Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	MATH1617-WE01

Title: Statistics I

Time:	2 hours	
Additional Material provided:	Tables: Normal distribution, t-distribution.	
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. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

Q1 A new test has been developed to detect a new strain of *human papillomavirus* (HPV), a major cause of cervical cancer. The following data classify 577 patients according to presence or absence of the HPV strain as diagnosed by a "gold standard" (an expensive and time consuming lab procedure) and by the results of the new fast, but less accurate swab test.

		Has disease? Yes D^+	No D^-	Total
Test result positive	T^+	276	17	293
Test result negative	T^-	15	269	284
Total		291	286	577

(a) Define and calculate the sensitivity and specificity of the test.

(b) A person from the general UK population, who is selected at random, is tested and receives a positive test result. It is thought that about 1/250 people in the UK have this strain of HPV. Calculate the probability the person has the disease given they received a positive test result, $P(D^+|T^+)$. Comment briefly on your answer.

(c) The patient actually had the test done twice and received two positive results, represented by the event T^{++}. Assuming the test results are *conditionally independent given disease status*, calculate the probability that they have the disease after receiving two positive tests, $P(D^+|T^{++})$. Comment briefly on your answer.

(d) Consider a general test with sensitivity x and specificity y, where again the test results are conditionally independent given disease status. Denoting n positive test results as the event T^{n+}, derive an expression for the posterior probability $P(D^+|T^{n+})$.

(e) In the limit of a large number of positive tests, we would hope that the posterior probability $P(D^+|T^{n+})$ would tend to 1. Derive a constraint on x and y that ensures this is true. Comment on your answer.

Q2 A set of n Bernoulli trials $X_1, \ldots, X_n$ are performed in a nuclear physics accelerator experiment, with probability p of success in each trial, where success relates to the creation of a new kind of exotic nucleus. The total number of successes over the n trials is summed and represented by $X = \sum_{i=1}^{n} X_i$.

(a) Give the name of the distribution of X.

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

(c) Derive the maximum likelihood estimate of p, and evaluate it in the case where $n = 20$ and $x = 6$.

(d) The scientist running the experiment has prior beliefs about p, based on previous experiments with related heavy nuclei. They specify a prior p.d.f. for p as

$$f(p) \;=\; 6p(1-p), \qquad 0 \leq p \leq 1$$

Derive the posterior p.d.f. for p, up to a proportionality constant, given $n = 20$ and $x = 6$.

(e) Show that, for general n and x, the prior specification of $p \sim Beta(a, b)$ is conjugate to the likelihood you gave in **Q2**(b). Your answer should clearly explain your reasoning, and contain an expression for the full posterior p.d.f. for p, including an expression for the proportionality constant. **Hint**: *You may wish to refer to the definition of the Beta distribution given at the end of this question.*

(f) Use your results from **Q2**(e) to specify the proportionality constant in **Q2**(d).

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

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

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 you plan to take a simple random sample $X_1, \ldots, X_n$ of size n from a population with mean μ and variance σ^2 but with unknown distribution.

(a) Derive expressions for the expectation of the sample mean $\mathrm{E}\big[\bar{X}\big]$, and the variance of the sample mean $\mathrm{Var}\big[\bar{X}\big]$.

(b) Assume the sample size is $n > 20$. By considering the distribution of the sample mean $\bar{X}$, derive a formula for a general Confidence Interval for the unknown population mean μ, assuming the population variance σ^2 is known. Clearly name any theorems you use.

(c) State the definition of the sample variance s^2 and derive its alternative form:

$$s^2 \;=\; \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2\right]$$

(d) A conservationist is studying the effects of pollution on the development of *Fraxinus*, more commonly known as the ash tree. They gather a random sample of size $n = 24$ of trunk diameters (in cm), summaries of which are given by:

$$\sum_{i=1}^{n} x_i \;=\; 831.6 \qquad\qquad \sum_{i=1}^{n} x_i^2 \;=\; 29011.1$$

The conservationist at first assumes $\sigma = 2.9$cm based on previous studies. Calculate a 99% Confidence Interval for the population mean μ.

(e) The conservationist now wishes to drop this assumption and to view σ as unknown. Calculate a 99% Confidence Interval for the population mean μ in this scenario. You should mention any assumptions that your answer relies upon, and how in principle you would check these assumptions (however you do not need to perform such checks).

(f) Test the hypothesis that $\mu = 36.3$ at the 1% significance level, both when σ is assumed to be $\sigma = 2.9$cm, and when σ is assumed to be unknown.

(g) Further examination of the ash tree sample shows significant departures from Normality, with a long tail to the left. Discuss how this finding impacts your results found in questions **Q3**(d) and **Q3**(e).

Q4 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 $X|\theta \sim U[0, \theta]$.

(a) Sketch a plot of the conditional pdf of X given theta, that is plot $f(x|\theta)$ against x, clearly labelling key points on both axes.

(b) 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.]

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

(d) If instead, four waiting times observations x_1, x_2, x_3, x_4 were made, show that the likelihood $\ell(\theta)$ is now given by:

$$\ell(\theta) \;=\; \begin{cases} \frac{1}{\theta^4} & \text{for } m < \theta, \\ 0 & \text{otherwise,} \end{cases}$$

where you should find an expression for the constant m.

(e) Suppose that the prior pdf of θ is as follows:

$$f(\theta) = \begin{cases} \frac{50}{\theta^3} & \text{for } 5 < \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Find the posterior pdf of θ given the observations x_1, x_2, x_3, x_4, including determining the normalisation constant. Your answer should also involve the constant $c = \max\{m, 5\}$.

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

Q5 Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from a $N(\mu, 1/\tau)$ distribution, where the precision $\tau = 1/\sigma^2$ is assumed known (and hence not a parameter of interest). The pdf for a generic random variable $Y \sim N(\mu, 1/\tau)$ is given by

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

and the corresponding likelihood for μ for the i.i.d. sample $X_1, \ldots, X_n$ is given by

$$\ell(\mu) \;=\; f(x_1, \ldots, x_n|\mu, \tau) \;\propto\; \exp\left\{-\frac{n\tau}{2}(\bar{x}-\mu)^2\right\}.$$

where terms depending only on the data or the known τ have been absorbed into the proportionality sign.

(a) If the prior for μ is judged to be a normal distribution such that $\mu \sim N(m, 1/t)$, show that the posterior for μ is also normal such that

$$\mu|x_1, \ldots, x_n \;\sim\; N(m_1, \frac{1}{t_1}),$$

$$\text{where} \qquad t_1 = t + n\tau, \qquad \text{and} \qquad m_1 = \frac{tm + n\tau\bar{x}}{t_1}.$$

(b) An ornithologist is interested in the value of the mean wing-span, μ, of a newly discovered species of hummingbird. The wing-spans, X, of individual hummingbirds are thought to have a normal distribution for which the value of the mean wing-span μ is unknown but the standard deviation is assumed to be $\sigma = 2$ cm. The ornithologist represents her prior beliefs about μ (based upon previous experience of similar species) by a normal distribution with a mean of $m = 9$ cm and a standard deviation of $v = 1.5$ cm. A sample of $n = 7$ adult hummingbirds are captured at random from the population, measured, and released, and their average wing-span is found to be $\bar{x} = 11.1$ cm. What is the ornithologist's posterior distribution for μ given the data?

(c) Explain why an equal-tailed (EQT) posterior credible interval and a highest posterior density (HPD) credible interval would give the same result for this posterior distribution.

(d) For *both the prior and posterior distribution of* μ, find the 95% EQT credible interval for μ. Comment on the difference between these two intervals.

(e) Consider the general case where the number of hummingbirds sampled, n, becomes very large, but let the mean of the data still be denoted as $\bar{x}$. What is the limiting form of the posterior for μ given the data? Comment on your answer.

(f) Derive the 95% EQT posterior credible interval for μ in this large n limit, and comment on its form in relation to the results of a corresponding Frequentist analysis.