The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2013–2014**

Sampling, Design, Medical Statistics **3 hours**

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 only for the best ***FIVE*** *answers. All questions are worth 20 marks. Total marks 100.*

Please leave this exam paper on your desk
Do not remove it from the hall

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

Blank

1 A 2-treatment, 2-period crossover trial has been appropriately powered and conducted to compare a new drug A with a placebo B. Low values of the response are good. 8 patients receive the drugs in each order and the trial data are organized in the following manner:

Variable name	V1	V2	V3	V4
Group	1	1	2	2
Period	1	2	1	2
Treatment	A	B	B	A

The following table shows the data (adapted from Altman(1995)).

Variable name	V1	V2	V3	V4
	1.6	1.2	1.8	1.2
	2.6	1.9	1.2	0.4
	0.8	2.0	4.6	3.7
	3.7	4.4	5.1	5.8
	0.9	2.5	2.8	0.2
	4.1	3.6	2.9	1.8
	5.2	3.6	5.1	4.4
	1.0	1.1	4.6	1.4
	1.1	2.0	1.8	3.0
	3.0	2.7	4.4	0.4

Additional variables are constructed in R as follows:

```
> V5<-0.5*(V1+V2)
> V6<-0.5*(V3+V4)
> V7<-V1-V2
> V8<-V3-V4
> V9<--V8
```

(i) Write down the standard model for such a trial, explaining your notation clearly. ***(3 marks)***

(ii) Use three of the four tests reported in the following R output to provide an appropriate analysis of the trial. Explain your reasoning at each stage by reference to the model in (i). Ensure you check for:

(a) a possible carryover effect from Period 1 to Period 2 ***(4 marks)***

(b) a possible treatment effect ***(4 marks)***

(c) a possible period effect. ***(4 marks)***

(d) For the test given in the R output but which is NOT used above, explain the circumstances in which it might be a useful test to conduct. ***(1 mark)***

1 (continued)

```
> t.test(V1,V3)

        Welch Two Sample t-test

data:  V1 and V3
t = -1.5052, df = 17.971, p-value = 0.1496
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -2.4677776  0.4077776
sample estimates:
mean of x mean of y
     2.40      3.43

> t.test(V5,V6)

        Welch Two Sample t-test

data:  V5 and V6
t = -0.6125, df = 17.384, p-value = 0.5481
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -1.6867372  0.9267372
sample estimates:
mean of x mean of y
     2.45      2.83
```

1 (continued)

```
> t.test(V7,V8)

        Welch Two Sample t-test

data:  V7 and V8
t = -2.1534, df = 14.79, p-value = 0.04821
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -2.5883311 -0.0116689
sample estimates:
mean of x mean of y
     -0.1       1.2
```

```
> t.test(V7,V9)

        Welch Two Sample t-test

data:  V7 and V9
t = 1.8221, df = 14.79, p-value = 0.08872
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -0.1883311  2.3883311
sample estimates:
mean of x mean of y
     -0.1      -1.2
```

(iii) Do your findings suggest that the new drug A should
a) definitely be introduced
b) definitely not be introduced
c) be introduced if it passes further checks?
Explain your reasoning, including, if you select option c), what further checks are needed. *(4 marks)*

	Type A		Type B	
	Time (hrs)	Status	Time (hrs)	Status
	5.10	1	5.20	1
	16.30	1	14.20	1
	0.30	0	7.20	0
	4.30	1	4.90	0
	10.00	1	18.40	1
	0.10	0	7.80	1
Total	36.1	4	57.7	4

2 A car manufacturer makes two different types of engine components (A and B) and wants to see if there is any difference between their lifetimes. The components were fitted to an engine that was run until failure. The length of time (in hrs) that the engines functioned with the components were recorded and can be seen below. Here the status variable records whether the component was still operating when something else on the engine broke (status = 0) or whether the engine broke due to the component itself failing (status = 1).

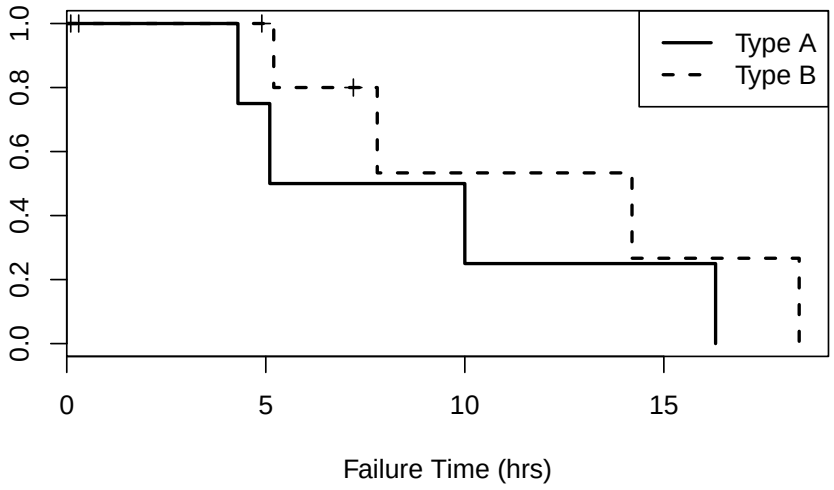
Given below are the results of a Kaplan-Meier preliminary graphical analysis of the data

```
> Comp.sv <- Surv(time, status, type = "right")
> summary(survfit(Comp.sv ~ type))
Call: survfit(formula = Comp.sv ~ type)

                type=A
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  4.3      4       1     0.75   0.217       0.4259            1
  5.1      3       1     0.50   0.250       0.1877            1
 10.0      2       1     0.25   0.217       0.0458            1
 16.3      1       1     0.00     NaN           NA           NA

                type=B
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  5.2      5       1    0.800   0.179       0.5161            1
  7.8      3       1    0.533   0.248       0.2142            1
 14.2      2       1    0.267   0.226       0.0507            1
 18.4      1       1    0.000     NaN           NA           NA
```

2 (continued)



(i) Without making any model assumptions, estimate the median failure times for the two component types. ***(3 marks)***

(ii) It is suggested that the survival times for components A and B are Exponentially distributed with rates λ_A and λ_B respectively. Under this assumption:

(a) Estimate λ_A and λ_B and hence the mean failure times with approximate 95% confidence intervals (note that selected standard normal percentage points are given below). ***(4 marks)***

p	qnorm(p)
0.025	-1.96
0.05	-1.64

(b) Use the likelihood ratio test to assess whether there is a difference in the failure time distribution of the two components (note that selected chi-squared percentage points are given below). Why is the LRT more suitable than an MLE test in this case? ***(4 marks)***

p	qchisq(p,1)
0.9	2.71
0.95	3.84

(c) Do the assumptions of Exponential survival distributions seem plausible? Explain your answer. ***(2 marks)***

(iii) By copying and completing the partially filled table on the next page, perform a non-parametric comparison of the two survival distributions (note that selected chi-squared percentage points are given on the next page). Compare your findings with the parametric test in part (ii)(b).

2 (continued)

p	qchisq(p,1)
0.9	2.71
0.95	3.84

		Number at risk			Number of deaths			Expected number of deaths	
i	t_i	r_{Ai}	r_{Bi}	r_i	d_{Ai}	d_{Bi}	d_i	e_{Ai}	e_{Bi}
1	4.3								
2	5.1								
3	5.2								
4	7.8								
5	10.0								
6	14.2								
7	16.3								
8	18.4								
Total					$O_A = 4$	$O_B = 4$		$E_A = 2.8$	$E_B = 5.2$

(7 marks)

3 (i) A cohort of live born babies was followed up for one year to explore the effect of birth weight and sex on survival. The results from fitting a proportional hazards regression model are shown in the Table below. The model relates survival to birth weight (grouped into four categories) and the sex of the baby, and the age of the mother (grouped into two categories).

Variable	Coefficient	Standard Error
Sex of Baby		
Male	Reference	—
Female	-0.39	0.07
Birthweight		
$\geq$ 4000 g	0.12	0.08
3500–3999g	Reference	—
3000–3499g	0.27	0.15
$<$ 3000g	1.61	0.23
Age of mother		
$>$ 40 yrs	0.65	0.33
$\leq$ 40 yrs	Reference	—

Table: Log hazard ratio of infant mortality with standard errors

(a) Specify the form of the hazard model used for this analysis, carefully defining each predictor variable. ***(3 marks)***

(b) Describe in detail the effects of these variables on infant survival (note that selected standard normal and chi-squared percentage points are given on the next page). ***(6 marks)***

3 (continued)

p	qnorm(p)	p	qchisq(p,3)
0.001	-3.09	0.95	7.81
0.005	-2.58	0.99	11.34
0.025	-1.96		
0.05	-1.64		

(c) How would you assess whether the proportional hazards model was appropriate for these data? ***(2 marks)***

(d) Using the model, calulate the estimate of the hazard ratio comparing the following two babies:

- A female baby with birth weight 3000–3499g and a 35 yr old mother,
- A male baby with birth weight 3500–3999g and a 41 yr old mother.

(3 marks)

(ii) A key task for radiologists is to interpret mammograms to see if they provide evidence of breast cancer. As part of his final examination, a student radiologist is asked to assess 85 mammograms and assign each to one of the 4 categories: 'Normal', 'Benign Disease', 'Suspected Cancer', 'Cancer'. A senior consultant radiologist has already categorized the mammograms. The cross-classified categorizations (adapted from Altman (1995)) are shown below. Do you think the student should pass the examination? Justify your answer. ***(6 marks)***

		Consultant				
		Normal	Benign	Suspected Cancer	Cancer	Total
Student	Normal	21	12	0	0	33
	Benign	4	17	1	0	22
	Suspected Cancer	3	9	15	2	29
	Cancer	0	0	0	1	1
	Total	28	38	16	3	85

4 A small experiment is being conducted to compare two new treatments against a placebo. There are six participants in the study. Two participants are given the placebo, two are given treatment A, and two are given treatment B. Each observation is subject to a measurement error with mean 0 and variance σ^2. The following model is proposed.

$$EY_{ij} = \mu + \tau_i,$$

for $i = 1, 2, 3$, $j = 1, \ldots, n_i$, with $n_1 = n_2 = n_3 = 2$. The constraint $\tau_1 + \tau_2 + \tau_3 = 0$ is applied.

(i) Write down the design matrix for this design. ***(3 marks)***

4 (continued)

(ii) Describe which parameters or groups of parameters are orthogonal.
(2 marks)

(iii) Suppose instead that four participants are given the placebo, one is given treatment A and one participant is given treatment B. Describe which parameters or groups of parameters are orthogonal for this design.
(2 marks)

(iv) Suppose instead that there are 10 participants in total of which $10-2t$ participants are given the placebo, t are given treatment A and t are given treatment B. The following model is proposed

$$EY_{ij} = \tau_i,$$

for $i = 1, 2, 3$, $j = 1, \ldots, n_i$, with $n_1 = 10 - 2t$, $n_2 = n_3 = t$.

(a) Explain why there is no need to impose constraints in this model.
(1 mark)

(b) Show that

$$(X^T X)^{-1} = \begin{pmatrix} 1/(10-2t) & 0 & 0 \\ 0 & 1/t & 0 \\ 0 & 0 & 1/t \end{pmatrix}.$$

Find the value of $t \in \{1, 2, 3, 4\}$ that minimizes $Var(\hat{\tau_2} - \hat{\tau_1})$ and give the minimum value of the variance. ***(7 marks)***

(v) Suppose we still want to compare the effect of placebo and the two treatments (labelled P, A, B respectively) but there are now nine participants. In addition there are also three different experimental conditions (labelled $1, 2, 3$). The nine participants are to be allocated to blocks based on whether they are a current smoker, ex-smoker or have never smoked (labelled $1, 2, 3$ respectively). Give 2 orthogonal Latin square designs for this set up. State your design by completing the following table for each design. ***(4 marks)***

participant	smoking status $(1, 2, 3)$	experimental condition $(1, 2, 3)$	treatment (P, A, B)
1			
$\vdots$			
9			

(vi) Does a third orthogonal Latin square exist? Justify your answer.
(1 mark)

5 (i) An experiment is to be carried out to investigate the effect of four teaching methods on chemistry exam scores. There are 12 randomly selected students in the study, who will each be taught using one of the four methods. After the course finishes, each participant will be given a chemistry test, and their exam scores will be recorded. The experimenter decides to organise the students into blocks, according to their abilities.

(a) If the four teaching methods are labelled A, B, C, D, explain why the following design satisfies the requirements of a balanced incomplete block design with block sizes of 3.

Block 1: ABC
Block 2: ABD
Block 3: BCD
Block 4: ACD

(1 mark)

(b) For the design in (a), write out the observation vector, parameter vector and design matrix in full and specify any parameter constraints. ***(5 marks)***

(c) Suppose instead that the experimenter is only interested in the proportion of students who would get a mark of 60% with each method. The experimenter proposes to randomly allocate students to methods (with equal numbers of students per method), and observe the sample proportions of students getting 60% or higher for each of the four methods. How many students would be needed **in total**, such that the width of a 90% confidence interval for any single proportion was no wider than 0.2? You may ignore the finite population correction. Part of the following R output will help you.

```
> qnorm(c(0.9,0.95,0.975),0,1)
[1] 1.281552 1.644854 1.959964
```

(5 marks)

(ii) An investigator is studying the dependence of a variable Y on two continuous explanatory variables x_1 and x_2, which have been scaled to lie between -1 and 1. Each observation is subject to a measurement error with mean 0 and variance σ^2. The following model is proposed.

$$EY = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The investigator proposes to take four observations of (x_1, x_2) given by $(-1, 1), (1, 1), (0, 0)$ and $(0, -1)$. Denote the response for the four observations as Y_1, Y_2, Y_3, Y_4 respectively.

(a) Show that this design is neither D-optimal nor G-optimal, by using the General Equivalence Theorem. You may use the fact that

$$\begin{pmatrix} 4 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 3 \end{pmatrix}^{-1} = \frac{1}{22} \begin{pmatrix} 6 & 0 & -2 \\ 0 & 11 & 0 \\ -2 & 0 & 8 \end{pmatrix}.$$

(6 marks)

5 (continued)

(b) Suggest an alternative design with four observations of (x_1, x_2) that is both D-optimal and G-optimal. ***(3 marks)***

6 (i) An ecologist wishes to estimate the population of rainbow trout in a lake. 50 trout are caught, tagged, and released to mix within the population. Another 100 trout are then caught, and 4 are observed to have tags. Estimate the population of trout, and give a standard error for your estimate. State one assumption that you have made when calculating your estimate.

(3 marks)

(ii) A survey is conducted to estimate the mean income of graduates from a university, five years after graduation.

(a) If stratified sampling is to be used, suggest a suitable choice of strata

(1 mark)

(b) A previous survey produced the following data

Stratum	Population size	Sample size	std. dev. (£)	mean (£)
1	5000	50	2000	30000
2	2000	50	500	22000
3	2000	20	1000	25000
4	1000	10	5000	40000

Estimate the mean income for all graduates. ***(2 marks)***

(c) If a new survey is to be conducted with a sample size of 500, suggest a sample size for each stratum using each of the following methods:

- proportional allocation; ***(2 marks)***
- Neyman allocation; ***(2 marks)***
- minimising the variance of the stratified sample mean, subject to unequal costs: a graduate from stratum 4 is four times as expensive to sample as a graduate from a different stratum; costs for the other three strata are the same. ***(3 marks)***

(You may round so that the total sample size is either 499 or 501, if necessary).

(iii) A computer model is given by the function

$$Y = X_1 + X_2 X_3.$$

The true values of the three inputs are uncertain, with $X_1 \sim N(1, 2)$, $X_2 \sim N(0, 6)$ and $X_3 \sim N(2, 1)$. The model user can choose to learn the true value of one of the inputs. The model user wishes to reduce the variance of the output Y as much as possible.

6 (continued)

(a) Calculate the main effect indices of each input, and hence state which is the best input to learn. You may leave your answers in terms of $Var(Y)$.

(b) Suppose, instead, the model user can choose to learn the value of two of the inputs. Is it best to choose the two with the largest main effect indices from part (a)? Justify your answer.

(7 marks)

End of Question Paper