

M. PHIL. IN STATISTICAL SCIENCE

1.30pm, Monday 9 June to 4.30pm, Thursday 12 June 2008

APPLIED STATISTICS

*Attempt at most **THREE** questions.
There are **FOUR** questions in total.
The questions carry equal weight.*

This is an 'Open Book' examination, involving the use of the Statistical Laboratory's network of workstations. Candidates will receive this paper at 1.30pm on 9th June, and must hand in their scripts to the Chairman of Examiners by 4.30pm on 12th June. The data sets will be emailed to candidates on 9th June. (The Statistical Laboratory Computer Officer and an Examiner will normally be available for consultation if required between 9.00am and 4.30pm 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. Candidates are advised to state models algebraically and to discuss formally the details of their statistical analyses.

STATIONERY REQUIREMENTS

*Cover sheet
Treasury Tag
Script paper*

SPECIAL REQUIREMENTS

None

<p>You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.</p>

1 The Ibrahim Index of African Governance for 48 sub-Saharan countries in Africa is obtained by averaging the five indices shown as the last five columns below. The first few lines are shown for the data from 2005. The quantities are:

II: the Ibrahim Index

SafetySec: Index A, an index of safety and security;

RuleOfLaw: Index B, an index of rule of law, transparency and corruption;

HumanR: Index C, an index of participation and human rights;

EconOpp: Index D, an index of sustainable economic opportunity;

HumanDev: Index E, an index of human development.

	II	SafetySec	RuleOfLaw	HumanR	EconOpp	HumanDev
Mauritius	86.2	91.7	85.2	88.7	75.5	90.0
Seychelles	83.1	83.3	74.2	79.3	80.5	98.3
Botswana	73.0	75.0	88.3	75.5	58.1	67.9
CapeVerde	72.9	84.0	80.4	74.6	52.8	72.8
SAfrica	71.1	61.1	75.2	81.1	67.4	70.5

The first few lines of the 2000 scores are given below with the same indices (00 at the end of each index name shows it for the year 2000).

	II00	SafetySec00	RuleOfLaw00	HumanR00	EconOpp00	HumanDev00
Mauritius	80.7	91.7	81.2	79.0	63.0	88.6
Seychelles	81.5	83.3	75.5	78.9	72.0	97.8
Botswana	72.0	75.0	87.3	78.1	52.2	67.1
CapeVerde	69.9	84.0	77.5	65.5	47.0	75.6
SAfrica	70.9	61.1	76.5	87.4	60.6	69.0

- Investigate the 2005 data and the 2000 data with suitable plots, tables and summaries.
- Using appropriate tests, investigate whether there are improvements between 2000 and 2005 for any of the six indices.
- Investigate how the index of sustainable economic opportunity depends on the year and indices A, B, C and E.

2 The table below shows the number of male and female applicants to UK universities for entry in 2005 and 2006 from various non-UK regions, together with the number of successful applicants in each category (source: UCAS). App is the number of applicants, and Succ is the number who are successful. The region FarEast1 is the Far East, except for Hong Kong and Malaysia. The region Europe(nonEU) includes European countries that are not in the European Union. The category Unknown includes applicants who are stateless or whose nationality is unknown.

	2005				2006			
	Men		Women		Men		Women	
	App	Succ	App	Succ	App	Succ	App	Succ
Africa	8662	4136	4337	2051	5376	2983	3437	1872
Americas	1838	836	2673	1278	1901	932	2879	1442
Australasia	141	59	193	72	150	54	156	63
Europe(nonEU)	1248	807	1571	949	1379	902	1805	1174
FarEast1	10997	6719	8008	4967	10420	6652	7550	4860
HongKong	1555	1115	1457	1030	1577	1163	1651	1197
Malaysia	1545	997	1319	881	1285	894	1221	804
MiddleEast	1872	1082	972	486	2077	1210	1041	587
Unknown	254	243	175	170	160	151	111	106

- (a) Use χ^2 tests to determine whether or not the data are consistent with the proportions of applications from women being the same for both years.
- (i) for the region FarEast1;
 - (ii) for all regions combined.
- (b) For the *success rates* in 2006, is there a significant **Gender** effect in an appropriate model that also contains **Region**? How do the 2006 success rates depend on Region?
- (c) Investigate how success rates depend on Region, Year and Gender.

3 The number of operations `nop` for various categories of surgical procedure at various levels of risk, together with the number `nssi` of those operations where there was a surgical site infection, between 1997 and 2005, are shown below (source: The Health Protection Agency). The risk index `Risk` is a standard measure of risk for surgical procedures, and has four levels 0, 1, 2, 3, of increasing risk, and one further category `U` where the risk level is unknown.

SurgProc	Risk	nssi	nop
AbdomHyst	0	138	8731
AbdomHyst	1	46	1313
AbdomHyst	2	9	120
AbdomHyst	3	1	5
AbdomHyst	U	45	2402
CorBypass	0	12	738
CorBypass	1	732	19413
CorBypass	2	116	1505
CorBypass	3	1	4
CorBypass	U	287	6657
Gastric	0	0	19
Gastric	1	2	26
Gastric	2	2	11
Gastric	3	1	3
Gastric	U	7	114
Hip	0	497	40472
Hip	1	441	20376
Hip	2	145	3520
Hip	3	9	86
Hip	U	275	15212
Knee	0	257	35852
Knee	1	220	15990
Knee	2	48	1931
Knee	3	1	25
Knee	U	129	14039
LBowel	0	264	4744
LBowel	1	456	4871
LBowel	2	306	1926
LBowel	3	58	236
LBowel	U	241	2589
Amputation	0	16	179
Amputation	1	78	679
Amputation	2	96	583
Amputation	3	23	123
Amputation	U	55	494
OpenRed	0	81	4773
OpenRed	1	132	3925
OpenRed	2	41	537

OpenRed	3	3	31
OpenRed	U	62	2082
SBowel	0	24	297
SBowel	1	43	637
SBowel	2	66	393
SBowel	3	13	52
SBowel	U	42	415
Vascular	0	44	1570
Vascular	1	200	3825
Vascular	2	193	1854
Vascular	3	12	35
Vascular	U	82	1586
HipHem	0	227	6689
HipHem	1	550	11467
HipHem	2	96	1521
HipHem	3	3	7
HipHem	U	232	5117

The surgical procedures are:

AbdomHyst: Abdominal Hysterectomy

CorBypass: Coronary Artery Bypass Surgery

Gastric: Gastric Surgery

Hip: Hip Prosthesis

Knee: Knee Prosthesis

LBowel: Large Bowel Surgery

Amputation: Limb Amputation

OpenRed: Open Reduction of Long Bone Fracture

SBowel: Small Bowel Surgery

Vascular: Vascular Surgery

HipHem: Hip Hemiarthroplasty

- (a) Carry out a preliminary investigation of the data giving suitable plots and summaries.
- (b) Fit and interpret an appropriate Poisson model where the logarithm of the rate of surgical site infection depends *additively* on the type of surgical procedure and the risk level. Give the fitted expected number of surgical site infections given by this model for the 20376 hip prosthesis operations at risk level 1, together with an approximate confidence interval.
- (c) In the additive model in (b), test whether the rates of surgical site infection for the two risk groups 1 and U are the same.

4 Shown below is a subset of a dataset on 648 patients from a study on Parkinson's disease (PD), where the age at diagnosis of PD, age at entry, death status, type of PD drug used at entry and the disease severity at diagnosis (measured through the Unified Parkinson's Disease Rating Scale, where a score of 0 corresponds to no disability and a score of 199 the most severe disability) of the patients were recorded. In addition the observed age at death or age at end of 2007 if alive were also recorded.

Ageatdiag	Ageatentry	Age	Death	Sex	Drug	UPDRS
52.87	59.42	63.344	1	1	2	0
82.67	89.32	92.938	1	0	3	49
62.92	68.52	72.82	1	1	1	47
43.03	51.78	55.69	1	0	2	0
.
40.49	46.15	51.13	1	0	3	0
79.59	86.18	93.01	1	0	1	8
79.97	86.14	87.488	1	1	2	58
55.3	63.86	68.206	1	0	1	12

Ageatdiag = Age of patient at diagnosis (in years)

Ageatentry = Age of patient at study entry (in years)

Age = Observed age at death or age at the end of 2007 if still alive (in years)

Death = Death status (0 corresponds to alive; 1 to death)

Sex = Sex of patient (0 corresponds to female; 1 to male)

Drug = Drug treatment for Parkinson's

(1 = Levodopa, 2 = Dopamine agonist, 3 = MAO-B inhibitors + Levodopa)

UPDRS = Unified Parkinson's Disease Rating Scale (0-199)

Recruitment of patients to the study took place in the year 2000, where patients diagnosed with PD during the period 1990 to 1995 were invited to participate. The researchers in the study followed up the PD patients until the end of 2007, where patients still alive were administratively censored at that time for the purposes of analysis.

The researchers are interested in the survival experiences of the patients, in particular, the 5-year and 10-year overall survival prospect, and the prognostic effects of the variables collected on survival. They however do not know how to analyse the data and approach you with their data.

By the appropriate modelling of the data and use of plots, analyse the data in order to answer the questions of interest to the researchers, paying particular attention to the most clinically appropriate time scale to use for measuring the survival experiences of patients. Interpret the results and test any assumptions made (graphically and/or otherwise).

END OF PAPER