Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2023	MATH1617-WE01

Title: Statistics I

Time:	2 hours	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Credit will be given for your answers to each question. All questions carry the same marks. Students must use the mathematics specific answer book.

Revision:	

ED01/2023

Q1 A new blood test is developed for a rare, but often fatal disease. The new test is quicker and cheaper than the existing gold standard method (which is essentially perfect), but the new test is sometimes inaccurate. The following data classify 441 patients according to presence or absence of the disease as diagnosed by the gold standard and by the results of the new fast but less accurate blood test.

		Has disease?		
		Yes	No	Total
		D^+	D^-	
Test result positive	T^+	120	29	149
Test result negative	T^-	4	288	292
Total		124	317	441

1.1 Define and calculate the sensitivity and specificity of the test.

1.2 Define and calculate the false positive and false negative rates. Briefly discuss which one is more of a problem in this example and why.

1.3 A patient is randomly selected from the UK population and receives a positive test result. She reads that the average incidence of this disease in the UK population is approximately 1 in 1000. Calculate the probability that she has the disease given the positive test result, $P(D^+|T^+)$. Comment briefly on your answer.

1.4 There are plans for widespread screening of the UK population using this test (the plans involve testing every adult in the UK). Comment on whether you think this is a good idea.

1.5 The patient then reads that the incidence of the disease for women aged between 55-60, of which she is a member, is suspected to be much higher, but is currently unknown. How high would her prior probability $P(D^+)$ have to be before there was a 75% posterior probability that she had the disease?

1.6 Comment briefly on whether the prior found in **Q1.5** is reasonable.

Q2 A set of n Bernoulli trials $X_1, \ldots, X_n$ are performed in a molecular biology experiment, with probability p of success in each trial, where success relates to the creation of a new type of biological enzyme molecule under controlled conditions. The total number of successes over the n trials is summed and represented by $X = \sum_{i=1}^{n} X_i$.

2.1 If data is measured to be $X = x$, give the likelihood as a function of the parameter of interest p.

2.2 Show that the maximum likelihood estimate $\hat{p}_{MLE} = x/n$.

2.3 For $X_i \sim Bernoulli(p)$ we have $\mathrm{E}[X_i] = p$ and $\mathrm{Var}[X_i] = p(1-p)$. Use this to show that

$$\mathrm{E}[\hat{p}_{MLE}] = p \qquad \text{and} \qquad \mathrm{Var}[\hat{p}_{MLE}] = \frac{p(1-p)}{n}$$

2.4 Show that the usual margin of error interval defined as

$$\hat{p}_{MLE} \pm 2\,\mathrm{SD}[\hat{p}_{MLE}]$$

can be bounded by the interval

$$\hat{p}_{MLE} \pm \frac{1}{\sqrt{n}}$$

and comment on the importance of this result.

2.5 The scientist running the experiment has prior beliefs about p, based on previous experiments with related enzymes. They wish to represent their prior beliefs in the form $p \sim Beta(a, b)$. Derive the full posterior pdf for p given data x, in terms of general n, x, a and b. Your answer should include an expression for the proportionality constant and clearly explain your reasoning.

Hint: Let $Y \sim Beta(a, b)$ for $a, b > 0$ known. Then Y has a *Beta distribution* with pdf

$$f(y) \;=\; \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad 0 \le y \le 1$$

and 0 otherwise, where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function, and $\Gamma(a)$ is the Gamma function with $\Gamma(a) = (a-1)!$ for $a \in \mathbb{N}$.

Q3 Suppose that the number of minutes X that a person must wait for a bus each morning has a uniform distribution on the interval $[0, \theta]$, where the value of the endpoint θ is unknown. So we have $X|\theta \sim U[0, \theta]$.

3.1 Sketch a plot of the conditional pdf of X given theta, that is plot $f(x|\theta)$ against x, clearly labelling θ.

3.2 If a single waiting time observation x_1 is obtained, write down the likelihood $\ell(\theta)$ as a function of θ, remembering to specify clearly the values of θ for which the likelihood will be zero. [Hint: Uniform distributions over arbitrary ranges still have to integrate to 1.]

3.3 Sketch a plot of the likelihood $\ell(\theta)$ against θ, carefully labelling the location of the single waiting time x_1.

3.4 If instead, three waiting times observations x_1, x_2, x_3 were made, show that the likelihood $\ell(\theta)$ is now given by:

$$\ell(\theta) \;=\; \begin{cases} \frac{1}{\theta^3} & \text{for } m < \theta, \\ 0 & \text{otherwise,} \end{cases}$$

where you should find an expression for the constant m.

3.5 Suppose that the prior pdf of θ is as follows:

$$f(\theta) = \begin{cases} \frac{6}{\theta^2} & \text{for } 6 < \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Find the posterior pdf of θ given the observations x_1, x_2, x_3, including the normalisation constant. Your answer should involve the constant $c = \max\{m, 6\}$.

3.6 Find the exact $(1 - \alpha)$ HPD and EQT Credible Intervals corresponding to this posterior, and comment on which of these two intervals you think is most appropriate to use here.

Q4 Data on radioactive counts is gathered, where the non-negative, integer data $X_1, \ldots, X_n$ are i.i.d. and known to be Poisson distributed with parameter λ, that is $X_i \sim Po(\lambda)$. The p.m.f. for an individual X_i is therefore given by:

$$f(x_i|\lambda) \;=\; \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \qquad \text{for } x_i = 0, 1, 2, \ldots$$

4.1 Derive the likelihood $\ell(\lambda)$ corresponding to the full set of data $X_1, \ldots, X_n$.

4.2 Find the sufficient statistic(s) for estimating λ, quoting any relevant theorems.

4.3 A random variable Y is said to have a Gamma distribution with parameters $\alpha, \beta > 0$, written $Y \sim Gamma(\alpha, \beta)$, if it has pdf

$$f(y|\alpha, \beta) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \qquad y > 0$$

where $\Gamma(\alpha)$ is the Gamma function with $\Gamma(a) = (a-1)!$ for $a \in \mathbb{N}$. Show that the particular choice of prior $\lambda \sim Gamma(\alpha, \beta)$ is conjugate for the above Poisson likelihood.

4.4 Derive the MAP estimate $\hat{\lambda}_{MAP}$ and compare with the maximum likelihood estimate $\hat{\lambda}_{MLE}$ in the large n limit.

4.5 In what situation does $\hat{\lambda}_{MAP}$ equal $\hat{\lambda}_{MLE}$ exactly? Comment on whether this makes intuitive sense.

Q5 Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from a $N(\mu, \sigma^2 = 1/\tau)$ distribution with mean μ, variance σ^2 and precision $\tau = 1/\sigma^2$. The pdf for a generic random variable $Y \sim N(\mu, \sigma^2 = 1/\tau)$ is given in terms of the mean and precision by

$$f(y|\mu, \tau) \;=\; \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{1}{2}\tau(y-\mu)^2\right\}.$$

5.1 Assume that the population mean μ is unknown, but that $\tau > 0$ is known (and is hence currently not a parameter of interest). The sample is observed to be $x_1, \ldots, x_n$. Show that the likelihood function for μ is given by:

$$\ell(\mu) \;=\; \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\tau \sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

5.2 Using the identity

$$\sum_{i=1}^{n}(x_i-\mu)^2 \;=\; \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$$

show that $\bar{x}$ is sufficient for μ.

5.3 Find the MLE $\hat{\mu}_{MLE}$ of μ.

5.4 Suppose now that the precision τ is also an unknown parameter as well as μ, and therefore that the likelihood $\ell(\mu, \tau)$ is a function of both parameters. By finding the partial derivatives of the log-likelihood with respect to μ, and with respect to τ, find the MLEs $\hat{\mu}_{MLE}$ and $\hat{\tau}_{MLE}$ for μ and τ respectively.

5.5 Use $\hat{\tau}_{MLE}$ to find $\hat{\sigma}_{MLE}$, carefully justifying your argument and comment on the form of $\hat{\sigma}_{MLE}$.