DANA 4800
Midterm Test I
Thursday June 12th, 2025

Student Name (Please circle your last name): _____

Student Number: _____

Read the following instructions carefully:

- Please turn off your cell phone and other communication device (including but not limited to laptop computers and tablets etc.) Any noise (ringing or noise from vibration) any device of yours makes during the exam will result in an automatic zero in the midterm.
- This is a closed book examination, but you are allowed to use a hand-held calculator.
- Any form of communication (like peeking over neighbour's paper, reading the questions out loud to yourself etc. in any languages) is not permitted during the exam, including sharing of calculator or stationery etc. Offender's midterm paper will be given a mark of zero and the incident will be reported to the Student Conduct and Academic Integrity Office.
- You could use either pens or pencils as long as your handwriting is legible. Remember, if I cannot read it, I cannot mark it.
- The criteria of marking are based on what you have provided, not what the marker thinks you know about the material.
- Marks might be deducted from wrong spelling or incorrect/incomplete names of proper statistical terminologies/jargons.
- Full credits might not be given for unsupported answers.
- Please answer all questions in plain language and in the context of the question.
- The duration of this midterm examination is 60 minutes.

Q1	Q2	Q3	Q4	Q5	Total
					/30

1. **[6 Total Marks]** Jonas, a recent college graduate, wanted to use his statistical learning to help purchasing a used car. He went to autotrader.ca to collect the data and part of the results are summarized in the following table.

Make	Year	Mileage	Price	Transmission	Condition	Owners	MPG
Honda	2010	120,000	$ 7,000	Automatic	Good	2	27.6
Ford	2012	90,000	$ 8,000	Manual	Fair	3	15.6
Chevrolet	2015	80,000	$ 12,000	Automatic	Excellent	1	16
Toyota	2018	30,000	$ 15,000	Automatic	Good	2	29.1
Nissan	2017	50,000	$ 14,000	Manual	Fair	3	28.7
Honda	2011	110,000	$ 7,500	Automatic	Good	2	28.1
Chevrolet	2013	70,000	$ 11,000	Manual	Excellent	1	22.5
Toyota	2016	40,000	$ 16,000	Automatic	Good	2	29.4
Ford	2014	60,000	$ 13,000	Manual	Fair	3	16.3
Nissan	2019	20,000	$ 17,000	Automatic	Excellent	1	29.6
Chevrolet	2012	85,000	$ 9,000	Manual	Good	2	22.9
Toyota	2015	45,000	$ 18,000	Automatic	Excellent	1	40.5
Ford	2013	95,000	$ 10,000	Manual	Fair	3	23.1
Nissan	2016	55,000	$ 14,500	Automatic	Good	2	27.8
Chevrolet	2018	35,000	$ 20,000	Manual	Excellent	1	18.1

In each of the following situations, (1) identify the best graph to summarize the data and (2) identify the best statistic to summarize the data. No reasoning is required.

a) Jonas wanted to see how the Price varies with Year. **[1+1 marks]**

Graph: scatterplot
Statistic: correlation coefficient

b) Jonas wanted to examine the relationship between Transmission and Condition. **[1+1 marks]**

Graph: side-by-side bar graph
Statistic: chi-square statistic

c) Jonas wanted to look at the distribution of Mileage. **[1+1 marks]**

Graph: histogram
Statistic: average or mean

2. **[6 Total Marks]** An economist wants to see the average amount of grocery expense in the past 12 months (May 2024 to April 2025) among all British Columbian households. Use the following results to answer the questions below.

Month	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Expense	500	590	640	370	690	210	300	1160	510	670	360	630

a) Find the median. **[1 mark]**

First, we need to arrange the expense in ascending order.

Order	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Expense	**210**	**300**	**360**	**370**	**500**	**510**	**590**	**630**	**640**	**670**	**690**	**1160**

Location of median: i = (12+1)/2 = 6.5
→ Median = (510 + 590)/2 = 550

b) Find the first quartile and the third quartile. **[1+1 marks]**

In ascending order,

Order	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Expense	**210**	**300**	**360**	**370**	**500**	**510**	**590**	**630**	**640**	**670**	**690**	**1160**

Q1 is the median strictly to the left of the overall median = 550. It will have 6 numbers.
Location of Q1: i = (6+1)/2 = 3.5.
→ Q1 = (360 + 370)/2 = 365 [1]
Similar for Q3:
→ Q3 = (640 + 670)/2 = 655 [1]

c) Are there any outliers in the data set? Briefly justify your answer. **[3 marks]**
IQR = Q3 – Q1 = 655 – 365 = 290
LL = Q1 – 1.5*IQR = 365 – 1.5*290 = -70 (or 0 since it doesn't make sense to have negative values) [1]
Since no expense is less than 0 or -70, there are no outliers on the left side. [0.5]
UL = Q3 + 1.5*IQR = 655 + 1.5*290 = 1090 [1]
Since the maximum value 1160 is bigger than the UL of 1090, there is one outlier on the right side. [0.5]

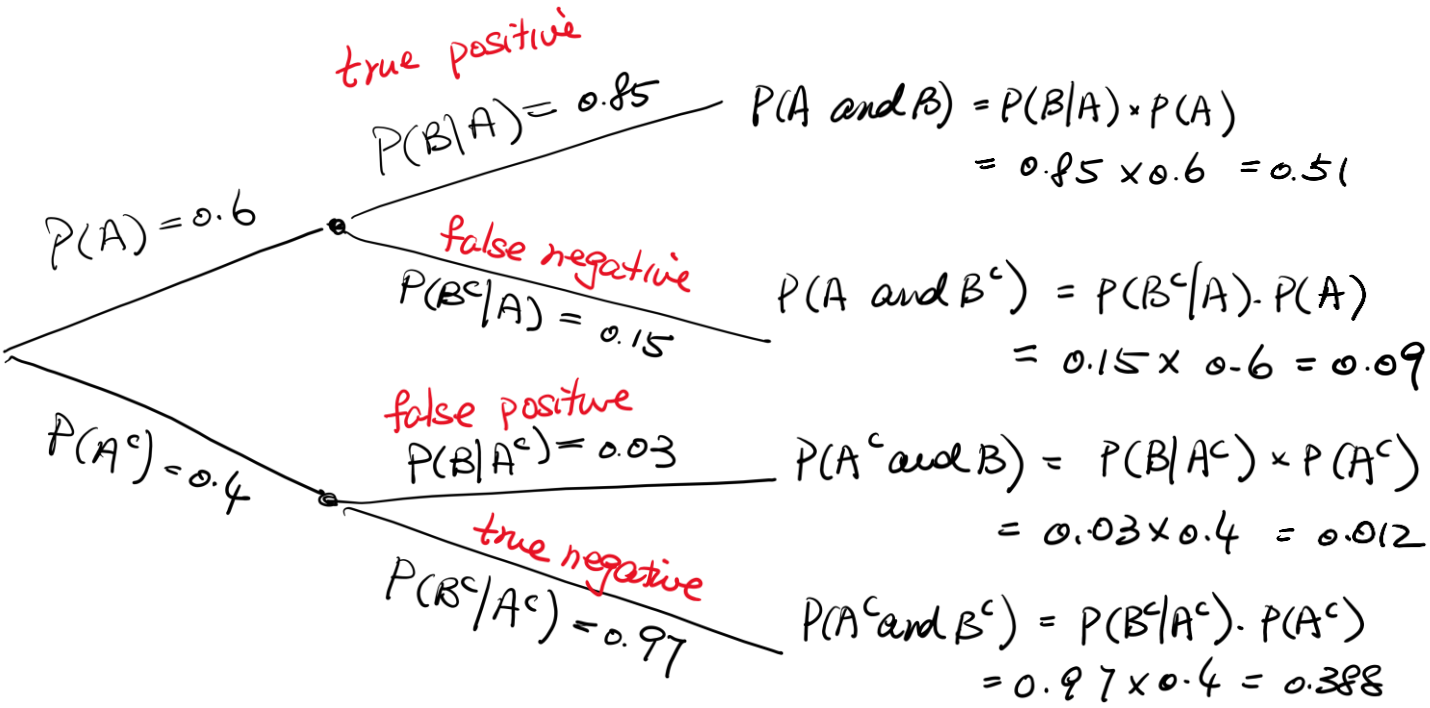
3. **[6 Total Marks]** Langara College is moving to a "centres of excellence" academic framework. As part of the consultation, the Senior Leadership Team (or SLT) of Langara College would want to know the opinion of all Langara employees (instructors and staff members). In each of the following situations,
   (1) identify the closest sampling method they adopted (no reasoning required),
   (2) critique the adopted sampling process, and
   (3) critique the adopted data collection process.
   a) SLT organized a "open house" in a room on a Friday 4:00 pm and employees were asked to drop by, look at the exhibit/presentation and give comments. In particular, every third (3rd) employee exiting the room will be asked in-person if they liked the presentation. **[1+1+1 marks]**

      1) Sampling method: voluntary or self-selected sampling method (Convenience sampling method is acceptable too)
      2) The major problem is selection bias. The target population should be "all Langara employees" but the sampling frame is likely to be "all Langara employees who would be on campus and are willing to take the time to go to the room".
      (Simply mentioning the inconvenient time and day is not enough.)
      3) The major problem is response bias. Employees are asked in-person and face-to-face. If an employee dislikes the presentation, it might be hard for them to say so.

   b) Suppose there are 62 departments in Langara College. Six (6) of the departments were randomly selected and all employees in those six departments were sent an online survey and given a two-day deadline. **[1+1+1 marks]**

      1) Sampling method: cluster sampling method
      2) The major problem is also selection bias, although it seems to be a good sampling method in general. Going to the "centres of excellence" is a college wide initiatives and it should not be left with only employees in the 4 randomly selected departments.
      3) The major problem is probably non-response bias. Employees are sent an online survey but needed to respond in a short time (in 2 days).

4. **[6 Total Marks]** A "Lie Detector" is often used in law enforcement agencies and sometimes in interviews of some high-security jobs. The true positive rate is 85% and the false positive rate is 3%. Suppose that 40% of all cases when using the lie detector are telling the truth. ~~Suppose that 40% of time among all cases in using the lie detector are telling the truth.~~
   a) Define an event A as a person is lying and define an event B as the lie detector gives a positive response (which means lying occurs). Draw a tree diagram showing the above situation. Please make sure you include a probabilities and also compute the four joint probabilities. **[4 marks]**



   b) Suppose the lie detector just showed a positive response (i.e. lying occurs), what is the probability that the person did not lie. **[2 marks]**

$$\boldsymbol{P(not\ lying \mid positive\ response) = P(A^c|B) = \frac{P(A^c\ and\ B)}{P(B)} = \frac{0.012}{0.012+0.51} = \frac{0.012}{0.522} = 0.023}$$

5. **[7 Total Marks]** As of early June 2025, it was reported that there were 69 active wildfires in the province of British Columbia and 41 were classified as "out of control". A reporter wanted to find out the percentage of British Columbians who are concern with the poor air quality due to the smoke particles from the forest fires. A random sample of 200 British Columbians was drawn to investigate this.
   a) Identify the subjects of interest. **[1 mark]**

      British Columbians (in June 2025)

   b) Provide a description of the variable of interest and identify its type. **[2 marks]**
      Note: Please do not forget any important words.

      Variable: Whether or not British Columbians are concerns with the poor air quality due to the smoke particles from the forest fires.

   c) Provide a description of the most appropriate statistic. **[2 marks]**

      Statistic: the proportion of 200 randomly selected British Columbians who are concerned with the poor air quality due to the smoke particles from the forest fires

   d) What is the best graph to summarize the above data? Briefly justify your choice. **[1+1 marks]**

      Graph: pie chart
      Justification: one categorical variable with nominal scale of measurement