Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2018	MATH2041-WE01

Title:

Statistical Concepts II

Time Allowed:	3 hours	
Additional Material provided:	Tables: Normal distribution, t-distribution, chi-squared distribution.	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio fx-83 GTPLUS or Casio fx-85 GTPLUS.
Visiting Students may use dictionaries: No		

Instructions to Candidates:	Credit will be given for: the best **FOUR** answers from Section A and the best **THREE** answers from Section B. Questions in Section B carry **TWICE** as many marks as those in Section A.

Revision:	

ED01/2018

SECTION A

1. (a) An independent and identically distributed sample of size n, $\underline{x} = (x_1, \dots, x_n)$, is drawn from a Poisson distribution with parameter λ.

   (The Poisson distribution, parameter λ, has probability mass function,

$$\mathrm{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, \dots)$$

   Show that the sample mean is the maximum likelihood estimator for λ.

   (b) The number of α particles emitted in one second by a certain radioactive substance is believed to follow a Poisson distribution. A sample of $n = 80$ one second intervals was monitored and for each interval the number of particles emitted was recorded. The data are as follows.

Number of particles emitted	0	1	2	3	4
Observed count	32	27	15	5	1

   Carry out an appropriate goodness of fit test for the Poisson hypothesis.

2. Suppose that a certain treatment is being tested on each of 100 patients. For each patient, the treatment is considered to be either a success or a failure. The quantity of interest is p, the probability that the treatment will be successful. The null hypothesis under test is that $p = 0.4$, and the alternative, for the purpose of this investigation, is that $p = 0.6$.

   (a) Find, approximately, the most powerful test of the hypothesis at significance level 0.01. (State, without proof, any results that you use to demonstrate that the test is most powerful.)

   (b) Find the power of this test.

3. A positive random quantity Y has a gamma distribution if it has probability density function

$$f(y \mid \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha y^{\alpha - 1} e^{-\lambda y}$$

   where the parameters α and λ are both positive.

   (a) Find the expectation of Y.

   (b) Show that the moment generating function of Y is given by

$$M_Y(t) = \left(\frac{\lambda}{\lambda - t} \right)^\alpha \qquad t < \lambda$$

   (c) Deduce that if $Y_1, \dots, Y_n$ is an independent random sample from such a gamma distribution then the sum $Y_1 + \dots + Y_n$ also has a gamma distribution, and give the parameters.

4. Consider the random quantities $X_1, \ldots, X_n$ which are independent and identically Normally distributed with mean μ and variance σ^2.

   (a) State the distribution of the random quantities $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

   (b) Show that the random quantity T given by

   $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

   has a t distribution with $n-1$ degrees of freedom. You should state, but not prove, any additional results that you may require.

   (c) A sample of 28 heart-attack patients had their cholesterol levels, X, measured two days after their heart attack. The sample was found to be approximately Normal, with a sample mean of $\bar{x} = 253.93$ mg/dL and a sample standard deviation of $s = 47.71$ mg/dL.

   Find 95% confidence intervals for the population mean cholesterol level, μ, and for the population variance of the cholesterol level, σ^2.

5. The random quantities X_1 and X_2 are independent and both have Gamma distributions such that $X_1 \sim \mathrm{Ga}(\alpha, \delta)$ and $X_2 \sim \mathrm{Ga}(\beta, \delta)$. A positive random quantity Z has a gamma distribution if it has probability density function

   $$f(z|a,b) = \frac{1}{\Gamma(a)} b^a z^{a-1} e^{-bz}$$

   where the parameters a and b are both positive.

   (a) Determine the joint pdf of X_1 and X_2.

   (b) Suppose we are interested in the transformations $Y_1 = \frac{X_1}{X_1+X_2}$ and $Y_2 = X_1+X_2$. What is the joint pdf of Y_1 and Y_2?

   For what values of Y_1 and Y_2 is this pdf non-zero?

   (c) Find the marginal pdf of Y_1, and hence identify the distribution of Y_1 and state its parameters.

6. (a) Suppose we have two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ that we wish to compare. Define the rank-sum test statistic for testing the null hypothesis of no difference between the two populations.

(b) An experiment is conducted to test whether the installation of cavity-wall insulation reduces the amount of energy consumed in houses. Twenty houses are selected from a housing estate, of which ten are selected at random for insulation. The total energy consumption over one winter is measured for each house. The data, in MWh, are as follows.

No insulation	12.64	11.85	12.82	11.37	14.42	12.32	11.47
	13.18	11.01	11.80				
Insulation	10.81	9.92	9.52	10.02	10.38	10.70	11.77
	7.46	11.82	10.13				

i. Use a nonparametric test to investigate the null hypothesis that the insulation has no effect on energy consumption. Use a Normal approximation to the test statistic and test at the 1% level of significance.

ii. Perform an appropriate t-test to test the same hypothesis and comment on your findings.

SECTION B

7. A certain quantity x can be measured for each member of a population of size N. The values of x in the population are $x_1, \ldots, x_N$, with population mean μ and population variance σ^2.

A random sample $Y_1, \ldots, Y_n$, of size n, is selected without replacement from the population.

(a) Show that $\sigma^2_{\overline{Y}}$, the variance of the sample mean, $\overline{Y}$, is given by

$$\sigma^2_{\overline{Y}} = \frac{(N-n)}{(N-1)} \frac{\sigma^2}{n}$$

(b) Show that the expected value of the sample variance, s^2 is given by

$$E(s^2) = \frac{N}{N-1}\sigma^2$$

(c) In a particular population, $N = 4$, the values of x are $\{1,3,6,8\}$ and the sample size is $n = 2$. Evaluate the sampling distribution of $\overline{Y}$ and of s^2. Hence directly evaluate the variance of $\overline{Y}$ and the expectation of s^2 for this example. Confirm that these answers agree with the general results of part (a), (b).

8. It has been suggested that the sexes of successive births within families may not be independent, and various models that allow for dependence have been proposed. One simple model states that there is a probability θ that successive births are of the same sex. In a study involving n three-child families, the frequencies a, b, c, d, e, f, g and h of the eight possible birth sequences and associated probabilities (based on a model in which the probability of a male on the first birth is 0.5 but subsequently depends on the sex of the previous child) are as follows:

MMM	MMF	MFM	MFF
a	b	c	d
$\frac{1}{2}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}(1-\theta)^2$	$\frac{1}{2}(1-\theta)\theta$
FMM	FMF	FFM	FFF
e	f	g	h
$\frac{1}{2}(1-\theta)\theta$	$\frac{1}{2}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}\theta^2$

(a) Show that the likelihood function for θ is proportional to $\theta^s(1-\theta)^{2n-s}$, where s is the total number of successive births of type MM and FF observed in the n families.

(b) Show that the maximum likelihood estimate of θ is $s/2n$.

(c) Show that the observed information is $8n^3/[s(2n-s)]$.

(d) In a particular study involving 6906 three-child families, the observed frequencies were $a = 953$, $b = 914$, $c = 846$, $d = 845$, $e = 825$, $f = 748$, $g = 852$ and $h = 923$.

Calculate an approximate large-sample 95% confidence interval for θ. State the large sample properties of the maximum likelihood estimator that you are using.

What do you infer from your confidence interval about the hypothesis that successive births of the same sex are more probable than successive births of opposite sexes, in the above model?

9. (a) An online shopping company is interested in the number of purchases made by its customers in 2017. Let $X_i \sim \text{Po}(\lambda)$ be the number of purchases made by customer i, for $i = 1, \ldots, n$. Given λ, each X_i is conditionally independent of X_j for $i \neq j$.

A Gamma distribution with parameters $a > 0$ and $b > 0$ is chosen to represent the prior for λ, with probability density function

$$f(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, \qquad \text{for } \lambda > 0$$

and 0 otherwise.

i. Show that the Gamma distribution is a conjugate prior for λ in this problem, and clearly state the parameters of the posterior distribution for $\lambda \mid x$.

ii. Show that the posterior expectation for λ can be written as a weighted average of the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the prior mean of λ, and give expressions for the weights.

iii. In 2017, the company records that 45 customers made a total of 182 purchases.

Using a prior distribution of $\lambda \sim \text{Ga}(5, 1)$, find $\mathbb{E}[\lambda]$ and $\mathbb{E}[\lambda \mid x]$, compute the values of the weights derived in part (ii), and interpret your findings.

(b) A regression model with a zero intercept term is known as *regression through the origin.* In such a regression, the linear model for y becomes:

$$y_i = \gamma x_i + \varepsilon_i,$$

where the regression errors, ε_i, are i.i.d. Normal with mean 0 and variance σ^2, and the x_i are treated as known.

i. Show that the maximum likelihood estimate, $\widehat{\gamma}$, of γ is given by

$$\widehat{\gamma} = \frac{T_{xy}}{T_{xx}},$$

where $T_{xy} = \sum_{i=1}^{n} x_i y_i$ and $T_{xx} = \sum_{i=1}^{n} x_i^2$.

[The probability density function for $X \sim N(\mu, \sigma^2)$ is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.]$$

ii. Prove that $\widehat{\gamma}$ is an unbiased estimate of γ, and that its variance is

$$\text{Var}[\widehat{\gamma}] = \frac{\sigma^2}{T_{xx}}.$$

iii. State briefly why the least-squares estimate for γ will be the same as the maximum likelihood estimate $\widehat{\gamma}$.

10. ***There is no question 10 on this paper.***