UNIVERSITI SAINS MALAYSIA

First Semester Examination
2013/2014 Academic Session

December 2013/January 2014

MST 567 – Categorical Data Analysis
[Analisis Data Berkategori]

Duration : 3 hours
[Masa : 3 jam]

Please check that this examination paper consists of **eight** pages of printed materials before you begin the examination.

*[Sila pastikan bahawa kertas peperiksaan ini mengandungi **lapan** 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 diguna pakai].

1. Discuss the scales of measurement and give examples in the discussion.

[10 marks]

1. Bincangkan skala pengukuran dan berikan contoh-contoh dalam perbincangan.

[10 markah]

2. More people are abandoning national brand products and buying store brand products to save money. The president of a company that produces national brand coffee claims that 40% of the people prefer to buy national brand coffee. A random sample of 700 people who buy coffee showed that 259 of them buy national brand coffee. Can you conclude that the percentage of people who buy national brand coffee is different from 40%?
   (a) Use the score test to compute the p-value of the hypothesis test
   (b) Are the results in (a) reliable?

[12 marks]

2. Lebih ramai orang meninggalkan produk jenama kebangsaan dan membeli produk jenama kedai untuk menjimatkan wang. Presiden sebuah syarikat yang menghasilkan kopi jenama kebangsaan mendakwa bahawa 40% daripada orang lebih suka untuk membeli kopi jenama kebangsaan. Satu sampel rawak 700 orang yang membeli kopi menunjukkan bahawa 259 daripada mereka membeli kopi jenama kebangsaan. Bolehkah anda membuat kesimpulan bahawa peratusan orang yang membeli kopi jenama kebangsaan adalah berbeza dari 40%?
   (a) Gunakan ujian skor untuk mengira nilai-p ujian hipotesis
   (b) Adakah keputusan di (a) boleh dipercayai?

[12 markah]

3. Results from a cross-sectional study examining the effect of smoking on lung function among employees of a MTI textile plant are provided in Table 1.

Table 1

Lung Function	**Never Smoker**	**Current Smoker**
Normal	577	682
Borderline	27	46
Moderately Impaired	7	11
Severely Impaired	0	0

   (a) The purpose of the study was to determine if there is an association between smoking status and lung function. The investigator chose to answer this question using a Pearson chi-square test with three degrees of freedom. Is this procedure appropriate? Provide a brief justification for your response.
   (b) Suggest two alternative methods which could be used to analyze these data. Give one advantage and one disadvantage for each approach.

[12 marks]

3. Hasil dari satu kajian keratan rentas memeriksa kesan merokok pada fungsi peparu dalam kalangan pekerja kilang tekstil MTI adalah seperti di Jadual 1.

Jadual 1

Fungsi Peparu	***Bukan Perokok***	***Perokok***
Normal	*577*	*682*
Pertengahan	*27*	*46*
Agak Terjejes	*7*	*11*
Teruk Terjejas	*0*	*0*

(a) Tujuan kajian ini adalah untuk menentukan sama ada terdapat kaitan antara status merokok dengan fungsi peparu. Penyiasat memilih untuk menjawab soalan ini menggunakan ujian khi-kuasa dua Pearson dengan tiga darjah kebebasan. Adakah prosedur ini sesuai? Berikan justifikasi ringkas untuk jawapan anda.

(b) Cadangkan dua kaedah alternatif yang boleh digunakan untuk menganalisis data ini. Berikan satu kelebihan dan satu kelemahan bagi setiap pendekatan.

[12 markah]

4. A case control study was carried out to determine whether welders were at increased risk of lung cancer. Fifty incident cases of lung cancer over the period 1999 to 2001 were recruited into the study. Contact was made within four months of the diagnosis. Fifty workers with no history of exposure to welding were randomly selected from a worker registry. Results are given in Table 1.

Table 1

Cancer	Welder	
	No	Yes
No	20	30
Yes	9	41

(a) Find the four measures of association. Comment on the results.

(b) Based on (a), construct a 99% confidence interval for each of the measures of association. Comment on the results.

[16 marks]

4. *Satu kajian kes kawalan telah dijalankan untuk menentukan sama ada pengimpal adalah berisiko tinggi mendapat kanser peparu. Lima puluh kes kejadian kanser peparu dalam tempoh 1999 hingga 2001 telah diambil ke dalam kajian ini. Pesakit dihubungi dalam tempoh empat bulan diagnosis. Lima puluh pekerja yang tidak mempunyai sejarah pendedahan kepada kimpalan telah dipilih secara rawak daripada daftar pekerja. Keputusan yang diberikan dalam Jadual 1.*

Jadual 1

Kanser	*Pengimpal*	
	Tidak	*Ya*
Tidak	*20*	*30*
Ya	*9*	*41*

(a) Cari empat pengukuran interaksi. Berikan komen terhadap keputusan.
(b) Berdasarkan (a), bina suatu selang keyakinan 99% bagi setiap pengukuran interaksi. Berikan Komen terhadap keputusan.

[16 markah]

5. The random variable Y has a distribution in the exponential family. Its probability density function (or probability mass function) can be written as

$$f_Y(y;\theta,\phi) = e^{\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\}}$$

for some specific function $a(\phi), b(\theta)$ and $c(y,\phi)$ where $a(\phi) > 0$ and $b(\theta)$ has up to twice derivatives. Let $l(\theta,\phi;y)$ denote the associated log-likelihood function. It is known that $E\left(\frac{\partial l}{\partial \theta}\right) = 0$ and $Var\left(\frac{\partial l}{\partial \theta}\right) = -E\left(\frac{\partial^2 l}{\partial \theta^2}\right)$. Show that $E(y) = b'(\theta)$ and $\text{Var}(y) = a(\phi)b''(\theta)$.

[10 marks]

5. *Pemboleh ubah rawak Y mempunyai taburan dalam keluarga exponen, fungsi ketumpatan kebarangkalian (atau fungsi jisim kebarangkalian) boleh ditulis sebagai*

$$f_Y(y;\theta,\varphi) = e^{\{(y\theta - b(\theta))/a(\varphi) + c(y,\varphi)\}}$$

untuk fungsi tertentu $a(\phi), b(\theta)$ and $c(y,\phi)$ *di mana* $a(\phi) > 0$ *dan* $b(\theta)$ *boleh dibeza dua kali. Biarkan* $l(\theta,\phi;y)$ *menunjukkan fungsi log-kebolehjadian yang berkaitan. Telah diketahui bahawa* $E\left(\frac{\partial l}{\partial \theta}\right) = 0$ *dan* $Var\left(\frac{\partial l}{\partial \theta}\right) = -E\left(\frac{\partial^2 l}{\partial \theta^2}\right)$. *Tunjukkan bahawa* $E(y) = b'(\theta)$ dan $\text{Var}(y) = a(\phi)b''(\theta)$.

[10 markah]

…5/-

6. Come up with an example of a study in which the outcome variable is likely to be Binomial distributed. Explain what the random component, systematic component, and link function would be in this case.

[8 marks]

6. Berikan contoh satu kajian di mana pembolehubah hasil adalah taburan Binomial. Jelaskan apakah komponen rawak, komponen sistematik, dan fungsi jaringan untuk kajian kes ini.

[8 markah]

7. Subjects participating in a survey in 2002 were classified as being alive/dead by the year 2012. We will limit attention to three potential mortality risk factors each of which was assessed in 2002: age (years); self-reported smoking history; and subject's gender. Age is modelled as a continuous variable while gender and smoking are modelled using dummy variables. Smokers were classified as being never smokers, current smokers or former smokers.

$$\text{Gender}=\begin{cases}1 \text{ for women}\\ 0 \text{ for men}\end{cases}$$

$$\text{Smoke1}=\begin{cases}1 \text{ for former smokers}\\ 0 \text{ otherwise}\end{cases}$$

$$\text{Smoke2}=\begin{cases}1 \text{ for current smokers}\\ 0 \text{ otherwise}\end{cases}$$

Results obtained by fitting logistic regression models are provided in Table 2 and Table 3.

(a) Can we conclude that there is a statistically significant interaction between subjects' gender and their smoking history? Provide a brief explanation for your answer.

(b) Is it a coincidence that the log-likelihood values provided in Table 2 get progressively larger as you move down the table from model 3 to model 6? Provide a brief explanation for your answer.

(c) Could these data have been modelled using Poisson regression? Provide a brief explanation for your answer.

[14 marks]

Table 2

Model	Covariates	Log-Likelihood value
1	Gender	-1701.77
2	Smoke1, Smoke2	-1700.76
3	Age	-1079.39
4	Gender, Age, Smoke1, Smoke2	-1051.24
5	Gender, Age, Smoke1, Smoke2, Gender×Smoke1, Gender×Smoke2	-1046.03
6	Gender, Age, Age^2, Smoke1, Smoke2, Gender×Smoke1, Gender×Smoke2	-1018.15

Table 3

Variables	Estimated Regression Coefficient	Estimated Standard Error
Intercept	-2.4959	0.7235
Gender	0.1163	0.2644
Age	-0.1049	0.0301
Age^2	0.0023	0.0003
Smoke1	0.7430	0.2875
Smoke2	1.3557	0.2536
Gender x Smoke1	-0.8148	0.3990
Gender x Smoke2	-0.7954	0.3045

7. *Subjek yang menyertai kaji selidik dalam tahun 2002 telah diklasifikasikan sebagai hidup / mati pada tahun 2012. Perhatian kepada dihadkan kepada tiga faktor yang berpotensi sebagai risiko kematian bagi setiap yang dinilai dalam tahun 2002: umur (tahun); sejarah merokok dilaporkan sendiri dan jantina subjek. Umur dimodelkan sebagai pembolehubah selanjar manakala jantina dan merokok adalah dimodelkan menggunakan pembolehubah pepatung. Perokok telah diklasifikasikan sebagai tidak perokok, perokok atau bekas perokok*

$$\text{Jantina}=\begin{cases}1 \text{ untuk wanita}\\ 0 \text{ untuk lelaki}\end{cases}$$

$$\text{Perokok1}=\begin{cases}1 \text{ untuk bekas perokok}\\ 0 \text{ selainnya}\end{cases}$$

$$\text{Perokok2}=\begin{cases}1 \text{ untuk perokok}\\ 0 \text{ selainnya}\end{cases}$$

Keputusan yang diperolehi oleh penyuaian model regresi logistik adalah seperti di Jadual 2 dan Jadual 3.

(a) Bolehkah kita membuat kesimpulan bahawa terdapat interaksi statistik yang signifikan antara jantina subjek dan sejarah merokok mereka? Berikan penjelasan yang ringkas bagi jawapan anda.

(b) Adakah ia satu kebetulan bahawa nilai-nilai log-kebolehjadian diberikan dalam Jadual 2 secara progresif lebih besar apabila bergerak daripada model 3 ke model 6? Berikan penjelasan yang ringkas bagi jawapan anda.

(c) Bolehkah data ini dimodel menggunakan model regresi Poisson? Berikan penjelasan yang ringkas bagi jawapan anda.

[14 markah]

Jadual 2

Model	*Kovariat*	*Nilai Log-Kebolehjadian*
1	*Jantina*	*-1701.77*
2	*Perokok1, Perokok2*	*-1700.76*
3	*Umur*	*-1079.39*
4	*Jantina, Umur, Perokok1, Perokok2*	*-1051.24*
5	*Jantina, Umur, Perokok1, Perokok2, Jantina×Perokok1, Jantina×Perokok2*	*-1046.03*
6	*Jantina, Umur, Umur^2, Perokok1, Perokok2, Jantina×Perokok1, Jantina×Perokok2*	*-1018.15*

Jadual 3

Pemboleh ubah	*Anggaran Pekali regresi*	*Anggaran Ralat Piawaian*
Intercept	*-2.4959*	*0.7235*
Jantina	*0.1163*	*0.2644*
Umur	*-0.1049*	*0.0301*
Umur^2	*0.0023*	*0.0003*
Perokok1	*0.7430*	*0.2875*
Perokok2	*1.3557*	*0.2536*
Jantina×Perokok1	*-0.8148*	*0.3990*
Jantina×Perokok2	*-0.7954*	*0.3045*

8. Table 4 depicts data that was obtained from the 2002 General National Survey on respondents' gender, whether separated from spouse/partner, ability to afford needed medical care, and condition of home. Fit and interpret the most parsimonious log-linear model to this data. Describe the process you went through to determine the best fitting and the most parsimonious model. What types of associations exist? Demonstrate how the fitted values obtained from the model depict these associations.

Table 4

Gender	Separated from spouse/partner	Unable to afford needed medical care	Home in poor condition	
			Yes	No
Male	Yes	Yes	1	16
		No	1	41
	No	Yes	15	40
		No	24	481
Female	Yes	Yes	7	12
		No	1	35
	No	Yes	10	58
		No	25	559

[18 marks]

8. Jadual 4 menggambarkan data yang telah diperolehi daripada Soalselidik Umum Kebangsaan 2002 terhadap jantina responden, sama ada yang berpisah daripada pasangan / berpasangan, tidak mampu membayar rawatan perubatan diperlukan dan keadaan rumah. Suai dan tafsirkan model log-linear paling pasimoni kepada data ini. Jelaskan proses yang anda lalui untuk menentukan model tersuai terbaik dan model yang paling pasimoni. Apakah jenis-jenis interaksi yang wujud? Tunjukkan bagaimana nilai-nilai tersuai yang diperolehi daripada model menggambarkan interaksi.

Jadual 4

Jantina	*Berpisah daripada pasangan / berpasangan*	*Tidak mampu membayar rawatan perubatan diperlukan*	*Rumah dalam keadaan miskin*	
			Ya	*Tidak*
Lelaki	*Ya*	*Ya*	*1*	*16*
		Tidak	*1*	*41*
	Tidak	*Ya*	*15*	*40*
		Tidak	*24*	*481*
Wanita	*Ya*	*Ya*	*7*	*12*
		Tidak	*1*	*35*
	Tidak	*Ya*	*10*	*58*
		Tidak	*25*	*559*

[18 markah]

-ooo000ooo-