Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2026	MATH1617-WE01

Title: Statistics I

Time:	2 hours	
Additional Material provided:	Tables for Normal and t distributions	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. The indicative marks shown in brackets for the main parts of each question are given as a guide to the weighting the markers expect to apply. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

1. A fast nasal swab test has been developed to detect colonisation with a drug-resistant strain of *Staphylococcus aureus* (*Staphylococcus aureus* is a common bacterium found on the skin and in the nose of many healthy people, and 'colonisation' means carrying it without symptoms, but it can spread to others and sometimes cause serious wound or bloodstream infections in hospital patients). The following data classify 600 individuals according to presence or absence of the *Staphylococcus aureus* strain as diagnosed by a "gold standard" (a perfect but expensive and time consuming lab procedure) and by the results of the swab test.

		Has disease? Yes D^+	No D^-	Total
Test result positive	T^+	108	36	144
Test result negative	T^-	12	444	456
Total		120	480	600

(a) Define and calculate the sensitivity and specificity of the test. **[4]**

(b) Define and calculate the false positive and false negative rates. Briefly discuss which one is more of a problem in this example and why. **[4]**

(c) It is known that the prevalence of the strain for incoming patients to the hospital is 1 in 50. Suppose that the hospital screens a randomly selected incoming patient for the strain, and that they test positive. What is the probability that this patient has the disease given that they received this positive test result? Comment briefly on your answer. **[5]**

(d) The hospital were understandably concerned by the first test result, and have the test carried out on the patient a second time, but this time receive a negative result. The two results are represented by the combined event $T^{+-} = \{T_1^+, T_2^-\}$, where T_1^+ represents the first test (a positive result), and T_2^- represents the second test (a negative result). Assuming the test results are *conditionally independent given disease status*, calculate the probability that the person has the disease after receiving these two test results: $P(D^+|T^{+-})$. Comment briefly on your answer. **[4]**

(e) Without calculation, explain why $P(D^+|T^{+-})$ must lie between $P(D^+|T^+)$ and $P(D^+|T^-)$. **[3]**

---

2. A set of n Bernoulli trials $X_1, \ldots, X_n$ are performed in a microchip fabrication line, with probability p of success in each trial, where success relates to the chip passing a high-stress test. 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. **[1]**

(b) Suppose after observing the n trials we find $X = x$ of them are successes. Give the likelihood as a function of the parameter of interest p. **[2]**

(c) Derive the maximum likelihood estimate of p, and evaluate it in the case where $n = 40$ and $x = 31$. **[5]**

(d) An engineer expresses prior beliefs about p, based on previous stress tests on similar microchips. They wish to represent their prior beliefs in the form $p \sim \text{Beta}(a, b)$. Derive the full posterior pdf for p given data x in terms of general a, b, n, x. Your answer should include an expression for the proportionality constant and clearly explain your reasoning. Note that your answer may make appropriate use of Beta or Gamma functions, such as utilised in the hint below.

Hint: Let $Y \sim \text{Beta}(a, b)$ 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 positive integer a. **[5]**

(e) Now suppose that the engineer instead provides prior information in the form:

$$\text{E}[p] = 0.8, \qquad \text{SD}[p] = 0.1.$$

What choice of prior Beta distribution is consistent with this specification? What would the corresponding posterior expectation and standard deviation be when $n = 40$ and $x = 31$? **[7]**

---

3. 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 $\mathrm{E}\left[\bar{X}\right]$, and the variance of the sample mean $\mathrm{Var}\left[\bar{X}\right]$. **[4]**

(b) Assume the sample size is $n > 20$. By considering the distribution of the sample mean $\bar{X}$, derive a formula for an approximate Confidence Interval for the unknown population mean μ, assuming the population variance σ^2 is known. Clearly name any theorems you use. **[4]**

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

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

(d) A psychologist is studying how quickly people detect a sudden change in their surroundings, such as a new object appearing in a virtual room (a "change detection" task). They gather a random sample of size $n = 22$ participants and record each participant's detection time (in ms), the summaries of which are given by:

$$\sum_{i=1}^{n} x_i \;=\; 612 \qquad\qquad \sum_{i=1}^{n} x_i^2 \;=\; 21108$$

The psychologist at first assumes $\sigma = 7.2$ ms based on previous studies. Calculate a 99% Confidence Interval for the population mean μ. **[3]**

(e) The psychologist 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). **[3]**

(f) Test the hypothesis that $\mu = 34$ at the 1% significance level, both when σ is assumed to be $\sigma = 7.2$ ms, and when σ is assumed to be unknown. **[2]**

(g) Further examination of the reaction times sample shows significant departures from Normality, with a long tail to the right. Discuss how this finding impacts your results found in **Q3**(d) and **Q3**(e). **[2]**

---

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

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

(a) Derive the likelihood $\ell(\lambda)$ corresponding to the full set of data $X_1, \ldots, X_n$. **[4]**

(b) Find the sufficient statistic(s) for estimating λ, quoting any relevant theorems. **[3]**

(c) A random variable Y is said to have a Gamma distribution with parameters $\alpha, \beta > 0$, written $Y \sim \text{Gamma}(\alpha, \beta)$, 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 positive integer a. Show that the particular choice of prior $\lambda \sim \text{Gamma}(\alpha, \beta)$ is conjugate for the above Poisson likelihood. **[4]**

(d) Derive the maximum a-posteriori (MAP) estimate $\hat{\lambda}_{MAP}$ and compare with the maximum likelihood estimate (MLE) $\hat{\lambda}_{MLE}$ in the large n limit. **[5]**

(e) In what situation does $\hat{\lambda}_{MAP}$ equal $\hat{\lambda}_{MLE}$ exactly? Comment on whether this makes intuitive sense. **[4]**

---

5. Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from a $\mathrm{N}(\mu, 1/\tau)$ distribution, where the precision $\tau = 1/\sigma^2$ is assumed known (and hence not a parameter of interest).

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

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

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

(a) An environmental scientist is interested in the value of the mean concentration (in mg/L), μ, of a dissolved nutrient in a remote freshwater lake during the summer. Daily measurements are thought to have a normal distribution for which the value of the mean concentration μ is unknown but the standard deviation is assumed to be $\sigma = 0.8$ mg/L. The environmental scientist represents her prior beliefs about μ (based upon previous experience of similar lakes) by a normal distribution with a mean of $m = 6.5$ mg/L and a standard deviation of $v = 1.2$ mg/L. Measurements from a sample $n = 10$ days over the summer are taken at random, and the sample mean concentration is found to be $\bar{x} = 7.3$ mg/L. What is her posterior distribution for μ given the data? **[5]**

(b) 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. **[2]**

(c) For both the prior and posterior distributions of μ, find the 95% HPD credible interval for μ. Comment on the difference between these two intervals. **[4]**

(d) Great interest lies in whether this freshwater lake has a larger mean concentration μ of the dissolved nutrient than some of the surrounding lakes, which are known to have a mean concentration of 6.8 mg/L. Calculate the probability of this being true, using first the prior distribution and then the posterior distribution. Comment on your answer. **[3]**

(e) For general values of the parameters τ, m, t, n and the data $\bar{x}$, find the limiting form of the posterior distribution of μ in the following situations and give a brief intuitive explanation in each case:

(a) τ, m, n and $\bar{x}$ all fixed, but $t \to \infty$,

(b) τ, m, n and $\bar{x}$ all fixed, but $t \to 0$,

(c) t, m, n and $\bar{x}$ all fixed, but $\tau \to 0$. **[6]**
