Question	Marks
1	5
2	5
3	8
4	10
5	7
Total	35

Aids: Only question pages will be marked. Final numerical answers must be rounded to exactly 3 decimal places, unless the answers are exact to less than 3 decimal places. For intermediate steps, please be sure to keep all decimal places to avoid round off errors.

1. Circle the letter corresponding to your choice. [1 mark for each]

a) Suppose that you wanted to generate 50 observations from a Normal distribution with a mean of 100 and a variance of 2500, rounded to 2 decimal places. Which of the following R commands would you use?

A. round(rnorm(50, 100, 2500), digits=2)

B. round(rnorm(50, 100, 50), digits = 2)

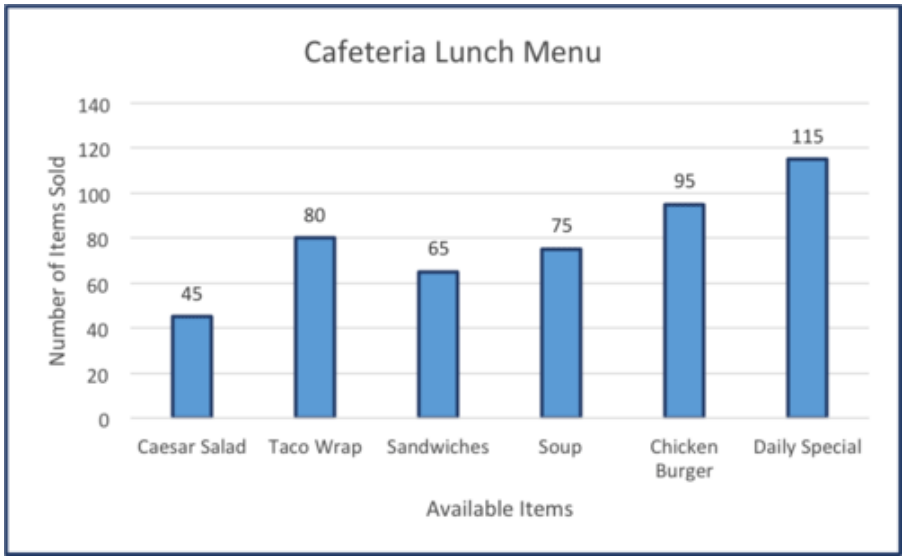
C. round(rnorm(2, 100, 2500), digits = 2)

D. round(rnorm(digits = 2), 50, 100, 50)

E. More than one of these is correct

- **Code similar to this was used in the first R assignment**

b) The following is a bar graph illustrating the lunch options, and the number sold at a local restaurant in a one-week period:



Which of the following is true based on the above bar graph?

A. The median choice was soup

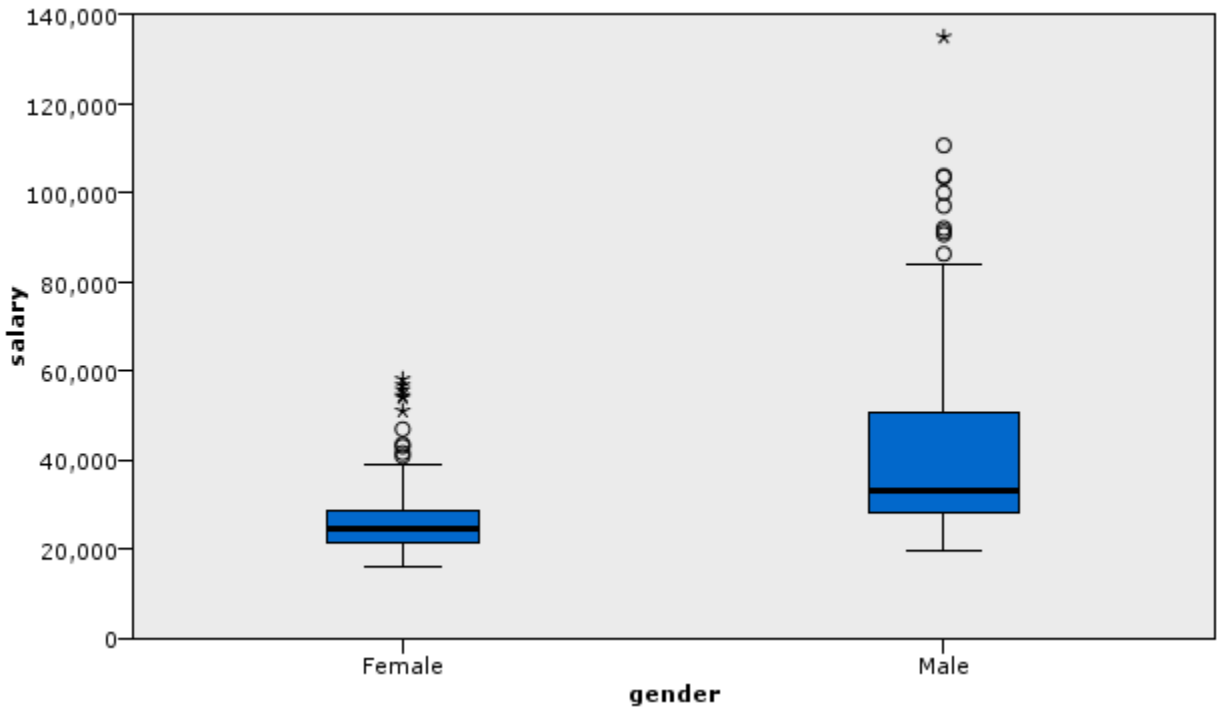
B. The average option was a taco wrap

C. The distribution of lunch options is skewed to the left

D. The most popular lunch item was the Daily Special

E. More than one of the above statements is true

c) You are given the following side-by-side boxplots for 2 different data sets:



Which of the following statements is / are true?

(I) The Interquartile Range (IQR) for females is greater than the Interquartile Range (IQR) for males

(II) Both data sets are skewed to the right / positively skewed

A. Both statements are true

B. (I) only

C. (II) only

D. Neither statements are true

d) Which of the following data sets has the largest range?

A. 6, 12, 15, 24, 35, 10 Ordered: 6, 10, 12, 15, 24, 35 -> Range = 29
B. 15, 5, 30, 60, 25, 8 Ordered: 5, 8, 15, 25, 30, 60 -> Range = 55
C. 3, 1, 3, 2, 5, 70 Ordered: 1, 2, 3, 3, 5, 70 -> Range = 69
D. 10, 40, 20, 50, 60, 30 Ordered: 10, 20, 30, 40, 50, 60 -> Range = 50

e) Which of the following statements is **FALSE**?

A. Pie charts and bar charts are not suitable for representing numerical / quantitative data
B. The sample IQR (Interquartile Range) of a data set cannot be determined from a boxplot
C. A run chart is a good way to summarize data collected over time
D. A sample median is not affected by outliers, so it is a more robust numerical summary of location.

- **The IQR is equal to the length of the box in the boxplot.**

2. [5 x 1=5 marks] For data $\{(x_i, y_i)$: i=1,2,…,30$\}$ we have:

$\sum_{i=1}^{30} x_i = 90.30$, $\sum_{i=1}^{30} x_i^2 = 449.27$, $\sum_{i=1}^{30} x_i y_i = 5090$, $\sum_{i=1}^{30} y_i = 1230.6$, $\sum_{i=1}^{30} y_i^2 = 63915$

range of $\{x_1, x_2, \ldots, x_{30}\} = 40$, range of $\{y_1, y_2, \ldots, y_{30}\} = 75$

Using the above information, answer the following questions. (Please give answers to 3 decimal places, wherever necessary).

a) The sample standard deviation of the data y=$\{y_1, y_2, \ldots, y_{30}\}$ is _____**21.524**_____

- $s_y = \sqrt{\frac{1}{30-1}\left[\sum_{i=1}^{30} y_i^2 - n(\bar{y})^2\right]} = \sqrt{\frac{1}{29}\left[63915 - 30\left(\frac{1230.6}{30}\right)^2\right]} = 21.524$

b) If all of the *y*-values were decreased by 2, the new range of the *y*-values would be _____**75**_____

- **If all y-values were decreased by 2, the new range = old range = 75**

c) The sample correlation for the data $\{(x_i, y_i)$: i=1,2,…,30$\}$ is _____**0.898**_____

- **Sample correlation, r =** $\frac{\sum_{i=1}^{30} x_i y_i - 30 * (\sum_{i=1}^{30} x_i/30) * (\sum_{i=1}^{30} y_i/30)}{\sqrt{\sum_{i=1}^{30} x_i^2 - (30) * \left(\frac{\sum_{i=1}^{30} x_i}{30}\right)^2 * \sum_{i=1}^{30} y_i^2 - (30) * \left(\frac{\sum_{i=1}^{30} y_i}{30}\right)^2}}$

d) Suppose another observation $(x_{31}, y_{31}) = (\bar{x}, \bar{y}) = (3.01, 41.02)$ is added to the data set. The new sample correlation for the data $\{(x_i, y_i)$: i=1,2,…,31$\}$ is _____**0.898**_____

- **Note that the observation added corresponds to the mean of the x-values, and the mean of the y-values.**
- **Adding this observation will leave the sample correlation unchanged.**
- **This can be verified by doing the new calculation.**

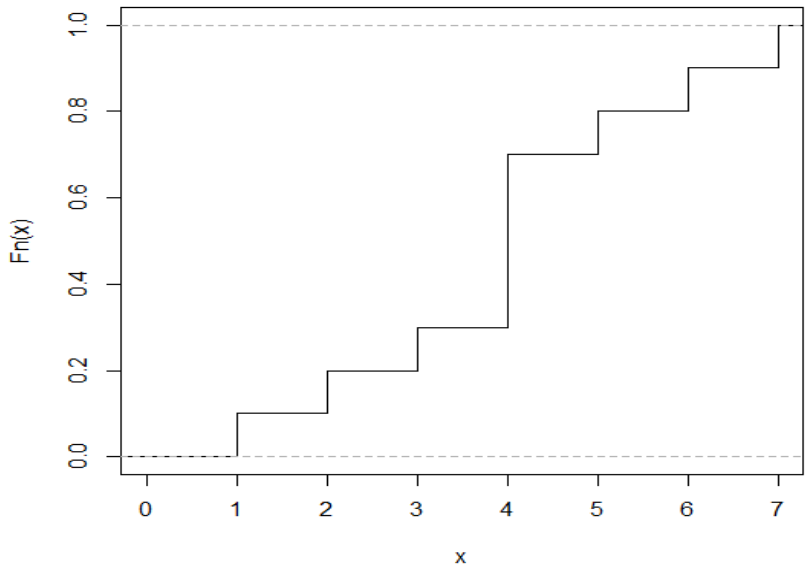
Please indicate whether the following statement is True or False:

e) Both of the R commands plot(x, y) and plot(y, x) would produce the same scatterplot of the original data $\{(x_i, y_i)$: i=1,2,…,30$\}$ _____**False**_____

- **When using the plot command in R, the placement of the arguments, x, and y, is important. Reversing them will produce a different scatterplot in this case.**

3. [8 marks] SHOW ALL WORK / JUSTIFY YOUR ANSWER where necessary. Answers to calculation questions without justification WILL NOT receive full credit.

Please use the ecdf plot below to answer the following questions:



a. There are 10 observations in the data set. List their values, from lowest to highest. ***[2 marks]***

- **From the ecdf, the data are: 1, 2, 3, 4, 4, 4, 4, 5, 6, 7**

b. Calculate the mean of the data set. ***[2 marks]***

- **The sample mean = $\frac{1+2+3+4+4+4+4+5+6+7}{10}$ = 4**

c. Calculate the median of the data set. ***[2 marks]***

- **With 10 observations, the median is given by the average of the 5th and 6th ordered values. So, the median is =(4+4)/2 = 4**

d. Are the data symmetric, or skewed? Justify your answer. ***[2 marks]***

- **The data are symmetric. The mean and median are equal.**
- **You can actually visualize the symmetry of the distribution by looking at the cdf.**

4. **[10 marks]** A sample of 24 test scores is given below. The data have already been ordered:

34 38 56 58 60 61 63 66 66 66 66 70 72 72 81 82 85 86 90 90 90 90 94 95

a. Are there any outliers in the data set? Justify your answer doing any necessary calculations. ***[3 marks]***

- **q(0.75): m = 25*0.75 = 18.75. So, the 75th percentile is given by the average of the 18th and 19th ordered observations. So, q(0.75) = (86+90)/2 = 88**
- **q(0.25): m = 25*0.25 = 6.25. So, the 25th percentile is given by the average of the 6th and 7th ordered observations. So, q(0.25) = (61+63)/2 = 62**

- **We need to check for outliers using: q(0.25) – 1.5IQR and q(0.75) + 1.5IQR**
- **IQR = 88-62 = 26; 1.5IQR = 39**
- **62 – 39 = 23 and 88 + 39 = 127**
- **There are no observations less than 23. There are no observations greater than 127. So there are no outliers.**

b. The above data set has either no mode, one mode, or more than one mode. State one of these choices and justify your answer. ***[2 marks]***

- **This data set has two modes. Observations 66 and 90 each appear most often, 4 times.**

c. Suppose that the test was considered too long, and an adjustment was made. In this case, 5 marks were added to all test scores. Due to this adjustment, state whether each of the following measures would be affected or unaffected. If unaffected, state that it would be unaffected. If affected, please state how (For example, the sample mean would increase by 4, or the sample median would decrease by 6, etc.). You can simply state the effect. No need to justify your answer / show work here. You can do so off to the right of each blank if you wish, but we are marking your answers only. ***[1 mark each; 5 marks]***

The sample mean would _____**increase by 5**_____

The sample median would _____ **increase by 5**_____

The sample standard deviation would _____**be unaffected**_____

The sample IQR would _____**be unaffected**_____

The range of the sample would _____**be unaffected**_____

5. ***[7 marks]*** A Waterloo-based public opinion research firm was hired by the Ontario Ministry of Education to investigate whether the financial worries of Ontario university students varied by sex. To reduce costs, the research firm decided to study only university students living in the Kitchener-Waterloo region in September 2012. An associate with the research firm randomly selected 300 university students attending a Laurier-Waterloo football game. The students were asked whether they agreed or disagreed with the statement: "I have significant trouble paying my bills." Their sex was also recorded. The results are given in the table below:

	Agreed	**Disagreed**	**Total**
Male	77	86	163
Female	43	94	137
Total	120	180	300

Please answer the questions below based on this article.

a. What are the units in this study? ***[1 mark]***

- **University student.**

b. **Answer True or False:** Is this an experimental study? Justify your answer. ***[2 marks]***

- **False. No treatment is being imposed. This is a survey.**

c. What are two variates in this study, and what type are they? ***[4 marks]***

- **One variate is sex (male / female). This is categorical. We would also accept explanatory variate.**
- **Another variate is whether they agree or disagree with the statement. This is categorical. We would also accept response variate.**