

MAS2905 Specimen paper

NEWCASTLE UNIVERSITY
SCHOOL OF MATHEMATICS, STATISTICS & PHYSICS

SEMESTER 2 2020/21

MAS2905 Specimen paper
Probability and Bayesian Methods with R
(This document contains solutions)

Time allowed: 24 hours

Your completed work must be submitted by 9:30am
(UK time, GMT+1) on 27/05/2021

You are not expected to work on this assessment for the full 24-hour period. The assessment corresponds in length to a **1 hour** exam under normal exam conditions.

You should submit a single PDF file containing your work via the Canvas 24 hour exam area, taking note of all instructions provided there.

When answering the exam you should observe normal exam conditions and must not communicate/collude with other students taking the paper. The work that you submit must be your own work.

Should you encounter IT difficulties that significantly affect your ability to complete this exam or when uploading your answer(s) you may submit a PEC detailing these.

Candidates should attempt all questions. Marks for each question are indicated. However you are advised that marks may be adjusted in accordance with the University's Moderation and Scaling Policy.

There are THREE questions in Section I and FIVE questions in Section II. There are 100 marks on this exam, 20 marks for Section I and 80 marks for Section II.

Beta distribution: If $X \sim Be(\alpha, \beta)$ then it has density

$$f(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1, \quad \alpha > 0, \beta > 0.$$

Also, $E(X) = \alpha/(\alpha + \beta)$, $Var(X) = \alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$ and $Mode(X) = (\alpha - 1)/(\alpha + \beta - 2)$. Note also that $B(a, b) = (a - 1)!(b - 1)!/(a + b - 1)!$, $a, b \in \mathbb{Z}$.

Beta-Binomial distribution: If $X \sim BetaBin(n, a, b)$ then it has density

$$f(x|n, a, b) = \binom{n}{x} \frac{B(x + a, b + n - x)}{B(a, b)}, \quad x = 0, 1, \dots, n, \quad n \in \mathbb{Z}_+, a > 0, b > 0.$$

Also, $E(X) = na/(a + b)$ and $Var(X) = nab(a + b + n)/(a + b)^2(a + b + 1)$.

Binomial distribution If $X|\theta \sim Bin(k, \theta)$ then it has probability function

$$f(x|\theta) = \binom{k}{x} \theta^x (1 - \theta)^{k-x}, \quad x = 0, 1, \dots, k, \quad k \in \mathbb{Z}_+, 0 < \theta < 1.$$

Also, $E(X) = k\theta$ and $Var(X) = k\theta(1 - \theta)$.

Exponential distribution: If $X \sim Exp(\lambda)$ then it has density

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0.$$

Also, $E(X) = 1/\lambda$ and $Var(X) = 1/\lambda^2$.

Gamma distribution: If $X \sim Ga(\alpha, \lambda)$ then it has density

$$f(x|\alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0, \quad \alpha > 0, \lambda > 0.$$

Also, $E(X) = \alpha/\lambda$, $Var(X) = \alpha/\lambda^2$ and $Mode(X) = (\alpha - 1)/\lambda$ when $\alpha \geq 1$. Note also that $\Gamma(a) = (a - 1)!$, $a \in \mathbb{Z}$.

Normal distribution: If $X \sim N(\mu, 1/\tau)$ then it has density

$$f(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x - \mu)^2\right\}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \tau > 0.$$

Also, $E(X) = \mu$ and $Var(X) = 1/\tau$.

Note: If $X \sim N(0, 1)$ then the following quantiles apply:

x	1.96	2.326	2.366	2.409	2.457	2.484	2.512	2.543	2.576
$\Pr(X < x)$	0.975	0.99	0.991	0.992	0.993	0.9935	0.994	0.9945	0.995

Poisson distribution: If $X \sim Po(\lambda)$ then it has probability function

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \dots, \quad \lambda > 0.$$

Also, $E(X) = \lambda$ and $Var(X) = \lambda$.

Continuous Uniform distribution: If $X \sim U(a, b)$, then it has density

$$f(x|a, b) = \frac{1}{b - a}, \quad a < x < b, \quad -\infty < a, b < \infty.$$

Also, $E(X) = (a + b)/2$ and $Var(X) = (b - a)^2/12$.

SECTION I

I.1. Suppose $X \sim Ga(4, 1)$ (see page 2 of this paper).

(a) Find the probability density function (PDF) for $Y = X/5$.

Answer:

Start with the CDF for Y :

$$\begin{aligned} P(Y \leq y) &= P(X/5 \leq y) \\ &= P(X \leq 5y) \\ &= F_X(5y). \quad (2 \text{ marks}) \end{aligned}$$

Then

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} (F_X(5y)) \\ &= f_X(5y) \times 5 \\ &= 5 \frac{(5y)^3 e^{-(5y)}}{\Gamma(4)} \\ &= \frac{5^4 y^3 e^{-5y}}{\Gamma(4)}, \quad y > 0. \quad (3 \text{ marks}) \end{aligned}$$

(b) Find $E(Y)$.

Answer:

From (a) we can see that $Y \sim Ga(4, 5)$, and so $E(Y) = 0.8$.
(3 marks)

(c) Suppose $Y = \sum_{i=1}^4 Z_i$, where Z_i are identically distributed. Find $E(Z_2)$.

Answer:

Students could perform some simulations in R and approximate, or could realise from earlier parts of this question that if $Y \sim Ga(4, 5)$ then $Z_i \sim Exp(5)$. Therefore, $E(Z_2) = 0.2$. (2 marks)

[10 marks]

I.2. The random variable X has cumulative distribution function (CDF)

$$F(x|\mu, \sigma) = e^{-e^{-(x-\mu)/\sigma}}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \sigma > 0.$$

Suppose we observe $u_1 = 0.201$ and $u_2 = 0.993$, both uniformly distributed on $(0, 1)$. Using u_1 and u_2 , generate two realisations from X with $\mu = 25$ and $\sigma = \mu/2$.

Answer:

Suppose $y = e^{-e^{-(x-\mu)/\sigma}}$. Then we can show that $x = \mu - \sigma \log(-\log y)$, and so

$$F_X^{-1}(x) = \mu - \sigma \log(-\log x). \quad \textbf{(3 marks)}$$

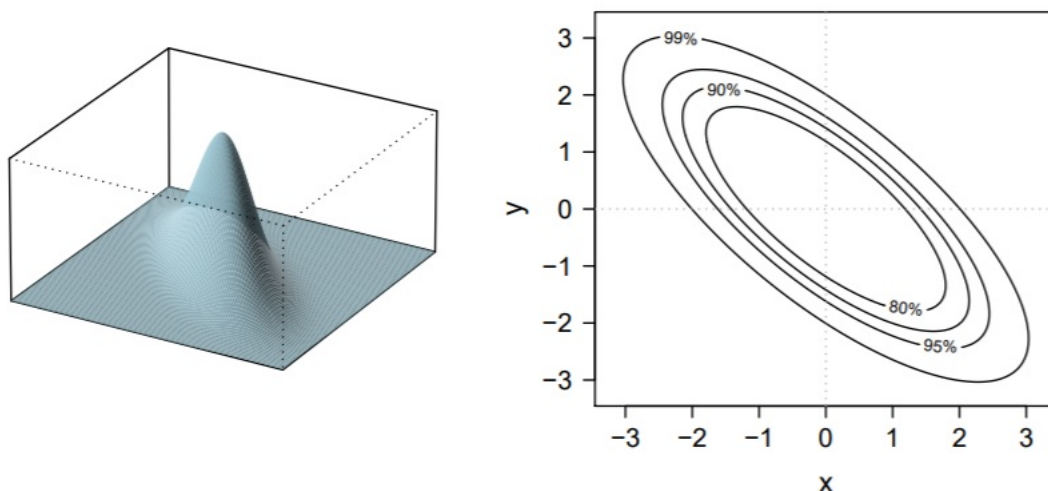
Then, with $u_1 = 0.201$, $u_2 = 0.993$ and $\mu = 25$ and $\sigma = 12.5$, we have

$$x_1 = F_X^{-1}(0.201) = 19.0902 \quad \text{and} \quad \textbf{(1 mark)}$$

$$x_2 = F_X^{-1}(0.993) = 86.9792 \quad \textbf{(1 mark)}$$

[5 marks]

- I.3.** Suppose $(X, Y)^T \sim N_2(\underline{\mu}, \Sigma)$. The figure below shows a plot of the joint density function (left) and a contour plot (right). Look at these plots, and then answer the questions underneath.



- (a) Write down the values of the marginal means and standard deviations.

Answer:

Clearly, we have

$$\mu_x = \mu_y = 0 \quad (1 \text{ mark})$$

and

$$\sigma_x = \sigma_y = 1. \quad (1 \text{ mark})$$

- (b) Given that $\rho = \text{Corr}(X, Y) = -3/4$, find $P(X < 1.4575 | Y = 2)$.

Answer:

We can find that $X|Y = 2 \sim N(-1.5, 1.5625)$ using standard results. Therefore

$$\begin{aligned} P(X < 1.4575 | Y = 2) &= P\left(Z < \frac{1.4575 + 1.5}{\sqrt{1.5625}}\right) \\ &= P(Z < 2.366) = 0.991 \quad (3 \text{ marks}) \end{aligned}$$

[5 marks]

SECTION II

II.4. The manufacturing process for producing *Lego* bricks can suffer from two intermittent faults: plastic overheating (θ_1) and incorrect pressure injection into the moulds (θ_2). Experience suggests that these occur with probabilities 0.3 and 0.7 respectively.

During quality inspections, bricks occasionally exhibit one of the following problems:

x_1 : Incorrect dimensions (measured in thousandths of millimetres)

x_2 : Unsuitable rigidity (measured in Newtons per metre)

Discussions with a *Lego* employee suggest these problems occur with probabilities that depend on the fault, according to the table below:

	x_1	x_2
θ_1	0.25	0.75
θ_2	0.85	0.15

(a) Find the posterior probabilities $\Pr(\theta_i|x_j)$, $i = 1, 2$ and $j = 1, 2$.

Answer:

Using Bayes' Theorem gives:

$$\begin{aligned}
 \Pr(\theta_1|x_1) &= \frac{\Pr(X = x_1|\theta_1)\Pr(\theta_1)}{\Pr(X = x_1|\theta_1)\Pr(\theta_1) + \Pr(X = x_1|\theta_2)\Pr(\theta_2)} \\
 &= \frac{0.25 \times 0.3}{0.25 \times 0.3 + 0.85 \times 0.7} \\
 &= 0.112. \quad \textbf{(3 marks)}
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 \Pr(\theta_1|x_2) &= \frac{\Pr(X = x_2|\theta_1)\Pr(\theta_1)}{\Pr(X = x_2|\theta_1)\Pr(\theta_1) + \Pr(X = x_2|\theta_2)\Pr(\theta_2)} \\
 &= \frac{0.75 \times 0.3}{0.75 \times 0.3 + 0.15 \times 0.7} \\
 &= 0.682. \quad \textbf{(3 marks)}
 \end{aligned}$$

Thus,

$$\Pr(\theta_2|x_1) = 1 - 0.112 = 0.888 \quad \textbf{(1 mark)}$$

and

$$\Pr(\theta_2|x_2) = 1 - 0.682 = 0.318. \quad \textbf{(1 mark)}$$

- (b) Suppose a *Lego* brick displays unsuitable rigidity during a quality inspection. How have our beliefs about any likely faults changed in light of this observation?

Answer:

Prior to this quality inspection, our beliefs were in favour of θ_2 : incorrect pressure (probability 0.7). Having observed x_2 , our beliefs are now in favour of θ_1 : plastic overheating (probability 0.682).

(2 marks)

A nice starter question – similar to one in class and in an assignment.

[10 marks]

- II.5.** (a) Consider a statistical model for data which depends on an unknown parameter μ . Describe the main difference in interpretation of C_F , a 90% frequentist confidence interval for μ , and C_B , a 90% Bayesian confidence interval for μ .

Answer:

For C_B , we have $\Pr(\mu \in C_B) = 0.9$. For C_F , μ is a fixed, but unknown constant and so $\Pr(\mu \in C_F) = 0$ or 1 . C_F has a *repeated sampling* interpretation. **(5 marks)**

This might be described as bookwork. However, this description is not given in the notes; rather, general differences between frequentist and Bayesian inference were *discussed*, verbally, in a preface to the course.

- (b) There were 4 “significant” flooding events in the UK in 2020. The posterior predictive distribution for Y , the number of significant flooding events in 2021, is shown below:

y	0	1	2	3	4	5	6	7	8	9	10
$f(y x=4)$	0.00	0.01	0.04	0.10	0.13	0.45	0.17	0.05	0.03	0.01	0.01

Find the most plausible 90% prediction interval for Y .

Answer:

Using points of greatest probability, we should include $\{5, 6, 4, 3, 7\}$ to obtain 90% coverage; hence, the 90% prediction interval is $(3, 7)$. **(5 marks)**

Simple.

[10 marks]

II.6. The aerobic capacity of 50 randomly chosen athletes was measured, giving a mean of $\bar{x} = 73.5$ ml/kg/min. From a previous study, it is assumed that these measurements are Normally distributed with mean μ and known standard deviation 6.5 ml/kg/min.

(a) The log-likelihood is

$$\log f(\mathbf{x}|\mu) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2.$$

Hence, or otherwise, derive the asymptotic posterior distribution for μ .

Answer:

We have

$$\frac{\partial}{\partial \mu} \log f(\mathbf{x}|\mu) = n\tau\bar{x} - n\tau\mu = 0 \quad \implies \quad \hat{\mu} = \bar{x}.$$

Now

$$J(\mu) = -\frac{\partial^2}{\partial \mu^2} \log f(\mathbf{x}|\mu) = n\tau,$$

and so

$$\begin{aligned} \implies J(\hat{\mu}) &= n\tau \\ \implies J(\hat{\mu})^{-1} &= \frac{1}{n\tau}. \end{aligned}$$

Therefore, for large n , the (approximate) posterior distribution for μ is

$$\mu|\mathbf{x} \sim N(\bar{x}, 1/n\tau).$$

We have

$$\bar{x} = 73.5 \quad \text{and} \quad 1/n\tau = \frac{1}{50 \times 1/6.5^2} = 0.845,$$

giving $\mu|\mathbf{x} \sim N(73.5, 0.845)$. Alternatively (although most students will probably not recognise this), as $n \rightarrow \infty$ we have

$$B \rightarrow \bar{x} \quad \text{and} \quad 1/D \sim 1/n\tau,$$

giving the same results as above. **(9 marks)**

Many students will find this a challenge.

- (b) A sports scientist suggests a prior mean of 65 ml/kg/min with a standard deviation of 4 ml/kg/min. Find the posterior distribution for μ incorporating the sports scientist's beliefs.

Answer:

We have $b = 65$, $d = 1/16$, $n = 50$, $\bar{x} = 73.5$ and $\tau = 1/6.5^2$, giving

$$D = 1/16 + 50 \times 1/6.5^2 = 1.245932$$

and

$$B = \frac{1/16 \times 65 + 50 \times 1/6.5^2 \times 73.5}{1.245932} = 73.074.$$

Thus,

$\mu|\mathbf{x} \sim N(73.074, 1/1.245932 = 0.803)$, as required. **(6 marks)**.

Just substitution, but some might struggle with the *correct* substitution.

- (c) Most top athletes have an aerobic capacity exceeding 75.3ml/kg/min. For each of your posterior distributions in parts (a) and (b), determine the probability that the mean aerobic capacity will exceed 75.3ml/kg/min, and comment.

Answer:

Assuming the asymptotic posterior, we have

$$\begin{aligned} \Pr(\mu > 75.3|\mathbf{x}) &= 1 - \Pr\left(Z \leq \frac{75.3 - 73.5}{\sqrt{0.845}}\right) \\ &= 1 - \Pr(Z \leq 1.96) \\ &= 1 - 0.975 = 0.025. \quad \mathbf{(2 \text{ marks})} \end{aligned}$$

Similarly, assuming the elicited prior gives

$$\begin{aligned} \Pr(\mu > 75.3|\mathbf{x}) &= 1 - \Pr\left(Z \leq \frac{75.3 - 73.074}{\sqrt{0.803}}\right) \\ &= 1 - \Pr(Z \leq 2.484) \\ &\approx 1 - 0.9935 = 0.0065. \quad \mathbf{(2 \text{ marks})} \end{aligned}$$

Both probabilities are small, but *much* smaller under the posterior formed using the elicited prior. This is to be expected as the mean is lower and precision greater, giving a much lighter tail. **(1 mark)**

[20 marks]

II.7. Suppose that the lifetimes X years of patients receiving *Tarceva*, a biological therapy used to treat some types of cancer, follow a Gamma $Ga(3, \theta)$ distribution.

- (a) Derive the likelihood function for θ , given a random sample of lifetimes $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$.

Answer:

We have

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n \frac{\theta^3 x_i^2 e^{-\theta x_i}}{\Gamma(3)} = \frac{\theta^{3n} \prod x_i^2 e^{-\theta n\bar{x}}}{2^n}. \quad \text{4 marks}$$

- (b) You have been working closely with an oncologist to specify a suitable prior distribution for θ . Suppose you use $\theta \sim \text{Exp}(g)$. Determine your posterior distribution for θ using the likelihood you obtained in part (a).

Answer:

We have

$$\pi(\theta) = g e^{-g\theta};$$

therefore

$$\pi(\theta|\mathbf{x}) \propto e^{-g\theta} \times \theta^{3n} e^{-\theta n\bar{x}} = \theta^{3n} e^{-(g+n\bar{x})\theta},$$

and so $\theta|\mathbf{x} \sim Ga(3n + 1, g + n\bar{x})$. **4 marks**

- (c) The oncologist provides you with some data on two patients who have received *Tarceva*. Between them, they have an average lifetime of 1.5 years from the time they started taking the drug.

- (i) Assuming $E(\theta) = 1$, show that, for a 95% Bayesian confidence interval (a, b) for θ ,

$$\int_a^b \theta^6 e^{-4\theta} d\theta = \frac{171}{4096}.$$

[Do not attempt to solve this integral]

Answer:

With $E(\theta) = 1$ we have $g = 1$; we also have $n = 2$ and $\bar{x} = 1.5$, giving $\theta|\mathbf{x} \sim Ga(7, 4)$. Therefore

$$\begin{aligned}\int_a^b \frac{4^7 \theta^6 e^{-4\theta}}{\Gamma(7)} d\theta &= \frac{95}{100} \\ \frac{16384}{720} \int_a^b \theta^6 e^{-4\theta} d\theta &= \frac{95}{100} \\ \int_a^b \theta^6 e^{-4\theta} d\theta &= \frac{95 \times 720}{100 \times 16384} = \frac{171}{4096},\end{aligned}$$

as required. **8 marks**, many students will find this challenging.

- (ii) Using the information in part (i), it can be shown that the 95% highest density interval for θ is $(0.59, 3.07)$. Use this, and the other information obtained throughout this question, to comment on how your beliefs about θ have been updated in light of the data.

Answer:

We can summarise our beliefs as follows:

	Mean	St. Dev.	95% HDI
Prior	1	1	(0, 3)
Posterior	1.75	0.661	(0.59, 3.07)

4 marks can be achieved for any sensible comments relating to the information in the above table.

[20 marks]

- II.8.** Scientists are interested in the migratory path of the rare pink pelican. At various locations the number of birds X , counted before the first pink pelican is observed, is assumed geometric distributed; that is,

$$X_i|\theta \sim \text{Geom}(\theta), \quad i = 1, 2, \dots, n \text{ (independent)}.$$

The probability mass function is

$$f(x|\theta) = (1 - \theta)^x \theta, \quad x = 0, 1, \dots,$$

where θ is the probability of observing the pink pelican.

- (a) Use the Factorisation Theorem to show that $T = \sum_{i=1}^n X_i$ is sufficient for θ .

Answer:

The likelihood is

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^n (1 - \theta)^{x_i} \theta \\ &= 1 \times (1 - \theta)^{\sum_{i=1}^n x_i} \theta^n \\ &= h(\mathbf{x}) g\left(\sum_{i=1}^n x_i, \theta\right), \end{aligned}$$

where $h(\mathbf{x}) = 1$ and $g(t, \theta) = (1 - \theta)^t \theta^n$. Therefore, by the FT, $T = \sum_{i=1}^n X_i$ is sufficient for θ . **(7 marks)**

Not seen for this distribution, but most students should be fine with this.

- (b) Explain, *very briefly*, the role of sufficiency in Bayesian Statistics.

Answer:

It can be shown that $\pi(\theta|\mathbf{x}) = \pi(\theta|T = t)$. Thus, the sufficient statistic is all that's needed from the data to obtain the posterior distribution. **(3 marks)**

Cannot be taken directly from lecture notes - requires some thought. Simple copying from Google will not work here, as a simple, very brief explanation is required.

- (c) A $U(0, 1)$ prior for θ , the probability of observing a pink pelican, is assumed.

- (i) Show that $\pi(\theta|\mathbf{x}) = \pi(\theta|T = t)$, and fully specify this posterior distribution.

Hint: You may use the fact that $T = \sum_{i=1}^n X_i$ follows a negative binomial $NegBin(n, \theta)$ distribution, with probability mass function

$$f_T(t|\theta) = \binom{t+n-1}{t} (1-\theta)^t \theta^n.$$

Answer:

$$\begin{aligned} \pi(\theta|t) &\propto \pi(\theta) f_T(t|\theta) \\ &\propto 1 \times (1-\theta)^{(t+1)-1} \theta^{(n+1)-1}. \end{aligned}$$

Thus, $\theta|t \sim \text{Beta}(n+1, t+1)$. **(3 marks)**

From (a), we know that

$$f_X(x|\theta) = (1-\theta)^{\sum_{i=1}^n x_i} \theta^n,$$

giving

$$\pi(\theta|\mathbf{x}) \propto 1 \times (1-\theta)^{(\sum_{i=1}^n x_i + 1) - 1} \theta^{(n+1) - 1},$$

which is identical to $\pi(\theta|t)$ as given above. **(3 marks)**

Trivial – although students have seen a few examples like this, they’ve not seen this particular example. I anticipate quite a few being caught out here, or not attempting the question at all.

- (ii) The following birds were recorded at two locations, listed in order of appearance:

Location	Type of bird						
A	Egret	Egret	Heron	Brown pelican	Heron	Pink pelican	Egret
B	Bittern	Spoonbill	Pink pelican	Egret	Pink pelican		

Use your answer to part (i) to write down the posterior distribution for θ , and briefly describe how beliefs about θ have changed in light of these data.

Answer:

We have $T = 5 + 2 = 7$ at $n = 2$ locations: $\theta|t \sim \text{Beta}(2+1, 7+1) = \text{Beta}(3, 8)$. **(2 marks)**

We have $E[\theta] = 0.5$ and $E[\theta|t] = 3/11 = 0.272$, suggesting we now believe we are less likely to see the pink pelican at one of these locations. **(2 marks)**

Simple, but I think this application might catch some students out.

[20 marks]

THE END