Two hours

Statistical tables to be provided

THE UNIVERSITY OF MANCHESTER

INTRODUCTION TO STATISTICS

?? June 2016

??.?? – ??.??

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. A sample of size $n = 20$ from a particular distribution has been obtained. The data have been entered into R and stored in a variable x. Some R output is recorded below:

```
> x
 [1]   1.49   1.67   2.20   3.23   7.32   9.28  10.35
 [8]  11.85  13.67  19.79  22.95  29.07  36.80  36.81
[15]  38.20  43.81  59.00  92.17 111.62 113.32

> quantile(x,type=6)
      0%      25%      50%      75%     100%
  1.4900   7.8100  21.3700  42.4075 113.3200
```

(a) (i) Use the output above to draw a box plot of the data.

(ii) Comment on the shape of the distribution, and any other features in the data.

(iii) Is it appropriate to fit a $N(\mu, \sigma^2)$ model to these data? If not, suggest a transformation that may enable a normal distribution to be fitted.

(b) Suppose now that the interval $[0, 120]$ is divided into bins of equal length $h = 20$, which are used to create a density histogram.

(i) At $x = 9$, compute the value of the function $\text{Hist}(x)$ defining the height of the histogram.

[10 marks]

A2. Let $X_1, \ldots, X_n$ be a random sample from $\text{Po}(\lambda)$.

(i) Show that the likelihood function for the sample can be written as

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}\,.$$

(ii) Show that the maximum likelihood estimator of λ is $\hat{\lambda} = \bar{X}$.

(iii) Compute $\text{bias}(\hat{\lambda})$ and $\text{Var}(\hat{\lambda})$.

(iv) If $n = 100$ and $\lambda = 10$, what is the approximate probability that $9.9 < \hat{\lambda} < 10.1$? Comment on any results used.

[10 marks]

A3. Let $X_{11}, \ldots, X_{1n}$ be a random sample from $N(\mu_1, \sigma^2)$, and $X_{21}, \ldots, X_{2m}$ be a random sample from $N(\mu_2, \sigma^2)$, where μ_1, μ_2 and the common variance σ^2 are all unknown. It is desired to test the following hypotheses at the $100\alpha\%$ significance level:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0\,.$$

(i) Write down a suitable unbiased estimator, $\hat{\sigma}^2$, for the common variance σ^2. What can be said about the distribution of a suitably scaled version of $\hat{\sigma}^2$?

(ii) Write down an appropriate statistic for testing H_0 vs H_1. What is the sampling distribution of your test statistic under the null hypothesis?

(iii) What is an appropriate rejection region for the test?

(iv) Suppose that a data set is obtained with $n = m = 10$, and

$$\bar{x}_1 = 46.0\,, \quad s_1^2 = 2.04^2\,,$$
$$\bar{x}_2 = 48.1\,, \quad s_2^2 = 1.92^2\,.$$

Do you reject H_0 when $\alpha = 0.05$? What about if $\alpha = 0.01$? Show your working.

[10 marks]

A4. In this question, the population of interest is the set of all UK adults who are eligible to vote. Suppose that an independent simple random sample of size $n = 1000$ is obtained from the population, and that 30% of individuals in the sample support Labour.

(i) Give general formulae for the end-points of an approximate $100(1-\alpha)\%$ confidence interval for the parameter p given a random sample of size n from $\mathrm{Bi}(1, p)$. Define any notation used in your answer.

Comment on any distributional results underlying the derivation of your confidence interval.

(ii) Use the above sample results to calculate a 95% confidence interval for the proportion of individuals supporting Labour in the population.

(iii) Suppose that in fact the true proportion supporting Labour in the population is $p = 0.28$. What is the approximate probability that, in a new sample of size 500, at least 150 will support Labour? Comment on the validity of any approximations used.

[10 marks]

SECTION B

Answer **TWO** of the three questions

B5.

(a) Let $X_1, \ldots, X_n$ be a random sample of size n from $N(\mu, \sigma^2)$, where μ and σ^2 are both unknown, and let

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

  (i) Show that $\mathrm{E}(S^2) = \sigma^2$.

  (ii) What is the distribution of $(n-1)S^2/\sigma^2$?

(b) Suppose now that we have observed data with $n = 10$, and

$$\sum_{i=1}^{n} x_i = 113.20 , \qquad \sum_{i=1}^{n} x_i^2 = 1474.5 .$$

  (i) Calculate s^2, and compute a 99% confidence interval for σ^2.

  (ii) Compute a 99% confidence interval for the population mean μ.

  (iii) Using these data, estimate the probability that the mean of a future sample of size $n = 10$ from the same population will satisfy $\bar{X} > 11.0$.

[Total 20 marks]

B6. A researcher conducts a clinical trial on $n = 60$ patients to investigate a new treatment for Rhinovirus. It is known that the recovery time (in days) of an untreated patient is randomly distributed as $N(6, 2^2)$.

The researcher considers a possible study design in which the data collected is the proportion of patients who recover within 7 days.

Let p denote the probability that an individual patient recovers within 7 days under the new treatment, and p_0 denote the probability that an individual patient recovers within 7 days under no treatment.

(i) Show that $p_0 = \Phi(\frac{1}{2})$, and calculate its numerical value.

(ii) Write down an appropriate test statistic for testing

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

(iii) Write down an approximate distribution for your test statistic under the null hypothesis, commenting on any assumptions you make.

(iv) Write down an appropriate rejection region for the test to achieve significance level $\alpha = 0.05$. Would H_0 be rejected if 52 out of 60 patients recovered within 7 days?

Suppose that in fact, unknown to the researcher, the recovery time (in days) for the ith patient under the new treatment is $X_i \sim N(5, 2^2)$, $i = 1, \ldots, n$, independently.

(v) Use this fact to calculate the probability that a patient recovers within 7 days under the new treatment. Hence find the approximate probability of rejecting the null hypothesis under the test in (ii).

[Total 20 marks]

B7.

(a) Suppose that $X_1, \ldots, X_n$ are a random sample of size n from $N(\mu, \sigma^2)$. Consider the following two estimators of μ:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \qquad \tilde{\mu} = \frac{1}{2n}\sum_{i=1}^{n} X_i\,.$$

(i) Compute the bias and variance of $\hat{\mu}$.

(ii) Compute the bias and variance of $\tilde{\mu}$.

(b) Let ϵ be a quantity specified by the experimenter. Moreover suppose that, unknown to the experimenter, $\mu = 0.2\sigma$ and $\epsilon = 0.1\sigma$.

(i) Show that $\hat{\mu}$ is within ϵ of the true value of μ with probability

$$p_1(n) = 2\Phi(0.1\sqrt{n}) - 1\,.$$

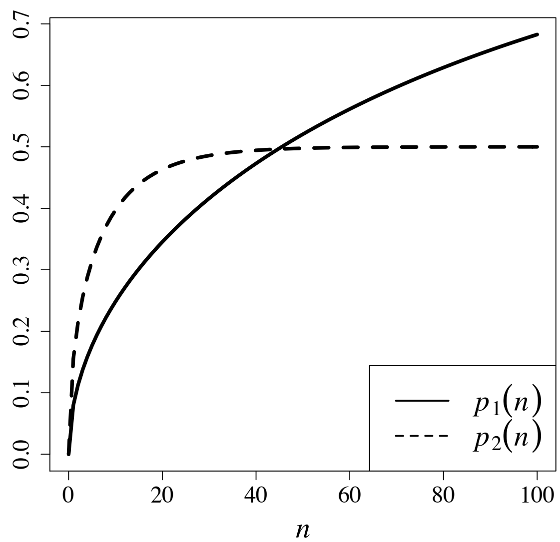
(ii) Show that $\tilde{\mu}$ is within ϵ of the true value of μ with probability

$$p_2(n) = \Phi(0.4\sqrt{n}) - 0.5\,.$$

(iii) Which of the estimators $\hat{\mu}$ and $\tilde{\mu}$ has the greatest probability of being within ϵ of the true value of μ when $n = 10$? Justify your answer with calculations.

(iv) The investigator decides that the experiment will be considered a success if and only if the estimate of μ is within ϵ of the true value. Which of the estimators $\hat{\mu}$ and $\tilde{\mu}$ is preferable, and under what circumstances?

[*Hint: use the graphs of* $p_1(n)$ *and* $p_2(n)$, *which are plotted below.*]



[Total 20 marks]

END OF EXAMINATION PAPER