Data provided: Formulae and Critical Values Sheet

MAS6012

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2015–2016**

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

Candidates may bring to the examination 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 Withdrawals can cause problems in clinical trials.

(i) Using a 'Worst Case Scenario' (WCS) approach to attribution of unknown outcomes where appropriate, provide both

(a) per protocol *(4 marks)*

and

(b) intention to treat *(4 marks)*

analyses of the following (artificial) data from a trial comparing Drug with Placebo.

Assessment	Drug	Placebo	TOTAL
Success	20	20	40
Failure	4	16	20
Total Assessed	24	36	60
Withdrawn	16	4	20
Total Randomized	40	40	80

(c) Explain why the WCS approach is not easy when the primary endpoint is quantitative. *(2 marks)*

(d) Which analysis do you feel is more appropriate and why? *(2 marks)*

(e) What is your overall conclusion about the value of the Drug? *(2 marks)*

(ii) If the trial coordinator adopted conventional values for many of the parameters in his sample size calculations, namely

20% allowance for dropouts
5% probability of a type I error
90% power

and the success rate on Placebo was only anticipated to be 50%, what Clinically Relevant Difference was the trial sized to detect? *(6 marks)*

2 (i) A randomized, double-blind comparative trial of drugs A and B is to be conducted. A statistician proposes that the small (16 patient) study might be conducted according to the following schematic diagram

Drug A		Drug B	
before	after	before	after
X	$\longleftrightarrow X$	X	$\longleftrightarrow X$
X	$\longleftrightarrow X$	X	$\longleftrightarrow X$
$\vdots$		$\vdots$	
X	$\longleftrightarrow X$	X	$\longleftrightarrow X$

where 'X' indicates that the quantitative response is measured and measurements on the same patient are linked. Thus measurments are taken both before and after application of the drug.

The following summary statistics for the original responses (A before, A after, B before, B after), and some derived 'variable', are available:

	n	mean	s.d.
A before	8	96.1	17.53
A after	8	92.4	17.14
B before	8	89.5	17.63
B after	8	91.1	17.26
A after - A before	8	−3.7	2.19
B after - B before	8	1.6	4.34
B before - A before	8	−6.6	5.45
B after - A after	8	−1.3	5.01

(a) Some of the 'variables' quoted above are not legitimate; they have been assigned arbitrary values and hence summary statistics. By considering the study design carefully, state whether each of the 4 derived 'variables' is legitimate or not. Explain your reasoning.
(2 marks)

(b) Suggest how you might test whether or not the drugs had the same effect using the legitimate summary statistics available.
(3 marks)

(c) State the distributional assumptions underpinning your suggestion in (b). ***(1 mark)***

(d) Use appropriate values from the summary statistics given to perform the test you suggested in (b). ***(3 marks)***

2 (continued)

(e) Explain why such a study might be preferable to the simpler trial (of the same size) represented in the following schematic diagram.
(2 marks)

Drug A	Drug B
X	X
X	X
$\vdots$	$\vdots$
X	X

(ii) 24 patients with multiple myeloma (a type of cancer arising from plasma cells) were allocated to one of two forms of treatment and followed up. 11 patients were given drug A and 13 patients drug B. The primary outcome of interest was mortality. The table below shows the data collected. Here the status variable records whether the patient was observed to die (status = 1) or whether they were lost to follow-up beforehand (status = 0).

2 (continued)

	Drug A		Drug B	
	Time (mths)	Status	Time (mths)	Status
	18.70	1	8.90	0
	1.40	1	14.90	1
	0.40	1	14.80	0
	9.20	1	16.50	0
	12.40	1	14.10	1
	23.70	1	13.80	1
	12.10	1	12.80	1
	3.10	0	11.10	0
	26.40	0	19.50	1
	33.50	0	17.30	0
	6.50	1	9.30	1
			19.90	1
			10.50	1
Total	147.4	8	183.4	8

Some R analysis is shown below:

```
> t1 <- c(18.7,1.4,0.4,9.2,12.4,23.7,12.1,3.1,26.4,33.5,6.5)
> t2 <- c(8.9,14.9,14.8,16.5,14.1,13.8,12.8,11.1,19.5,17.3,9.3,19.9,10.5)
> time <- c(t1, t2)
>
> c1 <- c(1,1,1,1,1,1,1,0,0,0,1)
> c2 <- c(0,1,0,0,1,1,1,0,1,0,1,1,1)
> status <- c(c1, c2)
>
> type <- rep(c("A", "B"), c(11, 13))
>
> Comp.sv <- Surv(time, status, type = "right")
```

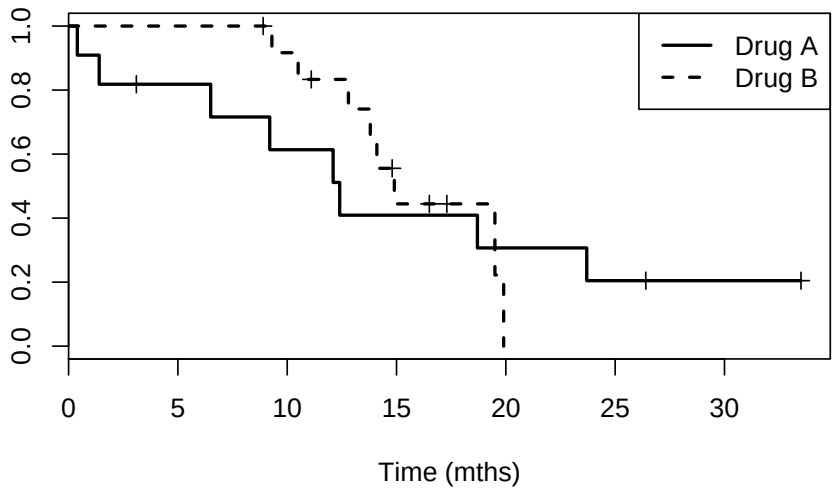
2 (continued)

```
> 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
  0.4     11       1    0.909  0.0867       0.7541        1.000
  1.4     10       1    0.818  0.1163       0.6192        1.000
  6.5      8       1    0.716  0.1397       0.4884        1.000
  9.2      7       1    0.614  0.1526       0.3769        0.999
 12.1      6       1    0.511  0.1578       0.2793        0.936
 12.4      5       1    0.409  0.1559       0.1939        0.863
 18.7      4       1    0.307  0.1467       0.1202        0.783
 23.7      3       1    0.205  0.1286       0.0597        0.701

                type=B
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  9.3     12       1    0.917  0.0798       0.7729        1.000
 10.5     11       1    0.833  0.1076       0.6470        1.000
 12.8      9       1    0.741  0.1295       0.5259        1.000
 13.8      8       1    0.648  0.1426       0.4211        0.998
 14.1      7       1    0.556  0.1493       0.3281        0.941
 14.9      5       1    0.444  0.1554       0.2240        0.882
 19.5      2       1    0.222  0.1753       0.0474        1.000
 19.9      1       1    0.000     NaN           NA           NA

> plot(survfit(Comp.sv ~ type), lty = c(1,2), lwd = 2, xlab = "Time (mths)
```



2 (continued)

(a) It is suggested that the survival times are Exponentially distributed with rates λ_A and λ_B respectively. Under this assumption, estimate λ_A and λ_B and hence the mean mortality times with approximate 95% confidence intervals. ***(4 marks)***

(b) Use the partially complete R output

```
> survdiff(Comp.sv ~ as.factor(type))
Call:
survdiff(formula = Comp.sv ~ as.factor(type))

                    N Observed Expected   ???
as.factor(type)=A 11        8     7.55   ???
as.factor(type)=B 13        8     8.45   ???
```

to perform a non-parametric test assessing whether there is a difference between the two drugs ***(3 marks)***

(c) Discuss the impact on your analysis if it were known that some of the patients who were lost-to-follow up had been withdrawn from the study as their condition was seen to be deteriorating or they showed side effects which need alternative treatment. ***(2 marks)***

3 A clinical trial was conducted to compare two treatments for breast cancer among women. After removal of the tumour, patients were randomly allocated to either treatment A or treatment B and followed up for 5 years for cancer recurrence. The data are stored in `breast` and coding for the different variables is shown below:

Coding:
Treat: treatment (0 = treatment A; 1 = treatment B)
Obese: indicator if patient is obese (0 = non-obese; 1 = obese)
Age: age of patient centred on 50 years (i.e. 55yr yr old is +5)
Time: time until cancer recurrence (yrs)
Status: indicator of relapse (1) or censoring (0)

(i) Some R analysis was performed with the edited output shown below:

```
> Breast.sv <- Surv(Time, Status)
>
> Br.fit <- coxph(Breast.sv ~Obese+Treat+Age)
> summary(Br.fit)
Call:
coxph(formula = Breast.sv ~ Obese + Treat + Age)

  n= 200, number of events= 170

           coef exp(coef)  se(coef)      z Pr(>|z|)
Obese  0.221694  1.248189  0.155460  1.426 0.153855
Treat -0.537459  0.584231  0.158781 -3.385 0.000712 ***
Age    0.023620  1.023902  0.007944  2.973 0.002945 **
---
```

(a) Specify the form the model used for this analysis in terms of the baseline hazard function $h_0(t)$ and the covariates. ***(3 marks)***

(b) Describe in detail the effects of these variables on the time to recurrence. ***(4 marks)***

(c) Using the model, calculate the estimate of the hazard ratio comparing

- An *obese* 56yr old woman on treatment A
- A *non-obese* 43yr old woman on treatment B

(3 marks)

3 (continued)

(ii) An alternative analysis is shown below

```
> Br.fit1 <- survreg(Breast.sv ~Obese+Treat+Age, dist="exponential")
> summary(Br.fit1)

Call:
survreg(formula = Breast.sv ~ Obese + Treat + Age, dist = "exponential")
              Value Std. Error     z        p
(Intercept)  0.6274    0.14495  4.33 0.000015
Obese       -0.2094    0.15447 -1.36 0.175219
Treat        0.5825    0.15563  3.74 0.000182
Age         -0.0251    0.00769 -3.27 0.001085

Scale fixed at 1

Exponential distribution
Loglik(model)= -266.1   Loglik(intercept only)= -281
Chisq= 29.91 on 3 degrees of freedom, p= 1.4e-06
Number of Newton-Raphson Iterations: 5
n= 200
```

(a) What type of analysis has been performed here and write down the fitted model for T, the time to recurrence? ***(4 marks)***

(b) Comment on the effects of the different covariates with this approach. Would you prefer to be on treatment A or B? ***(3 marks)***

(c) Estimate the expected time to recurrence for a 56yr old obese individual who is taking treatment B ***(3 marks)***

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 2 different designs, both using 4 observations:

Design	Design points
A	$x_1 = \{-1, -0.5, 0.5, 1\}$
B	$x_1 = \{-1, -1, 1, 1\}$

(i) Show that β_1 and β_{11} are orthogonal to each other in design B. ***(4 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 for design B. ***(2 marks)***

(iii) Experiments are performed and the response (Y) is measured at the 4 points in both designs A and then separately at the 4 points in design B. Describe the geometric shapes of the 95% confidence regions for $(\beta_1, \beta_{11})^T$ for both designs A and B. Justify whether the centres of these shapes coincide. ***(4 marks)***

(iv) Justify whether design B is G-optimal. ***(5 marks)***

(v) A design is called A-optimal if it minimises the sum of the diagonal elements of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Consider all **orthogonal** designs for the model $EY = \beta_1 x_1 + \beta_{11} x_1^2$ with 2 designs points such that $-1 \leq x_1 \leq 1$. Show that the design $x_1 = \{-1/2, 1/2\}$ is not A-optimal among all the **orthogonal** designs for this model with 2 design points and specify a design that is A-optimal. ***(5 marks)***

5 Consider a fractional factorial design with 4 factors (x_1, x_2, x_3, x_4) each of which occurs at two levels, denoted $+1$ and -1. Only 4 design points can be used.

(i) Specify the alias structure when the design generators are $x_1x_2 = 1$ and $x_2x_3 = 1$. ***(5 marks)***

(ii) Which parameters are confounded with the main effect parameter for x_1? ***(2 marks)***

(iii) What is the resolution of this fractional factorial design? Justify your answer. ***(2 marks)***

(iv) Suppose that 8 design points are now available with 4 factors (x_1, x_2, x_3, x_4). State the single design generator that allows the intercept, all main effects and the 3 pairwise interactions x_1x_2, x_1x_3 and x_1x_4 to be included in the linear model without confounding. ***(3 marks)***

(v) If 8 design points are still available, construct a fractional factorial design with the design generator $1 = x_1x_2x_3$. ***(3 marks)***

(vi) Two organisations have independently attempted to estimate the number of civilian casualties following a civil war. Each organisation has produced a list of named civilian casualties. The two lists are then cross-checked, to see which names appear on both lists. The following counts are observed.

Number of names on both lists	213
Number of names on first list only	802
Number of names on second list only	410

Estimate the total number of civilian casualties, stating any assumptions you have made. Give an approximate confidence 95% interval for the total number of civilian casualties. ***(5 marks)***

6 (i) A factory has manufactured 100 steel items. A simple random sample of 10 items are selected, and a property known as the Brinell hardnesses is measured for each item. Summary statistics for the ten items are below (where the unit of measurement is the Brinell hardness number, HB).

$$\sum_{i=1}^{10} x_i = 4551, \qquad \sum_{i=1}^{10} x_i^2 = 2,080,125.$$

Suppose the mean hardness is to be estimated for a second batch of 100 items. Another simple random sample is to be taken. Suggest a suitable sample size such that the width of an approximate 95% confidence interval is no more than 20HB, justifying your answer carefully. What assumptions have you made? ***(7 marks)***

(ii) A stratified sample has been taken to estimate annual household expenditure on energy bills in a town. Two strata are chosen based on postcodes. Summary data are below.

Stratum	Stratum size	Sample size	Sample mean (£)	Sample std. dev. (£)
1	20000	50	609.0	245.8
2	10000	50	438.8	116.6

(a) Estimate the mean annual household expenditure on energy bills for the town. ***(1 mark)***

(b) Give an approximate 95% confidence interval for the mean expenditure. ***(3 marks)***

(c) Suppose the survey is to be repeated next year, with the same total sample size. Assuming that costs of sampling within each stratum are the same, suggest sample sizes within each stratum for the new survey. ***(2 marks)***

6 (continued)

(iii) A survey has been conducted to estimate the use of essay-writing services in a cohort of History undergraduate students. Each participant was asked to toss a coin twice. If the two coin tosses produced the same result, the participant answered the question, "Did you observe two heads?". If the two coin tosses produced different results, the participant answered the question, "Have you ever used a commercial essay-writing service for a coursework assignment?"

(a) If there are 30 "yes" responses from 100 participants, estimate the proportion of students who have used an essay-writing service, showing clearly how your estimate has been obtained. *(2 marks)*

(b) Calculate the estimated variance of your estimator, ignoring finite population corrections and assuming that the number of "yes" responses is binomially distributed. *(2 marks)*

(c) Suggest how you might reduce the variance of your estimator, by modifying the experiment (assuming the same number of participants). You do not need to derive the variance for your modified experiment, or prove that it is smaller. Give one potential drawback of your modification. *(3 marks)*

End of Question Paper