UNIVERSITI SAINS MALAYSIA

First Semester Examination
2009/2010 Academic Session

November 2009

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 material 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 ten** [10] questions.

*[**Arahan:** Jawab **semua sepuluh** [10] 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. The rate of heart disease in 50-59 year old disease-free women is pproximately 2 per 1000 per year or 10 per 1000 over 5 years. Suppose that 3 heart diseases are reported over 5 years among 1000 women initially disease-free who have been taking postmenopausal hormones.
   (a) Use the binomial distribution to see if this experience represents unusual small number of events based on the overall rate.
   (b) Solve (a) using the Poisson approximation to the binomial distribution.
   (c) Compare answers from (a) and (b).

[6 marks]

2. Differences of proportions, relative risk and odds ratios are three ways in measuring the association in 2 x 2 tables;
   (a) Let π_1 and π_2 represent the probabilities of success, show the relationship between odds ratios and relative risk
   (b) If π_1 and π_2 are small, what will be the relationship between odds ratios and relative risk?

[6 marks]

3. Consider a multinomial distribution with probability parameters π_1, π_2 and π_3. Suppose we wish to test the hypothesis

$$H_0 : \frac{\pi_1}{\pi_2} = \frac{\pi_2}{\pi_3} = \frac{1}{2},$$

show that H_0 is equivalent to

$$\pi_1 = \frac{4}{7}, \quad \pi_2 = \frac{2}{7}, \quad \pi_3 = \frac{1}{7}.$$

[6 marks]

4. Among the 1,820 subjects in a study, 30 suffered from tuberculosis and 1,790 did not. Chest x-rays were administered to all individuals; 73 had a positive x-ray, indicating the significant presence of inflammatory disease, and 1,747 had a negative x-ray. The data for the study are displayed in table below. We note that the prevalence of the disease (tuberculosis) in the population is 2.25%.

	X-ray Result	
Disease Status	Positive	Negative
Present	22	8
Absent	51	1739

   (a) What is the sensitivity and specificity of this study?
   (b) Find the probability that a subject who is positive on the x-ray test has tuberculosis.
   (c) Find the probability that a subject who is negative on the x-ray test is tuberculosis free.

[12 marks]

1. Kadar penyakit jantung di kalangan wanita berumur 50-59 tahun yang tidak berpenyakit adalah kira-kira 2 daripada 1000 wanita setahun dan 10 daripada 1000 wanita untuk tempoh lima tahun. Jika 3 pesakit jantung dilaporkan selepas 5 tahun daripada 1000 wanita tidak berpenyakit yang mengambil hormon "postmenopausal";

   (a) Guna taburan binomial untuk melihat sama ada eksperimen ini menunjukkan bilangan peristiwa luarbiasa yang kecil berdasarkan kadar keseluruhan.

   (b) Selesaikan (a) menggunakan penghampiran Poisson ke taburan binomial.

   (c) Bandingkan jawapan daripada (a) dan (b).

[6 markah]

2. Tiga kaedah untuk mengukur pertalian bagi jadual 2 x 2 adalah perbezaan kadaran, risiko relatif dan nisbah peluang.

   (a) Biarkan π_1 dan π_2 mewakili kebarangkalian kejayaan, tunjukkan perhubungan antara nisbah peluang dan risiko relatif.

   (b) Jika π_1 and π_2 bernilai kecil, apa akan berlaku pada hubungan antara nisbah peluang dan risiko relatif?

[6 markah]

3. Pertimbangkan taburan multinomial dengan parameter kebarangkalian π_1, π_2 dan π_3. Jika kita ingin menguji hipotesis

$$H_0 : \frac{\pi_1}{\pi_2} = \frac{\pi_2}{\pi_3} = \frac{1}{2}$$

tunjukkan H_0 adalah setara dengan

$$\pi_1 = \frac{4}{7}, \quad \pi_2 = \frac{2}{7}, \quad \pi_3 = \frac{1}{7}.$$

[6 markah]

4. Di kalangan 1820 subjek dalam suatu kajian, 30 orang mengalami "tuberculosis" dan 1,790 orang tiada. Setiap subjek juga telah diambil sinaran-x dada dan 73 orang memperolehi sinaran-x positif yang menunjukkan kehadiran signifikan penyakit peparu serta 1,747 orang memperolehi sinaran-x negatif. Data untuk kajian ini diberikan dalam jadual di bawah. Juga diketahui penyebaran penyakit ini dalam populasi adalah 2.25%.

Status Penyakit	Keputusan X-ray	
	Positif	Negatif
Hadir	22	8
Tiada	51	1739

   (a) Apakah kepekaan dan ketepatan untuk kajian ini?

   (b) Dapatkan kebarangkalian subjek yang mendapat keputusan positif untuk ujian sinaran-x dan mengalami "tuberculosis".

   (c) Dapatkan kebarangkalian subjek yang mendapat keputusan negatif untuk ujian sinaran-x dan tidak mengalami "tuberculosis".

[12 markah]

5. The data in the table are taken from the National Survey of Children. The event of interest here is whether a teenager (15 and 16 years old) reported ever having had sexual intercourse (yes/no) by the time of survey.

Race	Sex	Intercourse	
		Yes	No
White	Male	43	134
	Female	26	149
Black	Male	29	23
	Female	22	36

The following are estimated parameters of logit and probit models for the above data

Variable	Logit Model Estimated	Probit Model Estimated
	$\hat{\beta}$	$\hat{\beta}$
White	1.314	0.789
Female	0.648	0.377
Constant	0.192	0.106

(a) Using the estimated parameters from logit and probit models, find the following probabilities for each model

(i) The probability of having had intercourse for white males

(ii) The probability of having had intercourse for black females

(b) Discuss the answers you get from (a).

[12 marks]

6. The negative binomial distribution is often used to model the number of trials until the *r*th success. The probability function for the negative binomial distribution is

$$p(y) = \binom{y-1}{r-1} p^r (1-r)^{y-r}$$

(a) Show that this distribution belongs to the exponential family.
(b) Derive the mean and the variance for this distribution.
(c) Derive the canonical link for this distribution.

[12 marks]

7. Discuss about Poisson regression models based on the following conditions.
(a) Type of response variable.
(b) Three examples of response variable.
(c) The equation of Poisson regression models.
(d) The overdispersion of Poisson distribution.

[10 marks]

5. *Data dalam jadual didapati daripada Soal selidik Kebangsaan untuk Kanak-kanak. Peristiwa yang mendapat perhatian adalah laporan sama ada remaja (15 dan 16 tahun) pernah mengadakan hubungan kelamin (Ya/Tidak) semasa soal selidik dijalankan.*

		Hubungan Kelamin	
Keturunan	*Jantina*	*Ya*	*Tidak*
Kulit Putih	*Lelaki*	*43*	*134*
	Wanita	*26*	*149*
Kulit Hitam	*Lelaki*	*29*	*23*
	Wanita	*22*	*36*

Anggaran parameter model logit dan probit bagi data di atas adalah seperti berikut;

Pemboleh Ubah	*Anggaran Model Logit*	*Anggaran Model Probit*
	$\hat{\beta}$	$\hat{\beta}$
Kulit Putih	*1.314*	*0.789*
Wanita	*0.648*	*0.377*
Pemalar	*0.192*	*0.106*

(a) *Dapatkan kebarangkalian berikut untuk model probit dan logit;*

(i) *Kebarangkalian lelaki kulit putih telah mengadakan hubungan kelamin.*

(ii) *Kebarangkalian wanita kulit hitam telah mengadakan hubungan kelamin.*

(b) *Bincangkan jawapan yang didapati di (a).*

[12 markah]

6. *Taburan binomial negatif selalu digunakan untuk pemodelan bilangan ujikaji untuk mencapai kejayaan ke r. Fungsi kebarangkalian bagi taburan binomial negatif adalah*

$$p(y)=\binom{y-1}{r-1}p^r(1-r)^{y-r}$$

(a) *Tunjukkan taburan ini daripada keluarga taburan eksponen.*

(b) *Terbitkan min dan varians untuk taburan ini.*

(c) *Terbitkan jaringan "canonical" untuk taburan ini.*

[12 markah]

7. *Bincangkan model regresi Poisson berdasarkan syarat-syarat berikut.*

(a) *Jenis pemboleh ubah sambutan.*

(b) *Tiga contoh pemboleh ubah sambutan.*

(c) *Persamaan model regresi Poisson.*

(d) *Lebihan serakan (overdispersion) taburan Poisson.*

[10 markah]

8. A sample of subjects were asked their opinion about current laws legalizing abortion (support, oppose). For the explanatory variables gender (female, male), religious affiliation (Protestant, Catholic, Jewish), and political party affiliation (Democrat, Republican, Independent), the model for the probability π of supporting legalized abortion

$$\text{logit}\left(\pi\right)=\alpha+\beta_h G+\beta_i \text{R}+\beta_j P$$

has reported parameter estimates (setting the parameter for the last category of a variable equal to 0)

$\hat{\alpha}=-0.01, \hat{\beta}_1 G=0.16, \hat{\beta}_2 G=0, \hat{\beta}_1 R=-0.57, \hat{\beta}_2 R=-0.66, \hat{\beta}_3 R=0$,

$\hat{\beta}_1 P=0.84, \hat{\beta}_2 P=-1.67, \hat{\beta}_3 P=0$.

(a) Interpret how the odds of supporting legalized abortion depend on gender and party affiliation?

(b) If we define parameters such that the first category of a variable has value 0, then what would $\beta_2 G$ equal?

[12 marks]

9. A study is conducted on adult male cancer patients to determine whether there is any association between the kinds of work they perform and the kinds of cancer they have. The data are classified by the two categories as below:

Occupation	Kinds of Cancer		
	Skin	Stomach	Prostate
Professional	25	58	37
Managerial	34	90	36
Laborer	41	52	27

(a) Test for independence between the kinds of work they perform and the kinds of cancer they have

(b) Show the portioning of chi square into several 2 x 2 tables and comment on the results

[14 marks]

10. Consider the following table describing health opinion by gender and information opinion. Based on given output in Appendix, discuss log-linear model that fits these data well.

Gender	Information Opinion	Health Opinion	
		Support	Oppose
Male	Support	76	160
	Oppose	6	25
Female	Support	114	181
	Oppose	11	48

[10 marks]

8. *Suatu sampel subjek telah ditanya pendapat mereka tentang undang-undang membenarkan pengguguran (menyokong/menentang). Bagi pemboleh ubah penerang pula adalah jantina (wanita, lelaki), agama ikutan (Protestan, Katholik, Yahudi), dan parti politik diwakili (Demokrat, Republikan, Bebas), model yang mewakili kebarangkalian* π *menyokong membenarkan pengguguran adalah*

$$\text{logit}\left(\pi\right) = \alpha + \beta_h G + \beta_i \text{R} + \beta_j P$$

dengan anggaran parameter seperti berikut (menetapkan nilai parameter bersamaan 0 untuk kategori terakhir bagi setiap pemboleh ubah) $\hat{\alpha} = -0.01, \hat{\beta}_1 G = 0.16, \hat{\beta}_2 G = 0, \hat{\beta}_1 R = -0.57, \hat{\beta}_2 R = -0.66, \hat{\beta}_3 R = 0$, $\hat{\beta}_1 P = 0.84$, $\hat{\beta}_2 P = -1.67$, $\hat{\beta}_3 P = 0$.

(a) *Jelaskan bagaimana peluang menyokong membenarkan pengguguran bersandar kepada jantina dan pewakilan parti?*

(b) *Jika kita mentakrif kategori pertama suatu pemboleh ubah bernilai 0, maka apakah nilai* $\beta_2 G$ *?*

[12 markah]

9. Suatu kajian telah dijalankan ke atas pesakit kanser lelaki dewasa untuk menentukan sama ada wujud pertalian antara jenis pekerjaan dan jenis kanser yang dihadapi. Data diklasifikasikan kepada dua kategori seperti dalam jadual di bawah.

Pekerjaan	*Jenis kanser*		
	Kulit	*Perut*	*Prostat*
Profesional	*25*	*58*	*37*
Pengurusan	*34*	*90*	*36*
Buruh	*41*	*52*	*27*

(a) *Jalankan ujian bagi ketaksandaran antara jenis pekerjaan dan jenis kanser yang dihadapi.*

(b) *Tunjukkan pembahagian khi kuasadua kepada beberapa jadual 2 x 2 dan komen keputusannya.*

[14 markah]

10. Pertimbangkan jadual berikut yang menjelaskan tentang pendapat kesihatan mengikut jantina dan pendapat maklumat. Berdasarkan output yang diberikan dalam Lampiran, bincangkan penyuian model log-linear yang terbaik untuk data.

Jantina	*Pendapat Maklumat*	*Pendapat Kesihatan*	
		Menyokong	*Menentang*
Lelaki	*Menyokong*	*76*	*160*
	Menentang	*6*	*25*
Wanita	*Menyokong*	*114*	*181*
	Menentang	*11*	*48*

[10 markah]

APPENDIX

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and High Order Effects[a]	1	7	445.823	.000	412.417	.000	0
	2	4	16.318	.003	15.701	.003	2
	3	1	.302	.583	.308	.579	2
K-way Effects[b]	1	3	429.505	.000	396.716	.000	0
	2	3	16.016	.001	15.393	.002	0
	3	1	.302	.583	.308	.579	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Step Summary

Step[b]			Effects	Chi-Square[a]	df	Sig.	Number of Iterations
0	Generating Class[c]		Gender*Informatin*Health	.000	0	.	
	Deleted Effect	1	Gender*Informatin*Health	.302	1	.583	2
1	Generating Class[c]		Gender*Informatin, Gender*Health, Informatin*Health	.302	1	.583	
	Deleted Effect	1	Gender*Informatin	3.825	1	.050	2
		2	Gender*Health	2.081	1	.149	2
		3	Informatin*Health	11.364	1	.001	2
2	Generating Class[c]		Gender*Informatin, Informatin*Health	2.383	2	.304	
	Deleted Effect	1	Gender*Informatin	3.198	1	.074	2
		2	Informatin*Health	10.737	1	.001	2
3	Generating Class[c]		Informatin*Health Gender	5.581	3	.134	
	Deleted Effect	1	Informatin*Health	10.737	1	.001	2
		2	Gender	12.229	1	.000	2
4	Generating Class[c]		Informatin*Health Gender	5.581	3	.134	

a. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted fro

b. At each step, the effect with the largest significance level for the Likelihood Ratio Cha the significance level is larger than .050.

c. Statistics are displayed for the best model at each step after step 0.

- ooo O ooo -