MAS5052

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS — Spring Semester 2015–2016

Basic Statistics — 2 hours

RESTRICTED OPEN BOOK EXAMINATION.
Candidates may bring to the examination lecture notes and associated lecture material (including set textbooks) plus a caluclator that conforms to University regulations.
*Candidates should attempt **ALL** questions.*
The maximum marks for the various parts of the questions are indicated.
The paper will be marked out of 80.

1 A major retailer wanted to test whether the amount of discount of a product had an effect its sales. The discounts were 2%, 3%, 4%, 5% and 6% and were applied to the product on consecutive weekdays and then similarly the following week and the number of items sold was recorded for each of the ten days.

Percent Discount, p	2	3	4	5	6
Units sold (week 1), u	320	381	411	460	482
Units sold (week 2), u	344	402	440	471	495

(i) Display the data in an appropriate plot. ***(3 marks)***

(ii) Find the (simple linear) regression line of units sold on percent discount and add this to your plot. ***(9 marks)***

(iii) Test the hypothesis that the slope of the line is zero; explain the meaning of this hypothesis and the results of your test. ***(6 marks)***

(iv) How could you improve the design of the experiment? ***(2 marks)***

2 Suppose that independent observations X_1 and X_2 are taken from Poisson $P(a\lambda)$ and Poisson $P(b\lambda)$ distributions respectively, where a and b are known and positive.

(i) Find the maximum likelihood estimator of λ. ***(5 marks)***

(ii) Compare the sampling distribution of the maximum likelihood estimator with the sampling distributions of the estimators:

$$T_1 = (X_1 - X_2)/(a - b)$$
$$T_2 = \frac{1}{2}\left(\frac{X_1}{a} + \frac{X_2}{b}\right)$$

and hence recommend an estimator for λ. ***(11 marks)***

3 A random sample of 100 people were monitored at a stall at a fun-fair. The game consists of trying to throw a table-tennis ball into a cup (and it staying in the cup). Each person has 3 attempts and the number of successful outcomes is recorded.

Number of successful throws	0	1	2	3	Total
Observed frequencies	16	35	31	18	100

A 'null' model suggested is that each throw has probability 0.5 of being successful, and that the outcomes of each throw for each person are independent.

(i) If the null model holds, what would the distribution be of the number of successful throws a person has when throwing 3 times? ***(2 marks)***

(ii) Explain some possible reasons why this model might not hold. ***(2 marks)***

(iii) Are the data above consistent with the null model? ***(10 marks)***

4 Six sets of identical twins were divided at random into two groups, each group containing one twin from each set. The first group was taught some basic statistics by method A and the second by method B. At the end of a fixed period of training, all the children were given a statistics test. Their test scores are given below.

	Twin set number					
	1	2	3	4	5	6
Method A	92	65	57	48	60	54
Method B	90	73	50	52	53	52

(i) Is there evidence of a difference in statistical ability between the two groups? ***(6 marks)***

(ii) State clearly the assumptions underlying the method you have performed and explain how you might be able to verify their acceptability [you need not perform the checks you suggest]. What assumption about the twins are we making when doing this study? ***(4 marks)***

(iii) Construct a 95% confidence interval for the population mean difference in test scores for the two groups. ***(4 marks)***

5 Let $X_1, \ldots, X_n$ denote independent observations and suppose that X_i has a $N(\mu_i, 1)$ distribution, $i = 1, \ldots, n$. Show that, according to the Neyman-Pearson Lemma, the most powerful test (with size 0.05) of

$$H_0: \quad \mu_i = 0 \quad 1 \le i \le n$$

against

$$H_1: \quad \mu_i = \left\{ \begin{array}{ll} 0 & 1 \le i \le r \\ 1 & r+1 \le i \le n, \end{array} \right\}$$

for known r with $1 < r < n$, has critical region of the form

$$\left\{ x : \sum_{r+1}^{n} x_i \ge 1.645\sqrt{(n-r)} \right\}.$$

(8 marks)

6 To compare 4 detection machines and also the performance of 3 scientists, the amount of gold within a chemical substance (of fixed weight) was determined by each scientist using each machine. Two independent determinations of gold content were made in each case, giving the following data, recorded in milligrams.

	Machine			
	A	B	C	D
Scientist A	4.1	3.2	1.5	3.6
	3.5	3.5	2.6	3.8
Scientist B	3.1	2.3	1.8	1.4
	3.3	3.4	2.1	2.0
Scientist C	3.5	3.1	2.9	2.7
	3.2	3.6	3.0	3.1

Use the R output below to identify the model initially fitted and the most appropriate model for the data. Explain your reasoning.

```
aov(formula = weight ~ scientist * machine, data
= chemicaldata, na.action = na.exclude)
                   Df  Sum of Sq   Mean Sq  F Value       Pr(F)
        scientist   2   3.080833  1.540417  9.455243  0.00342336
          machine   3   4.424583  1.474861  9.052856  0.00208485
scientist:machine   6   2.909167  0.484861  2.976130  0.05100618
        Residuals  12   1.955000  0.162917
```

(8 marks)

End of Question Paper