UNIVERSITI SAINS MALAYSIA

First Semester Examination
2010/2011 Academic Session

November 2010

MST 567 – Categorical Data Analysis
[Analisis Data Berkategori]

Duration : 3 hours
[Masa : 3 jam]

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

[Sila pastikan bahawa kertas peperiksaan ini mengandungi SEPULUH 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. (a) Given the multinomial probability model $p(n_1, n_2, ..., n_{k-1}) = \frac{n!\pi_1^{n_1}\pi_2^{n_2}...\pi_k^{n_k}}{n_1!n_2!...n_k!}$ for the one-way classification having k categories, find the maximum likelihood estimation of π_j, its variance and covariance.

   (b) According to a genetics theory, an experiment of a certain cross will result in red, black, and white offspring in the ratio 8:4:4. Find the probability that among 8 offsprings 5 will be red, 2 black, and 1 white.

   [10 marks]

2. In a trial comparing a new drug to a standard, π denotes the probability that the new one is judged better. It is desired to estimate π and test $H_0 : \pi = 0.5$ against $H_a : \pi \neq 0.5$. In 100 independent observations, the new drug is found better in 60 observations.

   (a) Conduct a score test using *P*-value method and construct a 95% score confidence interval for π .

   (b) The researchers wanted a sufficiently large sample to be able to estimate the probability of preferring the new drug to be within 0.06, with confidence 0.98. If the actual probability is 0.75, how large a sample is needed to achieve this accuracy?

   [12 marks]

3. A 97% confidence interval for a population odds ratio is given as (1.25, 5.21) with sample variance of log odds ratio as 0.1090.

   (a) Find the sample odds ratio.

   (b) If the cell frequencies $n_{11} = 23$, $n_{12} = 34$ and $n_{21} = 35$, find n_{22} .

   (c) Based on (b) construct a 99% confidence interval for the population relative risk and a 94% confidence interval for the differences of population proportions.

   [12 marks]

1. (a) *Diberikan model kebarangkalian multinomial* $p(n_1, n_2, ..., n_{k-1}) = \frac{n!\pi_1^{n_1}\pi_2^{n_2}...\pi_k^{n_k}}{n_1!n_2!...n_k!}$ *untuk klasifikasi sehala dengan kategori k, cari anggaran kebolehjadian maksimum untuk* π_j, *varians dan kovarians.*

(b) *Menurut teori genetik, eksperimen persilangan tertentu akan menghasilkan keturunan merah, hitam, dan putih dengan nisbah 8:4:4. Cari kebarangkalian bahawa dalam kalangan 8 keturunan ada 5 akan menjadi merah, 2 hitam, dan 1 putih.*

[10 markah]

2. *Dalam ujikaji membandingkan ubat baru dengan piawaian,* π *mewakili kebarangkalian bahawa ubat baru dikatakan lebih baik. Hal ini dikehendaki untuk menganggar* π *dan uji* $H_0 : \pi = 0.5$ *terhadap* $H_a : \pi \neq 0.5$ *. Dalam 100 pemerhatian tak bersandar, ubat baru ini didapati lebih baik dalam 60 cerapan.*

(a) *Lakukan ujian skor dengan menggunakan kaedah nilai-P dan binakan selang keyakinan 95% untuk nilai* π *.*

(b) *Para penyelidik menghendaki satu sampel yang cukup besar yang mampu untuk menganggar kebarangkalian lebih menyukai ubat baru dalam lingkungan 0.06, dengan keyakinan 0.98. Jika kebarangkalian sebenar adalah 0.75, berapa besarkah sampel yang diperlukan untuk mencapai kejituan ini?*

[12 markah]

3. *Suatu selang keyakinan 97% untuk nisbah peluang populasi diberikan sebagai (1.25, 5.21) dengan varians sampel nisbah log peluang adalah 0.1090.*

(a) *Cari nisbah peluang sampel.*

(b) *Jika frekuensi sel* $n_{11} = 23$, $n_{12} = 34$ *and* $n_{21} = 35$, *cari* n_{22} *.*

(c) *Berdasarkan (b) bina suatu selang keyakinan 99% untuk risiko relatif populasi dan suatu selang keyakinan 94% untuk perbezaan kadaran populasi.*

[12 markah]

4. The table below shows a cross-tabulation between level of education and attitude toward premarital sex. The data are drawn from the US 1987-1991 pooled General Social Survey.

	Premarital Sex is			
Education	Always Wrong	Almost Always Wrong	Sometimes Wrong	Not Wrong At All
Less than high school	332	99	141	311
High school	313	129	258	480
Some college	199	87	218	423
College and above	176	71	208	359

(a) Obtain adjusted residuals and interpret.

(b) Suppose the two variables are ordinal, discuss the measure of association between the two variables

[15 marks]

5. If the binomial distribution for the proportion Y of success in n independent binary trials with the probability of success μ has the probability function

$$p(y) = \binom{n}{ny} \mu^{ny} (1-\mu)^{n(1-y)}$$

(a) Show that this probability function belongs to the exponential family of distributions.

(b) Based on (a), find the mean and the variance of the distribution.

[10 marks]

6. (a) Describe the purpose of the link function of a GLM.

(b) What is the identity link? Explain why it is not often used with the Binomial and Poisson probability models?

(c) Which statistic, -2log(LR) or Deviance, is better for choosing between alternative nested generalized linear models.

(d) Is Deviance an appropriate measure of goodness of fit for continuous predictors, that is, for ungrouped data? Explain.

[12 marks]

4. *Jadual di bawah menunjukkan tabulasi silang antara tahap pendidikan dan sikap terhadap seks sebelum kahwin. Data diambil dari Tinjauan Sosial Umum di Amerika Syarikat dari 1987-1991.*

	Seks Sebelum Kahwin			
Pendidikan	*Selalu salah*	*Hampir selalu salah*	*Kadang-kadang salah*	*Tidak salah langsung*
Kurang daripada sekolah tinggi	*332*	*99*	*141*	*311*
Sekolah Tinggi	*313*	*129*	*258*	*480*
Kolej sedikit	*199*	*87*	*218*	*423*
Kolej dan atas	*176*	*71*	*208*	*359*

(a) Dapatkan reja disesuaikan dan tafsirkannya.

(b) Jika kedua-dua pemboleh ubah adalah ordinal, bincangkan ukuran pertalian antara kedua-dua pemboleh ubah tersebut.

[15 markah]

5. *Jika taburan binomial untuk kadaran Y kejayaan dalam n percubaan binari dengan kebarangkalian kejayaan μ mempunyai fungsi kebarangkalian*

$$p(y) = \binom{n}{ny} \mu^{ny} (1-\mu)^{n(1-y)}$$

(a) Tunjukkan fungsi kebarangkalian ini adalah dari keluarga taburan eksponen.

(b) Berdasarkan (a), cari min dan varians taburan.

[10 markah]

6. *(a) Jelaskan tujuan fungsi jaringan untuk GLM.*

(b) Apakah jaringan identiti? Jelaskan mengapa hal ini tidak sering digunakan dengan model kebarangkalian Binomial dan Poisson?

(c) Statistik yang manakah,-2log (LR) atau Penyimpangan, adalah lebih baik untuk memilih antara model linear teritlak alternatif tersarang.

(d) Adakah Penyimpangan suatu ukuran kebagusan penyuaian model yang sesuai untuk peramal selanjar iaitu bagi data tak berkumpulan? Jelaskan.

[12 marks]

7. In a study to evaluate effectiveness of the drug Timolol (Samuels & Witmer, 1999) in preventing angina attacks, patients were randomly allocated to receive a daily dosage of either Timolol or placebo for 28 weeks. The numbers of patients who became free of angina attacks are displayed in table below

		Response	
		Free	Not Free
Treatment	Timolol	44	116
	Placebo	19	128

(a) How many model parameters are there for the above 2 x 2 table log-linear model? Explain each of the parameters.

(b) How does PROC CATMOD in SAS overcome the overparameterization in 2 x 2 table log-linear model?

(c) Write down the SAS code using PROC CATMOD that will produce the estimated parameters for the above 2 x 2 table log-linear model.

[14 marks]

8. Records of 53 prostate cancer patients described in Brown (1980) are analyzed using logistic model. Prostate cancer is more serious disease if it spreads to the lymph nodes. The objective of this analysis is to predict the nodal involvement of the cancer from a small number of covariate values (X-ray, stage, grade, age, acid level).

Responses Variable	**Interpretations**
Nodal Involvement (Binary Response)	• 1 represents nodal involvement found • 0 represents nodal involvement not found

Explanatory Variables	**Interpretations**
age (Patients' age)	• Continuous variable: 45 – 68 years old
xray (X-ray reading)	• 0 represents less serious finding • 1 represents more serious finding
stage (Size and location of tumor)	• 0 represents less serious finding • 1 represents more serious finding
grade (Pathology reading)	• 0 represents less serious finding • 1 represents more serious finding
acidlevel (Level of serum acid phosphates)	• Continuous variable: 40 – 187

It is desired to find the risk factors for nodal involvement so that unnecessary surgery can be avoided. Results of fitting a logistic model using SAS are given below.

...7/-

7. *Dalam kajian untuk menilai keberkesanan ubat Timolol (Samuels & Witmer, 1999) dalam mencegah serangan angina, pesakit secara rawak diperuntukkan untuk menerima dos harian Timolol atau plasebo selama 28 minggu. Jumlah pesakit yang menjadi bebas dari serangan angina dipaparkan dalam jadual di bawah ini*

		Kesan	
		Bebas	*Tak Bebas*
Rawatan	*Timolol*	*44*	*116*
	Plasebo	*19*	*128*

(a) Berapakah bilangan parameter dalam model log-linear bagi jadual 2 x 2 di atas? Jelaskan setiap parameter.

(b) Bagaimanakah PROC CATMOD dalam SAS mengatasi lebihan parameter pada model log-linear bagi jadual 2 x 2?

(c) Tuliskan kod SAS dengan menggunakan PROC CATMOD yang akan menghasilkan anggaran parameter model log-linear bagi jadual 2 x 2 di atas.

[14 markah]

8. *Rekod dari 53 pesakit kanser prostat yang diperihalkan di dalam Brown (1980) dianalisis dengan menggunakan model logistik. Kanser prostat adalah penyakit yang lebih serius jika menyebar ke kelenjar getah bening. Tujuan dari analisis ini adalah untuk meramal penglibatan nodal kanser dari sejumlah kecil nilai kovariat (X-ray, Peringkat, Kelas, Umur, Tahap Asid).*

Pemboleh ubah Sambutan	***Terjemahan***
Penglibatan nodal (Sambutan Biner)	• *1 merupakan penglibatan nodal dijumpai* • *0 merupakan penglibatan nodal tidak dijumpai*

Pemboleh ubah penerang	***Terjemahan***
age (umur pesakit)	• *Pemboleh ubah selanjar: 45 – 68 tahun*
xray (Bacaan X-ray)	• *0 merupakan penemuan kurang serius* • *1 merupakan penemuan lebih serius*
stage (Saiz and lokasi tumor)	• *0 merupakan penemuan kurang serius* • *1 merupakan penemuan lebih serius*
grade (bacaan patologi)	• *0 merupakan penemuan kurang serius* • *1 merupakan penemuan lebih serius*
acidlevel (Tahap serum asid fosfat)	• *Pemboleh ubah selanjar: 40 – 187*

Adalah dikehendaki untuk mencari faktor risiko untuk penglibatan nodal supaya pembedahan boleh dielakkan. Keputusan penyuaian model logistik menggunakan SAS diberikan di bawah ini.

...8/-

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	72.252	60.126
SC	74.222	71.948
-2 Log L	70.252	48.126

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0618	3.4599	0.0003	0.9857
xray	1	2.0453	0.8072	6.4208	0.0113
stage	1	1.5641	0.7740	4.0835	0.0433
grade	1	0.7614	0.7708	0.9759	0.3232
age	1	-0.0693	0.0579	1.4320	0.2314
acidlevel	1	0.0243	0.0132	3.4230	0.0643

(a) Interpret the parameter estimates.

(b) Construct and interpret the 98% confidence interval for the odds ratios of xray and stage.

(c) Conduct a Likelihood Ratio Test.

(d) Based on the output what is a possible best fitted model?

[15 marks]

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	72.252	60.126
SC	74.222	71.948
-2 Log L	70.252	48.126

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0618	3.4599	0.0003	0.9857
xray	1	2.0453	0.8072	6.4208	0.0113
stage	1	1.5641	0.7740	4.0835	0.0433
grade	1	0.7614	0.7708	0.9759	0.3232
age	1	-0.0693	0.0579	1.4320	0.2314
acidlevel	1	0.0243	0.0132	3.4230	0.0643

(a) Tafsirkan anggaran-anggaran parameter.

(b) Bina dan tafsirkan selang keyakinan 98% untuk nisbah peluang xray dan stage.

(c) Jalankan ujian nisbah kebolehjadian.

(d) Berdasarkan output apakah model tersuai terbaik?

[15 markah]

APPENDIX

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \qquad \theta = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad \tilde{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)} \qquad se(\ln\hat{\theta}) = \left(\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right)^{1/2}$$

$$U = -\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij}\log\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right)}{\sum_{j=1}^{J}\pi_{+j}\log\pi_{+j}} \qquad \hat{\gamma} = \frac{C-D}{C+D} \qquad X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \qquad se(\hat{\pi}_1-\hat{\pi}_2) = \left[\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}\right]^{1/2}$$

$$e_{ij} = \frac{n_{ij}-\hat{\mu}_{ij}}{[\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})]^{1/2}} \qquad e_{ij} = \frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}} \qquad G^2 = -2\sum\sum n_{ij}\ln\frac{n_{ij}}{\hat{\mu}_{ij}} \qquad se(\ln r) = \left[\frac{1-\pi_1}{n_1\pi_1}+\frac{1-\pi_2}{n_2\pi_2}\right]^{1/2}$$

$$f_{Y(y;\theta,\phi)} == e^{\left\{\left(\frac{y\theta-b(\theta)}{a(\phi)}\right)+c(y,\phi)\right\}}$$

- ooo O ooo -