Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2022	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. Students must use the mathematics specific answer book.

Revision:	

ED01/2022

Q1 A fast swab test has been developed for the "Omicron Variant" strain of Covid-19. The following data classify 585 patients according to presence or absence of the Omicron Variant 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^+	207	54	261
Test result negative	T^-	18	306	324
Total		225	360	585

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

(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.

(c) A person from the general UK population who is selected at random, is tested and receives a positive test result. In the region of the UK the person is from, it is known that about 1 in 70 people have the disease. Calculate the probability the patient has the disease given they received a positive test result, $P(D^+|T^+)$. Comment briefly on your answer.

(d) 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.

(e) Examine the coherence of Bayes Theorem for this example, by checking if you obtain the same value for $P(D^+|T^{++})$ if you calculate it by updating by both results T^{++} at once, or by updating by the individual results sequentially. Comment briefly on the importance of your conclusion.

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

(a) What is 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 $n = 15$ and $x = 5$.

(d) The scientist running the experiment has prior beliefs about p, based on previous experiments with related plant seeds. They wish to represent their prior beliefs in the form $p \sim Beta(a, b)$. Derive the full posterior pdf for p given $n = 15$ and $x = 5$, in terms of general 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 $a \in \mathbb{N}$.

(e) The scientist now completes the prior specification by choosing values $a = 2$ and $b = 2$. Give the prior expectation and the posterior expectation for p. Comment on your answer in relation to your answer to question (c).

(f) For general values of n, x, a and b, show that the posterior expectation can be written as a linear combination of the prior expectation and the maximum likelihood estimate you derived in question (c).

(g) Hence show, for general values of n, x, a and b, that if say $\hat{p}_{\text{MLE}} < \text{E}[p]$, then the posterior expectation is bounded such that $\hat{p}_{\text{MLE}} < \text{E}[p|x] < \text{E}[p]$.

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) A conservationist is studying the effects of pollution on the development of *Betula pendula*, more commonly known as the silver birch tree. They gather a random sample of size $n = 22$ of trunk diameters (in cm), summaries of which are given by:

$$\sum_{i=1}^{n} x_i \;=\; 796.4 \qquad\qquad \sum_{i=1}^{n} x_i^2 \;=\; 28960.9$$

The conservationist at first assumes $\sigma = 2.2$cm based on previous studies. Calculate a 99% Confidence Interval for the population mean μ.

(d) 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).

(e) Test the hypothesis that $\mu = 37.5$ at the 1% significance level, both when σ is assumed to be $\sigma = 2.2$cm, and when σ is assumed to be unknown.

(f) Further examination of the silver birch tree sample shows significant departures from Normality, with a long tail to the left. Discuss how this finding impacts your results found in questions (c) and (d).

Q4 A non-negative continuous random quantity X, has probability density function given by

$$f(x|\alpha) = 3\alpha^3 x^2 e^{-(\alpha x)^3}$$

with parameter $\alpha > 0$.

(a) Assume that n i.i.d. observations $x_1, \ldots, x_n$ are sampled from this distribution. Derive the likelihood function for α based on these data.

(b) Find the sufficient statistic(s) for estimating α.

(c) Derive the corresponding maximum likelihood estimate of α.

(d) Evaluate the maximum likelihood estimate of α for observed data:

$$\{x_1 = 2.1, x_2 = 4.2, x_3 = 1.7\}$$

(e) Show that for general data $x_1, \ldots, x_n$, conjugate priors for α can be specified by using the probability density function

$$f(a) \propto a^{3\nu} e^{-\tau a^3}$$

for $a > 0$ and with $\nu > 0$ and $\tau > 0$. Your answer should include (up to proportionality constant) the corresponding posterior density function.

(f) Explain carefully how one could interpret the parameters ν and τ of this prior distribution in question (e), in relation to sufficient statistics of the data.

(g) Suppose that, in addition to the n observations $x_1, \ldots, x_n$, for one further measurement it was observed that the corresponding random value $X_{n+1} > c$, for some $c > 0$, so its actual value was not observed but only that it exceeds c. Is it important to take this information into account? If so, find the new form of the likelihood and subsequent posterior. If you consider it not important that this information is taken into account, briefly explain why not.

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 $\mu \sim N(m, 1/t)$, shown 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) A marine biologist is interested in the value of the mean leg-span, μ, of a newly discovered species of deep sea spider crab. The leg-spans, X, of individual spider crabs are thought to have a normal distribution for which the value of the mean leg-span μ is unknown but the standard deviation is assumed to be $\sigma = 8$ cm. The marine biologist represents her prior beliefs about μ (based upon previous experience of similar species) by a normal distribution with a mean of $m = 80$ cm and a standard deviation of $v = 6$ cm. A sample of $n = 8$ spider crabs are captured at random from the population, measured, and released, and their average leg-span is found to be $\bar{x} = 89$ cm. What is her 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 spider crabs 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.