University of New South Wales
School of Mathematics and Statistics

MATH5905 Statistical Inference
Term One 2025

Assignment One

Given: Friday 28 February 2025 **Due date: Sunday 16 March 2025**

Instructions: This assignment is to be completed **collaboratively** by a group of **at most** 3 students. Every effort should be made to join or initiate a group. (Only in a case that you were unable to join a group you can present it as an individual assignment.) The same mark will be awarded to each student within the group, unless I have good reasons to believe that a group member did not contribute appropriately. This assignment must be submitted no later than 11:59 pm on Sunday, 16 March 2025. The first page of the submitted PDF should be **this page**. Only one of the group members should submit the PDF file on Moodle, **with the names, student numbers and signatures of the other students in the group clearly indicated on this cover page**. By signing this page you declare that:

I/We declare that this assessment item is my/our own work, except where acknowledged, and has not been submitted for academic credit elsewhere. I/We acknowledge that the assessor of this item may, for the purpose of assessing this item reproduce this assessment item and provide a copy to another member of the University; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking). I/We certify that I/We have read and understood the University Rules in respect of Student Academic Misconduct.

Name	Student No.	Signature	Date

Problem One

a) Suppose that the X and Y are two components of a continuous random vector with a density

$$f_{X,Y}(x,y) = 12xy^3, 0 < x < y, 0 < y < c$$

(and zero else). Here c is unknown.

i) Find c.

ii) Find the marginal density $f_X(x)$ and $F_X(x)$.

iii) Find the marginal density $f_Y(y)$ and $F_Y(y)$.

iv) Find the conditional density $f_{Y|X}(y|x)$.

v) Find the conditional expected value $a(x) = E(Y|X = x)$.

Make sure that you show your working and do not forget to always specify the support of the respective distribution.

b) In the zoom meeting problem from the lecture, show that the probability that if there are 40 participants in the meeting then the chance that two or more share the same birthday, is very close to 90 percent.

Problem Two

A certain river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark X has a distribution function

$$F_X(x) = P(X \leq x) = 1 - \frac{1}{x^3}, 1 \leq x < \infty$$

1. Verify that $F_X(x)$ is a cumulative distribution function
2. Find the density $f_X(x)$ (specify it on the whole real axis)
3. If the (same) low-water mark is reset at 0 and we use a unit of measurement that is $\frac{1}{10}$ of that used previously, express the random variable Z for the new measurement as a function of X. Find the cumulative distribution function and the density of Z.

Problem Three

a) A machine learning model is trained to classify emails as spam or not spam based on certain features. The probability that an email is spam is 0.3. The probability that the model predicts spam given that the email is actually spam, is 0.9. The probability that the model predicts spam given that the email is not spam, is 0.15. If a randomly received email is classified as spam by the model, what is the probability that the email is actually spam?

b) In a Bayesian estimation problem, we sample n i.i.d. observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a population with conditional distribution of each single observation being the geometric distribution

$$f_{X_1|\Theta}(x|\theta) = \theta^x(1-\theta), x = 0, 1, 2, \ldots; 0 < \theta < 1.$$

The parameter θ is considered as random in the interval $\Theta = (0, 1)$ and is interpreted as a probability of success in a success-failure experiment.

i) Interpret in words the conditional distribution of the random variable X_1 given $\Theta = \theta$.

ii) If the prior on Θ is given by $\tau(\theta) = 30\theta^4(1-\theta), 0 < \theta < 1$, show that the posterior distribution $h(\theta|\mathbf{X} = (x_1, x_2, \dots, x_n))$ is also in the Beta family. Hence determine the Bayes estimator of θ with respect to quadratic loss.

Hint: For $\alpha > 0$ and $\beta > 0$ the beta function $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ satisfies $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ where $\Gamma(\alpha) = \int_0^\infty exp(-x)x^{\alpha-1}dx$. A Beta (α, β) distributed random variable X has a density $f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1$, with $E(X) = \alpha/(\alpha+\beta)$.

iii) Seven observations form this distribution were obtained: 2, 3, 5, 3, 5, 4, 2. Using zero-one loss, what is your decision when testing $H_0 : \theta \leq 0.80$ against $H_1 : \theta > 0.80$. (You may use the `integrate` function in `R` or any favourite programming package to answer the question.)

Problem Four

A manager of a large fund has to make a decision about investing or not investing in certain company stock based on its potential long-term profitability. He uses two independent advisory teams with teams of experts. Each team should provide him with an opinion about the profitability. The random outcome X represents the number of teams recommending investing in the stock to their belief (based on their belief in its profitability).

If the investment is not made and the stock is not profitable, or when the investment is made and the stock turns out profitable, nothing is lost. In the manager's own judgement, if the stock turns out to be not profitable and decision is made to invest in it, the loss is equal to four times the cost of not investing when the stock turns out profitable.

The two independent expert teams have a history of forecasting the profitability as follows. If a stock is profitable, each team will independently forecast profitability with probability 5/6 (and no profitability with 1/6). On the other hand, if the stock is not profitable, then each team predicts profitability with probability 1/2. The fund manager will listen to both teams and then make his decisions based on the random outcome X.

a) There are two possible actions in the action space $\mathcal{A} = \{a_0, a_1\}$ where action a_0 is to invest and action a_1 is not to invest. There are two states of nature $\Theta = \{\theta_0, \theta_1\}$ where $\theta_0 = 0$ represents "profitable stock" and $\theta_1 = 1$ represents "stock not profitable". Define the appropriate loss function $L(\theta, a)$ for this problem.

b) Compute the probability mass function (pmf) for X under both states of nature.

c) The complete list of all the non-randomized decisions rules D based on x is given by:

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
$x = 0$	a_0	a_1	a_0	a_1	a_0	a_1	a_0	a_1
$x = 1$	a_0	a_0	a_1	a_1	a_0	a_0	a_1	a_1
$x = 2$	a_0	a_0	a_0	a_0	a_1	a_1	a_1	a_1

For the set of non-randomized decision rules D compute the corresponding risk points.

d) Find the minimax rule(s) among the **non-randomized** rules in D.

e) Sketch the risk set of all **randomized** rules $\mathcal{D}$ generated by the set of rules in D. You might want to use `R` (or your favorite programming language) to make the sketch precise.

f) Suppose there are two decisions rules d and d'. The decision d strictly dominates d' if $R(\theta, d) \le R(\theta, d')$ for all values of θ and $R(\theta, d) < (\theta, d')$ for at least one value θ. Hence, given a choice between d and d' we would always prefer to use d. Any decision rule that is strictly dominated by another decisions rule is said to be inadmissible. Correspondingly, if a decision rule d is not strictly dominated by any other decision rule then it is admissible. Indicate on the risk plot the set of randomized decisions rules that correspond to the fund manager's admissible decision rules.

g) Find the risk point of the minimax rule in the set of randomized decision rules $\mathcal{D}$ and determine its minimax risk. Compare the two minimax risks of the minimax decision rule in D and in $\mathcal{D}$. Comment.

h) Define the minimax rule in the set $\mathcal{D}$ in terms of rules in D.

i) For which prior on $\{\theta_1, \theta_2\}$ is the minimax rule in the set $\mathcal{D}$ also a Bayes rule?

j) Prior to listening to the two teams, the fund manager believes that the stock will be profitable with probability 1/2. Find the Bayes rule and the Bayes risk with respect to his prior.

k) For a small positive $\epsilon = 0.1$, illustrate on the risk set the risk points of all rules which are ϵ-minimax.

Problem Five

The length of life T of a computer chip is a continuous non-negative random variable T with a finite expected value $E(T)$. The survival function is defined as $S(t) = P(T > t)$.

a) Prove that for the expected value it holds: $E(T) = \int_0^\infty S(t)dt$.

b) The hazard function $h_T(t)$ associated with T is $h_T(t) = \lim_{\eta \to 0} \frac{P(t \le T < t+\eta | T \ge t)}{\eta}$.

(In other words, $h_T(t)$ describes the rate of change of the probability that the chip survives a little past time t given that it survives to time t.)

i) Denoting by $F_T(t)$ and $f_T(t)$ the cdf and the density of T respectively, show that

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{dt}\log(1 - F_T(t)) = -\frac{d}{dt}\log(S(t)).$$

ii) Prove that $S(t) = e^{-\int_0^t h_T(x)dx}$.

iii) Verify that the hazard function is a constant when T is exponentially distributed, i.e., $f_T(t) = \beta e^{-t\beta}, t > 0$ where $\beta > 0$ implies that $h_T(t) = \beta$.

iv) Let now T be Pareto distributed with parameter $\theta > 0$, i.e., with a density

$$f_T(t|\theta) = \frac{\theta}{t^{\theta+1}} \text{ when } t \ge 1$$

(and zero else). Find its hazard function.