STATISTICS 200 - Section 202
PRACTICE FE
APRIL, 2012

Student Name (Please print): ....

Student Number: ....

Lab Section (please, circle the one you are registered in):

L2C/L2D (Tue 1pm)	**L2E/L2F (Wed 12 pm)**	**L2G/L2H (Fri 12pm)**
L2I/L2J (Fri 1pm)	**L2K/L2L (Fri 3pm)**	**L2M/L2N (Fri 10am)**

Please, read the following instructions carefully:

1. This exam is 2.5 hour long.
2. This exam is closed book. Two handwritten 8.5" by 11" (double-sided) cheat sheet are allowed.
3. Scientific calculators are allowed. Graphing calculators, programmable calculators and cellular phones are not permitted.
4. Answer all questions in the space provided. If more space is required, use the back of the previous page and clearly indicate that you are doing so.
5. For full marks, all appropriate work must be shown unless stated otherwise.
6. There are a total of 10 pages (including cover page) and 8 questions.
7. Total mark = 150.

Questions

1. A sample of 200 UBC students answered the following questions

Question	Answer	Variable Name
- Where do you leave?	(a) Vancouver (b) Richmond (c) Burnaby (d) Surrey (e) Other	X
- How old are you?	Age in years	Y
- How many people leave in your household?	Number of people (including the student himself)	V

   (a) Variable X can be well summarized using its sample mean and standard deviation **[2 mark]**

   T F

   (b) Variable X can be displayed using a (choose the appropriate ones) **[2 mark]**

   barplot histogram boxplot pie chart

   (c) Variable Y can be displayed using a (choose the appropriate ones) **[2 mark]**

   barplot histogram boxplot pie chart

   (d) The Person correlation between X and Y is 0.5 **[2 mark]**

   T F

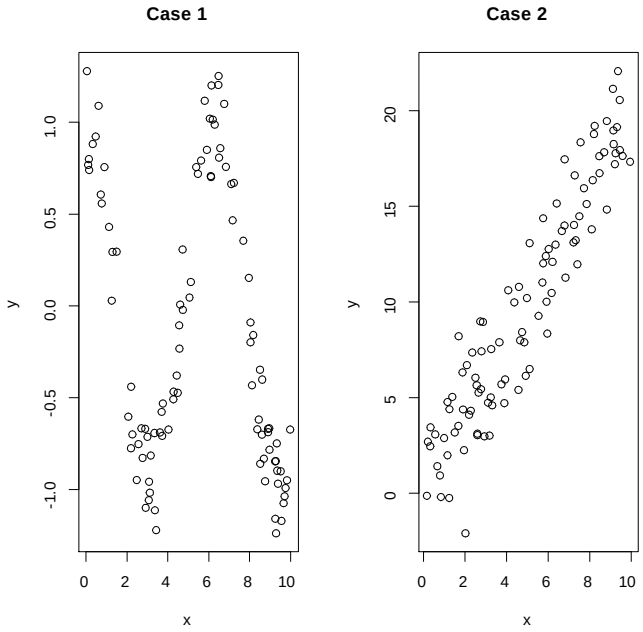
   (e) Variable Y and V can be described using their (choose the appropriate ones) **[2 mark]**

   scatter plot frequency table list of univariate summaries Pearson correlation

   (f) If the Pearson Correlation coefficient between Y and V is near zero we can conclude that these two variables are not associated **[2 mark]**

   T F

2. Consider the following plots. The datasets sizes in both cases are equal.



(a) Circle your choice:

(i) A linear regression line provides a reasonable fit to the data. **[2]**

Case 1 only Case 2 only Both cases Neither case

(ii) The corresponding residual plot shows a random pattern. **[2]**

Case 1 only Case 2 only Both cases Neither case

(iii) There is a very weak association between the explanatory and response variables. **[2]**

Case 1 only Case 2 only Both cases Neither case

(b) Suppose that least squares lines are fitted to both datasets and residual are computed for both fits. Let's denote by e_i the residuals from Case 1 and by r_i the residuals from Case 2.

Circle your choice:

(i) **[2]**

$\sum e_i = 0$ and $\sum r_i \neq 0$	$\sum e_i = 0$ and $\sum r_i = 0$	$\sum e_i \neq 0$ and $\sum r_i = 0$	$\sum e_i \neq 0$ and $\sum r_i \neq 0$

(ii) **[2]**

$\sum e_i^2 > \sum r_i^2$	$\sum e_i^2 < \sum r_i^2$	$\sum e_i^2 = \sum r_i^2$	Cannot say from available information

(iii) The residual plot seems random, with no pattern **[2]**

Case 1 only Case 2 only Both cases Neither case

(c) Fill in the blank. **[2 marks]**

The original number of household occupants, denoted by V includes the interviewed person. A new variable W is defined as W=V-1 to exclude the respondent. The Pearson correlation coefficient between V and W is equal to _____.

(d) Fill in the blanks. **[2 marks]**

The regression line of W on V has slope equal to _____ and intercept equal to _____.

3. The election campaign manager for a Richmond Mayor candidate (called Mr. Doe here) wished to investigate the popularity level of Mr. Doe halfway in the election campaign. **Six hundred and fifty** adult Richmond residents were randomly chosen and asked the question "Do you plan to vote for Mr. Doe in the upcoming election?". The poll results are given in the following table

	Before TV Ad	Right After a 2 Weeks TV Ad
Answered YES	200	50
Answered NO	300	100

(a) Is this study observational or experimental? Briefly justify your answer **[4 marks]**

(b) What are the response and explanatory variables? **[4 marks]**

(c) Compute the conditional frequencies given "Before Ad" and given "After Ad" **[4 marks]**

(d) Are the two variable associated? Why? **[4 marks]**

4. **The percentage of junk mail arriving to a given exchange is 25%.** The number X of "*suspicious words*" in a randomly chosen junk-mail message has Geometric distribution with probability $p = 1/5$:

$$P(X = x|\text{ Junk-Mail}) = \left(1 - \frac{1}{5}\right)^x \frac{1}{5}, \quad x = 0, 1, 2, ... \quad \text{(pmf)}$$

$$P(X \leq x|\text{ Junk-Mail}) = 1 - \left(1 - \frac{1}{5}\right)^{x+1}, \quad x = 0, 1, 2, ... \quad \text{(CDF)}$$

The corresponding conditional distribution when the message is not junk mail has a Geometric distribution with $p = 5/6$:

$$P(X = x|\text{ Not Junk-Mail}) = \left(1 - \frac{5}{6}\right)^x \frac{5}{6} \quad x = 0, 1, 2, ...\text{(pmf)}$$

$$P(X \leq x|\text{ Not Junk-Mail}) = 1 - \left(1 - \frac{5}{6}\right)^{x+1}, \quad x = 0, 1, 2, ... \text{ (CDF)}$$

(a) Calculate $P(X > 1|\text{Junk-Mail})$ **[5 marks]**

(b) Calculate $P(X > 1 \mid \text{Not Junk-Mail})$ **[5 marks]**

(c) A new mail has arrived and it has 1 suspicious words. Should it be called "*Junk Mail*"? Why?. **[5 marks]**

(d) A new mail has arrived and it has 2 suspicious word. Should it be called "*Junk Mail*"? Why?. **[5 marks]**

5. The continuous random variable X is used to model the lifetime of a certain machine. It has the following cumulative distribution function (CDF)

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - \frac{1}{1+x^2} & x > 0 \end{cases}$$

(a) Calculate $P(X > 1)$ **[3 marks]**

(b) Calculate $P(X > 2 \mid X > 1)$ **[6 marks]**

(c) Does this random variable have the "memoryless property"? Why? **[5 marks]**

(d) What is the corresponding probability density function (pdf) ? **[4 marks]**

6. Circle your choice

(a) The sample space contains all the possible outcomes of a random experiment. **[3 marks]**

T F

(b) If A and B are independent events, then

$$P(A \cup B) = 1 - P(A^c) P(B^c) \quad \textbf{[3 marks]}$$

T F

(c) If the events A is a subset of B then A and B are not independent. **[3 marks]**

T F

(d) If the events A and B are mutually exclusive then they must be independent. **[3 marks]**

T F

(e) A geometric random variable is expected to be larger and less variable when p is smaller. **[3 marks]**

T F

(f) An exponential random variable is expected to be smaller and less variable when its rate is larger. **[3 marks]**

T F

(g) If X and Y share the same expected value and variance then they cannot be independent. **[3 marks]**

T F

7. An oil corporation wishes to add a new well. Suppose that each time they drill a well the probability that they strike oil is 0.20. Suppose that building the geological map pointing to the best drilling areas costs \$300,000 and that drilling a well costs \$250,000. Wells are drilled one at the time until a successful well is found.

   (a) What is the expected number of drilled wells? What is the standard deviation? **[3 marks]**

   (b) What is the expected cost for the entire drilling project? **[8 marks]**

   (c) What is the probability that the corporation will drill at least 2 dry wells before reaching oil? **[6 marks]**

   (d) Suppose now timing is very important and that the corporation chooses a different strategy. They decide to drill a total of 4 wells simultaneously. What is the probability that they will meet the required demand? **[8 marks]**

8. A drunkard leaves his bar (located at a corner) and try to walk back to his home. His home is several blocks away on the same street (which has a North-South orientation). It takes him 5 minutes to walk one block (100 meters). Each time he reaches a corner he goes North with prob 5/6 and South with prob 1/6 (he is not sure about going home or going back to his bar). Unfortunately, he is so drunken that he will not recognize neither his home nor his bar when (if) he reaches them.

   (a) Define a random variable that represents his stepwise movement. **[5 marks]**

   (b) Calculate the mean and variance for the variable defined in **Part a**. **[5 marks]**

   (c) Define a random variable that represents the "distance from the bar" after 2 hours and 5 minutes (that is, after 25 steps) **[10 marks]**

   (d) What is the approximated probability that the drunckard will be less than 100 meters away from his bar after 2 hours and 5 minutes? **[10 marks]**