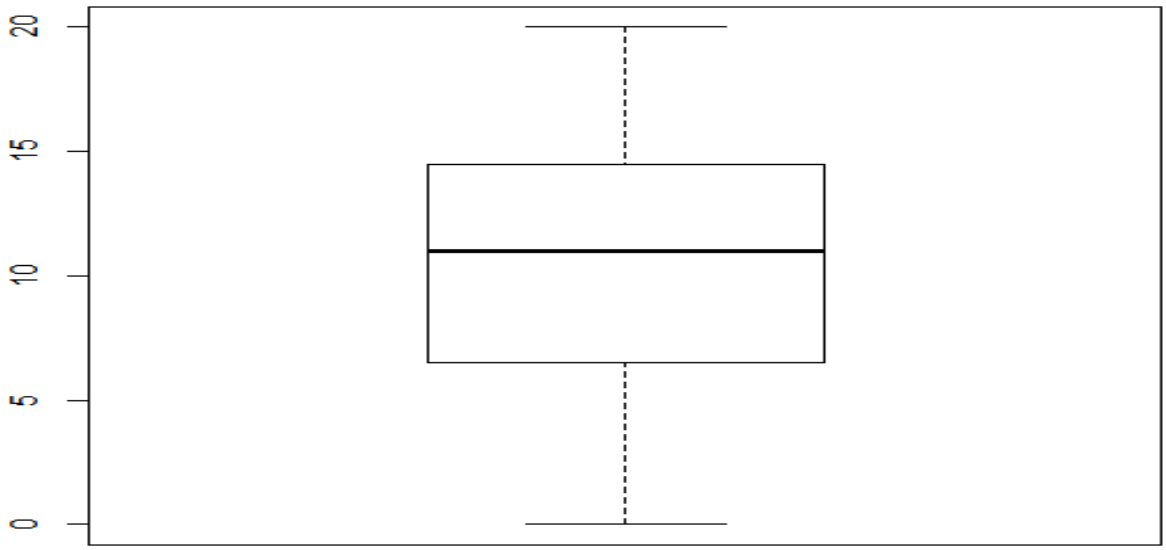
   (c) [2] Obtain a histogram of the new prices in c10 and comment on the shape of the distribution of these prices Skewed to the left

Part II Comprehension questions

1. Identify each of the following variables as either quantitative discrete, quantitative continuous, or qualitative.
   (a)[2] The brands of ice cream that you purchase
   (qualitative)
   (b)[2] The daily high temperature for the last four weeks,
   (quantitative continuous)
   (c)[2] The amount of sugar consumed by Americans in one year,
   (quantitative continuous)
   (d)[2] Number of brothers and sisters you have,
   (quantitative discrete)

2. Construct a boxplot for these data and identify any outliers:[17]

   6, 19, 0, 2, 11, 12, 13, 12, 5, 16, 2, 7, 13, 20, 18, 19, 9, 9, 9
   min = 0, max = 20
   position of the first quartile=$0.25 * (19 + 1) = 5 \rightarrow Q_1 = 6$
   position of the second quartile=$0.5 * (19 + 1) = 10 \rightarrow$ median $= Q_2 = 11$
   position of the third quartile=$0.75 * (19 + 1) = 15 \rightarrow Q_3 = 16$



3. Tornadoes cause many deaths each year in the United States. The order values for the yearly number of deaths for the 54 years 1950 through 2003 are
15, 24, 25, 27, 28, 30, 30, 31, 32, 33, 34, 34, 36, 39, 39, 39, 40, 43, 44, 46, 50, 51, 52, 53, 53, 55, 58, 59, 60, 64, 66,67, 67, 69, 70, 73, 73, 83, 84, 89, 94, 94, 98, 114, 122, 129, 130, 131, 159, 193, 230, 301, 366, 519
(a)[8] Determine the intervals $\bar{x} \pm s, \bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

$\bar{X} = \frac{\sum X_i}{n} = \frac{4645}{54} = 86.02$ $s^2 = \frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1} = \frac{825199 - \frac{(4645)^2}{54}}{53} = 8031$ $s = \sqrt{s^2} = \sqrt{8031} = 89.62$

$\bar{X} \pm s$	$\bar{X} \pm 2s$	$\bar{X} \pm 3s$
86.02 ± 89.62	$86.02 \pm 2(89.62)$	$86.02 \pm 3(89.62)$
(-3.6, 175.64)	(-93.22, 265.26)	(-182.84, 354.88)

(b)[6] Find the proportion of the measurements that lie in each of these intervals.

$\bar{X} \pm s$	$\bar{X} \pm 2s$	$\bar{X} \pm 3s$
49/54=0.91	51/54=0.94	52/54=0.96

(c)[6] How do the percentages obtained in part (a) compare with those given by the Empirical Rule? Should they be approximately the same? Explain.

The proportions for Empirical rule are:

$\bar{X} \pm s$	$\bar{X} \pm 2s$	$\bar{X} \pm 3s$
0.68	0.95	0.997

The percentages in part (a) are not similar to Empirical Rule since the distribution of data is not symmetric unimodal.

4. The following is a stem and leaf plot for data on the costs (in dollars) of a sample of 30 postal mailings by a company.

85	2
86	
87	
88	
89	
90	
91	3 4 8
92	0 0 1 3
93	1 1 2 5 7
94	0 1 3 3 4 8 9
95	2 3 4 5 5
96	2
97	0 4 7
98	
99	
100	0

Leaf Unit = 0.01

(a) [8] What are the values of the median, lower quartile (Q1), and upper quartile (Q3) for this data set?

(b) [5] What is approximate value for standard deviation?

The data set is 8.52,9.13,9.14,9.18,9.2,9.2,9.21,9.23,9.31,9.31,9.32,9.35,9.37,9.4,9.41,9.43,9.43, 9.44,9.48,9.49,9.52,9.53,9.54, 9.55,9.55,9.62,9.7,9.74,9.77,10

(a)

The position of $Q_1 = 0.25 * (31) = 7.75$, so $Q_1 = 9.21 + 0.75(9.23 - 9.21) = 9.225$

Median=The position of $Q_2 = 0.5 * (31) = 15.5$, so median$= 9.41 + 9.43/2 = 9.42$

The position of $Q_3 = 0.75 * (31) = 23.25$, so $Q_3 = 9.54 + 0.25(9.55 - 9.54) = 9.5425$

(b)

$S \approx \frac{R}{4} = \frac{Max - Min}{4} = \frac{10 - 8.52}{4} = 0.37$