The University Of Sheffield.

**SCHOOL OF MATHEMATICS AND STATISTICS** **Spring Semester 2020–2021**

**Basic Statistics - Solutions** **3 hours 30 minutes**

*This is an open book exam.*

*Attempt ALL questions*

*You can work on the exam during the 24 hour period starting from 10am (BST), and you must submit your work within 3 hours and 30 minutes of accessing the exam paper or by the end of the 24 hour period (whichever is earlier).*

***Late submission will not be considered without extenuating circumstances.*** *Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a university-approved calculator). To gain full marks, you will need to show your working.*

*By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged, and that no unfair means have been used.*

*The maximum marks for the various parts of the questions are indicated.*
*The paper will be marked out of 60.*

**1** The Probability and Statistics section of the School of Mathematics at the University of Sheffield used to have regular 10-pin bowling tournaments over a number of years. Data from this was collected and two of the participants, Jeremy and Jonathan, were interested in which of them was the best player. Scores from their last 6 games were taken (they did not always play at the same time).

| | | | | | | |
|---|---|---|---|---|---|---|
| Jeremy | 116 | 111 | 106 | 145 | 120 | 119 |
| Jonathan | 112 | 150 | 115 | 146 | 156 | 130 |

(i) Find the sample mean and an unbiased estimate of the variance of each player *(4 marks)*

(ii) Test whether the variances of the scores for each player are equal. Do not use `R` directly for this. After calculating the test statistic please write a command in `R` to assess its significance. After running this command what are your conclusions? *(8 marks)*

(iii) If there was no reason to believe there was a difference in variances, perform a suitable test to determine if the mean scores for the two players are equal. Again, do not use `R` directly to do this. After calculating the test statistic please write a command in `R` to assess its significance and give your conclusions. *(8 marks)*

(iv) What assumptions have you made in performing the tests in parts (ii) and (iii)? Are they reasonable? What might you do to test the assumptions? *(4 marks)*

**2** During the Covid 19 pandemic the air industry suffered a great deal. The relationship between passenger arrivals at UK airports in the UK, $Y$, was assumed to be linearly related to the number people in hospital with Covid 19, $X$. The data were collected for 10 random days between March and July 2020. The sources were the NHS and the Home Office.

Using R the output of fitting a simple linear regression, $y_i = \alpha + \beta x_i + \varepsilon_i$, is shown below.

```
Call:
lm(formula = PassengerArrivals ~ HospitalCases)

Residuals:
   Min     1Q Median     3Q    Max
-21130 -12201  -3598   2999  49749

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   37133.105  12171.585   3.051   0.0158 *
HospitalCases    -2.367      1.168  -2.027   0.0773 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21950 on 8 degrees of freedom
Multiple R-squared:  0.3392,Adjusted R-squared:  0.2567
F-statistic: 4.107 on 1 and 8 DF,  p-value: 0.07725
```

(i) From the output what can you say about the relationship between passenger arrivals and hospital numbers? ***(3 marks)***

(ii) It has been suggested that for each hospital case that the number of arrivals reduces by 4. Test the hypothesis $\beta = -4$ vs $\beta \neq -4$, using a test size of 0.05. ***(4 marks)***

(iii) Provide a 99% confidence interval for $\beta$ and comment on how this relates to the result in the first part of the question. ***(5 marks)***

Some of the following R output may help with parts (ii) and (iii).

```
 > qt(0.95,8)  > qt(0.95,9)  > qt(0.95,10)   > qnorm(0.95)
[1] 1.859548  [1] 1.833113  [1] 1.812461    [1] 1.644854
 > qt(0.975,10) > qt(0.975,9)   > qt(0.975,8)   > qnorm(0.975)
[1] 2.228139   [1] 2.262157    [1] 2.306004    [1] 1.959964
 > qt(0.995,8) > qt(0.995,9)  > qt(0.995,10)  > qnorm(0.995)
[1] 3.355387  [1] 3.249836   [1] 3.169273    [1] 2.575829
 > qt(0.999,10) > qt(0.999,9)    > qt(0.999,8)    > qnorm(0.999)
[1] 4.143700   [1] 4.296806     [1] 4.50079      [1] 3.090232
```

**3** A local council employs 500 people, of these 150 are women and 350 are men. During the Covid pandemic 120 of the women and 100 of the men worked from home. The rest were required to work as normal at council offices. The council wishes to carry out a survey of their employees to determine job satisfaction during the pandemic. An overall sampling fraction of 10% has been decided on and they have decided to stratify by gender and location of work (home or office).

(i) Why might the company have decided to stratify by gender and location of work? ***(2 marks)***

(ii) Suppose that the company wishes to use a larger sampling fraction of female employees. If they wish to sample 15% of the women in each location category, while still keeping the overall sampling fraction the same at 10%, how many employees will be sampled in each category? ***(4 marks)***

(iii) They propose to sample people at the Christmas party in 2021 (assuming Covid restrictions allow this) as they think that it will be an easy opportunity to find interviewees. Give three drawbacks of this proposal. ***(3 marks)***

(iv) After consultation they decide to instead conduct the survey over email. After the survey was sent out they had quite a large number of non-responses from people who said they were too busy. Should the company be worried about these non-responses or can they just ignore them? Explain your answer. ***(3 marks)***

**4** The Office of National Statistics conduct a regular survey on household expenditure. The latest one was for the financial year ending April 2019. In the table below, the spending categories are 1: Food, 2: Non-alcoholic drinks, 3: Alcoholic drinks and 4: Tobacco and narcotics. The amounts are average weekly spend in £s.

| | | Product | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Age | 30-49 | 62.9 | 6.2 | 8.9 | 4.7 |
| | 50-64 | 61.3 | 5.9 | 10.8 | 4.6 |
| | 65-74 | 55.8 | 4.6 | 11.7 | 3.1 |

(i) Use the R output given below to establish whether there is a difference in the spend across the products. ***(4 marks)***

```
Call:
   aov(formula = Spend ~ Age + Product)

Terms:
                      Age  Product Residuals
Sum of Squares     9.252 6452.849    25.628
Deg. of Freedom        2        3         6

Residual standard error: 2.066734
Estimated effects may be unbalanced

     Df Sum Sq Mean Sq F value   Pr(>F)
Age           2      9     4.6   1.083    0.397
Product       3   6453  2150.9 503.572 1.35e-07 ***
Residuals     6     26     4.3
```

(ii) Is there evidence of a difference in the spend between the age groups? ***(3 marks)***

(iii) If we did not have information on the age split, how would the appropriate analysis differ? ***(2 marks)***

(iv) If one wished to establish whether a specific age group had a higher spend on a specific product group, how would one have to change the design and analysis of the experiment? ***(3 marks)***

## End of Question Paper