The University Of Sheffield.

PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.

Data Provided:
Neaves Tables
Graph Paper

SCHOOL OF MATHEMATICS AND STATISTICS MAS6012

Session 2010–2011 **3 Hours**

Medical Statistics, Sampling, Design

RESTRICTED OPEN BOOK EXAMINATION.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.
*All answers will be marked but credit will be given for only the best **FIVE** answers.*
All questions carry equal marks. Total marks 100.

Registration number from U-Card (9 digits) – to be completed by student

(This page is left blank)

1 In a clinical trial to compare two devices designed to inhibit snoring, 794 men aged between 40 and 65 (all notorious snorers) were allocated either to device A or to device B. Of the 396 who used device A, 233 reported that they had had no episodes of snoring during the week they used the device whereas 203 of those using B reported no snoring.

(a) Do these data provide sufficient evidence that switching men in this age group from device B to device A would reduce the overall incidence of snoring?

(4 marks)

(b) Further investigation revealed that 403 men in the trial were aged under 50. In this age group 28 of the 103 who used device A reported no snoring episodes, whereas 121 of those using device B were relieved of snoring during the week of the study. Do these data give sufficient evidence to suggest that switching men aged under 50 from device B to A would increase or decrease the incidence of snoring in this age group?

(5 marks)

(c) Do the data provided in part (b) give sufficient evidence to suggest that switching men aged between 50 and 65 from device B to A would increase or decrease the incidence of snoring in this age group?

(5 marks)

(d) Using a Mantel-Haenszel test to allow for the differences in snoring rates and in allocation rates to devices between the two age groups, assess the evidence for a difference in effectiveness between devices A and B. Which device should be recommended for use by any notorious male snorer aged between 40 and 65?

(6 marks)

2 Given below is a record (edited in places) of an **R** analysis of the results of a two period crossover trial to investigate the effects of two treatments A (standard) and B (new) for an allergic coughing reaction. The figures represent the numbers of coughing incidents during a three-day period. Because of doubts about normality of the data, nonparametric methods of analysis have been used. The new treatment will be introduced provided that there is good evidence that there is a reduction of at least six from the level of the standard treatment in the median number of coughing incidents in this period. Patients were randomly allocated to two groups: group 1 received treatment A in period 1 and B in period 2. Group 2 received the treatments in the opposite order.

(a) Plot the treatment medians for each group for each period.

(2 marks)

(b) Assess all of the evidence that there is a carryover effect from period 1 to period 2.

(4 marks)

(c) Do the data provide evidence that there is a difference in average response between periods 1 and 2?

(4 marks)

(d) Assess whether the treatments differ in effect, taking into account the results of your assessments of carryover and period effects.

(5 marks)

(e) Describe what further analysis, if any, would be required to assess whether there is evidence to support the introduction of the new treatment.

(5 marks)

Question 2 continued on next page

Question 2 (continued)

```
     ***  Summary Statistics for data in:  coughs ***
group:1
          period1 period2
    Mean:   13.83    7.00
  Median:   15.00    6.00
 Total N:   12.00   12.00
Std Dev.:    5.13    3.59
--------------------------
group:2
          period1 period2
    Mean:    6.73   10.40
  Median:    6.00    8.00
 Total N:   15.00   15.00
Std Dev.:    3.37    7.39
> attach(coughs)
> totalresponse <- period1 + period2
> perioddiffs <- period1 - period2
> treatorder <- 3 - 2 * group
> treatdiffs <- perioddiffs * treatorder
> wilcox.test
+ (totalresponse [group==1], totalresponse [group==2])

     Wilcoxon rank-sum test

data:  totalresponse[group == 1] and totalresp[group ==
2]
rank-sum normal statistic with correction Z = 1.615,
p-value = 0.1062
alternative hypothesis:  mu is not equal to 0

>wilcox.test(perioddiffs[group==1],perioddiffs[group==2])

     Wilcoxon rank-sum test

data:  perioddiffs[group==1] and perioddiffs[group==2]
rank-sum normal statistic with correction Z = 4.091,
p-value = 0
alternative hypothesis:  mu is not equal to 0

> wilcox.test(treatdiffs[group==1],treatdiffs[group==2])

     Wilcoxon rank-sum test

data:  treatdiffs[group == 1] and treatdiffs[group == 2]
rank-sum normal statistic with correction Z = 1.645,
p-value = 0.0999
alternative hypothesis:  mu is not equal to 0

>
```

3 The data given below represent survival times in days of 26 patients randomized to one of two forms of chemotherapy following surgery for ovarian cancer, where status records whether the observation is censored (status = 0) or complete (status =1),
(Source: Collett, 2003):

	Treatment A		Treatment B	
	time	status	time	status
	59	1	353	1
	115	1	365	1
	156	1	377	0
	268	1	421	0
	329	1	464	1
	431	1	475	1
	448	0	563	1
	477	0	744	0
	638	1	769	0
	803	0	770	0
	855	0	1129	0
	1040	0	1206	0
	1106	0	1227	0
Totals	***6725***	***7***	***8863***	***5***

Given below is a record (edited in places) of some initial analyses of these data performed in **R**. Careful examination of this record reveals that the data have been entered into **R** incorrectly: the censoring indicators for time 855 for treatment A and 563 for treatment B have been entered incorrectly.

(a) Compute the Kaplan-Meier product limit estimates of the survivor functions for treatments A and B for the corrected data as given in the table above and provide estimates of the median survival times based upon the Kaplan Meier estimates.

(7 marks)

(b) Assuming that the survival times are exponentially distributed with rates λ_j, j=A, B, estimate λ_A and λ_B and hence the median survival times and provide approximate 95% confidence intervals for the median survival times for each group.

(7 marks)

(c) By using a parametric test, assess the evidence for a difference in the mean survival times between the two stages.

(6 marks)

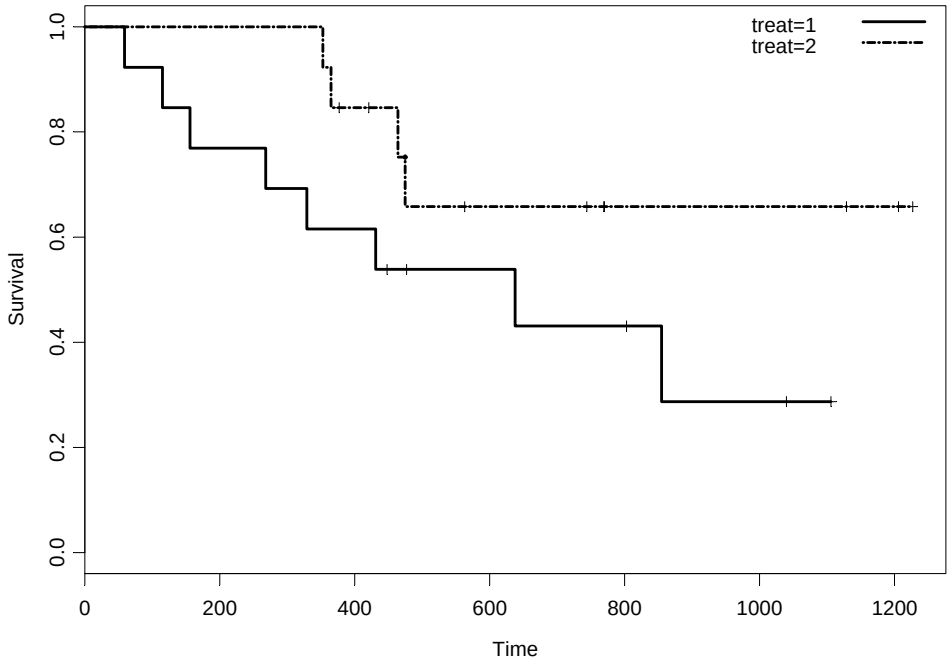
Question 3 continued on next page

Question 3 continued

Analysis of Survival Times of Ovarian Cancer

```
> library(survival)
> attach(ovarian)
> ovarian.sv<-Surv(time, status)
> survfit(ovarian.sv~treat)
Call: survfit(formula = ovarian.sv ~ treat)

        records n.max n.start events median 0.95LCL 0.95UCL
treat=0      13    13      13      7    638     268      NA
treat=1      13    13      13      5     NA     475      NA>
summary(survfit(ovarian.sv~treat))
                treat=1                          treat=2
time n.risk n.event survival    time n.risk n.event survival
  59     13       1    0.923     353     13       1    0.923
 115     12       1    0.846     365     12       1    0.846
 156     11       1    0.769     377     11       0    0.846
 268     10       1    0.692     421     10       0    0.846
 329      9       1    0.615     464      9       1    0.752
 431      8       1    0.538     475      8       1    0.658
 448      7       0    0.538     563      7       0    0.658
 477      6       0    0.538     744      6       0    0.658
 638      5       1    0.431     769      5       0    0.658
 803      4       0    0.431     770      4       0    0.658
 855      3       1    0.287    1129      3       0    0.658
1040      2       0    0.287    1206      2       0    0.658
1106      1       0    0.287    1227      1       0    0.658
```



4 An investigator is studying the dependence of a variable Y on one continuous explanatory variable x_1, which has been scaled to lie between -1 and 1. It is known that $EY = 0$ when $x_1 = 0$, and the following model is proposed.

$$EY = \beta_1 x_1 + \beta_{11} x_1^2.$$

The investigator proposes to take four observations, at $x_1 = -1, -0.5, 0.5, 1$.

(i) Show that β_1 and β_{11} are orthogonal to each other. ***(3 marks)***

(ii) If each observation is subject to a measurement error with mean 0 and variance σ^2, give the variances of the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_{11}$ in terms of σ^2. ***(1 mark)***

(iii) If a constant β_0 were to be added to the model, explain why β_0 would not be orthogonal to β_{11} for any choice of design points (with at least one design point not at 0). ***(2 marks)***

(iv) Show that this design (for the model $EY = \beta_1 x_1 + \beta_{11} x_1^2$) is neither D-optimal nor G-optimal, by using the General Equivalence Theorem. ***(4 marks)***

(v) Suggest an alternative design, with four observations, that is D-optimal. Justify your suggestion. ***(5 marks)***

(vi) The investigator is interested in predicting Y at values of x close to its upper limit, and wishes to consider V-optimality with the weight function $w(x) = 1 + x$. Show that your design in part (v) is V-optimal out of all possible orthogonal designs with 4 observations. ***(5 marks)***

5 (i) In a small study, four treatments A, B, C and D are to be compared. There are 12 females in the study, all of different ages. It is suspected that age may have an effect on the response variable Y. Each person in the study can only receive one treatment.

(a) Explain how a balanced incomplete block design with 6 blocks could be used in the design of the study. How would you choose the blocks, and which treatments would you used in each block? What would you randomise in your design? ***(6 marks)***

(b) The following model is proposed for the data.

$$Y_{it} = \gamma_i + \delta_t + \varepsilon_{it},$$

with Y_{it} the observed response variable on treatment t in block i, with $t = A, B, C$ or D, and $\varepsilon_{it} \sim N(0, \sigma^2)$. If $\delta_A + \delta_B + \delta_C + \delta_D = 0$, give an expression for the estimator of δ_C in terms of Y_{it}, and state the variance of the estimator. ***(2 marks)***

(c) Suppose instead there were 20 females. Why would a balanced incomplete block design not be possible? ***(3 marks)***

(d) As an alternative to blocking, it is suggested to allocate each treatment to three patients at random, and include terms to represent the effect of age in the linear model. Give one disadvantage of this approach. ***(2 marks)***

(ii) An alloy is being developed from a mixture of four different metals A, B, C and D. A single observation consists of choosing proportions of each metal, for example, 50% metal A, 20% metal B, 20% metal C and 10% metal D, constructing the alloy, and then measuring how much weight the alloy can bear before breaking.

For a design with 10 observations, state the most complex linear model that would be appropriate to fit to the data, justifying your answer, and suggest a suitable design. ***(7 marks)***

6 (i) An opinion poll is to be conducted to estimate the proportion of voters who intend to vote for the Liberal Democrats at the next UK General Election. If a simple random sample is to be used, how large would the sample need to be to ensure that a 99% confidence interval for the true proportion was no wider than 0.05. You may ignore the finite population correction. For your choice of n, would you expect the observed 99% confidence interval to be narrower than 0.05? Explain your answer. ***(5 marks)***

(ii) A survey is conducted to estimate the proportion of current cannabis users in a population. 100 members of the population have been selected using simple random sampling. Each participant first rolls a die, but does not reveal the outcome to the interviewer. If the outcome of rolling the die is a 1, the participant responds "true" or "false" to the following statement

"I have never used cannabis".

If the outcome of rolling the die is 2,3,4,5 or 6, the participant responds "true" or "false" to the following statement

"I have used cannabis at least once".

(a) If r is the proportion of participants who respond "true", derive an unbiased estimator of the proportion of the population who have used cannabis at least once. Calculate the variance of your estimator, as a function of the true proportion of cannabis users.
(5 marks)

(b) Suppose, instead of rolling a die, the participant tosses a coin, with the outcome of heads or tails determining which statement the participant responds to. What would the flaw be with this version of the survey?

(2 marks)

6 (continued)

(iii) A company wishes to estimate the average expenditure on foreign holidays last year by members of a population. The organization has decided to use a stratified random sample with proportional allocation from 3 strata. Information on spending from surveys in the previous year is given in the following table.

Stratum	Size (1,000's)	Est. std. dev.	Est. mean
1	50	800	5000
2	200	300	1000
3	100	100	800

Members of stratum 1 are more difficult to access compared to the other strata, so that there is an extra 20% cost of surveying each individual in stratum 1, compared to strata 2 and 3.

If the aim is to minimise the variance of the estimator of the population mean, within a fixed overall cost of conducting the survey, in what ratio should the three strata be sampled? If, for your chosen ratio, the company calculates that it can afford a total sample size of 1000, calculate the variance of your estimator. ***(8 marks)***

End of Question Paper