1 Instructions:

- Please do all the questions.
- This is a fully open book assessment.
- This Final Assessment has to be entirely your own effort, and it will be considered a breach of academic integrity if otherwise.
- Please upload your answers in 1 PDF or in 1 Word document onto the Final Assessment Link in the course website by 25th of November Friday.
- If you are handwriting, then scanning, please use blue or black pen and write on white background paper.
- If needed, you may use a standard Normal tables, or use R for any other related computations.
- If you use R or any other mathematical software in any part of your answers, please specify the software used in those sections.

Question 1 [20 marks]

Parts (a), (b) and (c) are separate and unrelated parts of the questions, and can be completed independently of each other.

(a) (8 marks) Let A_1, A_2 and A_3 be events that are not necessarily disjoint. Show with the help of Venn diagrams that

$$P[A_1 \cup A_2 \cup A3] \leq P[A_1] + P[A_2] + P[A_3].$$

(b) (4 marks) Show that $(A \cup B)^c = A^c \cap B^c$ and that $(A \cap B)^c = A^c \cup B^c$.

(c) (6 marks) Assume that A and B are independent events and do not have zero probabilities. Show that
i. $P[A \cap B^c] = P[A]P[B^c]$.
ii. $P[A^c \cap B^c] = P[A^c]P[B^c]$

Question 2 [20 marks]

(a) (20 marks) The probability that an email spam filter incorrectly identifies spam email is 0.01. Suppose we receive 100 spam emails and each occurs independently from email to email. Assume that at most one error (if any) is present in each job. Let the random variable X be the total number of misidentifications from the spam filter.

(i) (2 marks) What would be a suitable distribution for X? State the parameters for the distribution.

(ii) (3 marks) Calculate the probability of no misidentifications.

(iii) (4 marks) Calculate the probability of at least 2 errors.

(iv) (6 marks) If X is approximated by a suitable alternative distribution, recalculate the probability in parts (ii) and (iii). State the most appropriate alternative distribution and give reasons.

(v) (5 marks) Suppose there is a third party app which runs another check independently after to determine if the email is spam, it has a misidentification probability of 0.005. Given the first spam misidentifies an email, then what is the probability the second spam filter identifies the email as spam? Also, what is the probability that the first spam filter misidentifies given the second spam filter correctly identifies?

PLEASE TURN OVER FOR QUESTION 3.

Question 3 [20 marks]

A Las Vegas casino has a card game called Deuce Poker. The game consists of the standard 52 deck of cards where the player is dealt a hand of five cards. A winning hand must consist of a pair, 3 of a kind, 4 of a kind or all 5 cards are of the same suit. However, you are able to discard up to two cards and then draw two more cards in an attempt to achieve a winning hand.

(a) (4 marks) What is the probability that all five cards are *Hearts*?

(b) (2 marks) What distribution properly describes the scenario? State the distribution and parameters.

(c) (4 marks) Produce a random sample of five cards. State them down as '2D' for a 2 of *Diamonds*. Calculate the probability you got this particular hand? *This must be a different hand from part (a).*

(d) (4 marks) Suppose you discard 2 cards from your hand in part (b) and pick up 2 new cards which are both *Hearts*. What is the probability of this hand?

(e) (6 marks) In a poker hand, Nick has a very strong hand and decides to bet 5 dollars. The probability that you have a better hand is 0.04. If you had a better hand you would raise with probability 0.9, but with a poorer hand you would only raise with probability 0.1. If you decide to raise, what is the probability that you have a better hand than Nick?

PLEASE TURN OVER FOR QUESTION 4.

Question 4 [20 marks]

Consider an online store where a number of customers visit and buy a product every hour. Let X be the number of people who enter the store per hour. The store is active for 14 hours per day, every day of the week. It is calculated from data collected that the average number of customers per hour is 10.

(a) (3 marks) When is it appropriate to approximate a Poisson Distributed random variable with a Normal Distribution? State the appropriate parameters for the Normal Distribution.

(b) (7 marks) Calculate the probability that 6 or less customers enter the store. *Hint: Remember to use the correction for continuity and* **R** *to help calculate the probability.*

(c) (3 marks) The manager of the store decides to open up a chain of five stores and it is observed that the average number of customers entering the stores is the same. What is the mean and variance of the distribution for the average of all the five stores?

(d) (7 marks) Calculate the probability for the average of the five stores when 6 or less customers enter?

PLEASE TURN OVER FOR QUESTION 5.

Question 5 [20 marks]

A new exciting wine event is coming to Adelaide since it has been gaining popularity in other major cities. The event has a capacity of 100 people and the number of people each hour is Poisson distributed with $\lambda = 80$ and runs for 3 hours. The patrons purchase bottle of wines for consumption at the event and the total amount spent per patron per hour is assumed to be Normally distributed with mean \$30 and standard deviation \$5.

(a) (4 marks) Based on the information, what is the prior distribution of the mean total amount spent by patrons μ_A per hour?

(b) (6 marks) A concern to the event management is when the average number of patrons exceeds 100, what is the probability that this could occur?

(c) (4 marks) The following is a sample generated from the last city in anticipation of the Adelaide event. What is the distribution which describes the mean total amount spent by patrons per hour based on this sample?

```
> numpatron <- rpois(3,80)
> numpatron
[1] 97 92 71
>  totalamount <-cumsum(rnorm(97,30,5))[97];
+ totalamount2 <-cumsum(rnorm(92,30,5))[92];
+ totalamount3 <-cumsum(rnorm(71,30,5))[71];
>  totalamount;totalamount2;totalamount3
[1] 2887.344
[1] 2786.754
[1] 2103.666
> mean(c(totalamount,totalamount2,totalamount3))
[1] 2592.588
> sd(c(totalamount,totalamount2,totalamount3))
[1] 426.3952
```

(d) (6 marks) By using part (c), compute the posterior distribution of the mean μ_A for the total average spent per hour at the event. *Hint: The following formulae can be used to update the mean and standard deviation.*

$$\tilde{\mu}_A = \frac{\mu_A(s/\sqrt{n_1})^2 + \bar{x}(\sigma_A/\sqrt{n})^2}{(s/\sqrt{n_1})^2 + (\sigma_A/\sqrt{n})^2} \quad \textit{and} \quad \tilde{\sigma}_A = \sqrt{\frac{(s/\sqrt{n_1})^2 \times (\sigma_A/\sqrt{n})^2}{(s/\sqrt{n_1})^2 + (\sigma_A/\sqrt{n})^2}}.$$

PLEASE TURN OVER FOR QUESTION 6.

Question 6 [20 marks]

Suppose you are working at a bank in an analytics department inside a Credit Risk Team. You have been tasked with the challenge to model whether someone defaults on a loan or not. The notion of *default* is the event that a person does not make their next payment obligation for a loan. Your manager requires you analyse the output from a data set 'defaults.csv', which consists of a person's *'Risk_Score'* and whether they default or not. Note that the output is below and you do not need the dataset.

(a) (2 marks) What is the most appropriate model to use for predicting whether a customer defaults or not?

(b) (2 marks) State the independent and dependent variables.

(c) (4 marks) Write a piece of **R** code to fit an appropriate model to this problem and state the family distribution. Give reasons.

(d) (4 marks) Interpret the following confidence interval output from **R**.

```
> confint(default_glm,parm="Risk_Score")
     2.5 %     97.5 %
0.2336791  1.6566246
```

(e) (4 marks) What conclusions can you deduce from the model in relation to Risk Score and whether someone defaults?

(f) (4 marks) By using the output generated by **R** for the model, calculate the corresponding probability to *default* given a person's Risk Score is 4.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9298  -0.5399  -0.4382  -0.3356   2.4794

Coefficients:
             Estimate Std.    Error  z value  Pr(>|z|)
(Intercept)       -6.8451    2.7703  -2.471    0.0135 *
Risk_Score         1.8271    0.9009   2.028    0.0425 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for family taken to be 1)

Null deviance: 30.885  on 61  degrees of freedom
Residual deviance: 24.840  on 60  degrees of freedom
```