DANA 4800
Homework #1

Due: 11:59 pm on May 25th, 2025
Submitted to: BrightSpace Homework Folder

Please read these instructions before submitting your homework.

1. **Format**: Please convert all work to PDF.
2. **Excel Output**: If you have Excel Output, please copy and paste the Excel output to the corresponding part of the Word document.
3. **Submission**: No email submission will be accepted regardless of the circumstances.
4. **Number of Submission**: You are allowed to submit your homework as many times as you want. But only the latest version of your submission will be marked.
5. **Late submission**: You can still submit the homework till 6:00 am the next day with 50% penalty. Submission after 6:00 am will not be accepted.
6. **Order of Work**: All work must be submitted in the same order of the questions. It is suggested that you use a brand new page for a different question, especially true for hand-written work.
7. **Filename**: Please name your file in the following format: "DANA4800_Lastname_Firstname_HW#.pdf". For example, my first assignment would have the name "DANA4800_Lo_Michael_HW1.pdf".
8. **Last Advice**: Internet connections do go off, computers do break down, or other unexpected events do occur when they are least expected. The onus is squarely on you to submit it on time. No excuses.

Note: Failure to comply with any of the above will result in loss of marks.

1. A counselor at ABC University wanted to find out the amount of outstanding tuition (in CAD$) ABC University international students carry in the Spring semester. A random sample of 50 ABC University international students was drawn in March to investigate this.
   a) Identify the researcher. **[1 mark]**
   b) Provide a description of the objective. **[1 mark]**
   c) Identify the subjects of interest. **[2 marks]**
   Note: Make sure you also include the when and where, if available.
   d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. **[1+1 marks]**
   e) Identify the type of the variable and the corresponding scale of measurement. **[1+1 marks]**
   Note: Marks will be deducted with missing words (like "variable" or "scale").
   f) Provide a description of the population of interest. **[1 mark]**
   g) Identify the most appropriate sampling method. **[1 mark]**
   h) Provide a description of the sample. **[1 mark]**
   i) Is there an issue with the selection bias? Briefly justify your answer by comparing the (target) population and the sampling frame. **[0+2 marks]**

2. Marisol lives in the City of Vancouver. She was planning to buy an electric vehicle (EV) this summer and wondering what percentage of EV owners have installed a Level 3 Supercharger at home. A sample of 20 EV owner living in her neighbourhood was drawn in June to investigate this.
   a) Identify the researcher. **[1 mark]**
   b) Provide a description of the objective. **[1 mark]**
   c) Identify the subjects of interest. **[2 marks]**
   Note: Make sure you also include the when and where, if available.
   d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. **[1+1 marks]**
   e) Identify the type of the variable and the corresponding scale of measurement. **[1+1 marks]**
   Note: Marks will be deducted with missing words (like "variable" or "scale").
   f) Provide a description of the population of interest. **[1 mark]**
   g) Identify the most appropriate sampling method. **[1 mark]**
   h) Provide a description of the sample. **[1 mark]**
   i) Is there an issue with the selection bias? Briefly justify your answer by comparing the (target) population and the sampling frame. **[0+2 marks]**

3. The director overseeing all senior homes in the Fraser Health Authority (FHA) wanted to know how many falls seniors have in 2023 that leads to major hip or lower body surgery. Ten seniors were randomly selected from each senior home in FHA region to form the sample.
   Note: This is a tricky question. Make sure you spend some time thinking about how to differentiate between a categorical variable from a numerical variable.
   a) Identify the researcher. **[1 mark]**
   b) Provide a description of the objective. **[1 mark]**
   c) Identify the subjects of interest. **[2 marks]**
   Note: Make sure you also include the when and where, if available.

d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. **[1+1 marks]**
e) Identify the type of the variable and the corresponding scale of measurement. **[1+1 marks]**
Note: Marks will be deducted with missing words (like "variable" or "scale").
Note: I suggest you provide some reasoning to justify your choice of the type of variable.
f) Provide a description of the population of interest. **[1 mark]**
g) Identify the most appropriate sampling method. **[1 mark]**

4. Students are generally confused with the stratified random sampling method and the cluster sampling method. I hope students have a better understanding of them after completing this question.
There are two situations in the questions: one method is used in each of the two situations. In each of the following two situations, (1) identify the sampling method used, (2) describe how you would define the non-overlapping groups, and (3) outline the three main steps.
Note: In practice, geographic factor should not be used to do the stratification. It is used here only for the sake of illustration.
   a) To get an idea of how much detached houses in the city of Vancouver cost these days, Joe went to a real estate web site to collect some information. He divided the city of Vancouver into 20 communities; randomly selected 5 detached houses from each community; and the 100 (20 communities times 5 houses per community) detached houses formed the sample. **[1+1+3 marks]**
   b) To get an idea of how much detached houses in the city of Vancouver cost these days, Joe went to a real estate web site to collect some information. He divided the city of Vancouver into 20 communities; randomly selected 5 of the communities; and all detached houses (that are listed on the web site) within those 5 communities were chosen in the sample. **[1+1+3 marks]**

5. Consider a movie theatre with 30 rows of 20 seats in each row. There are some prize giveaways before the movie starts. Identify the most appropriate sampling method used below.
Note: In practice, geographic factor should not be used to do the stratification. It is used here only for the sake of illustration.
   a) Two random rows will be drawn (from among all 30 rows). Everybody in the two selected rows (40 of them) will get a prize. **[1 mark]**
   b) Each person entering the theatre will have to write down their names on a piece of paper and then put the paper in a big bag. Twenty (20) names were randomly drawn for the prize. **[1 mark]**
   c) Two random winners will be drawn in each row, for all 30 rows. **[1 mark]**
   d) The first 30 movie goers who enter the theatre get a prize. **[1 mark]**
   e) A random person is chosen among the first five who enter the theatre and prizes are given to every 5^{th} person thereafter. **[1 mark]**

6. General public underestimate the power of statistics. In particular, misuse of statistics could have a devastating impact on individuals or organizations.
   Your task is to do a Google search about the 1936 US Presidential Election between the incumbent Democratic candidate Franklin D. Roosevelt and Republican candidate Alf Landon. You want to pay special attention to the pre-election prediction between the two organizations – *Literary Digest* and *Gallup Poll*. Answer the following questions.
   Note: You might get slightly different values from different websites. So, there are no standard answers here.
   a) What was the success rate (predicting the correct election outcome) of *Literary Digest* prior to 1936? **[1 mark]**
   b) What were the 1936 pre-election prediction by both *Literary Digest* and *Gallup Poll*? Please focus on your answer as 1) who would win the 1936 Presidential election and by what percentage of popularity vote. **[2 marks]**
   c) What was the official result of the 1936 US Presidential Election? **[1 mark]**

   Now let us focus on the *Literary Digest* only from this point on.
   d) In relation to selection bias, what did the *Literary Digest* do (or not do) to wrongly predict the election results? Please provide as much details as possible. **[2 marks]**
   e) In relation to non-response bias, what did the *Literary Digest* do (or not do) to wrongly predict the election results? Please provide as much details as possible. **[2 marks]**
   f) In relation to response bias, what did the *Literary Digest* do (or not do) to wrongly predict the election results? Note that The Great Depression started in 1929 and last till around 1939. So, general public did not know when the Great Depression would end. So, imagine to whom you would point your fingers when time is tough. **[2 marks]**
   g) Now focus on modern days. What are your thoughts about 2016 US Presidential Election between Democratic candidate Hilary Clinton and Republican candidate Donald Trump, in relation to the three biases that we have learned? **[2 marks]**

From this point on, you are expected to use R (no python, no Excel and no other tools).

7. A police officer from Vancouver Police Department wanted to find out the percentage of drivers who were distracted (defined as using their phone while driving or waiting at the traffic lights) during the day. The officer took a random sample of 80 drivers to investigate this. The results can be found in "DANA4800_HW1_Q07_Data.xlsx" on BrightSpace.
   a) Create a Frequency Table of the variable "Distracted" using the *table()* function. **[1 mark]**
   Note: When copy-and-pasting text output from R to Word document, for example, make sure you use "fixed-width fonts", like **`Courier New`**. Otherwise, the output does not look right or aligned properly.
   b) Create a Probability Table of the variable "Distracted" using the *proportions()* function. **[1 mark]**
   c) Provide a description of the parameter of interest. **[2 marks]**
   d) Provide a description of the corresponding statistic. **[2 marks]**
   e) Calculate the value of the most appropriate statistic. **[1 mark]**
   f) Produce a Pie Chart using the *pie()* function, with the "clockwise" arguments set as TRUE, and "Yes" goes before "No". Please also submit the code to produce such graph using fixed-width fonts. **[2+1 marks]**

Note: Please make sure you personalize the pie chart by including the Main Title, and add labels to axes (if applicable) etc.

g) Provide a description of the above pie chart. **[1 mark]**

h) Produce a Bar Graph using the *barplot()* function, "Yes" goes to the left of "No". Please also submit the code to produce such graph. Please use fixed-width fonts. **[2+1 marks]**
Note: A graph directly copied from Excel without any annotation will get a zero.

i) Which graph is better to use here? Briefly justify your answer using statistical reasoning. **[0+1 mark]**

Question #8 is meant to be for you to practice the calculations manually. That said, feel free to use R (or Excel) to double-check your work before submission.

8. Whenever there is a major concert in a city, the hotel rate during that time normally go up. A random sample of 20 hotels in downtown Vancouver was drawn during the time of major concert and the rate of a hotel room per night (based on two double-bed rooms) was recorded.

286	378	245	292	244	314	298	282	281	317
319	237	289	275	285	227	270	322	274	293

a) Identify the subjects of interest. **[1 mark]**

b) Calculate the average hotel room rate manually. Please show all work. **[1 mark]**

c) Provide a description of the statistic in part (b). **[2 marks]**

d) Find the median hotel room rate manually. Please show all work. **[2 marks]**

e) Use the method in our notes to find the first quartile and the third quartile. Then find the interquartile range. **[1+1+1 marks]**

f) Determine if there is/are any outlier(s), using IQR. **[3 marks]**

The following two parts require the use of R.

g) Use the *summary()* function to find the five-number summary of the hotel rates. **[1 mark]**

9. A first-year Langara student wanted to know the weekly expenses (in CAD$) of typical Langara students. To investigate this, she got a random sample of 100 students this term. The data set is in "Expense" worksheet of the file "DANA4800_HW1_Q09_Data.xlsx" on BrightSpace.

a) Use the *summary()* function to find the Five-Number Summary. **[1 mark]**

b) Use the *hist()* functions to produce a Histogram with proportion on the y-axis with 6 bars only. Specifically, please make sure there are 6 bars (1-10, 11-20, …, 51-60), chart title included and axes labelled properly. **[4 marks]**
Note: Make sure you label your graph and axes appropriately.

c) Provide a description of the above graph. **[2 marks]**

d) Use the following segment of code to produce a Density Curve. Provide a description of the shape (only). **[2+1 marks]**

e) Use the *boxplot()* function to make a horizontal Boxplot. **[2 marks]**
Note: Make sure you label your graph and axes appropriately.

10. A dietician wanted to find out how the total fat content (**Fat**; measured in grams per serving) is dependent on the amount of calories (**Calories**; measured in calories) among chicken burgers made from different fast food chains in Canada. A random sample of 20 chicken burgers was collected from different fast food chains (one burger per fast food chain) and the information was recorded. The data set is from "DANA4800_HW1_Q10_Data.xlsx" on BrightSpace.
    a) Provide a description of the subjects of interest. **[1 mark]**
    b) Identify the role (or use) of the two variables. **[1 mark]**
    c) Use the *plot()* function to produce a scatterplot, based on the roles you defined in the above part. **[2 marks]**
    Note: Make sure you label your graph and axes appropriately.
    d) Provide a description of the above Scatterplot. **[2 marks]**
    Note: Please make sure the title and axes are properly labeled.
    e) Use the *cov()* function to find the Variance-Covariance matrix. Please keep one decimal place only and identify which number is the covariance and which numbers are the variances of what. **[1+1 marks]**
    f) Use the cor() function to find the Correlation Coefficient. Please keep four decimal places and identify the value of the correlation coefficient. **[1 mark]**
    g) Provide the description the above correlation coefficient (or most appropriate statistic in this situation). **[2 marks]**

11. Trying to accurately allocate labour hours in a moving job, the manager of a moving company would like to develop a method of predicting the labour hours (**Labour**; measured in hours) based on the size of the high-rise apartment (**Size**; measured in cubic feet). A random sample of 25 high-rise apartment moves was randomly selected in downtown Vancouver in the previous calendar year. The data set is in "DANA4800_HW1_Q11_Data.xlsx" on BrightSpace.
    a) Provide a description of the subjects of interest. **[1 mark]**
    b) Identify the role (or use) of the two variables. **[1 mark]**
    c) Use the *plot()* function to produce a Scatterplot, based on the roles you defined in the above part. **[2 marks]**
    Note: Make sure you label your graph and axes appropriately.
    d) Provide a description of the above scatter diagram. **[2 marks]**
    e) Use the cor() function to find the correlation coefficient. Please keep four decimal places and identify the value of the correlation coefficient. **[1 mark]**
    f) Upon seeing the above correlation coefficient, an assistant reported it to the manager and said the following. Identify two major flaws of the statements. Briefly justify your answers. **[2+2 marks]**
    "Because the correlation coefficient 0.8857 cubic feet per hour is close to one, the reason of working long work hour is because of the high-rise apartment size only. "
    g) Provide the description the most appropriate statistic used in this situation (about apartment moving). **[2 marks]**

12. The PopularKids data set was about opinions of a group of primary school students, who were stratified by their origin (rural, suburban and urban). More information about the data set can be in the following link: https://www.openml.org/search?type=data&sort=runs&id=1100&status=active. The data set is in "DANA4800_HW1_Q12_Data.xlsx" on BrightSpace.
*Note: In this question, let us only use **Gender** (boy and girl) as the row variable and **Goal** (Grades, Popular, and Sports) as the column variable. Every subsequent mentioning of "row" and "column" refer to this definition.*
    a) Use the *table()* function to produce a Two-Way Table (or Contingency Table) with frequency. **[1 mark]**
    b) Use the frequency table from part (a), calculate and enter expected frequencies in the following table. Please keep one decimal place in all entries. **[2 marks]**

		Goals			
		Grade	**Popular**	**Sports**	
Gender	**Boy**	**117 ( )**	**50 ( )**	**60 ( )**	**227**
	Girl	**130 ( )**	**91 ( )**	**30 ( )**	**251**
		247	**141**	**90**	**478**

    c) Manually calculate the χ^2-statistic. **[3 marks]**

Please use the above two-way table with frequency to answer the following 3 questions.
Hint: You are expected to do this manually. But you could also use the margin.table() function to find the marginal totals first. There is an argument called MARGIN with three options. Please look up the R documentation for details.

    d) Find the percentage of students who are boys and their main goal is being popular. **[1 mark]**
    e) Find the percentage of boys whose main goal is being popular. **[1 mark]**
    f) Among the students whose main goal is being popular, find the percentage of them who are boys. **[1 mark]**

Note: There are the same 3 MARGIN options in the proportion() function. Please look up the R documentation for details.

    g) Use the *proportions()* function to produce a two-way table with Table Percentages. Please keep only two decimal places. **[1 mark]**
    h) Use the *proportions()* function to produce a two-way table with Row Percentages. Please keep only two decimal places. **[1 mark]**
    i) Use the *proportions()* function to produce a two-way table with Column Percentages. Please keep only two decimal places. **[1 mark]**
    j) Use the *barplot()* function to produce a Side-by-Side Bar Graph, with the variable Goals on the x-axis, column percentages on the y-axis, and including a legend. The title of the graph should say "Side-by-side Bar Graph of Goals by Gender". **[2 marks]**
    k) Provide a description of the above graph. **[1 mark]**