Two hours

Statistical tables to be provided

THE UNIVERSITY OF MANCHESTER

INTRODUCTION TO STATISTICS

21 May 2018

14.00 – 16.00

Answer **ALL FOUR** questions in Section A (10 marks each) and **TWO** of the **THREE** questions in Section B (20 marks each). If more than **TWO** questions from Section B are attempted, then credit will be given for the best **TWO** answers.

Electronic calculators may be used, provided that they cannot store text.

SECTION A

Answer **ALL** four questions

A1. Suppose that we have a sample of observations $x_1, \ldots, x_n$ obtained by random sampling from a continuous distribution with probability density function $f(x)$.

Suppose that $x_1, \ldots, x_n \in (a_1, a_{K+1}]$ and that the range $(a_1, a_{K+1}]$ is divided evenly into K subintervals, or *bins*, $B_k = (a_k, a_{k+1}]$, $k = 1, \ldots, K$, of width h. Recall that the density histogram based on these bins is defined as

$$\text{Hist}(x) = \frac{\nu_k}{nh}\,, \qquad \text{for } x \in B_k\,,$$

and 0 otherwise.

(i) Define what is meant by the notation ν_k, and state an equation for h in terms of a_k and a_{k+1}, for $k = 1, \ldots, K$.

[2 marks]

Recall that the distribution mean (or population mean) is

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx\,,$$

One way of estimating μ is to substitute $\text{Hist}(x)$ as an estimate of $f(x)$, giving

$$\hat{\mu}_{\text{Hist}} = \int_{a_1}^{a_{K+1}} x\, \text{Hist}(x)\, dx\,.$$

(ii) Show that

$$\hat{\mu}_{\text{Hist}} = \frac{1}{2n} \sum_{k=1}^{K} \nu_k (a_k + a_{k+1})\,.$$

[5 marks]

A data set is obtained which has the following frequency table:

Interval	(5,10]	(10,15]	(15,20]	(20,25]	(25,30]	(30,35]
Frequency	1	11	39	38	10	1

(iii) Use the result given in part (ii) to estimate the mean of the distribution that generated these data.

[3 marks]

[Total 10 marks]

A2. Suppose that $\hat{\theta}$ is an estimator of a parameter θ.

(i) Define what it means to say that $\hat{\theta}$ is unbiased.

[3 marks]

Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, and $Y_1, \ldots, Y_m \sim N(\mu, \tau^2)$ independently of each other and of the X_i.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$.

(ii) Show that $\bar{X}$ and $\bar{Y}$ are both unbiased estimators of the common population mean, μ, and calculate $\operatorname{Var}\bar{X}$ and $\operatorname{Var}\bar{Y}$.

[3 marks]

(iii) Show that if $\hat{\mu} = w\bar{X} + (1-w)\bar{Y}$, with $0 \leq w \leq 1$, then $v(w) = \operatorname{Var}(\hat{\mu})$ is minimized when

$$w = \frac{n\tau^2}{m\sigma^2 + n\tau^2}\,.$$

[4 marks]

[Total 10 marks]

A3. Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently. It is desired to test the hypotheses

$$H_0 : \mu = 13.5 \qquad \text{vs.} \qquad H_1 : \mu \neq 13.5\,.$$

(i) Define what is meant by the following terms: Type I error, Type II error, and the significance level, α, of the test.

[3 marks]

Some data are obtained with $n = 8$, $\sum_{i=1}^{n} x_i = 113.6627$, and $\sum_{i=1}^{n} x_i^2 = 1621.391$.

For the following questions, show your working by clearly stating the general formulae for the appropriate test statistic and the critical values of the rejection region, as well as their numerical values in this specific case.

(ii) Do you reject the null hypothesis at the 5% significance level if it is known that $\sigma^2 = 1$?

[3 marks]

(iii) Do you reject the null hypothesis at the 5% significance level if σ^2 is unknown?

[4 marks]

[Total 10 marks]

A4. Let $\mathbf{X} = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n$ an independent random sample from a distribution F_X with unknown parameter θ. Let $I(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$ denote an interval estimator for θ.

(i) Define what it means to say that $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence interval.

[2 marks]

Suppose now that $X_{11}, \ldots, X_{1n} \sim N(\mu_1, \sigma_1^2)$ independently and $X_{21}, \ldots, X_{2m} \sim N(\mu_2, \sigma_2^2)$ independently of each other and of the first sample. It is desired to estimate the parameter $\theta = \mu_1 - \mu_2$. One possible estimator is $\hat{\theta} = \bar{X}_1 - \bar{X}_2$, where $\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$ and $\bar{X}_2 = \frac{1}{m}\sum_{j=1}^{m} X_{2j}$.

(ii) Write down expressions for $\mathrm{E}(\hat{\theta})$ and $\mathrm{Var}(\hat{\theta})$. What is the distribution of $\hat{\theta}$?

[4 marks]

Some data are obtained with

$$n = 10, \quad \sum_{i=1}^{10} x_{1i} = 96.08\,,$$
$$m = 20, \quad \sum_{j=1}^{20} x_{2j} = 237.09\,.$$

(iii) Calculate a 95% confidence interval for $\theta = \mu_1 - \mu_2$ if it is known that $\sigma_1^2 = 2$ and $\sigma_2^2 = 4$. Is it plausible that $\mu_1 = \mu_2$?

[4 marks]

[Total 10 marks]

SECTION B

Answer **TWO** of the three questions

B5. Suppose that $X_1, \ldots, X_n$ is an independent random sample from a distribution with cumulative distribution function F_X, population mean μ and population variance σ^2. The sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denotes the sample mean.

(i) Show that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\,.$$

[6 marks]

(ii) Hence show that S^2 is an unbiased estimator of σ^2.

[4 marks]

Now suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, and let $\mathbf{X} = (X_1, \ldots, X_n)$.

(iii) Show that the interval estimator

$$I(\mathbf{X}) = \left[\frac{(n-1)S^2}{\chi^2_{n-1;\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2. Above, $\chi^2_{k;\alpha}$ denotes the upper α point of a $\chi^2(k)$ distribution, i.e. if $Y \sim \chi^2(k)$, then $\mathrm{P}(Y > \chi^2_{k;\alpha}) = \alpha$.

[6 marks]

Some data are obtained with $n = 10$, $\sum_{i=1}^{n} x_i = 104.334$, and $\sum_{i=1}^{n} x_i^2 = 1132.207$.

(iv) Using these data, calculate a 95% confidence interval for σ^2.

[4 marks]

[Total 20 marks]

B6. Suppose that $X_1, \ldots, X_n$ is an independent random sample from the discrete distribution with probability mass function

$$p_X(x) = (1-p)^x p\,, \qquad x = 0, 1, 2, \ldots\,,$$

where $p \in [0, 1]$ is an unknown parameter.

(i) Show that the likelihood function is

$$L(p) = (1-p)^{\sum_{i=1}^n X_i} p^n\,.$$

[2 marks]

(ii) Hence show that the maximum likelihood estimator of p is

$$\hat{p} = \frac{1}{1+\bar{X}}\,.$$

[8 marks]

(iii) Show that, for $a, b \in (0, 1)$,

$$\mathrm{P}(a < \hat{p} < b) = \mathrm{P}\left(\frac{1-b}{b} < \bar{X} < \frac{1-a}{a}\right)\,.$$

[5 marks]

(iv) Use the result given in part (iii) to calculate the approximate probability that $0.22 < \hat{p} < 0.26$ when $p = 0.25$ and $n = 100$.

[*Hint: you may use without proof the fact that* $\mathrm{E}(X_1) = \frac{1-p}{p}$ *and* $\mathrm{Var}(X_1) = \frac{1-p}{p^2}$.]

[5 marks]

[Total 20 marks]

B7. To investigate the effectiveness of a new treatment for Rhinovirus, a clinical trial is conducted with n patients. Suppose that a patient receiving the new treatment recovers with probability p independently of other patients. For $i = 1, \ldots, n$, let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th patient recovers} \\ 0 & \text{if the } i\text{th patient does not recover} \end{cases} .$$

Given $p_0 \in (0, 1)$, it is desired to test the hypotheses

$$H_0 : p = p_0 \qquad \text{vs.} \qquad H_1 : p > p_0 \, .$$

(i) A natural estimator of p is $\hat{p}$, the sample proportion of patients that recover. State an expression for $\hat{p}$ in terms of the X_i. What is the approximate distribution of $\hat{p}$ if H_0 is true and n is large?

[5 marks]

Define the following notation:

$$Z_1 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \quad \text{and} \quad Z_2 = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \, .$$

The usual way of testing H_0 vs H_1 is to reject H_0 if $Z_1 > z_\alpha$, with z_α the upper α point of a $N(0, 1)$ distribution. This gives a test with significance level α_1, with $\alpha_1 \approx \alpha$ for large n.

Another way of testing H_0 vs H_1 is to reject H_0 when $Z_2 > z_\alpha$. This gives a test with significance level α_2 with $\alpha_2 \approx \alpha$ for large n. Usually α_1 approximates α more closely than α_2 does.

It is known that under the testing procedure based on Z_2, H_0 is rejected if and only if $\hat{p} > \gamma$ where

$$\gamma = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \, , \quad \text{with } a = 1 + \frac{z_\alpha^2}{n} \, , \quad b = -\left(2p_0 + \frac{z_\alpha^2}{n}\right) \, , \quad c = p_0^2 \, .$$

(ii) Using the fact given above, show that if $p > p_0$ and n is large then the probability of rejecting H_0 under the procedure using Z_2 is approximately

$$1 - \Phi\left(\frac{\sqrt{n}(\gamma - p)}{\sqrt{p(1-p)}}\right) \, ,$$

where Φ denotes the c.d.f. of a $N(0, 1)$ distribution.

[5 marks]

(iii) Hence compute the approximate probability of rejecting H_0 when using the procedure based on Z_2 if $\alpha = 0.05$, $n = 200$, $p_0 = 0.3$ and $p = 0.35$.

[5 marks]

(iv) Show that the fact given above is true, i.e. that under the testing procedure based on Z_2, H_0 is rejected if and only if $\hat{p} > \gamma$, with γ defined as above.

[*Hint: consider the equation $Z_2^2 = z_\alpha^2$. You may use without proof the fact that Z_2 is a strictly increasing function of $\hat{p}$ for fixed $p_0 \neq 0, 1$.*]

[5 marks]

[Total 20 marks]

END OF EXAMINATION PAPER