Two hours

Statistical tables to be provided

THE UNIVERSITY OF MANCHESTER

INTRODUCTION TO STATISTICS

30 May 2017

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 $f(x)$.

We wish to calculate a density histogram. 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.

(i) Write down an expression for the function $\text{Hist}(x)$ defining the density histogram based on bins B_k. Explain any notation you use.

[2 marks]

One possible way of estimating the distribution mean,

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

is to substitute $\text{Hist}(x)$ as an estimate of the p.d.f. $f(x)$, i.e.

$$\hat{\mu}_{\text{Hist}} = \int_{-\infty}^{\infty} x \, \text{Hist}(x) dx \,.$$

(ii) Evaluate this integral to obtain an expression for $\hat{\mu}_{\text{Hist}}$.

[5 marks]

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

Interval	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]
Frequency	7	19	38	23	12	1

(iii) Using your answer to part (ii), or otherwise, estimate the mean of the distribution that generated these data.

[3 marks]

[Total 10 marks]

A2. Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, and let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ denote the sample variance.

(i) Show that an equivalent expression for the sample variance is

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

[5 marks]

(ii) State the distribution of $(n-1)S^2/\sigma^2$, and give formulae for the end-points of a $100(1-\alpha)\%$ confidence interval for σ^2.

[3 marks]

(iii) A data set is obtained with $n = 10$, $\sum_{i=1}^{10} x_i = 859$, and $\sum_{i=1}^{10} x_i^2 = 74227$. Calculate a 95% confidence interval for σ^2.

[2 marks]

[Total 10 marks]

A3. Let $X_1, \ldots, X_n$ be a random sample from $U\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, where θ is unknown.

(i) Show that the estimator

$$\hat{\theta} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of θ with variance $1/(12n)$.

[4 marks]

(ii) If n is large, what is the approximate sampling distribution of $\bar{X}$? State any results you use.

[2 marks]

(iii) Suppose that a sample of size $n = 75$ is obtained, for which $\sum_{i=1}^{n} x_i = 1147$. Use your answer to part (ii) to estimate the approximate probability that, for a future sample of size $m = 80$ from the same distribution, the value of $\bar{X}_m$ is greater than 15.25.

[4 marks]

[Total 10 marks]

A4. Suppose that $X_1, \ldots, X_n \sim N(\mu_1, \sigma^2)$ independently and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2)$ independently of each other and of $X_1, \ldots, X_n$, with μ_1, μ_2 and σ^2 all unknown. Let $\theta = \mu_1 - \mu_2$ denote the difference of the two population means. It is desired to test the null hypothesis $H_0 : \theta = \theta_0$ versus the two-sided alternative hypothesis $H_1 : \theta \neq \theta_0$.

(i) Write down an appropriate unbiased estimator, $\hat{\sigma}^2$, of the common variance σ^2. Write down a suitably scaled version of $\hat{\sigma}^2$ and state its sampling distribution.

[3 marks]

(ii) Suggest a suitable test statistic for testing H_0 vs H_1, and state its exact distribution in the case that H_0 is true. Give an appropriate rejection region to achieve significance level α.

[3 marks]

(iii) A data set is obtained satisfying

$$\begin{array}{lll} n = 10 & \bar{x} = 23.7 & s_1^2 = 1.25 \\ m = 12 & \bar{y} = 21.6 & s_2^2 = 1.16 \end{array}$$

Do you reject H_0 if $\theta_0 = 1$ and $\alpha = 0.05$? How about if $\alpha = 0.01$? Show your working.

[4 marks]

[Total 10 marks]

SECTION B

Answer **TWO** of the three questions

B5. Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, with μ unknown and σ^2 known. It is of interest to test the null hypothesis $H_0 : \mu = \mu_0$ versus the one-sided alternative $H_1 : \mu > \mu_0$ at significance level α.

(i) Define the following terms: Type I error, Type II error, and the significance level of the test (also known as the size).

[3 marks]

(ii) Write down an appropriate test statistic and rejection region, making sure to define any notation used. Show that this choice does indeed achieve a significance level of α.

[4 marks]

(iii) Suppose for this part that $\mu_0 = 10$, $\sigma^2 = 1$ and a data set is obtained with $n = 10$ and $\sum_{i=1}^{n} x_i = 106$. Do you reject H_0 at significance level $\alpha = 0.05$? What if $\alpha = 0.01$?

[3 marks]

(iv) Suppose for this part that a data set is obtained with $\sum_{i=1}^{n} x_i = n\mu_0 + 0.2n\sigma$. What is the minimum value of n for which H_0 is rejected at (i) a 5% significance level, and (ii) a 1% significance level?

[3 marks]

(v) Show that for $\mu > \mu_0$ the probability of correctly rejecting H_0 is equal to

$$1 - \Phi\left(z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right).$$

Compute the probability of rejecting H_0 if $\mu_0 = 10$, $\mu = 10.5$, $\sigma = 1$, $\alpha = 0.05$ and $n = 10$.

[7 marks]

[Total 20 marks]

B6. A random variable X has a *negative binomial distribution* $\mathrm{NegBi}(r, p)$, with $p \in (0, 1)$ and r a positive integer, if it has probability mass function (p.m.f.)

$$p_X(x) = \mathrm{P}(X = x) = \binom{x + r - 1}{x} (1 - p)^r p^x \,, \quad x = 0, 1, 2, \ldots \,,$$

where above the notation $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ denotes a binomial coefficient.

(i) Show that if $X_1, \ldots, X_n \sim \mathrm{NegBi}(r, p)$ independently, with r known and p unknown, then the likelihood function is

$$L(p) = \left[\prod_{i=1}^{n} \binom{X_i + r - 1}{X_i} \right] (1 - p)^{nr} p^{\sum_{i=1}^{n} X_i} \,.$$

[2 marks]

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

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

[7 marks]

(iii) Show that

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

[6 marks]

(iv) Use the result given in part (iii) to calculate the approximate probability that $0.45 < \hat{p} < 0.55$ when $p = 0.5$, $r = 3$ and $n = 100$.

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

[5 marks]

[Total 20 marks]

B7. The population proportion of batteries of Type 1 that are defective is p_1, and the population proportion of batteries of Type 2 that are defective is p_2. Both p_1 and p_2 are unknown, and it is desired to estimate the difference $\theta = p_1 - p_2$.

To investigate this, an experimenter will examine n batteries of Type 1 and m batteries of Type 2, and record how many of each type are defective.

Let X_1 denote the number of defectives in the sample of Type 1 batteries, and let X_2 denote the number of defectives in the sample of Type 2 batteries.

(a) (i) Suggest an appropriate estimator, $\hat{\theta}$, for θ and define any notation used.

[2 marks]

(ii) Show that $\hat{\theta}$ is unbiased, and derive an expression for $v(n, m) = \operatorname{Var} \hat{\theta}$.

[6 marks]

(b) The investigator considers two possible experimental designs with total sample size $n+m = 300$:

Design 1: $n = 150$, $m = 150$
Design 2: $n = 100$, $m = 200$

(i) Show that $\operatorname{Var} \hat{\theta}$ is smaller under Design 1 than under Design 2 if and only if

$$\frac{p_1(1-p_1)}{p_2(1-p_2)} \geq \frac{1}{2}\,.$$

[4 marks]

(ii) Hence argue that, when $p_1 = 0.1$, the estimator $\hat{\theta}$ performs better if we use Design 1 rather than Design 2 provided that $p_2 < 0.235$ or $p_2 > 0.765$.

[4 marks]

(c) Now suppose that the experimenter finds that 13 out of $n = 150$ batteries of Type 1 are defective and 17 out of $m = 150$ batteries of Type 2 are defective.

(i) Calculate an approximate 95% confidence interval for θ.

[4 marks]

[Total 20 marks]

END OF EXAMINATION PAPER