Assignment one - part II

Total points : 50

Part I of the assignment :

- Must be done via MySytatlab
- Due-date: Monday, May 25, 11:59 P.M.

Part II of the assignment :

- Please don't forget to complete your signed statement of Academic Integrity within the body of your solution
- Submit a PDF of your type-written (i.e., not handwritten) solution (recall that a submission cannot be marked unless it is in PDF format).
- Submit your assignment via blackboard learn by due-date Monday, May 25, 11:59 P.M.
- Use MiniTab or Excel (or any other statistical tool) if you need to. However you should cut and paste your result and sufficient explanation.
- Make sure to provide details calculations and steps of how you arrived at your answer when appropriate . Providing only the software output is not acceptable.

Question one (10 points)

Here are advertised house power rating and expected gas mileage, in mile per gallon, for several 2010 vehicles. (www.kbb.com)

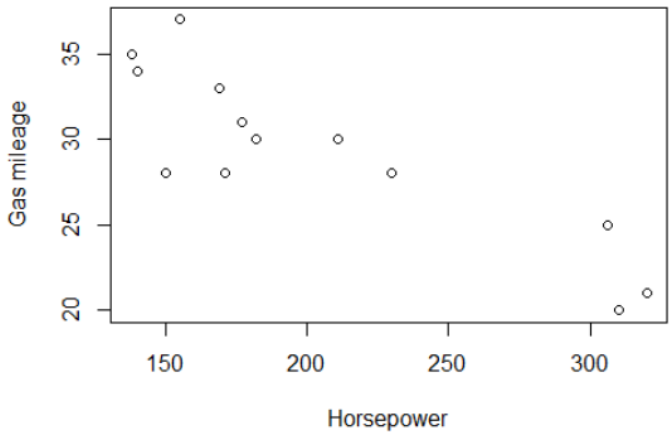
Car	hp	mpg
Audi A4	211	30
BMW 3 series	230	28
Buick LaCrosse	182	30
Chevy Cobalt	155	37
Chevy Suburban	320	21
Ford Expedition	310	20
GMC Yukon	320	21
Honda Civic	140	34
Honda Accord	177	31
Hyundai Elantra	138	35
Lexus IS 350	306	25
Lincoln Navigator	310	20
Mazda Tribute	171	28
Toyota Camry	169	33
Volkswagen Beetle	150	28

a) Which one of these variables is explanatory and which one is response variable ? (1 point)

The horsepower is explanatory and gas mileage is response variable.

b) Using an appropriate tool make a scatterplot for these data. (2 points)

The scatterplot of gas mileage versus horsepower is shown below:



c) Describe the direction, form, and strength of the association. (2 points)

There is a strong, negative, straight association between horsepower and mileage of the selected vehicles. There don't appear to be any outliers. All of the cars seem to fit the same pattern. Cars with more horsepower tend to have lower mileage. [NOTE: The plot seems to be missing two of the 15 models, but in fact there are two pairs of identical points].

d) Find the correlation between horsepower and miles per gallon. (2 points)

Since the relationship is linear, with no outliers, correlation is an appropriate measure of strength. The correlation between horsepower and mileage of the selected vehicles is –0.909.

e) Write a few sentences telling what the plot says about fuel economy. (1 point)

There is a strong linear relationship in the negative direction between horsepower and highway gas mileage. Lower fuel efficiency is associated with higher horsepower.

f) Convert the mileage to litres per 100 km (=235.2 ÷ mpg) and repeat part b) to e). Did the correlation change ? which scale for fuel consumption do you prefer and why ? (2 points)

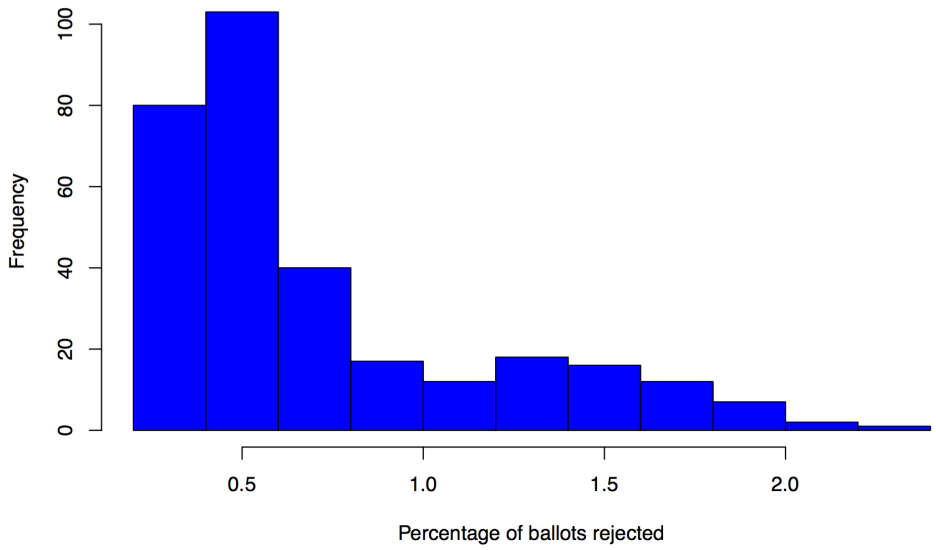
The scatterplot of fuel consumption versus horsepower is shown below. There is a strong, positive, straight association between fuel consumption and horsepower—more powerful cars burn more fuel. The correlation between the fuel consumption and horsepower is stronger than the correlation between mileage and horsepower, and the direction of association is reversed (correlation = 0.921). The strength of correlation changed here because the transformation from mileage to consumption is non-linear. The two presentations of the data are in some sense equivalent; mileage focuses on the

Question 2 (10 points)

For this question you will be examining the data-set labelled "***Election_2011***" which you can find in the assignment folder. In the dataset you will find the results from the 2011 Canadian federal election for the Maritime provinces, showing the riding name, winning candidate, and percentage of rejected ballots.

a) Using appropriate tool, make a histogram of the percentages of rejected ballots. (2 points)

A histogram is a suitable display as is shown below.



b) Find the mean and standard deviation. (1 point)

The mean is 0.665 and the standard deviation is 0.452.

c) Report the 5-number summary. (2 points)

The 5-number summary is given below.

Descriptive Statistics: Percentage of Ballots Rejected

Min.	*1st Qu.*	*Median*	*3rd Qu.*	*Max.*
0.20	*0.30*	*0.50*	*0.80*	*2.40*

d) Why do the mean and median differ here ? (1 point)

The mean and median differ here because the distribution of the percentage ballots rejected is skewed.

e) Which of (b) and (c) above does a better job of summarizing the distribution of percentage of rejected ballots? why? (2 points)

For skewed distributions (like the distribution of the percentage ballots rejected in this question) the 5-number summary is better than the mean and the standard deviation. From the histogram (and the stemplot above), we see that there are some possible outliers in the data. Mean and the standard deviation are measures not resistant to outliers. This also favours the use of the 5-number summary to summarize this data set.

f) Suppose the true percentage for Egmont was %1.8 and not %0.8. How would you expect the mean, median, standard deviation, and IQR to change? Explain your expectations for each (no computations, please) (1 point)

The mean will increase.
The current median is 0.66%. Both the 0.8 and 1.8 are on the same side of the median. Changing any value to a different value on the same side of the median does not change the median.
The standard deviation will increase.
The IQR will not change. The current third quartile is 0.8 and the value 0.8, which currently falls to the right of Q3 remains to the right in the new ordered data set. Thus Q3 is unchanged, as is the IQR.

g) Write a brief report about rejected ballots in the Maritimes. (1 point)

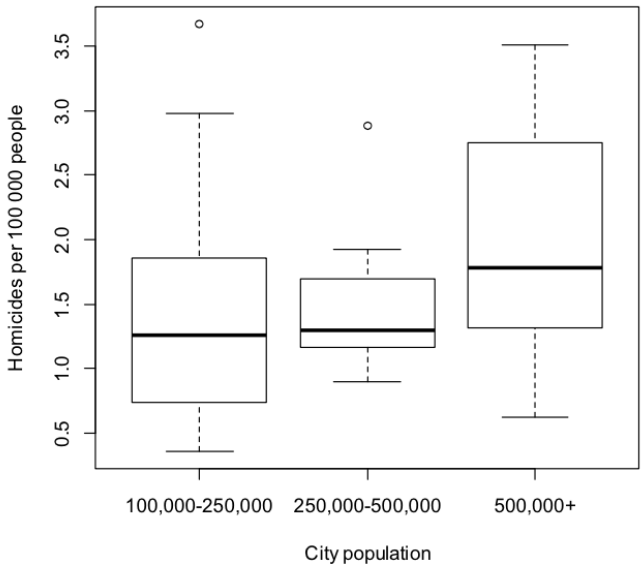
The distribution is right skewed. This means the proportion of ridings with high percentages of ballots rejected is relatively small. The median is 0.5%. The interquartile range is 0.8% – 0.3% = 0.5%, so about half the ridings have rejection percentages between 0.3% and 0.8%.

Question 3 (2 points)

For this question you will be examining the data-set labelled "***Homicides_2011***" which you can find on the assignment folder. The rate provided in the dataset is the average homicide rate per 100,000 population over the 11 years from 2001 to 2011. Cities were classified into three size categories. I also classified them by geographic region.

a) Use appropriate graphs to examine the relationship between homicide rate and city size (**category**). Describe any patterns you see. (3 points)

The side-by-side boxplots of homicide rates are shown below. Large cities have the highest median homicide rate. Medium-sized cities have the least variable homicide rate. There are two large outliers.

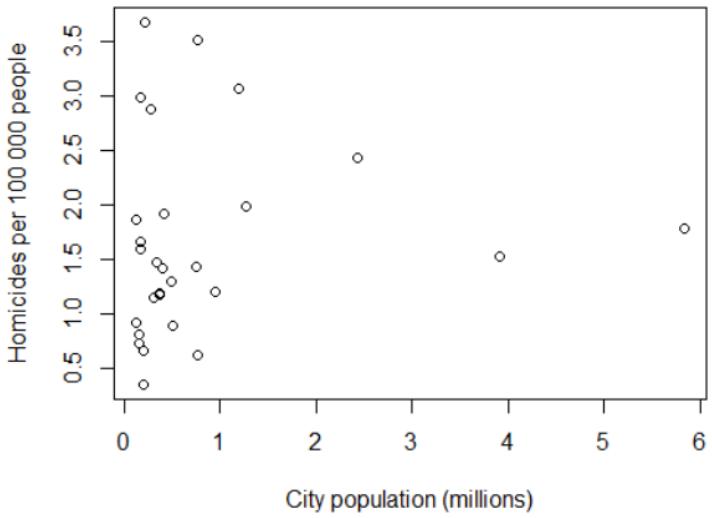


b) What type of variable is city size, if determined precisely by a count of residents? What type is in our analysis in part (a) ? (2 points)

City size is a quantitative variable. In our analysis in(a) above we used the variable city size to create a categorical variable (with three categories). This helps us get some idea about how the homicide rates are related to city size

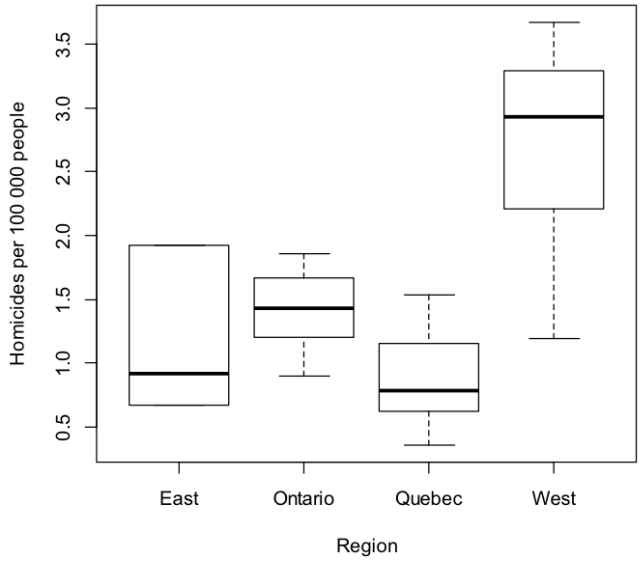
c) If you wanted to use the actually city sizes (populations), can you think of another type of graph to use to explore the relationship between crime rate and city size? if so produce this graph. (2 points)

A plot of homicide rate versus population size (a scatterplot) will show the relation between the two variables.



d) Produce appropriate graph to examine the relationship between homicide rate and geographical region. Describe any geographical pattern you see. (3 points)

The side-by-side boxplots of homicide rates are shown below. Quebec has the lowest homicide rate and the West has the highest.



Question 4 (10 points)

For this question you will be examining the data-set labelled "***Blood_Pressure***" which is provided on the assignment folder. The dataset contains data on blood pressure screening clinic for employees at a certain company.

a) Using pivot table in excel create contingency table that summarizes number of cases in the dataset by age group and blood pressure level. (you can also use minitab or other softwares to create this contingency table) (2 points)

Count of Blood pressure				
Row Labels	30 - 49	Over 50	Under 30	Grand Total
High	51	73	23	147
Low	37	31	27	95
Normal	91	93	48	232
Grand Total	**179**	**197**	**98**	**474**

b) Find the marginal distribution of blood pressure level. (2 points)

The marginal distribution of blood pressure for the employees of the company is the total column of the table, converted to percentages. 20% low, 49% normal, and 31% high blood pressure.

Count of Blood pressure				
Row Labels	30 - 49	Over 50	Under 30	Grand Total
High	10.76%	15.40%	4.85%	31.01%
Low	7.81%	6.54%	5.70%	20.04%
Normal	19.20%	19.62%	10.13%	48.95%
Grand Total	**37.76%**	**41.56%**	**20.68%**	**100.00%**

c) Find the conditional distribution of blood pressure level within each age group (2 points)

The conditional distribution of blood pressure within each age category is:

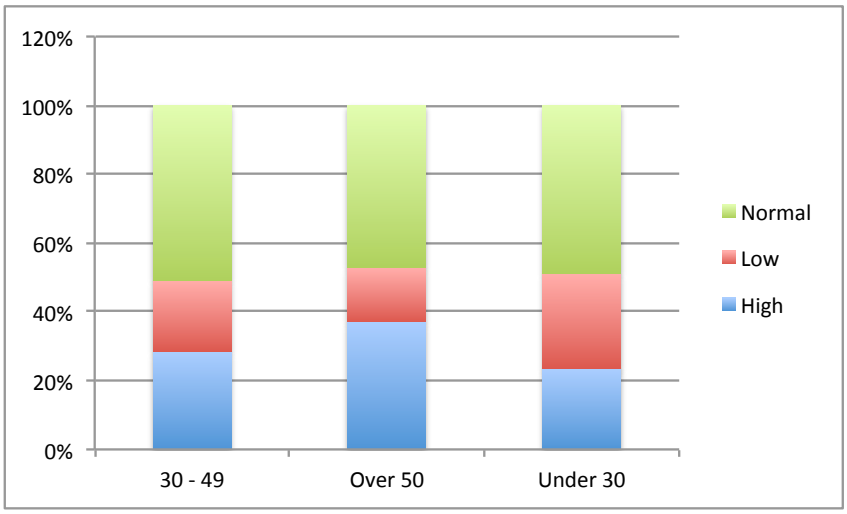
Under 30: 28% low 49% normal, 23% high

30–49: 21% low, 51% normal, 28% high

Over 50: 16% low, 47% normal, 37% high

Count of Blood pressure				
Row Labels	30 - 49	Over 50	Under 30	Grand Total
High	28%	37%	23%	31%
Low	21%	16%	28%	20%
Normal	51%	47%	49%	49%
Grand Total	**100%**	**100%**	**100%**	**100%**

d) Create a segmented bar graph for your result in part (b) to compare conditional distribution of blood pressure level within each age group. (You can create it simply in excel) (2 points)



e) Write a brief description of the association between age and blood pressure among these employees (1 point)

In this company, as age increases, the percentage of employees with low blood pressure decreases, and the percentage of employees with high blood pressure increases.

f) Do you think that this proves that people's blood pressure increases as they age? explain. (1 point)

No, this does not prove that people's blood pressure increases as they age. Generally, an association between two variables does not imply a cause-and-effect relationship. Specifically, these data come from only one company and cannot be applied to all people. Furthermore, there may be some other variable that is linked to both age and blood pressure.

Question 5 **(5 points)**

The following data represent the square footage of 10 three-bedroom condos for sale in Hilton Head, South Carolina.

1,559 1,625 1,167 1,264 1,676 1,300 2,058 1,126 1,858 1,321

Determine the interquartile range, upper limit and lower limit for this sample. Are there any outliers in this data set? (Show your steps)

In order to be consistent, Please use the **second method** discussed in lectures for finding percentiles (which also applies to quartiles).

1,126 1,167 1,264 1,300 1,321 1,559 1,625 1,676 1,858 2,058

Q_1: i = (25/100)(10) = 2.5
So Q_1 is the 3^{rd} position.

Q_1 = 1,264
Q_3: i = (75/100)(10) = 7.5
So Q_3 is the 8^{th} position.
Q_3 = 1,676

IRQ = 1,676 - 1,264 = 412
Upper Limit = 1,676 + 1.5(412) = 2,294
Lower Limit = 1,264 - 1.5(412) = 646

There are no outliers in this sample.

Question 6 **(5 points)**

The following data shows the number of minutes that seven customers waited for a table at a particular restaurant.

1 17 26 10 5 22 19 8

Manually calculate the coefficient of variation (CV) for this data. What does it mean? (Show your steps)

Answer:

$\overline{x} = 13.5$

x_i	$x_i 2$
1	1
17	289
26	676
10	100
5	25
22	484
19	361
8	64

$$\sum_{i=1}^{8} x_i = 108 \qquad \sum_{i=1}^{8} x_i^2 = 2,000$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = (108)^2 = 11,664$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{2,000 - \frac{11,664}{8}}{8-1}} = \sqrt{\frac{2,000 - 1,458}{7}} = 8.80$$

$CV = (8.8/13.5)(100) = 65.2\%$