UNIVERSITI SAINS MALAYSIA

First Semester Examination
2011/2012 Academic Session

January 2012

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 material 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. The four main distributions for categorical data are the hypergeometric distribution, binomial distribution, multinomial distribution and Poisson distribution.

   (a) Summarize in table form the process modeled and parameters for each of the distributions.
   (b) A company is interested in evaluating its current inspection procedure on shipments of 50 identical items. The procedure is to take a sample of 5 and pass the shipment if no more than 2 are found to be defective. What proportion of 20% defective shipments will be accepted?

   [13 marks]

1. Empat taburan utama bagi data kategori adalah taburan hipergeometri, taburan binomial, taburan multinomial dan taburan Poisson.

   (a) Ringkaskan dalam bentuk jadual proses yang dimodelkan dan parameter bagi setiap taburan.
   (b) Sebuah syarikat berminat untuk menilai prosedur pemeriksaan ke atas penghantaran 50 item yang serupa. Prosedur ini adalah untuk mengambil sampel 5 dan lulus penghantaran jika tidak lebih daripada 2 yang didapati rosak. Apakah bahagian penghantaran 20% rosak akan diterima?

   [13 markah]

2. Utusan Malaysia reported that about 47% of the general consumer population in Malaysia is loyal to the automobile manufacturer of their choice. Suppose that Proton did a study of a random sample of 1006 Proton owners and found that 490 said they would buy another Proton.

   (a) Construct a 99% Wald confidence interval for Proton owners who said they buy another Proton.
   (b) Does the confidence interval indicate that the population proportion of consumers loyal to Proton is more than 47%?

   [9 marks]

2. Utusan Malaysia melaporkan bahawa kira-kira 47% daripada populasi pengguna umum di Malaysia adalah setia kepada pengeluar automobil pilihan mereka. Katakan Proton melakukan kajian terhadap sampel rawak 1006 pemilik Proton dan mendapati bahawa 490 berkata, mereka akan membeli Proton lain.

   (a) Bina selang keyakinan Wald 99% untuk pemilik Proton yang berkata mereka membeli Proton lain.
   (b) Adakah selang keyakinan menunjukkan bahawa kadaran pengguna yang setia kepada Proton adalah lebih daripada 47%?

   [9 markah]

3. Table 1 shows data of cross classifying a person's perceived happiness with their family income and Figure 1 is the SAS output.

Table 1

Income	Happiness		
	Not too happy	Pretty Happy	Very Happy
Above average	21	159	110
Average	53	372	221
Below average	94	249	83

```
                    The FREQ Procedure

                Table of income by happiness

income          happiness

Frequency      ,
Expected       ,
Cell Chi-Square,Not_too_,Pretty_H,Very_hap,  Total
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Above_av       ,     21 ,    159 ,    110 ,    290
               , 35.771 , 166.08 ,  88.15 ,
               , 6.0994 , 0.3018 , 5.4161 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Average        ,     53 ,    372 ,    221 ,    646
               , 79.683 , 369.96 , 196.36 ,
               , 8.9351 , 0.0113 , 3.0916 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Below_av       ,     94 ,    249 ,     83 ,    426
               , 52.546 , 243.96 , 129.49 ,
               , 32.703 , 0.1039 ,  16.69 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total             168      780      414     1362
```

Figure 1

(a) Write a complete SAS program to produce the output in Figure 1.
(b) Find standardized residuals for each cell and discuss the results.

[13 marks]

3. *Jadual 1 menunjukkan data yang mengklasifikasikan tanggapan kebahagiaan seseorang dengan pendapatan keluarga mereka dan Rajah 1 adalah output SAS.*

Jadual 1

Pendapatan	*Kegembiraan*		
	Kurang gembira	*Agak gembira*	*Sangat gembira*
Atas sederhana	*21*	*159*	*110*
Sederhana	*53*	*372*	*221*
Bawah sederhana	*94*	*249*	*83*

```
                The FREQ Procedure

          Table of income by happiness

income          happiness

Frequency      ,
Expected       ,
Cell Chi-Square,Not_too_,Pretty_H,Very_hap,  Total
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Above_av       ,     21 ,    159 ,    110 ,    290
               , 35.771 , 166.08 ,  88.15 ,
               , 6.0994 , 0.3018 , 5.4161 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Average        ,     53 ,    372 ,    221 ,    646
               , 79.683 , 369.96 , 196.36 ,
               , 8.9351 , 0.0113 , 3.0916 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Below_av       ,     94 ,    249 ,     83 ,    426
               , 52.546 , 243.96 , 129.49 ,
               , 32.703 , 0.1039 ,  16.69 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total              168      780      414     1362
```

Rajah 1

(a) Tulis satu program SAS lengkap untuk menghasilkan output dalam Rajah 1.
(b) Cari reja terpiawai untuk setiap sel dan bincangkan keputusan.

[13 markah]

4. Refer to the previous question on the data cross classifying a person's perceived happiness with their family income.

Table 1

Income	Happiness		
	Not too happy	Pretty Happy	Very Happy
Above average	21	159	110
Average	53	372	221
Below average	94	249	83

(a) Calculate the likelihood-ratio statistic and test for independence between a person's perceived happiness with their family income.
(b) Partition Table 1 and discuss the results.

[15 marks]

4. Rujuk kepada soalan sebelumnya tentang data mengklasifikasikan tanggapan kebahagiaan seseorang dengan pendapatan keluarga mereka.

Jadual 1

Pendapatan	*Kegembiraan*		
	Kurang gembira	*Agak gembira*	*Sangat gembira*
Atas sederhana	*21*	*159*	*110*
Sederhana	*53*	*372*	*221*
Bawah sederhana	*94*	*249*	*83*

(a) Kirakan statistik nisbah kebolehjadian dan uji ketaksandaran antara tanggapan kebahagiaan seseorang dengan pendapatan keluarga mereka.
(b) Partisi Jadual 1 dan bincangkan keputusan.

[15 markah]

5. Table 2 is a three-way contingency table involving three variables gender, race and whether respondent uses a personal computer.

Table 2

Gender	Race	Computer Use	
		Yes	No
Male	Chinese	518	292
	Malay	55	64
Female	Chinese	563	403
	Malay	105	122

(a) Discuss whether there appears to be marginal independence or marginal dependence when controlling gender.
(b) Discuss whether there appears to be conditional independence or conditional dependence when controlling gender.
(c) Calculate Cochran-Mantel-Haenszel statistic and test for conditional independence when controlling gender.

[13 marks]

5. *Jadual 2 adalah jadual kontingensi tiga hala yang melibatkan tiga pembolehubah jantina, bangsa dan sama ada responden menggunakan komputer peribadi.*

Jadual 2

Jantina	*Bangsa*	*Guna Komputer*	
		Ya	*Tidak*
Lelaki	*Cina*	*518*	*292*
	Melayu	*55*	*64*
Perempuan	*Cina*	*563*	*403*
	Melayu	*105*	*122*

(a) Bincangkan sama ada terdapat ketaksandaran sut atau sandaran sut apabila mengawal jantina.
(b) Bincangkan sama ada terdapat ketaksandaran bersyarat atau sandaran bersyarat apabila mengawal jantina.
(c) Kirakan statistik Cochran-Mantel-Haenszel dan uji ketaksandaran bersyarat apabila mengawal jantina.

[13 markah]

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

[7 marks]

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

[7 markah]

7. Consider Table 3 which describes gender, education and attitude toward life from a General Social Survey. Based on the SAS output given in Figure 2, discuss a log-linear model that fits these data well.

Table 3

Gender	Education	Think life is exciting or dull?	
		Dull	Exciting
Males	No college degree	178	130
	College degree	38	70
Females	No college degree	245	182
	College degree	37	69

[13 marks]

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	7	358.124	.000	350.954	.000	0
	2	4	38.859	.000	39.005	.000	2
	3	1	.006	.936	.006	.936	2
K-way Effects[b]	1	3	319.265	.000	311.949	.000	0
	2	3	38.852	.000	38.999	.000	0
	3	1	.006	.936	.006	.936	0

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

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
Gender*Education	1	4.826	.028	2
Gender*Life	1	.008	.928	2
Education*Life	1	33.852	.000	2
Gender	1	14.461	.000	2
Education	1	302.475	.000	1
Life	1	2.329	.127	2

Step Summary

Step[a]			Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]		Gender*Education*Life	.000	0	.	
	Deleted Effect	1	Gender*Education*Life	.006	1	.936	2
1	Generating Class[b]		Gender*Education, Gender*Life, Education*Life	.006	1	.936	
	Deleted Effect	1	Gender*Education	4.826	1	.028	2
		2	Gender*Life	.008	1	.928	2
		3	Education*Life	33.852	1	.000	2
2	Generating Class[b]		Gender*Education, Education*Life	.014	2	.993	
	Deleted Effect	1	Gender*Education	4.909	1	.027	2
		2	Education*Life	33.935	1	.000	2
3	Generating Class[b]		Gender*Education, Education*Life	.014	2	.993	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than .050.
b. Statistics are displayed for the best model at each step after step 0.
c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Figure 2

7. *Pertimbangkan Jadual 3 yang menggambarkan jantina, pendidikan dan sikap terhadap hidup daripada Soal Selidik Umum. Berdasarkan output SAS yang diberikan dalam Rajah 2, bincangkan model log linear yang menyuaikan data ini dengan baik.*

Jadual 3

Jantina	*Pendidikan*	*Anggap kehidupan menarik atau bosan?*	
		Bosan	*Menarik*
Lelaki	*Tiada Ijazah*	*178*	*130*
	Berijazah	*38*	*70*
Perempuan	*Tiada Ijazah*	*245*	*182*
	Berijazah	*37*	*69*

[13 markah]

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	7	358.124	.000	350.954	.000	0
	2	4	38.859	.000	39.005	.000	2
	3	1	.006	.936	.006	.936	2
K-way Effects[b]	1	3	319.265	.000	311.949	.000	0
	2	3	38.852	.000	38.999	.000	0
	3	1	.006	.936	.006	.936	0

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

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
Gender*Education	1	4.826	.028	2
Gender*Life	1	.008	.928	2
Education*Life	1	33.852	.000	2
Gender	1	14.461	.000	2
Education	1	302.475	.000	1
Life	1	2.329	.127	2

Step Summary

Step[a]			Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]		Gender*Education*Life	.000	0	.	
	Deleted Effect	1	Gender*Education*Life	.006	1	.936	2
1	Generating Class[b]		Gender*Education, Gender*Life, Education*Life	.006	1	.936	
	Deleted Effect	1	Gender*Education	4.826	1	.028	2
		2	Gender*Life	.008	1	.928	2
		3	Education*Life	33.852	1	.000	2
2	Generating Class[b]		Gender*Education, Education*Life	.014	2	.993	
	Deleted Effect	1	Gender*Education	4.909	1	.027	2
		2	Education*Life	33.935	1	.000	2
3	Generating Class[b]		Gender*Education, Education*Life	.014	2	.993	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than .050.
b. Statistics are displayed for the best model at each step after step 0.
c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Rajah 2

8. Table 4 is from a study of factors influencing the primary food choice of alligators. The nominal response variable is the primary food type, in volume, found in an alligator's stomach. This has five categories: fish (1), invertebrate (2), reptile (3), bird (4) and other (5). Table 4 also classifies the alligators according to lake of capture (Trafford (1), George (2)), gender (male (1), female (2)) and size ($\leq$ 2.3 meters long (1), $>$ 2.3 meters long (2)). Figure 3 is the SPSS output showing the estimated model.

Table 4

Lake	Gender	Size	Food				
			Fish	Invertebrate	Reptile	Bird	Other
Trafford	Male	≤ 2.3	3	7	1	0	1
		> 2.3	8	6	6	3	5
	Female	≤ 2.3	2	4	1	1	4
		> 2.3	0	1	0	0	0
George	Male	≤ 2.3	13	10	0	2	2
		> 2.3	9	0	0	1	2
	Female	≤ 2.3	3	9	1	0	1
		> 2.3	8	1	0	0	1

Parameter Estimates

food		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
1	Intercept	1.504	.715	4.418	1	.036			
	[lake=1]	-1.740	.657	7.001	1	.008	.176	.048	.637
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	.799	.677	1.392	1	.238	2.223	.590	8.382
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	-.436	.639	.465	1	.495	.647	.185	2.263
	[size=2]	0	.	.	0	.	.	.	.
2	Intercept	.046	.808	.003	1	.955			
	[lake=1]	-.423	.639	.439	1	.507	.655	.187	2.290
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	.310	.662	.219	1	.640	1.363	.372	4.990
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	1.322	.675	3.836	1	.050	3.749	.999	14.072
	[size=2]	0	.	.	0	.	.	.	.
3	Intercept	-1.696	1.422	1.424	1	.233			
	[lake=1]	1.420	1.212	1.371	1	.242	4.136	.384	44.513
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	.262	1.093	.057	1	.811	1.299	.153	11.057
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	-.397	.976	.165	1	.684	.672	.099	4.554
	[size=2]	0	.	.	0	.	.	.	.
4	Intercept	-1.489	1.358	1.202	1	.273			
	[lake=1]	-.530	.992	.285	1	.593	.589	.084	4.113
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	1.381	1.243	1.234	1	.267	3.977	.348	45.452
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	-.132	.984	.018	1	.893	.877	.128	6.026
	[size=2]	0	.	.	0	.	.	.	.

Figure 3

(a) Write the estimated logistic models and find the predicted probability for each response categories.
(b) Is there a difference in the preferred food of large compared to smaller alligators?
(c) Does the lake in which the alligators live make a difference in which food is preferred?
(d) Do male compared to female alligators prefer different foods?

[17 marks]

8. *Jadual 4 menunjukkan kajian faktor-faktor yang mempengaruhi pilihan makanan utama aligator. Pembolehubah sambutan nominal adalah jenis makanan yang utama, dalam isipadu, yang ditemui dalam perut aligator. Ia mempunyai lima kategori iaitu ikan (1), invertebrata (2), reptilia (3), burung (4) dan lain-lain (5). Jadual 4 juga mengklasifikasikan aligator mengikut tasik penangkapan (Trafford (1), George (2)), jantina aligator (jantan (1), betina (2)) dan saiz (≤ 2.3 meter panjang (1), $>$ 2.3 meter panjang (2)). Rajah 3 adalah output SPSS menunjukkan anggaran model.*

Jadual 4

Tasik	*Jantina*	*Saiz*	*Makanan*				
			Ikan	*Invertebrata*	*Reptilia*	*Burung*	*Lain-lain*
Trafford	*Jantan*	≤ 2.3	*3*	*7*	*1*	*0*	*1*
		> 2.3	*8*	*6*	*6*	*3*	*5*
	Betina	≤ 2.3	*2*	*4*	*1*	*1*	*4*
		> 2.3	*0*	*1*	*0*	*0*	*0*
George	*Jantan*	≤ 2.3	*13*	*10*	*0*	*2*	*2*
		> 2.3	*9*	*0*	*0*	*1*	*2*
	Betina	≤ 2.3	*3*	*9*	*1*	*0*	*1*
		> 2.3	*8*	*1*	*0*	*0*	*1*

Parameter Estimates

food		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
1	Intercept	1.504	.715	4.418	1	.036			
	[lake=1]	-1.740	.657	7.001	1	.008	.176	.048	.637
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	.799	.677	1.392	1	.238	2.223	.590	8.382
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	-.436	.639	.465	1	.495	.647	.185	2.263
	[size=2]	0	.	.	0	.	.	.	.
2	Intercept	.046	.808	.003	1	.955			
	[lake=1]	-.423	.639	.439	1	.507	.655	.187	2.290
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	.310	.662	.219	1	.640	1.363	.372	4.990
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	1.322	.675	3.836	1	.050	3.749	.999	14.072
	[size=2]	0	.	.	0	.	.	.	.
3	Intercept	-1.696	1.422	1.424	1	.233			
	[lake=1]	1.420	1.212	1.371	1	.242	4.136	.384	44.513
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	.262	1.093	.057	1	.811	1.299	.153	11.057
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	-.397	.976	.165	1	.684	.672	.099	4.554
	[size=2]	0	.	.	0	.	.	.	.
4	Intercept	-1.489	1.358	1.202	1	.273			
	[lake=1]	-.530	.992	.285	1	.593	.589	.084	4.113
	[lake=2]	0	.	.	0	.	.	.	.
	[gender=1]	1.381	1.243	1.234	1	.267	3.977	.348	45.452
	[gender=2]	0	.	.	0	.	.	.	.
	[size=1]	-.132	.984	.018	1	.893	.877	.128	6.026
	[size=2]	0	.	.	0	.	.	.	.

Rajah 3

(a) Tuliskan anggaran model logistik dan cari kebarangkalian yang diramalkan bagi setiap kategori sambutan.
(b) Adakah terdapat perbezaan makanan pilihan antara aligator kecil dan besar?
(c) Adakah tasik tempat tinggal aligator membezakan pilihan makanan.
(d) Adakah aligator jantan berbanding betina memilih makanan berbeza?

[17 markah]

- ooo O ooo -