Data provided: Neave's Tables, Graph Paper

MAS6012

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2012–2013**

Sampling, Design, Medical Statistics **2 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 Obesity is a well-known risk factor in many diseases and a clinician is hoping to obtain a clear indication of its role in hypertension (high blood pressure) for males in the 'over 40 years' age range.

(i) A quick literature search finds 3 studies in which findings are presented for males in the relevant age range. The results are summarized below. The clinician is keen to quote the studies, but concerned that the small study sizes might mean their results are unreliable.

Study I

	Blood Pressure raised	not	
Obese	28	72	100
Not obese	18	82	100
	46	154	200

Study II

	Blood Pressure raised	not	
Obese	20	36	56
Not obese	23	101	124
	43	137	180

Study III

	Blood Pressure raised	not	
Obese	24	54	78
Not obese	30	140	170
	54	194	248

(a) The authors of Study I used a standard χ^2 test with a 5% significance level to assess whether raised blood pressure is associated with obesity. Assume that 20% of non-obese males over 40 years of age suffer from hypertension. If 30% of obese males over 40 suffer from hypertension, show that the power to detect this difference is actually less than 50%. ***(5 marks)***

(b) The clinician decides to combine all the studies using a meta-analysis. Explain under what conditions this is a sensible approach. Assuming these are satisfactory, carry out the analysis for him. ***(6 marks)***

(ii) In fact, he suspects that numerous factors are important in determining blood pressure, so is pleased to discover a larger and more sophisticated study in which an analysis based on logistic regression was conducted. The following data (from Altman, 1995) show the results. Note that Obesity, Smoking and Snoring are indicators with 1 indicating that a subject exhibits the factor.

1 (continued)

factor	regression coefficient b	s.e.(b)	z	p
Obesity	0.695	0.285	2.44	0.015
Constant	-2.378	0.380		
Smoking	-0.068	0.278	0.24	0.81
Snoring	0.872	0.398	2.199	0.028

(a) Explain briefly what is meant by 'logistic regression', ensuring your terminology and/or notation is clear. ***(4 marks)***

(b) Explain, in terms that might be understood by a non-statistician, how the risk of exhibiting hypertension differs for males in this age group who are obese and non-obese. How does the risk differ for smokers and non-smokers? ***(5 marks)***

2 The following tables (adapted from Matthews 1989) show data and derived statistics from a trial investigating the effects of two treatments (A and B) on asthma (values are FEV1, a measure of lung function, in litres; high values are good). The design is an AB/BA crossover.

Subject	Group	Period1	Period2	sum	diff
1	1	1.28	1.33	2.61	-0.05
2	1	1.60	2.21	3.81	-0.61
3	1	2.46	2.43	4.89	0.03
4	1	1.41	1.81	3.22	-0.40
5	1	1.40	0.85	2.25	0.55
6	1	1.12	1.20	2.32	-0.08
7	1	0.90	0.90	1.80	0.00
8	1	2.41	2.79	5.20	-0.38
9	2	2.68	2.10	4.78	0.58
10	2	2.60	2.32	4.92	0.28
11	2	1.48	1.30	2.78	0.18
12	2	2.08	2.34	4.42	-0.26
13	2	2.72	2.48	5.20	0.24
14	2	1.94	1.11	3.05	0.83
15	2	3.35	3.23	6.58	0.12
16	2	1.16	1.25	2.41	-0.09

2 (continued)

Group 1: A then B ($n_1 = 8$)				
	Period 1	Period 2	Sum (1+2)	Difference (1-2)
mean	1.5725	1.69	3.2625	-0.1175
s.d.	0.5717829	0.7311244	1.263914	0.3542295
Group 2: B then A ($n_2 = 8$)				
	Period 1	Period 2	Sum (1+2)	Difference (1-2)
mean	2.25125	2.01625	4.2675	0.235
s.d.	0.7215348	0.7381238	1.417954	0.3468223

(i) Plot the treatment means for each period and make a preliminary graphical assessment of the trial's findings. ***(5 marks)***

(ii) Assess whether there is any evidence of a carryover effect from Period 1 to Period 2. ***(4 marks)***

(iii) Assess whether there is any evidence of a difference in mean response between Periods 1 and 2. ***(3 marks)***

(iv) Assess whether there is any evidence of a Treatment effect, taking into account the results of your analyses in (ii) and (iii). ***(3 marks)***

(v) Suppose it was later discovered that Subjects 13 and 14 had not in fact completed their Period 2 treatment (i.e. Treatment A) correctly, because they found it had unpleasant side effects.

(a) How would this affect your confidence in the conclusions reached above? ***(3 marks)***

(b) Would you propose any revised action or analysis? Explain your answer. [NB You need not actually carry out any new analysis proposed.] ***(2 marks)***

3 (i) Consider the Cox Proportional Hazards Model.

(a) Write down the model, specifying your notation clearly. ***(2 marks)***

(b) Why is this model described as 'semi-parametric'? ***(1 mark)***

(c) What assumptions does it make and how might these be verified? ***(2 marks)***

3 (continued)

(ii) The table below gives the results of fitting a Cox Proportional Hazards model to survival time data (from Altman, 1995) from a trial of Azathioprine as a therapy for Primary Biliary Cirrhosis (a liver disease). The coding of the variables is explained in the lower table.

Variable	regression coefficient b	s.e.(b)	exp(b)
Serum bilirubin	2.0510	0.316	7.78
Age	0.00690	0.00162	1.01
Cirrhosis	0.879	0.216	2.41
Serum albumin	-0.0504	0.0181	0.95
Central cholestasis	0.679	0.275	1.97
Therapy	0.520	0.207	1.68

Variable	scoring
Serum bilirubin	$\log_{10}$ (value in μ mol/l)
Age	exp[(age in years - 20)/10]
Cirrhosis	0=No; 1=Yes
Serum albumin	value in g/l
Central cholestasis	0=No; 1=Yes
Therapy	0=Azathioprine; 1=Placebo

(a) Is there evidence that the therapy is beneficial? Explain your answer. ***(4 marks)***

(b) If the Serum bilirubin value (on original scale) increases by a factor of 10, all other variables remaining constant, how will the hazard change? ***(3 marks)***

(c) What can be said about the probabilities of surviving 3 years for two patients of the same age, same Serum bilirubin value and same Cirrhosis and Central cholestasis status, one of whom has a Serum albumin level of 40 g/l and is given Azathioprine, while the other has Serum albumin of 50.32g/l but is given the Placebo? ***(4 marks)***

(iii) Suppose a further study of the same Therapy and covariates is to be undertaken. Because of the large number of covariates, a dynamic randomization procedure is to be used to allocate incoming patients to either Azathioprine or Placebo in such a way as maintain as much balance as possible. To simplify handling of continuous covariates, these have all been dichotomized to 'High' or 'Low'. After 40 patients have been enrolled, the marginal counts are as in the table below. If the next patient has high Serum bilirubin, low Age, Cirrhosis, high Serum albumin and no Central cholestasis, should they be allocated to Azathioprine or Placebo?

3 (continued)

Variable	Level	Azathioprine	Placebo
Serum bilirubin	High	15	16
	Low	5	4
Age	High	11	12
	Low	9	8
Cirrhosis	0	14	12
	1	6	8
Serum albumin	High	14	13
	Low	6	7
Central cholestasis	0	7	8
	1	13	12

(4 marks)

4 A small experiment is being conducted to compare a new treatment against a placebo. There are four participants in the study. Three participants are given the placebo, and one the new treatment. 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 $i = 1, 2$, $j = 1, \ldots, n_i$, with $n_1 = 3$ and $n_2 = 1$. The constraint $\tau_1 + \tau_2 = 0$ is applied.

(i) Show that the estimators

$$\hat{\mu} = \frac{Y_{11} + Y_{12} + Y_{13} + 3Y_{21}}{6},$$
$$\hat{\tau}_1 = \frac{Y_{11} + Y_{12} + Y_{13} - 3Y_{21}}{6},$$

are unbiased, and give their variances in terms of σ^2 only. ***(3 marks)***

(ii) What condition must hold for μ and τ_1 to be orthogonal? Show whether this condition is satisfied or not. ***(3 marks)***

(iii) Calculate the standardised prediction variance for a new observation in each group. By considering the maximum of these two variances, explain why the design is not D-optimal. ***(7 marks)***

(iv) Suggest an alternative design with four participants such that μ and τ_1 are orthogonal, verifying orthogonality with a suitable calculation. ***(4 marks)***

4 (continued)

(v) If the response was believed to depend on the characteristics of the patient, explain how you would implement blocking to account for this, and state the model you would fit to the data to incorporate block effects. Specify any additional necessary parameter constraints. ***(3 marks)***

5 An experiment is to be carried out to investigate the effect of three methods on reading comprehension. There are 6 participants in the study, who will each be given instruction using one of the three methods. After instruction, each participant will be given a reading test, and their scores will be recorded.

(i) State a balanced incomplete block design, with block sizes of 2, that could be used in the experiment. ***(1 mark)***

(ii) For your design in (i), write down the model you would fit to the data in matrix notation, defining your notation carefully, specifying any parameter constraints, and writing out the observation vector, parameter vector and design matrix in full. ***(5 marks)***

(iii) Suggest how you could organise participants into blocks, assuming you could obtain suitable data before choosing the blocks. ***(2 marks)***

(iv) Suppose a larger study is to be conducted, with more participants. If the block size is to be fixed at 2, what is the smallest sample size that is greater than 6, that could be used in a balanced incomplete block design? Justify your answer. ***(2 marks)***

(v) Suppose now there are 9 participants, three methods of instruction, and three teachers. The aim is now to assess the effect of teacher as well as method of instruction.

(a) If the participants are to be organised into blocks of size 3, give a suitable design using a Latin square. State your design by completing the following table.

participant	block	method	teacher
1			
⋮			
9			

(4 marks)

(b) If it is suspected that some teachers may be more effective when using particular methods, without making any further assumptions, would your design in part (a) be suitable? Briefly justify your answer. What further assumption would you have to make in order to fit a suitable model to the data? ***(6 marks)***

6 (i) In a warehouse containing individual frozen beef lasagnes, it is suspected that a proportion of them may contain horse meat. A simple random sample will be taken, and the proportion will be estimated. How large does the sample need to be to ensure that a 95% confidence interval for the true proportion is no wider than 0.05. It is thought to be very unlikely that the proportion containing horse meat will exceed 0.1. You may ignore the finite population correction. Note that the 97.5th percentile of a standard normal random variable is 1.96.

(5 marks)

(ii) In a health study, the proportion of adults who are smokers is to be estimated. In a pilot survey, a stratified sample is taken using four strata, and the results are tabulated below.

Stratum	Population size (millions)	Sample size	Number of smokers
1	15	50	25
2	10	50	15
3	5	50	5
20	10	50	20

(a) Estimate the proportion of smokers in the population. ***(1 mark)***

(b) Using a normal approximation, calculate a 90% confidence interval for the population proportion of smokers. You may ignore the finite population correction. Note that the 5th percentile of a standard normal random variable is -1.645. ***(5 marks)***

(c) If a larger sample was to be taken, with 1000 participants, suggest a sample size for each stratum using Neyman allocation. ***(2 marks)***

(iii) A survey has been conducted to estimate the proportion of adults in a population who have driven a car within two hours of consuming alcohol. Each participant first tosses a coin and rolls a six-sided die in secret. If the outcome of the coin toss is a head, the participant answer question A: "Have you ever driven a car within two hours of consuming alcohol?". If the outcome of the coin toss is a tail, the participant answers question B: "Was the outcome of the die-roll an even number?" There are 100 participants in the survey.

(a) Out of the 100 participants, 30 answer "Yes". Estimate the proportion of adults in the population who have driven a car within two hours of consuming alcohol, and estimate the variance of your estimator. ***(5 marks)***

(b) Suggest a modification to question B that would reduce the variance of the estimator of the proportion, if the survey were to be repeated, and give the variance of the new estimator. ***(2 marks)***

End of Question Paper