First Semester Examination
2017/2018 Academic Session

January 2018

MST567 - CATEGORICAL DATA ANALYSIS
[ANALISIS DATA BERKATEGORI]

Duration : 3 hours
(Masa : 3 jam)

Please check that this examination paper consists of **TEN (10)** pages of printed material before you begin the examination.

*[Sila pastikan bahawa kertas peperiksaan ini mengandungi **SEPULUH (10)** muka surat yang bercetak sebelum anda memulakan peperiksaan ini].*

Instructions : Answer **all eight (8)** questions.

*[**Arahan :** Jawab **semua lapan (8)** soalan.]*

In the event of any discrepancies, the English version shall be used.

[Sekiranya terdapat sebarang percanggahan pada soalan peperiksaan, versi Bahasa Inggeris hendaklah digunapakai].

Question 1

Patients suffering from chest pain were assigned to receive each of a new (Drug A) and a standard therapy (Drug B) in random order. At the end of the treatment each patient was asked whether they felt better when receiving Drug A as compared to Drug B.

(a) Use data in Table 1 to construct an approximate test of whether or not there is any difference in how well the two drugs prevent chest pain. Clearly state the null and alternative hypotheses and provide a brief interpretation of your results.

(b) Would an exact test have been preferable? Provide a brief explanation supporting your answer.

Table 1

Patient Identification Number	Did the Patient feel Better on Drug A?
1	No
2	Yes
3	Yes
4	Yes
5	Yes
6	Yes
7	Yes
8	Yes
9	Yes
10	Yes

[10 marks]

Soalan 1

Pesakit yang menderita sakit dada dipastikan untuk menerima ubat baru (Ubat A) dan terapi piawai (Ubat B) secara rawak. Pada akhir rawatan setiap pesakit ditanya sama ada mereka berasa lebih baik apabila menerima Ubat A berbanding Ubat B.

(a) Gunakan data dalam Jadual 1 untuk membina ujian anggaran sama ada terdapat sebarang perbezaan dalam keberkesanan kedua-dua ubat mencegah sakit dada. Jelaskan dengan jelas hipotesis nol dan alternatif serta sediakan tafsiran ringkas tentang penemuan anda.

(b) Adakah ujian tepat lebih baik? Berikan penjelasan ringkas yang menyokong jawapan anda.

Jadual 1

Nombor Pengenalan Pesakit	*Adakah Pesakit Lebih Baik Menggunakan Ubat A?*
1	*Tidak*
2	*Ya*
3	*Ya*
4	*Ya*
5	*Ya*
6	*Ya*
7	*Ya*
8	*Ya*
9	*Ya*
10	*Ya*

[10 markah]

Question 2

Show that the Pearson chi-square test, X^2 and likelihood-ratio test, G^2 for a Binomial proportion are given by

$$X^2 = \frac{(r-np)^2}{np(1-p)} \text{ and } G^2 = 2r\log\left(\frac{r}{np}\right) + 2(n-r)\log\left(\frac{n-r}{n(1-p)}\right),$$

where r is the numbers of particular category occurs in n trials and p is the probability of occurrence of that category.

[8 marks]

Soalan 2

Tunjukkan bahawa ujian khi-kuasa dua Pearson, X^2 dan ujian nisbah kebolehjadian, G^2 untuk perkadaran Binomial diberikan oleh

$$X^2 = \frac{(r-np)^2}{np(1-p)} \textit{ dan } G^2 = 2r\log\left(\frac{r}{np}\right) + 2(n-r)\log\left(\frac{n-r}{n(1-p)}\right),$$

yang mana r adalah bilangan suatu kategori yang berlaku pada n percubaan and p adalah kebarangkalian berlakunya kategori tersebut.

[8 markah]

Question 3

A study in the School of Fisheries and Aquaculture Sciences at the Universiti Malaysia Terengganu sampled wild common carp fish from a wetland in central Terengganu. An analysis to investigate whether the fish muscle had lead pollutant and whether there was evident malformation in the fish were conducted. The findings show that of 25 fish without lead, 7 had malformation and of 14 with lead, 7 had malformation. Suppose you were a reporter assigned to write a news report about the statistical results. Write a short news report that would be appropriate for your city's newspaper. Remember that most readers are not familiar with the statistical measures, give explanation when writing the article.

[12 marks]

Soalan 3

Satu kajian di Pusat Pengajian Perikanan dan Sains Akuakultur di Universiti Malaysia Terengganu mengambil sampel ikan kap liar dari tanah paya di Terengganu tengah. Satu analisis untuk menyiasat sama ada dalam badan ikan mempunyai pencemaran plumbum dan sama ada terdapat mutasi pada ikan telah dijalankan. Keputusan mendapati daripada 25 ekor ikan tanpa plumbum, 7 mempunyai mutasi dan daripada 14 dengan plumbum, 7 mempunyai mutasi. Katakan anda seorang wartawan yang ditugaskan menulis laporan mengenai hasil statistik kajian ini. Tuliskan laporan pendek yang sesuai untuk akhbar bandar anda. Ingat bahawa kebanyakan pembaca tidak biasa dengan pengiraan statistik, beri penjelasan ketika menulis artikel tersebut.

[12 markah]

Question 4

Table 2 shows the application to the Master of Science program in the School of Mathematical Sciences, USM for the 2015 session. Admission decisions are presented by gender of applicant for the two main programs. Denote the three variables by A=admission status (Yes/No), G=gender and P= program. Without using SAS answer the following questions.

Table 2

	Admission status, male		Admission status, female	
Program	Yes	No	Yes	No
Statistics	40	32	32	60
Mathematics	75	120	24	96

(a) Compute and interpret the conditional odds ratios and the marginal odds ratio.

(b) Show whether Simpson's Paradox occurs.

(c) Test the hypothesis that gender is independent of admission status, controlling for program.

[14 marks]

Soalan 4

Jadual 2 menunjukkan permohonan program Sarjana Sains di Pusat Pengajian Sains Matematik, USM bagi sesi 2015. Keputusan kemasukan disampaikan berdasarkan jantina pemohon untuk kedua-dua program utama. Tiga pembolehubah dinyatakan sebagai A = Status diterima (Ya/Tidak), G = jantina dan P = program. Tanpa menggunakan SAS, jawab soalan-soalan berikut.

Jadual 2

	Status diterima, lelaki		*Status diterima, wanita*	
Program	*Ya*	*Tidak*	*Ya*	*Tidak*
Statistik	*40*	*32*	*32*	*60*
Matematik	*75*	*120*	*24*	*96*

(a) Kira dan tafsir nisbah odds bersyarat dan nisbah odds marginal.

(b) Tunjukkan sama ada Paradoks Simpson berlaku.

(c) Uji hipotesis bahawa jantina adalah tak bersandar kepada status diterima, apabila mengawal program.

[14 markah]

Question 5

Based on the Table 2 from question 4, use SAS and fit model (AG,AP,PG) to the data.

(a) Write down the fitted model. Report the deviance and comment on the goodness of fit.

(b) What is the odds ratio between A and G? Find and explain the 95% confidence interval for the odds ratio.

(c) What is the predicted number of female applicants admitted to Mathematics program?

(d) What is the estimated probability of being admitted to Mathematics program for a female applicant?

[14 marks]

Soalan 5

Berdasarkan Jadual 2 dari soalan 4, gunakan SAS dan suaikan model (AG,AP,PG) kepada data.

(a) Tuliskan model yang disuaikan. Laporkan penyimpangan dan komen tentang kebaikan penyuaian.

(b) Apakah nisbah odds antara A dan G? Cari dan terangkan selang keyakinan 95% untuk nisbah odds.

(c) Berapakah jumlah pemohon wanita yang diramal masuk ke program Matematik?

(d) Berapakah kebarangkalian diterima masuk ke program Matematik untuk pemohon wanita?

[14 markah]

Question 6

Data in Table 3 are measures taken in Hokkaido, Japan. Use the data below to determine whether the number of ice free days has changed over time and whether this is related to average temperature. Do these data support the claim that there is climate change? Report all analyses and write a short paragraph describing your findings.

Table 3

Year	Temp (^{0}C)	IceFreeDays	Year	Temp (^{0}C)	IceFreeDays
1981	-3.9	140	1993	-4.2	173
1982	-4.7	144	1994	-6.3	148
1983	-4.4	116	1995	-4.9	159
1984	-7.0	134	1996	-5.6	157
1985	-5.9	120	1997	-4.5	183
1986	-5.7	154	1998	-3.8	162
1987	-5.2	152	1999	-6.2	133
1988	-5.1	142	2000	-5.0	156
1989	-4.8	159	2001	-5.3	137
1990	-6.4	149	2002	-3.3	203
1991	-5.2	154	2003	-3.8	179
1992	-7.3	137			

[12 marks]

Soalan 6

Data dalam Jadual 3 adalah sukatan di ambil di Hokkaido Jepun. Gunakan data tersebut untuk menentukan sama ada bilangan hari bebas ais telah berubah dari masa ke masa dan sama ada ini berkaitan dengan suhu purata. Adakah data ini menyokong tuntutan bahawa terdapat perubahan iklim. Laporkan semua analisis dan tulis suatu perenggan pendek yang menerangkan penemuan anda.

Jadual 3

Tahun	*Suhu (^0C)*	*HariBebasAis*	*Tahun*	*Suhu (^0C)*	*HariBebasAis*
1981	*-3.9*	*140*	*1993*	*-4.2*	*173*
1982	*-4.7*	*144*	*1994*	*-6.3*	*148*
1983	*-4.4*	*116*	*1995*	*-4.9*	*159*
1984	*-7.0*	*134*	*1996*	*-5.6*	*157*
1985	*-5.9*	*120*	*1997*	*-4.5*	*183*
1986	*-5.7*	*154*	*1998*	*-3.8*	*162*
1987	*-5.2*	*152*	*1999*	*-6.2*	*133*
1988	*-5.1*	*142*	*2000*	*-5.0*	*156*
1989	*-4.8*	*159*	*2001*	*-5.3*	*137*
1990	*-6.4*	*149*	*2002*	*-3.3*	*203*
1991	*-5.2*	*154*	*2003*	*-3.8*	*179*
1992	*-7.3*	*137*			

[12 markah]

Question 7

Table 4 shows the results of a study about Y = whether a patient having surgery with general anesthesia experienced a sore throat on waking (0 = no, 1 = yes) as a function of the D = duration of the surgery (in minutes) and the T = type of device used to secure the airway (0 = laryngeal mask airway, 1 = tracheal tube).

Table 4

Patient	D	T	Y	Patient	D	T	Y	Patient	D	T	Y
1	45	0	0	13	50	1	0	25	20	1	0
2	15	0	0	14	75	1	1	26	45	0	1
3	40	0	1	15	30	0	0	27	15	1	0
4	83	1	1	16	25	0	1	28	25	0	1
5	90	1	1	17	20	1	0	29	15	1	0
6	25	1	1	18	60	1	1	30	30	0	1
7	35	0	1	19	70	1	1	31	40	0	1
8	65	0	1	20	30	0	1	32	15	1	0
9	95	0	1	21	60	0	1	33	135	1	1
10	35	0	1	22	61	0	0	34	20	1	0
11	75	0	1	23	65	0	1	35	40	1	0
12	45	1	1	24	15	1	0				

(a) Fit a logistic model for these data. For your model, interpret parameter estimates, and conduct inference about the effects.

(b) Fit a probit model for these data. For your model, interpret parameter estimates, and conduct inference about the effects.

(c) How does the probit model fit to the data compare to the logistic model?

[16 marks]

Soalan 7

Jadual 4 menunjukkan hasil kajian mengenai Y = sama ada pesakit yang menjalani pembedahan dengan anesthesia mengalami sakit tekak apabila sedar (0 = tidak, 1 = ya) sebagai fungsi D = tempoh pembedahan (dalam minit) dan jenis T = alat yang digunakan untuk membuka saluran udara (0 = topeng laluan laring, 1 = tiub trakea).

Jadual 4

Pesakit	*D*	*T*	*Y*	*Pesakit*	*D*	*T*	*Y*	*Pesakit*	*D*	*T*	*Y*
1	*45*	*0*	*0*	*13*	*50*	*1*	*0*	*25*	*20*	*1*	*0*
2	*15*	*0*	*0*	*14*	*75*	*1*	*1*	*26*	*45*	*0*	*1*
3	*40*	*0*	*1*	*15*	*30*	*0*	*0*	*27*	*15*	*1*	*0*
4	*83*	*1*	*1*	*16*	*25*	*0*	*1*	*28*	*25*	*0*	*1*
5	*90*	*1*	*1*	*17*	*20*	*1*	*0*	*29*	*15*	*1*	*0*
6	*25*	*1*	*1*	*18*	*60*	*1*	*1*	*30*	*30*	*0*	*1*
7	*35*	*0*	*1*	*19*	*70*	*1*	*1*	*31*	*40*	*0*	*1*
8	*65*	*0*	*1*	*20*	*30*	*0*	*1*	*32*	*15*	*1*	*0*
9	*95*	*0*	*1*	*21*	*60*	*0*	*1*	*33*	*135*	*1*	*1*
10	*35*	*0*	*1*	*22*	*61*	*0*	*0*	*34*	*20*	*1*	*0*
11	*75*	*0*	*1*	*23*	*65*	*0*	*1*	*35*	*40*	*1*	*0*
12	*45*	*1*	*1*	*24*	*15*	*1*	*0*				

(a) Suaikan model logistik untuk data ini. Untuk model anda, tafsirkan anggaran parameter dan jalankan pentaabiran tentang kesannya.

(b) Suaikan model probit untuk data ini. Untuk model anda, tafsirkan anggaran parameter dan jalankan pentaabiran tentang kesannya.

(c) Bagaimanakah model probit menyuaikan data berbanding dengan model logistik?

[16 markah]

Question 8

Data in Table 5 are from an investigation of satisfaction level with housing conditions in Penang. Residents in selected areas living in rented homes built between 1960 and 1968 were questioned about their satisfaction and the degree of contact with other residents. The data were tabulated by type of housing.

Table 5

Satisfaction with Housing Conditions						
	Low		Medium		High	
Contact with other residents	Low	High	Low	High	Low	High
Apartment	65	34	54	47	100	100
Condominium	130	141	76	116	111	191
House	67	130	48	105	62	104

(a) Use logistic regression to model associations between level of satisfaction and the other two variables. Obtain a parsimonious model that summarizes the patterns in the data.

(b) From the best fitted model you obtained in (a), calculate the standardized residuals and use them to find where the largest discrepancies are between the observed frequencies and expected frequencies estimated from the model.

[14 marks]

Soalan 8

Data dalam Jadual 5 adalah dari penyiasatan tahap kepuasan dengan keadaan perumahan di Pulau Pinang. Penduduk di kawasan terpilih yang tinggal di rumah sewa yang dibina antara 1960 dan 1968 disoal mengenai kepuasan mereka dan tahap hubungan dengan penduduk lain. Data tersebut disusun mengikut jenis perumahan.

Jadual 5

Kepuasan Dengan Keadaan Perumahan						
	Rendah		*Sederhana*		*Tinggi*	
Hubungan dengan penduduk lain	*Rendah*	*Tinggi*	*Rendah*	*Tinggi*	*Rendah*	*Tinggi*
Pangsapuri	*65*	*34*	*54*	*47*	*100*	*100*
Kondominium	*130*	*141*	*76*	*116*	*111*	*191*
Rumah	*67*	*130*	*48*	*105*	*62*	*104*

(a) *Gunakan regresi logistik untuk pemodelan perhubungan antara tahap kepuasan dengan dua pembolehubah lain. Dapatkan model parsimoni yang meringkaskan corak dalam data.*

(b) *Daripada model terbaik yang disuaipadan dalam (a), hitung reja piawai dan gunakannya untuk mencari di mana perubahan terbesar antara frekuensi yang diperhatikan dan jangkaan frekuensi daripada model yang dianggarkan.*

[14 markah]

-ooo00ooo-