EXAMINATION PAPER

Examination Session: May	**Year:** 2017	**Exam Code:** MATH2041-WE01

Title:

Stastical 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:	

ED01/2017

SECTION A

1. A quantity of interest, x, can be measured on the members of a population of size N, with population mean, μ, and variance, σ^2. The values of x in the population are $x_1, ..., x_N$. A random sample $Y_1, ..., Y_n$, of size n, is selected without replacement from the population.

   (a) State, without proof, the expectation and variance of the sample mean $\overline{Y}$.

   (b) State, without proof, the expected value of the sample variance s^2. Denoting the variance of $\overline{Y}$ by $\sigma^2_{\overline{Y}}$, derive an unbiased estimator, $s^2_{\overline{Y}}$, for $\sigma^2_{\overline{Y}}$.

   (c) In a particular population, $N = 4$, the values of x are $\{4,5,7,9\}$ and the sample size is $n = 2$. Evaluate the sampling distribution of $\overline{Y}$. Hence directly evaluate the mean and variance of $\overline{Y}$ for this example. Confirm that these answers agree with the general results of part (a).

2. In a batch of 2000 items from a manufacturer, each item is either faulty or not faulty. An independent sample of 200 items is tested and 40 items are found to be faulty. Find an approximate 95% confidence interval for q, the proportion of faulty items in the batch. State carefully each of the results and approximations that you use to construct the confidence interval.

3. Consider a sequence of independent trials, each of which will succeed with probability $p \in (0, 1)$ and fail with probability $1 - p$. Let Y be the number of failures in one sequence of trials before the first success, so that Y, given p, has a Geometric distribution

$$P(Y = y|p) = (1 - p)^y p \quad \text{for } y = 0, 1, 2, \ldots$$

   Suppose that p is unknown, and that you evaluate n such sequences of trials, each time counting the number of failures before the first success. Let y_i be the observed number of failures before the first success in sequence i $(i = 1, \ldots, n)$.

   (a) Show that the maximum likelihood estimator, $\hat{p}$, of p is given by

$$\hat{p} = \frac{n}{n + T}$$

   where $T = \Sigma_{i=1}^n y_i$.

   (b) Suppose that $n = 100$ and the following counts were observed.

Number of failures	0	1	2	3	4	5
Observed counts	45	32	10	6	5	2

   (so, in 45 of the 100 experiments, the first trial was a success, etc.) Suppose that you wish to check that the data were generated from a Geometric distribution. Carry out a chi square goodness of fit test of the hypothesis that this data was generated from a Geometric distribution and report your conclusions.

4. Suppose that the time in minutes, X, to serve a customer in Paul's shop has an exponential distribution with pdf

$$f(x|\tau) = \tau e^{-\tau x},$$

for $x \geq 0$. The value of the parameter $\tau > 0$ is unknown, so a prior for τ is chosen to be a Gamma distribution with pdf

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

where the parameters $a > 0$ and $b > 0$ are specified constants.

(a) A collection of n independent serving times $x_1, \ldots, x_n$ are observed. Show that the Gamma distribution is a conjugate prior for samples from an exponential distribution.

(b) Based on past experience, Paul specifies the mean and standard deviation of the Gamma prior as 0.2 and 0.04 respectively. If the average time required to serve a random sample of 20 customers is observed to be 3.2 minutes, what is Paul's posterior distribution for τ?

(c) Suppose that Paul's boss, Mary, instead specifies an improper prior for τ with "pdf" $1/\tau$, for $\tau > 0$. Find Mary's posterior distribution for τ, and show that her posterior mean of τ is $1/3.2$.

5. The following data was a small part of a study to investigate the amount of pain relief (in hours) provided by two analgesic drugs in 10 patients suffering from arthritis. The first row gives the length of pain relief for each patient after taking Drug A, and the second row gives the length of pain relief for each patient after taking Drug B.

Patient	1	2	3	4	5	6	7	8	9	10
Drug A	2.1	3.6	2.7	2.6	7.3	3.4	14.9	6.6	2.3	2.1
Drug B	3.5	5.7	2.9	2.3	9.9	3.3	16.7	6.0	3.8	4.0

(a) Define the Wilcoxon signed rank test statistic for comparing paired samples.

(b) Use the large-sample Wilcoxon signed rank test procedure to assess whether there is a difference between the levels of pain relief provided by drugs A and B.

(c) Assuming the differences follow a Normal distribution, compare your conclusions in part (b) with the results from the appropriate t-test, given that the average of the 10 differences (calculated as $B - A$) is 1.06, and the standard deviation of the 10 differences is 1.1276.

(d) Explain briefly how you would have modified your test procedure if there were ties in the observed differences.

6. Suppose the random quantities X_1 and X_2 are such that $X_1 \sim \text{Gamma}(\alpha, 1)$ and $X_2 \sim \text{Gamma}(\beta, 1)$, where the probability density function for a random quantity $\theta > 0$ following a Gamma distribution with parameters $a > 0$ and $b > 0$ is given by:

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

(a) Let the random quantities Y_1 and Y_2 be defined as:

$$Y_1 = X_1 + X_2, \qquad\qquad Y_2 = \frac{X_1}{X_1 + X_2}.$$

Show that the joint probability density function of Y_1 and Y_2 is given by:

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha+\beta-1} y_2^{\alpha-1} (1 - y_2)^{\beta-1} e^{-y_1},$$

for $y_1 > 0$ and $0 < y_2 < 1$.

(b) Explain whether or not Y_1 and Y_2 are independent.

(c) Show that marginally $Y_1 \sim \text{Gamma}(\alpha + \beta, 1)$ and $Y_2 \sim \text{Beta}(\alpha, \beta)$.

The probability density function for the Beta distribution with parameters v and w is, for $x \in [0, 1]$

$$f(x) = \frac{\Gamma(v + w)}{\Gamma(v)\Gamma(w)} x^{v-1} (1 - x)^{w-1}.$$

SECTION B

7. An independent and identically distributed sample of size n, $\underline{x} = (x_1, \ldots, x_n)$, is drawn from an exponential distribution with parameter λ.

(The exponential distribution, parameter λ has probability density function,

$$f(x|\lambda) = \frac{1}{\lambda}\mathrm{e}^{-x/\lambda},\ x > 0$$

and zero otherwise.)

For this sampling problem, answer the following questions.

(a) Find the maximum likelihood estimator $\hat{\lambda}$ for λ.

(b) Show that $\hat{\lambda}$ is an unbiased estimator for λ, and find the variance of $\hat{\lambda}$.

(c) Find Fisher's information for λ. Deduce, approximately, the variance of $\hat{\lambda}$ for n large.

(d) Compare the approximate assessment of the variance of $\hat{\lambda}$ based on the calculations in part (c) with the precise calculation in part (b). Explain carefully what the above analysis tells you about the optimality of the maximum likelihood estimator for this problem.

(e) Explain two general methods for estimating the variance of the maximum likelihood estimator, one based on estimated information and one based on observed information. Compare the two methods for this problem.

(f) Suppose, in this problem, that we wish to test the null hypothesis that $\lambda = 1$ against the alternative hypothesis that $\lambda \neq 1$. Derive the generalised likelihood ratio test statistic for this test. Explain how to determine critical values for the statistic for large n.

8. (a) Explain what is meant by a uniformly most powerful test of a simple null hypothesis H_0 against a composite alternative hypothesis H_1.

   (b) To check whether or not a die is biased towards 6, it is rolled 100 times. Let X denote the number of times 6 is rolled. Let p denote the probability of outcome 6 per roll of the die. Derive, by approximation, the critical value for the uniformly most powerful test at 5% level of significance for the null hypothesis $H_0 : p = 1/6$ against the alternative $H_1 : p \in (1/6, 1]$.

   (c) Derive the corresponding power function $\pi(p)$ which can be used to compute the power of the above test, for any value $p > 1/6$. Calculate the power of this test for the cases $p = 0.2$ and $p = 0.3$. Briefly discuss these results and, without further calculations, explain the behaviour of $\pi(p)$ for p increasing towards 1.

   (d) Explain what is meant by a monotone likelihood ratio for sampling from a probability distribution with parameter θ. Explain the relationship between a monotone likelihood ratio and uniformly most powerful tests for null hypotheses about the value of θ.

   (e) Show that the binomial distribution, with parameter p, has monotone likelihood ratio. Explain the relationship between this analysis and the results of part (b).

9. The multinomial distribution is the generalisation of the binomial to multiple categories. A vector of counts $\underset{\sim}{n} = (n_1, \ldots, n_k)$ has a multinomial distribution with parameter $\underset{\sim}{p} = (p_1, \ldots, p_k)$ if

$$\mathrm{P}[n_1, \ldots, n_k | p_1, \ldots, p_k] = \frac{N!}{n_1! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i}.$$

where $p_i > 0$, $\sum_{i=1}^{k} p_i = 1$, and $N = \sum_{i=1}^{k} n_i$.

(a) Explain what it means to say that a family of probability distributions is a conjugate prior family for sampling from a particular family of distributions.

Show that the Dirichlet distribution, written $(p_1, \ldots, p_k) \sim \mathrm{Dir}(a_1, \ldots, a_k)$, with pdf

$$f(p_1, \ldots, p_k | a_1, \ldots, a_k) = \frac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\Gamma(a_1) \ldots \Gamma(a_k)} \prod_{i=1}^{k} p_i^{a_i - 1} \propto \prod_{i=1}^{k} p_i^{a_i - 1}.$$

is a conjugate prior for the multinomial distribution, with posterior parameters $\tilde{a}_i = a_i + n_i$.

(b) Consider the case where $k = 3$. By exploiting the fact that $p_3 = 1 - p_1 - p_2$, show that the marginal distribution for p_1 from a Dirichlet distribution is $\mathrm{Beta}(a_1, a_2 + a_3)$.

Hint: The transformation $u = p_2/(1 - p_1)$ will be useful for the integration.

The probability density function for the Beta distribution with parameters v and w is, for $x \in [0, 1]$

$$f(x) = \frac{\Gamma(v + w)}{\Gamma(v)\Gamma(w)} x^{v-1} (1 - x)^{w-1}.$$

(c) In early November 2016, a poll was conducted by YouGov of 3677 likely voters in the USA to find out their voting intentions in the upcoming presidential election. Of the 3677 people, $n_1 = 1655$ supported Hillary Clinton, $n_2 = 1507$ supported Donald Trump, and $n_3 = 515$ either supported another candidate or had not yet chosen a candidate to support.

Assuming no other information on the respondents, simple random sampling, and a non-informative Dirichlet prior with $a_1 = a_2 = a_3 = 1$:

(i) State the posterior distribution for (p_1, p_2, p_3), and by using the result from (b) state the posterior marginal distribution for probability of support for Clinton, p_1, given this sample.

(ii) Find the large-sample Normal approximation to the Beta distribution and its parameters, and use this to construct an equal-tailed 95% credible interval for the posterior probability of support for Clinton, p_1.

(iii) State the form of the limiting posterior distribution for the parameter of a probability distribution, given iid samples from that distribution, as the sample size increases.
Find the limiting form for the posterior distribution for p_1. Compare the parameters of this distribution with those obtained in part (c)(ii), and comment on the adequacy of the limiting approximation.
You may use without proof the fact that Fisher's information for the sample of n observations from $\text{Bin}(n, p)$ is $I(p) = \frac{n}{p(1-p)}$.

10. As part of the Manhattan project in World War II, plutonium for use in atomic weapons was produced at a facility in Hanford, Washington, USA. Over the years significant quantities of radioactive waste has leaked into the nearby river. To investigate the health consequences of this contamination, Fadeley (1965) calculated an index of radioactive exposure for nine nearby counties over the period 1959-1964, with higher values representing higher levels of contamination. Over the same period, the cancer mortality rate was determined as cancer deaths per 100 000 people per year for each of these same counties.

Exposure, x	2.49	2.57	3.41	1.25	1.62	3.83
Mortality, y	147.10	130.10	129.90	113.50	137.50	162.30
Exposure, x	11.64	6.41	8.34			
Mortality, y	207.50	177.90	210.30			

(a) Consider the simple linear regression model for mortality on exposure, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Let the least-squares regression line be $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

i. State the assumptions of the usual simple linear regression model with normal errors.

ii. Give expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$. Prove that they are unbiased estimators of β_0 and β_1.

(b) A simple linear regression model was fitted to the above data using R, as shown on the next page. Use the data and the R output in the following.

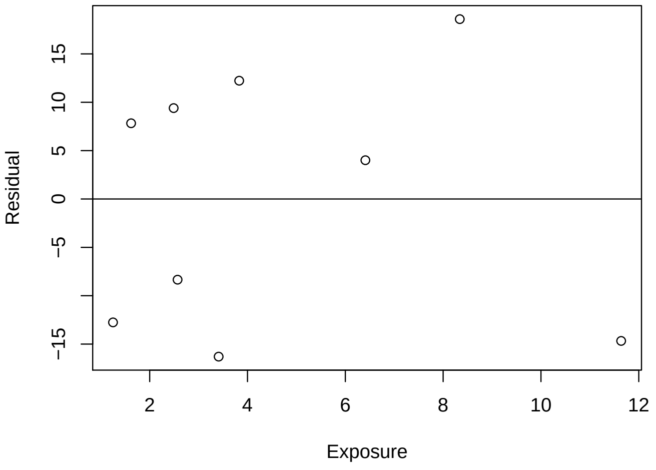
i. What are the estimates of the slope and intercept for the above data? Interpret the t-values for the coefficients. Give a 95% confidence interval for the slope.

ii. Explain how the 'Multiple R-squared' value is calculated, and interpret the corresponding value for this regression from the R output.

iii. For a region with exposure $x^* = 7.50$, estimate the expected value of mortality, y, and give a 95% confidence interval for the expected value.

iv. Explain how the residuals are calculated. Use the residual plot to assess informally the adequacy of the linear regression model as a summary description of these data.

```
> slr <- lm(mortality~exposure)
> summary(slr)

Call:
lm(formula = mortality ~ exposure)

Residuals:
    Min      1Q  Median      3Q     Max
-16.295 -12.755   4.011   9.398  18.594

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  114.716      8.046  14.258 1.98e-06 ***
exposure       9.231      1.419   6.507 0.000332 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.01 on 7 degrees of freedom
Multiple R-squared:  0.8581,Adjusted R-squared:  0.8378
F-statistic: 42.34 on 1 and 7 DF,  p-value: 0.0003321

> mean(exposure)
[1] 4.617778
> var(exposure)
[1] 12.18842
> plot(resid(slr)~exposure)
```


Residual
Exposure