`

UNIVERSITI SAINS MALAYSIA

First Semester Examination
2012/2013 Academic Session

January 2013

MST 567 – Categorical Data Analysis
[Analisis Data Berkategori]

Duration : 3 hours
[Masa : 3 jam]

Please check that this examination paper consists of ELEVEN pages of printed material before you begin the examination.

[Sila pastikan bahawa kertas peperiksaan ini mengandungi SEBELAS muka surat yang bercetak sebelum anda memulakan peperiksaan ini].

Instructions: Answer **all nine** [9] questions.

*[**Arahan:*** *Jawab **semua sembilan** [9] 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. Determine whether procedures for analyzing categorical data are needed to address each of the following research questions. Provide a rationale for each of your answers by identifying the dependent and independent variables as well as their scales of measurement.
   (a) A researcher would like to determine whether males who have suffered a heart attack have higher fat content in their diets than males who have not suffered a heart attack in the past 6 months.
   (b) A researcher would like to predict whether a man is likely to suffer a heart attack in the next 6 months based on the fat content in his diet.

   [8 marks]

1. Pastikan sama ada kaedah analisis data berkategori diperlukan untuk setiap soalan penyelidikan berikut. Sertakan rasional bagi setiap jawapan anda dengan mengenal pasti pemboleh ubah bersandar dan tak bersandar dan juga skala pengukuran setiap pemboleh ubah.
   (a) Seorang penyelidik ingin menentukan sama ada lelaki yang telah mengalami serangan jantung mempunyai kandungan lemak yang lebih tinggi dalam diet daripada lelaki yang tidak mengalami serangan jantung bagi tempoh 6 bulan yang lalu.
   (b) Seorang penyelidik ingin meramalkan sama ada seorang lelaki adalah berkemungkinan untuk mengalami serangan jantung dalam enam bulan seterusnya berdasarkan kandungan lemak dalam dietnya.

   [8 markah]

2. A random sample of 202 college accounting faculty members was questioned. Of these sample members, 140 felt there was a need for more ethics coverage in accounting courses.
   (a) Find the *p*-value for testing the null hypothesis that 75% of all college accounting faculty members hold this view by using score test. Interpret the results.
   (b) Construct a 94% Wald confidence interval for all college accounting faculty members hold this view and discuss the results.

   [10 marks]

2. Satu sampel rawak 202 ahli fakulti perakaunan kolej telah disoal. 140 daripada ahli dalam sampel merasakan terdapat keperluan penekanan etika dalam kursus-kursus perakaunan.
   (a) Cari nilai p untuk menguji hipotesis nol bahawa 75% daripada semua ahli fakulti kolej perakaunan yang berpandangan sedemikian dengan menggunakan ujian skor. Tafsirkan keputusan yang didapati.
   (b) Bina selang keyakinan Wald 94% untuk semua ahli kolej fakulti perakaunan yang berpandangan sedemikian dan bincangkan keputusan.

   [10 markah]

3. Consider the following Table 1 representing State Assembly and Political Party. Figure 1 is the SAS output.

Table 1

Political Party	State Assembly			
	Penanti	Komtar	Bayan Lepas	Telok Bahang
PR	221	160	360	140
BN	200	291	160	311
Independent	208	106	316	97

Observation Statistics

Observation	count	party	dun	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual
1	**221**	PR	Penanti	5.377821	0.3662348	0.364728	0.5176959	0.5198346
2	**160**	PR	Komtar	-30.94047	-2.239123	-2.304086	-3.211406	-3.12086
3	**360**	PR	Bayan	73.417899	4.3368801	4.16906	6.2608297	6.5128514
4	**140**	PR	Telok	-47.85525	-3.491547	-3.658182	-5.087367	-4.855631
5	**200**	BN	Penanti	-35.44669	-2.310093	-2.372029	-3.450621	-3.360522
6	**291**	BN	Komtar	82.50428	5.7138378	5.3882285	7.6968712	8.1619912
7	**160**	BN	Bayan	-152.9307	-8.645113	-9.550021	-14.69839	-13.30565
8	**311**	BN	Telok	105.87315	7.392215	6.8635212	9.7824246	10.53596
9	**208**	IND	Penanti	30.068872	2.2541933	2.1947962	2.9823039	3.0630131
10	**106**	IND	Komtar	-51.56381	-4.107871	-4.369513	-5.830175	-5.48107
11	**316**	IND	Bayan	79.51284	5.1705112	4.9151241	7.0661101	7.4332612
12	**97**	IND	Telok	-58.0179	-4.659841	-5.008279	-6.667582	-6.203702

Figure 1

(a) Test for independent between State Assembly and Political Party. Comments on the results.

(b) Discuss the results on Haberman residuals.

[12 marks]

3. *Pertimbangkan Jadual 1 berikut yang menunjukkan data yang mengklasifikasikan Dewan Undangan Negeri dan Parti Politik. Rajah 1 adalah output SAS.*

Jadual 1

Parti Politik	*Dewan Undangan Negeri*			
	Penanti	*Komtar*	*Bayan Lepas*	*Telok Bahang*
PR	*221*	*160*	*360*	*140*
BN	*200*	*291*	*160*	*311*
BEBAS	*208*	*106*	*316*	*97*

Observation Statistics

Observation	count	party	dun	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual
1	**221**	PR	Penanti	5.377821	0.3662348	0.364728	0.5176959	0.5198346
2	**160**	PR	Komtar	-30.94047	-2.239123	-2.304086	-3.211406	-3.12086
3	**360**	PR	Bayan	73.417899	4.3368801	4.16906	6.2608297	6.5128514
4	**140**	PR	Telok	-47.85525	-3.491547	-3.658182	-5.087367	-4.855631
5	**200**	BN	Penanti	-35.44669	-2.310093	-2.372029	-3.450621	-3.360522
6	**291**	BN	Komtar	82.50428	5.7138378	5.3882285	7.6968712	8.1619912
7	**160**	BN	Bayan	-152.9307	-8.645113	-9.550021	-14.69839	-13.30565
8	**311**	BN	Telok	105.87315	7.392215	6.8635212	9.7824246	10.53596
9	**208**	IND	Penanti	30.068872	2.2541933	2.1947962	2.9823039	3.0630131
10	**106**	IND	Komtar	-51.56381	-4.107871	-4.369513	-5.830175	-5.48107
11	**316**	IND	Bayan	79.51284	5.1705112	4.9151241	7.0661101	7.4332612
12	**97**	IND	Telok	-58.0179	-4.659841	-5.008279	-6.667582	-6.203702

Rajah 1

(a) Uji ketaksandaran antara Dewan Undangan Negeri dan Parti Politik. Berikan komen terhadap keputusan.

(b) Bincangkan keputusan reja Haberman.

[12 markah]

4. The data in Table 2 are obtained from a sample of 1237 men between the ages of 40 and 59 (who did not develop coronary heart attack) taken from the Framingham longitudinal study. The men were cross-classified according to their serum cholesterol and systolic blood pressure. Figure 2 is the SAS output.

Table 2

Cholesterol (in mg/100 cc)	Blood pressure (in mm Hg)			
	<127	127-148	149-166	≥167
<200	117	121	47	22
200-219	85	98	43	20
220-259	119	209	68	43
≥ 260	67	99	46	33

Statistics for Table of CHOLES by BLOOD

Statistic	DF	Value	Prob
Chi-Square	9	20.8522	0.0133
Likelihood Ratio Chi-Square	9	20.3777	0.0157
Mantel-Haenszel Chi-Square	1	12.8396	0.0003
Phi Coefficient		0.1298	
Contingency Coefficient		0.1288	
Cramer's V		0.0750	

Statistics for Table of CHOLES by BLOOD

Statistic	Value	ASE
Pearson Correlation	0.1019	0.0285
Spearman Correlation	0.1021	0.0287
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000
Uncertainty Coefficient C|R	0.0066	0.0029
Uncertainty Coefficient R|C	0.0061	0.0027
Uncertainty Coefficient Symmetric	0.0063	0.0028

Sample Size = 1237

Summary Statistics for CHOLES by BLOOD

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	12.8396	0.0003

Summary Statistics for CHOLES by BLOOD

Cochran-Mantel-Haenszel Statistics (Based on Ridit Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	12.8914	0.0003

Total Sample Size = 1237

Figure 2

(a) Discuss the SAS output in Figure 2.

(b) Calculate gamma and discuss how the serum cholesterol and systolic blood pressure are associated.

[16 marks]

4. *Data dalam Jadual 2 diperoleh daripada sampel 1237 lelaki di antara umur 40 dan 59 (yang tidak mengalami serangan jantung koronari) yang diambil daripada kajian longitud Framingham. Kesemua lelaki diklasifikasikan mengikut kolesterol serum dan tekanan darah sistolik. Rajah 2 adalah output SAS.*

Jadual 2

Kolesterol (in mg/100 cc)	*Tekanan darah (in mm Hg)*			
	<127	*127-148*	*149-166*	*≥167*
<200	*117*	*121*	*47*	*22*
200-219	*85*	*98*	*43*	*20*
220-259	*119*	*209*	*68*	*43*
≥260	*67*	*99*	*46*	*33*

```
                Statistics for Table of CHOLES by BLOOD

Statistic                     DF       Value      Prob
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Chi-Square                     9     20.8522    0.0133
Likelihood Ratio Chi-Square    9     20.3777    0.0157
Mantel-Haenszel Chi-Square     1     12.8396    0.0003
Phi Coefficient                      0.1298
Contingency Coefficient              0.1288
Cramer's V                           0.0750

               Statistics for Table of CHOLES by BLOOD

Statistic                              Value       ASE
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

Pearson Correlation                   0.1019    0.0285
Spearman Correlation                  0.1021    0.0287

Lambda Asymmetric C|R                 0.0000    0.0000
Lambda Asymmetric R|C                 0.0000    0.0000
Lambda Symmetric                      0.0000    0.0000

Uncertainty Coefficient C|R           0.0066    0.0029
Uncertainty Coefficient R|C           0.0061    0.0027
Uncertainty Coefficient Symmetric     0.0063    0.0028
                    Sample Size = 1237

            Summary Statistics for CHOLES by BLOOD

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
    1        Nonzero Correlation        1     12.8396    0.0003


            Summary Statistics for CHOLES by BLOOD

  Cochran-Mantel-Haenszel Statistics (Based on Ridit Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
    1        Nonzero Correlation        1     12.8914    0.0003

                Total Sample Size = 1237
```

Rajah 2

(a) Bincangkan output SAS di Rajah 2.

(b) Kira gamma dan bincangkan perkaitan antara kolesterol serum dan tekanan darah sistolik.

[16 markah]

5. Suppose that a researcher would like to predict the number of students who are absent from a given classroom per day based on several predictors such as family income, race, and gender. Explain your reasoning for each of the following:
   (a) What is the random component of the GLM used in this study?
   (b) What is the systematic component of the GLM used in this study?
   (c) What is the most appropriate link function for the GLM used in this study?

[8 marks]

5. Katakan seorang penyelidik mahu meramal bilangan pelajar yang tidak hadir ke kelas tertentu setiap hari berdasarkan beberapa peramal seperti pendapatan keluarga, bangsa, dan jantina. jelaskan hujah anda bagi setiap yang berikut:
   (a) Apakah komponen rawak GLM yang digunakan dalam kajian ini
   (b) Apakah komponen sistematik GLM yang digunakan dalam kajian ini
   (c) Apakah fungsi pautan yang paling sesuai bagi GLM yang digunakan dalam kajian ini?

[8 markah]

6. Let Y be the number of successes out of n trials. Using the exponential dispersion family of distributions, find E(Y) and Var(Y) for binomial distribution with the following parameters:
   (a) π is the probability of success on each trial.
   (b) $P=Y/n$ is the proportion of successes.

[10 marks]

6. Andaikan Y adalah bilangan kejayaan dari n per cubaan. Dengan menggunakan taburan penyebaran keluarga eksponen, cari E(Y) dan Var(Y) untuk taburan binomial dengan parameter seperti berikut
   (a) π ialah kebarangkalian kejayaan bagi setiap percubaan.
   (b) $P=Y/n$ ialah kadaran kejayaan.

[10 markah]

7. Assume that the dependent variable take three nominal scale values, 0, 1 and 2. Then find the conditional probabilities of Y=0, Y=l and Y=2 for given X and show the likelihood and log-likelihood functions.

[10 marks]

7. Andaikan pembolehubah bersandar boleh mengambil tiga nilai skala nominal, 0, 1 dan 2. Seterusnya dapatkan kebarangkalian bersyarat Y=0, Y=1 dan Y=2 diberikan X dan tunjukkan fungsi kebolehjadian dan fungsi log-kebolehjadian.

[10 markah]

8. In a study to investigate the effectiveness of a drug in killing cancer cells in rats, half of n = 78 female pregnant rats induced with cancer were randomized to the drug and the other half to a control. Then the number of baby rates (y in the data set) whose cancer cells were below certain limit was recorded for each litter (whose size is denoted by n). The weight in gram for each mother rat before pregnancy was also recorded (weight in the data set). The treatment indicator is drug, where drug = 1 for the drug and drug = 0 for the control. The following SAS program is used to fit the above data:

```
proc genmod;
model y/n = drug weight /dist=bin link=logit;
run;
```

and obtain the following output in Figure 3 with some information intentionally deleted:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	??	194.8490	??
Scaled Deviance	??	194.8490	??
Pearson Chi-Square	??	167.1993	??
Scaled Pearson X2	??	167.1993	??
Log Likelihood		-310.0266	
Full Log Likelihood		-159.4305	
AIC (smaller is better)		324.8610	
AICC (smaller is better)		325.1853	
BIC (smaller is better)		331.9311	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.3327	8.6207	-20.2290	13.5635	0.15	0.6991
drug	1	1.3000	0.2000	?	?	?	?
weight	1	0.0318	0.1079	-0.1797	0.2433	0.09	0.7684
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Figure 3

(a) What model did the SAS program fit? Write down the fitted model.

(b) Describe the drug effect in killing cancer cells. Construct a 95% confident interval for the parameter describing the drug effect assuming no overdispersion.

(c) Is there an overdispersion in the data? If there is, what is the reason for the overdispersion?

[14 marks]

8. *Dalam satu kajian untuk menyiasat keberkesanan dadah dalam membunuh sel-sel kanser pada tikus, separuh daripada n=78 tikus betina bunting telah disuntik dengan kanser dan diberi dadah secara rambang dan separuh lagi untukkawalan. Kemudian bilangan kadar kelahiran (y dalam set data) di mana sel-sel kanser adalah di bawah had tertentu dicatatkan bagi setiap kelahiran (yang saiz ditandakan oleh n). Berat dalam gram bagi setiap ibu tikus sebelum bunting juga direkodkan (berat dalam set data). Petunjuk rawatan adalah dadah di mana, dadah=1 diberi dadah dan dadah=0 untuk kawalan. Program SAS berikut telah digunakan untuk menyuaikan data di atas:*

```
proc genmod;
model y/n = drug weight /dist=bin link=logit;
run;
```

dan output berikut dalam Rajah 3 didapati dengan beberapa maklumat sengaja dipadam:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	??	194.8490	??
Scaled Deviance	??	194.8490	??
Pearson Chi-Square	??	167.1993	??
Scaled Pearson X2	??	167.1993	??
Log Likelihood		-310.0266	
Full Log Likelihood		-159.4305	
AIC (smaller is better)		324.8610	
AICC (smaller is better)		325.1853	
BIC (smaller is better)		331.9311	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.3327	8.6207	-20.2290	13.5635	0.15	0.6991
drug	1	1.3000	0.2000	?	?	?	?
weight	1	0.0318	0.1079	-0.1797	0.2433	0.09	0.7684
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Rajah 3

(a) Apakah model yang disuaikan program SAS ? Tuliskan model tersebut.

(b) Huraikan kesan dadah dalam membunuh sel-sel kanser. Bina suatu selang keyakinan 95% untuk parameter menerangkan kesan dadah dengan andaian tiada lebihan sebaran.

(c) Adakah terdapat lebihan sebaran dalam data? Jika ada, apakah sebab untuk lebihan sebaran?

[14 markah]

9. Let X with 2 categories (1 and 2), Y with 3 categories (1, 2, and 3) and Z with four categories (1, 2, 3 and 4). The variable named count contains the cell frequencies for all combinations of categories of X, Y and Z. Using proc catmod in SAS, the following output is obtained:

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
x	1	0.00	0.9847
y	2	7.14	0.0281
x*y	2	11.15	0.0038
z	3	9.66	0.0217
x*z	3	9.88	0.0196
y*z	6	8.71	0.1906
x*y*z	6	9.60	0.1427
Likelihood Ratio	0	.	.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
x	1	0.28	0.5967
y	2	7.95	0.0188
x*y	2	11.59	0.0030
z	3	9.65	0.0218
x*z	3	11.34	0.0100
y*z	6	7.72	0.2591
Likelihood Ratio	6	10.67	0.0990

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
x	1	0.25	0.6193
y	2	6.82	0.0330
z	3	8.47	0.0372
x*y	2	11.44	0.0033
x*z	3	11.21	0.0107
Likelihood Ratio	12	19.18	0.0844

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
x	1	-0.0303	0.0610	0.25	0.6193
y	1	0.2000	0.0796	6.30	0.0120
	2	-0.1636	0.0871	3.53	0.0603
z	1	-0.2071	0.1103	3.53	0.0603
	2	-0.1176	0.1058	1.23	0.2667
	3	0.1018	0.0991	1.05	0.3044
x*y	1 1	0.2470	0.0796	9.62	0.0019
	1 2	-0.2367	0.0871	7.39	0.0066
x*z	1 1	-0.2634	0.1103	5.71	0.0169
	1 2	-0.1212	0.1058	1.31	0.2521
	1 3	0.2434	0.0991	6.03	0.0141

Figure 4

Based on the SAS output given in Figure 4, discuss a log-linear model that fits these data well.

[12 marks]

9. *Andaikan X dengan 2 kategori (1 dan 2), Y dengan 3 kategori (1, 2, dan 3) dan Z dengan empat kategori (1, 2, 3 dan 4). Pemboleh ubah bernama count merupakan frekuensi sel untuk semua kombinasi kategori X, Y dan Z. Menggunakan proc catmod dalam SAS, output berikut diperolehi:*

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
x	1	0.00	0.9847
y	2	7.14	0.0281
x*y	2	11.15	0.0038
z	3	9.66	0.0217
x*z	3	9.88	0.0196
y*z	6	8.71	0.1906
x*y*z	6	9.60	0.1427
Likelihood Ratio	0	.	.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
x	1	0.28	0.5967
y	2	7.95	0.0188
x*y	2	11.59	0.0030
z	3	9.65	0.0218
x*z	3	11.34	0.0100
y*z	6	7.72	0.2591
Likelihood Ratio	6	10.67	0.0990

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
x	1	0.25	0.6193
y	2	6.82	0.0330
z	3	8.47	0.0372
x*y	2	11.44	0.0033
x*z	3	11.21	0.0107
Likelihood Ratio	12	19.18	0.0844

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
x	1	-0.0303	0.0610	0.25	0.6193
y	1	0.2000	0.0796	6.30	0.0120
	2	-0.1636	0.0871	3.53	0.0603
z	1	-0.2071	0.1103	3.53	0.0603
	2	-0.1176	0.1058	1.23	0.2667
	3	0.1018	0.0991	1.05	0.3044
x*y	1 1	0.2470	0.0796	9.62	0.0019
	1 2	-0.2367	0.0871	7.39	0.0066
x*z	1 1	-0.2634	0.1103	5.71	0.0169
	1 2	-0.1212	0.1058	1.31	0.2521
	1 3	0.2434	0.0991	6.03	0.0141

Rajah 4

Berdasarkan output SAS yang diberikan dalam Rajah 4, bincangkan model log linear yang menyuaikan data ini dengan baik.

[12 markah]

...12/-