

M. PHIL. IN STATISTICAL SCIENCE

---

9 am Monday 11 June until 1 pm Thursday 14 June 2001

---

APPLIED STATISTICS

*Attempt **three** of the four questions.*

*This is an 'Open-Book' examination, involving use of the Statistical Laboratory's network of workstations. Candidates will receive this paper at 9.00 a.m. on Monday 11 June, and must hand in their scripts to the Chairman of Examiners by 1.00 p.m. on Thursday 14 June.*

*The data-sets will be emailed to candidates on Monday 11 June.*

*(The Statistical Laboratory Computer Officer and an Examiner will normally be available for consultation if required between 9.00 a.m. and 4.30 p.m. on these four days.)*

*Each candidate should submit his/her script with a signed statement that the work has been carried out without any collaboration with others.*

*The scripts may be handwritten. Candidates are requested to submit at most 25 pages in total. They are advised that the total work set should take between 4 and 6 hours.*

*The questions carry equal weight.*

**You may not start to read the questions  
printed on the subsequent pages until  
instructed to do so by the Invigilator.**

1 The Independent, October 2 and 3, 2000, gives the following list of 78 Sydney medals winners G,S,B = gold, silver, bronze respectively, MMP= ‘medals in proportion to population’ and MBGDP= medals per \$billion GDP.

	G	S	B	total	MMP	MBGDP
USA	39	25	33	97	0.36	0.01
Russia	32	28	28	88	0.60	0.32
China	28	16	15	59	0.05	0.06
Australia	16	25	17	58	3.09	0.16
Germany	14	17	26	57	0.69	0.03
France	13	14	11	38	0.65	0.03
Italy	13	8	13	34	0.59	0.03
Nether'ds	12	9	4	25	1.59	0.07
Cuba	11	11	7	29	2.61	1.45
Britain	11	10	7	28	0.47	0.02
Romania	11	6	9	26	1.16	0.68
.....						
Chile	0	0	1	1	0.07	0.01
Iceland	0	0	1	1	3.65	0.13
India	0	0	1	1	0.01	0.00
Kyrgyzstan	0	0	1	1	0.21	0.59
Kuwait	0	0	1	1	0.54	0.04
Macedonia	0	0	1	1	0.50	0.40
Qatar	0	0	1	1	1.35	0.08
SriLanka	0	0	1	1	0.05	0.06

(i) Compute the variables pop and BGDP by

$$\text{pop} = \text{total}/\text{MMP} \text{ and } \text{BGDP} = \text{total}/(\text{MBGDP} + \cdot005)$$

(the  $\cdot005$  is introduced for obvious reasons)

Summarise the data, including pop and BGDP, by appropriate graphs and tables.

(ii) Now investigate a linear regression of score on  $\log(\text{pop})$  and  $\log(\text{BGDP})$ , where score is defined by

$$\text{score} = 3*\text{G} + 2*\text{S} + 1*\text{B}$$

Is this this best scale for your regression?

(iii) Give a brief discussion of principal components, as applied to the data matrix

$$(\log(\text{score}), \log(\text{pop}), \log(\text{BGDP})).$$

<b>2</b>	E	29.0	Ff	14.5	D	20.5	A	22.5	B	16.0	C	6.5
	B	17.5	A	29.5	E	12.0	C	9.0	D	33.0	Ff	12.5
	Ff	17.0	B	30.0	C	13.0	D	29.0	A	27.0	E	12.0
	A	31.5	D	31.5	Ff	24.0	E	19.5	C	10.5	B	21.0
	D	25.0	C	13.0	B	31.0	Ff	26.0	E	19.5	A	–
	C	12.2	E	13.0	A	34.0	B	20.0	Ff	–	D	–

The above table shows the outcome of an agricultural experiment to compare the weights of 6 varieties of turnips, labelled  $A, B, C, D, E, Ff$ . The  $6 \times 6$  grid shows the layout of the turnip varieties on the field, according to *rows* and *columns*. Different rows may have different fertilities, as may different columns. Unfortunately this latin square design was vandalised, perhaps by environmental protestors.

(i) Write down clearly an appropriate model for the above design, and analyse it carefully. Test whether there is a difference between the 6 varieties.

Please show your S-Plus5 commands.

(ii) Use the function

`multicomp()`

as a further comparison of the 6 varieties.

(iii) If you had to estimate the missing values in the lower right-hand plots, how would you do so?

**3** The data set ‘biopsy’, of which the first few lines appear in the Table below, is an edited version of the Venables and Ripley library (MASS) data-set: any row containing an NA has been deleted. As you can see from

```
? biopsy
```

this data frame, which results from biopsies of breast cancer tumours of 699 patients, contains the following measurements

ID Sample code number

V1 Clump thickness

V2 Uniformity of cell size

V3 Uniformity of cell shape

V4 Marginal adhesion

V5 Single epithelial cell size

V6 Bare nuclei

V7 Bland chromatin

V8 Normal nucleoli

V9 Mitoses

Class Class of tumor, “benign” or “malignant”.

We are interested in predicting *class* from  $V_1, \dots, V_9$ .

Table

	ID	V1	V2	V3	V4	V5	V6	V7	V8	V9	class
1	1000025	5	1	1	1	2	1	3	1	1	benign
2	1002945	5	4	4	5	7	10	3	2	1	benign
3	1015425	3	1	1	1	2	2	3	1	1	benign
4	1016277	6	8	8	1	3	4	3	7	1	benign
5	1017023	4	1	1	3	2	1	3	1	1	benign
6	1017122	8	10	10	8	7	10	9	7	1	malignant
7	1018099	1	1	1	1	2	10	3	1	1	benign
8	1018561	2	1	2	1	2	1	3	1	1	benign
9	1033078	2	1	1	1	2	1	1	1	5	benign
10	1033078	4	2	1	1	2	1	2	1	1	benign

(i) Summarise, with appropriate graphs,  $V_1, \dots, V_9$  for each of the 2 values of class.

(ii) How does  $P(\text{malignant})/P(\text{benign})$  depend on the 9 variables? (You may like to use `stepAIC( )` from library (MASS).)

How well does your best model predict malignant tumours?

(iii) Now try

```
tree(class ~ ...)
```

and compare the results with those of (ii).

(iv) If the physician can only afford to measure 7 of the 9 variables, which ones would you recommend that he use?

4 The data below is from a hypothetical study to assess whether a computerised scheduling intervention would be effective in reducing the waiting time for a certain knee operation after the decision to operate had been made. The study was designed as a randomised-controlled trial of the computerised intervention versus common practice in two centres: Centre A and Centre B. Centre A recruited 200 patients, while Centre B recruited 300 patients. The trial was conducted over a three-year period, with an eighteen-month recruitment period and an additional eighteen months follow-up at the end of this recruitment phase. Additional information on the patient’s age and sex was recorded. Patients under 18 years of age were excluded. It is also known from an audit that took place before the study began that patients aged between 35 and 45, in general, wait longer than patients who are aged 45 years or older or aged 35 years or younger. It was thought at the beginning of the study that the impact of treatment might possibly differ for males and females.

Note that patients may drop out of the study. For example, they may decide to go “private”. However, information on why they dropped out was not recorded. Note also that information on the severity of the knee injury was not recorded.

Using survival data analysis methods, find an appropriate way of modelling this data. Interpret the results and test any assumptions (graphically and/or otherwise) made. Give possible (plausible) reasons for your findings.

ID	Time	Op	Age	Sex	Intervention	Centre	
1	192	0	39	1	0	0	
2	101	1	46	1	0	0	
3	197	1	28	0	0	0	
4	66	1	49	0	0	0	
5	345	1	35	0	1	0	
6	113	1	45	1	0	0	
7	62	1	48	1	1	0	
8	6	0	30	1	1	0	
9	39	1	58	1	1	0	
10	174	0	43	0	1	0	
11	137	1	39	1	0	0	
12	73	1	35	1	0	0	
.....							
	488	34	1	43	1	1	1
	489	80	0	45	1	0	1
	490	24	1	22	1	1	1
	491	20	0	44	0	1	1
	492	33	1	43	1	1	1
	493	18	1	37	1	1	1
	494	27	1	40	1	1	1
	495	41	1	53	0	0	1
	496	43	1	39	1	0	1
	497	27	0	24	0	0	1
	498	27	1	21	0	1	1
	499	81	1	34	0	0	1
	500	57	1	37	1	0	1

ID = Patient's identification number

Time = Waiting time for operation (in days)

Op = Operation done? (0=No, 1=Yes)

Age = Age of patient at entry

Sex = Gender of patient (0=Female, 1=Male)

Intervention = What intervention was the patient randomised to receive?  
(0=Current, 1=Computerised)

Centre = Centre which study took place in (0=Centre A, 1=Centre B)

The above table gives only the beginning and end of the *full* data set, which will be emailed to you.