Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2024	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 fast swab test has been developed to detect a new, more aggressive, strain of Tuberculosis. The following data classify 379 patients according to presence or absence of the Tuberculosis 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?		Total
		Yes D^+	No D^-	
Test result positive	T^+	186	15	201
Test result negative	T^-	6	172	178
Total		192	187	379

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 person from the general population of India, who is selected at random, is tested and receives a positive test result. It is thought that about 1/500 people in India have this strain of Tuberculosis. Calculate the probability the person has the disease given they received a positive test result, $P(D^+|T^+)$. Comment briefly on your answer.

1.4 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.

1.5 Derive a constraint on the number n of such positive test results that they would have to receive in a row, in order to ensure the probability that they did indeed have the disease, given such test results, is greater than p_0, where $p_0 \in [0, 1]$ is an arbitrary constant.

1.6 Using the results of part **1.5** or otherwise, find the lowest possible n for $p_0 = 0.8$.

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

2.1 What is the distribution of X?

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

2.3 Derive the maximum likelihood estimate of p, and evaluate it in the case where $n = 20$ and $x = 5$.

2.4 The scientist running the experiment has prior beliefs about p, based on theoretical considerations. They wish to represent their prior beliefs in the form $p \sim Beta(a, b)$. Derive the full posterior pdf for p in the case where $n = 20$ and $x = 5$, but in terms of general a and b. Your answer should include an expression for the proportionality constant and clearly explain your reasoning. **Hint**: *You may wish to refer to the definition of the Beta distribution given at the end of this question.*

2.5 The scientist now specifies additional prior information in terms of the expectation and standard deviation of p in the form:

$$\mathrm{E}[p] = \frac{1}{2} \qquad \text{and} \qquad \mathrm{SD}[p] = \frac{1}{6}$$

What choice of prior Beta distribution is consistent with this specification?

2.6 Find the posterior expectation and variance, corresponding to the prior specified in question **Q2.5**. Discuss your answers in relation to the prior expectation and variance and in terms of the maximum likelihood estimate of p you derived in **Q2.3**.

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

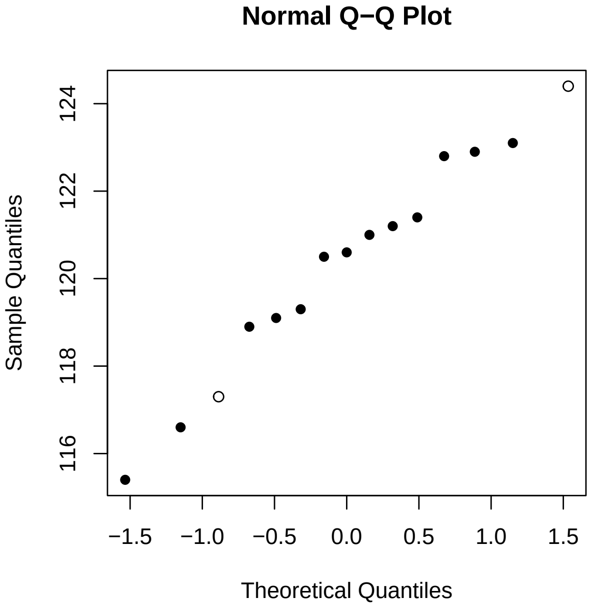
$$f(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad 0 \le y \le 1$$

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 $a \in \mathbb{N}$. We also have:

$$\mathrm{E}[Y] = \frac{a}{a+b}, \qquad \mathrm{Var}[Y] = \frac{ab}{(a+b)^2(a+b+1)}.$$

Q3 A well-known international car company claims that the average CO_2 emission for its new saloon model is 118 g/km. Industry regulators test 15 cars and measure their CO_2 emissions under controlled conditions, yielding data x_i with $i = 1, \ldots, n$. The CO_2 emission data in g/km is given below for the $n = 15$ cars, along with a corresponding normal quantile plot.

CO_2 emissions	121.4	119.3	124.4	120.5	122.9	121.0	123.1	120.6
x_i (g/km)	119.1	122.8	117.3	121.2	116.6	115.4	118.9	



The sum of the data is $\sum_{i=1}^{n} x_i = 1804.5$ while the sum of the squares of the data is $\sum_{i=1}^{n} x_i^2 = 217172.75$.

3.1 Construct a boxplot for the CO_2 emissions data.

3.2 Show how to calculate the coordinates of the 3rd and 15th points on the normal quantile plot (i.e. the two points shown as open circles).

3.3 Construct a 95% Confidence Interval for the mean CO_2 emission μ. State clearly, and justify where possible, any assumptions that you make.

3.4 The industry regulators are concerned that the new saloon model may emit more CO_2 than the car company claims. Test the hypothesis that the mean CO_2 emission for the new saloon model is 118 g/km at the 0.5% significance level using the Confidence Interval approach, and comment on your answer in relation to the regulators' concerns.

3.5 Calculate the corresponding p-value for this hypothesis, up to the level of precision allowed by the attached tables and comment on your result.

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

$$f(x|\alpha) \ = \ 4\alpha^4 x^3 e^{-(\alpha x)^4}$$

with parameter $\alpha > 0$.

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

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

4.3 Derive the corresponding maximum likelihood estimate of α.

4.4 Evaluate the maximum likelihood estimate of α for observed data:

$$\{x_1 = 3.2, x_2 = 4.1, x_3 = 2.5\}$$

4.5 Show that, for general data $x_1, \ \dots \ , x_n$, conjugate priors for α can be specified using the probability density function

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

for $a > 0$ and with $\nu > 0$ and $\tau > 0$, where b is a parameter that you must choose the value of to ensure conjugacy. Your answer should include (up to a proportionality constant) the corresponding posterior density function.

4.6 Explain carefully how one could interpret the parameters ν and τ of this prior distribution in question **Q4.5**, in relation to sufficient statistics of the data.

4.7 Suppose that, in addition to the n observations $x_1, \ \dots, \ x_n$, 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 is lower than c. Is it important to take this information into account? If so, find the new form of the likelihood. 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).

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

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

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

5.1 An entomologist is interested in the value of the mean wing-span, μ, of a newly discovered and extremely rare species of beetle, only found in the Kinabalu National Forest, Malaysia. The wing-spans, X, of individual beetles 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 $\sigma = 0.6$ cm. The entomologist represents her prior beliefs about μ (based upon previous experience of similar species) by a normal distribution with a mean of $m = 4.3$ cm and a standard deviation of $v = 0.4$ cm. A sample of $n = 9$ adult beetles are captured at random from the population, measured, and released, and their average wing-span is found to be $\bar{x} = 5.1$ cm. What is her posterior distribution for μ given the data?

5.2 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?

5.3 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.

5.4 Great interest lies in whether this new rare species of beetle has a larger mean wing-span μ, than a more common competing species which has a known mean wing-span of 4.9 cm. Calculate the probability of this being true, using first the prior distribution and then using the posterior distribution. Comment on your answer.

5.5 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 $\tau \to 0$.

END