RESTRICTED OPEN BOOK EXAMINATION (Not to be removed from the examination hall)
Data provided: "Statistics Tables" by H.R. Neave

MAS5052

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2016–2017

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 The height (in cm) and weight (in kg) of 9,039 of the 10,385 athletes who competed in the 2012 London Olympic Games was obtained from the Guardian newspaper (`https://www.theguardian.com/sport/datablog/2012/aug/07/olympics-2012-athletes-age-weight-height#data`)

(i) Using this data, we would like to try to predict the weight of athletes from their height by using a linear regression model. How would you assess the suitability of doing this before fitting the regression model? ***(3 marks)***

(ii) The R output from fitting a linear regression model to the data (outlined above) is given below. Use it to obtain the regresion equation. ***(2 marks)***

```
Call:
lm(formula = Weight.kg ~ Height..cm)

Residuals:
    Min      1Q  Median      3Q     Max
-58.439  -6.085  -1.451   4.001 136.915

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.207e+02  1.743e+00  -69.23   <2e-16 ***
Height..cm   1.091e+00  9.802e-03  111.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.44 on 9036 degrees of freedom
  (1424 observations deleted due to missingness)
Multiple R-squared:  0.578,Adjusted R-squared:  0.578
F-statistic: 1.238e+04 on 1 and 9036 DF,  p-value: < 2.2e-16
```

(iii) Would it be sensible to fit a model with no intercept? Give reasons. ***(2 marks)***

(iv) We have a missing value for weight for one competitor who is 170cm tall. Predict the weight of this person using the formula you obtained in part ii). How accurate do you think the prediction is? Give reasons. ***(6 marks)***

(v) What checks would you do to explore the validity of the model? Does the R output give any indication as to what you might find? ***(5 marks)***

(vi) Are there any other factors that we could look at that may influence the height and weight of a competitor? ***(3 marks)***

2 An exponential random variable, with parameter θ (denoted by $Exp(\theta)$), has density

$$f(x;\theta) = \frac{1}{\theta}e^{-x/\theta} \quad x \geq 0 \qquad \textit{(Note parameterization).}$$

Suppose that X_1 and X_2 are taken from independent $Exp(a^{1/2}\lambda)$ and $Exp(b^{1/2}\lambda)$ distributions respectively, where a and b are known and positive.

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

(ii) Calculate the bias and variance of the maximum likelihood estimator and do the same for the alternative estimators:

$$\begin{aligned} T_1 &= \frac{X_1 - X_2}{a^{1/2} - b^{1/2}} \\ T_2 &= \frac{X_1 + X_2}{a^{1/2} + b^{1/2}}. \end{aligned}$$

(8 marks)

(iii) If $a = 4$ and $b = 9$, which is the best estimator? ***(3 marks)***

3 In 2011, the BBC carried out a survey on UK school children aged 11 to 16 (inclusive). The number of respondents was 24,052. The Office for National Statistics (ONS) estimates that there were 4,493,965 11 to 16 years olds in the UK in 2011.

(i) Calculate the sampling fraction. ***(2 marks)***

(ii) Of the 24,052 pupils surveyed, 640 were attending schools in Wales. ONS estimate that there were 217,752 11 to 16 year olds in Wales in 2011. If we used the sampling fraction calculated in part i), how many 11 to 16 year olds would we actually survey from Wales? Comment on the difference. ***(4 marks)***

(iii) It is proposed to now re-survey all the respondents who were 11 years old at the time of the survey and ask the question " How has your internet use changed since you were 11 years old?" Comment on this choice of question. ***(3 marks)***

4 Britain's membership of the European Union has (for many) been a subject of some debate since the Common Market (as it was first called) was set up in the late 1950s. In 1975 a referendum was held to decide on whether to stay in the Common Market (Britain joined in 1973). The result was that 67% of people voted in favour of staying. A survey of 1,561 people was carried out after the referendum. The survey asked the following questions:

- Did you vote to remain in the Common Market?
- Are you a Conservative voter?

The data can be seen below:

	Stay in Common Market		
	In Favour	Not In Favour	
Conservative	429	133	562
Not Conservative	502	497	999
	931	630	1561

(i) Test if there was any association between voting Conservative and voting to remain in the Common Market. ***(10 marks)***

(ii) Suggest on how this research may be improved, updated or followed up, indicating on how you might implement the collection of relevant information and what problems there may be. ***(5 marks)***

5 $X_1, \ldots, X_n$ are a random sample from the $Be(\alpha, 2)$ distribution. We wish to test

$$H_0 : \alpha = \alpha_0$$
$$H_1 : \alpha = \alpha_1$$

where $\alpha_0 < \alpha_1$. Show that the Neyman-Pearson Lemma indicates that the most powerful test is to reject H_0 if

$$T = \prod_{i=1}^{n} (X_i) > k^\star$$

for some $k^\star$. ***(6 marks)***

6 The share prices (in pence) of companies in 3 different sectors, as defined by the Financial Times Share index (FTSE), were recorded at 11am on Friday 13th January 2017 and shown below:

Life Insurance	483.6, 361.5, 106.38, 247.25, 212.8, 743.75, 80.5, 351.7
Electricity	21.88, 376.9, 280.5, 414, 62.5, 14.88, 150.5, 63.25, 1.13
Industrial Transportation	201.5, 284.65, 303.13, 2219.5, 1601, 2.75, 1000, 174.38, 28, 246.13

Using any graphical methods and summary statistics which you think appropriate, compare the three distributions and describe any differences between the sectors in share price. ***(13 marks)***

End of Question Paper