UNIVERSITI SAINS MALAYSIA

First Semester Examination
2016/2017 Academic Session

December 2016 / January 2017

MST567 - Categorical Data Analysis
[Analisis Data Berkategori]

Duration : 3 hours
[Masa : 3 jam]

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

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

Instructions: Answer **EIGHT** (8) questions.

*[**Arahan:** Jawab **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].

Question 1

Do women receive equal pay for equal work? A survey of 725 women conducted between July 30 and August 3, 2015, by the Ministry of Woman, Family and Community Development found that 450 said that their current job provides equal pay for equal work. Construct a 90% confidence interval for the true probability that a randomly chosen woman feels that her current job provides equal pay for equal work. What assumptions are you making in constructing this interval? Do this confidence interval reasonable?

[8 marks]

Soalan 1

Adakah wanita menerima gaji yang sama bagi kerja yang sama? Satu kajian terhadap 725 wanita yang dijalankan antara 30 Julai dan 3 Ogos 2015, oleh Kementerian Pembangunan Wanita, Keluarga dan Masyarakat mendapati bahawa 450 menyatakan bahawa pekerjaan mereka memberi penggajian yang setara dengan pekerjaan. Bina selang keyakinan 90% bagi kebarangkalian benar bahawa seorang wanita yang dipilih secara rawak merasakan bahawa kerja semasa beliau menyediakan penggajian yang setara dengan pekerjaan. Apa andaian anda dalam membina selang ini? Adakah selang keyakinan ini munasabah?

[8 markah]

Question 2

Let $Y_1 \sim Poisson(\mu_1)$ and $Y_2 \sim Poisson(\mu_2)$. If $N = Y_1 + Y_2$, show that the conditional distribution of Y_1 given $N = n$ is a $Binomial(n, \pi)$ where $\pi = \dfrac{\mu_1}{\mu_1 + \mu_2}$.

[8 marks]

Soalan 2

Biarkan $Y_1 \sim Poisson(\mu_1)$ *dan* $Y_2 \sim Poisson(\mu_2)$. *Jika* $N = Y_1 + Y_2$, *tunjukkan taburan bersyarat* Y_1 *diberikan* $N = n$ *ialah* $Binomial(n, \pi)$ *yang mana* $\pi = \dfrac{\mu_1}{\mu_1 + \mu_2}$.

[8 markah]

Question 3

A researcher went to a university campus. During one hour he asked students who passed by to fill in a questionnaire on movie and television watching habits. Each question had three alternatives as to whether a person never (0), sometimes (1) or frequently (2) watched movies or programs related to a certain subject. A total of 102 students responded, and the results of watching a fantasy and a sports program are summarized in the following 3 x 3 table:

Fantasy	Sports		
	0	1	2
0	3	2	12
1	8	14	7
2	20	33	3

(a) Discuss the sampling distribution of the table.

(b) Test independence between the two programs.

(c) Investigate the direction of dependence.

[14 marks]

Soalan 3

Seorang penyelidik pergi ke kampus universiti. Dalam satu jam dia bertanya kepada pelajar yang lalu-lalang untuk mengisi soal selidik mengenai tabiat menonton filem dan televisyen. Setiap soalan mempunyai tiga alternatif sama ada orang yang tidak pernah (0), kadang-kadang (1) atau kerap (2) menonton filem atau program yang berkaitan dengan sesuatu subjek. Seramai 102 pelajar menjawab, dan keputusan menonton program fantasi dan sukan, diringkaskan dalam jadual 3 x 3 berikut:

Fantasi	*Sukan*		
	0	*1*	*2*
0	*3*	*2*	*12*
1	*8*	*14*	*7*
2	*20*	*33*	*3*

(a) Bincangkan taburan pensampelan bagi jadual.

(b) Uji ketakbersandaran kedua-dua soalan.

(c) Siasat arah bersandaran.

[14 markah]

Question 4

In a study on treatment for schizophrenic patients, 32 schizophrenic patients were treated by some antipsychotic drugs for a brief period of time, and their recidivism status was collected 12 months before and after the treatment. The figure below shows the SAS output.

The SAS System

The FREQ Procedure

Frequency

Table of AFTER by BEFORE

AFTER(12 month post-treatment)	BEFORE(12 month pre-treatment) NO	YES	Total
NO	12	2	14
YES	9	9	18
Total	21	11	32

Statistics for Table of AFTER by BEFORE

Statistic	DF	Value	Prob
Chi-Square	1	4.4527	0.0348
Likelihood Ratio Chi-Square	1	4.7469	0.0294
Continuity Adj. Chi-Square	1	3.0102	0.0827
Mantel-Haenszel Chi-Square	1	4.3135	0.0378
Phi Coefficient		0.3730	
Contingency Coefficient		0.3495	
Cramer's V		0.3730	

WARNING: Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	12
Left-sided Pr <= F	0.9950
Right-sided Pr >= F	0.0393
Table Probability (P)	0.0343
Two-sided Pr <= P	0.0608

(a) Write a complete SAS program to produce the output.

(b) Why SAS output indicate the Chi-square may an invalid test?

(c) Show the calculation of the one tailed and two tailed *p*-values for the Fisher's Exact Test. What is the conclusion of the test?

[14 marks]

Soalan 4

Dalam kajian mengenai rawatan untuk pesakit skizofrenia, 32 pesakit skizofrenia telah dirawat dengan beberapa ubat-ubatan antipsikotik untuk tempoh masa yang singkat, dan status pengulangan sakit mereka dikumpulkan 12 bulan sebelum dan selepas rawatan. Rajah di bawah menunjukkan output SAS.

The SAS System

The FREQ Procedure

Frequency — *Table of AFTER by BEFORE*

AFTER(12 month post-treatment)	*BEFORE(12 month pre-treatment)* *NO*	*YES*	*Total*
NO	12	2	14
YES	9	9	18
Total	21	11	32

Statistics for Table of AFTER by BEFORE

Statistic	*DF*	*Value*	*Prob*
Chi-Square	1	4.4527	0.0348
Likelihood Ratio Chi-Square	1	4.7469	0.0294
Continuity Adj. Chi-Square	1	3.0102	0.0827
Mantel-Haenszel Chi-Square	1	4.3135	0.0378
Phi Coefficient		0.3730	
Contingency Coefficient		0.3495	
Cramer's V		0.3730	

WARNING: Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	12
Left-sided Pr <= F	0.9950
Right-sided Pr >= F	0.0393
Table Probability (P)	0.0343
Two-sided Pr <= P	0.0608

(a) *Tulis satu program SAS lengkap untuk menghasilkan output tersebut.*

(b) *Mengapa output SAS menunjukkan Khi Kuasadua berkemungkinan ujian yang tidak sah?*

(c) *Tunjukkan pengiraan nilai p satu hujung dan dua hujung bagi ujian tepat Fisher. Apakah kesimpulan ujian?*

[14 markah]

Question 5

Let *Y* be the indicator of a certain type of cancer among smoking people (*Y*=1 means that the person has cancer and *Y*=0 that the person has no cancer). It is believed that the number of years that the person has smoked, *x* affects the risk of having the cancer.

(a) Determine the logistic model for the scenario described above.

(b) Suppose data from 30 patients were collected, and that the intercept was estimated by $\hat{\alpha} = -2.832$ and linear factor β was estimated as $\hat{\beta} = 0.012$. Use these values to estimate the risk of having this type of cancer for those who have smoked for 10 years, $\pi(10)$.

(c) Suppose the estimated covariance matrix was given by $\begin{pmatrix} 0.0091 & -0.00006 \\ -0.00006 & 0.000004 \end{pmatrix}$. Construct a 99% confidence interval for this risk, $\pi(10)$, by assuming that $\hat{\alpha}$ and $\hat{\beta}$ are approximately normally distributed.

[12 marks]

Soalan 5

Biar Y menjadi petunjuk kepada suatu jenis kanser tertentu di kalangan perokok (Y = 1 bermaksud bahawa orang itu mempunyai kanser dan Y = 0 orang itu mendapat kanser). Adalah dipercayai bahawa bilangan tahun orang itu telah merokok, x memberi kesan kepada risiko mendapat kanser.

(a) Tentukan model logistik untuk senario yang diterangkan di atas.

(b) Katakan data dari 30 pesakit telah dikumpul, pintasan yang dianggar adalah $\hat{\alpha} = -2.832$ *dan faktor linear* β *yang dianggar adalah* $\hat{\beta} = 0.012$. *Guna anggaran ini untuk menganggar risiko mendapat kanser bagi mereka yang merokok selama 10 tahun,* $\pi(10)$.

(c) Katakan anggaran matriks kovarians diberikan oleh $\begin{pmatrix} 0.0091 & -0.00006 \\ -0.00006 & 0.000004 \end{pmatrix}$. *Bina selang keyakinan 99% untuk risiko ini,* $\pi(10)$, *dengan mengandaikan* $\hat{\alpha}$ *dan* $\hat{\beta}$ *tertabur secara normal.*

[12 markah]

Question 6

A three-way table contains data of the binary categorical variables *X*, *Y* and *Z*. The number of observations $N_{ijk} \in Po\left(\mu_{ijk}\right)$ with *X* = *i*, *Y* = *j* and *Z* = *k* is Poisson distributed for $1 \le i, j, k \le 2$, and are independent for all different cells (*i*,*j*,*k*).

(a) Let (*XY*,*Z*) be the loglinear model where *X* and *Y* are jointly independent of *Z*. Express all expected cell counts μ_{ijk} in terms of the loglinear parameters, excluding those that are put to zero in order to avoid overparametrization.

(b) Use (a) to prove that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}}$$

where a plus sign denotes summation over the corresponding index.

(c) Use (b) and data n_{ijk} from the two partial tables below to find the estimates $\hat{\mu}_{ijk}$ of all μ_{ijk}.

Observed values n_{ij1}:

	j=1	*j*=2
i=1	65	42
i=2	29	38

Observed values n_{ij2}:

	j=1	*j*=2
i=1	20	32
i=2	19	15

[14 marks]

Soalan 6

Satu jadual tiga hala mengandungi data pemboleh ubah kategori binari X, Y and Z. Bilangan cerapan $N_{ijk} \in Po\left(\mu_{ijk}\right)$ *dengan X = i, Y = j and Z = k adalah bertaburan Poisson untuk* $1 \le i, j, k \le 2$, *dan tak bersandar untuk semua sel (i,j,k) yang berbeza.*

(a) Biarkan (XY,Z) adalah model loglinear yang mana X dan Y tak bersandar tercantum terhadap Z. Tuliskan semua jangkaan bilangan sel μ_{ijk} *berdasarkan parameter loglinear, tidak termasuk parameter yang diberikan nilai sifar bagi mengelakkan lebihan parameter.*

(b) Guna (a) untuk membuktikan bahawa

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}},$$

yang mana tanda tambah merujuk kepada penjumlahan indeks yang tertentu.

(c) Guna (b) dan data n_{ijk} daripada jadual separa di bawah bagi mencari anggaran $\hat{\mu}_{ijk}$ untuk semua μ_{ijk}.

Nilai tercerap n_{ij1}:

	j=1	*j=2*
i=1	*65*	*42*
i=2	*29*	*38*

Nilai tercerap n_{ij2}:

	j=1	*j=2*
i=1	*20*	*32*
i=2	*19*	*15*

[14 markah]

Question 7

The following 2 x 2 x 2-table summarize the finding from a study on the usage of alcohol (A), cigarettes (C) and marijuana (M) among 2276 students in American high schools.

Alcohol (A)	Cigarette (C)	Marijuana (M)	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

(a) Define the full loglinear model (ACM).

(b) Specify the loglinear model (A,CM).

(c) Find the best model.

[14 marks]

Soalan 7

Jadual 2 x 2 x 2 berikut, meringkaskan dapatan daripada kajian mengenai penggunaan alkohol (A), rokok (C) dan ganja (M) di kalangan 2276 pelajar di sekolah-sekolah tinggi Amerika.

Alkohol (A)	*Rokok (C)*	*Ganja (M)*	
		Ya	*Tidak*
Ya	*Ya*	*911*	*538*
	Tidak	*44*	*456*
Tidak	*Ya*	*3*	*43*
	Tidak	*2*	*279*

(a) Tentukan model loglinear penuh (ACM).

(b) Nyatakan model loglinear (A,CM).

(c) Cari model terbaik.

[14 markah]

Question 8

The usage of psychotropic drugs has increased over recent years in England. The patterns of psychotropic drug consumption in a sample from West London are collected and given in table below.

Gender	Age Group	Psychological case	On drugs	Total
M	1	No	**9**	**531**
M	2	No	**16**	**500**
M	3	No	**38**	**644**
M	4	No	**26**	**275**
M	5	No	**9**	**90**
M	1	Yes	**12**	**171**
M	2	Yes	**16**	**125**
M	3	Yes	**31**	**121**
M	4	Yes	**16**	**56**
M	5	Yes	**10**	**26**
F	1	No	**12**	**588**
F	2	No	**42**	**596**
F	3	No	**96**	**765**
F	4	No	**52**	**327**
F	5	No	**30**	**179**
F	1	Yes	**33**	**210**
F	2	Yes	**47**	**189**
F	3	Yes	**71**	**242**
F	4	Yes	**45**	**98**
F	5	Yes	**21**	**60**

Is Psychotropic drug use affected by gender (M=male, F=female), age group or psychological case and are there interactions among these effects? Fit an appropriate model and interpret the results.

(16 marks)

Soalan 8

Penggunaan ubat-ubatan psikotropik telah meningkat sejak beberapa tahun kebelakangan ini di England. Corak penggunaan dadah psikotropik dalam sampel dari Barat London dikumpul dan diberikan dalam jadual di bawah.

Jantina	***Kumpulan Umur***	***Kes psikotropik***	***Menguna Ubat***	***Jumlah***
M	*1*	*Tiada*	***9***	***531***
M	*2*	*Tiada*	***16***	***500***
M	*3*	*Tiada*	***38***	***644***
M	*4*	*Tiada*	***26***	***275***
M	*5*	*Tiada*	***9***	***90***
M	*1*	*Ya*	***12***	***171***
M	*2*	*Ya*	***16***	***125***
M	*3*	*Ya*	***31***	***121***
M	*4*	*Ya*	***16***	***56***
M	*5*	*Ya*	***10***	***26***
F	*1*	*Tiada*	***12***	***588***
F	*2*	*Tiada*	***42***	***596***
F	*3*	*Tiada*	***96***	***765***
F	*4*	*Tiada*	***52***	***327***
F	*5*	*Tiada*	***30***	***179***
F	*1*	*Ya*	***33***	***210***
F	*2*	*Ya*	***47***	***189***
F	*3*	*Ya*	***71***	***242***
F	*4*	*Ya*	***45***	***98***
F	*5*	*Ya*	***21***	***60***

Adakah penggunaan dadah psikotropik dipengaruhi oleh jantina (M=lelaki, F=perempuan), kumpulan umur atau kes psikologi serta adakah wujud interaksi antara kesan-kesan ini? Suaikan model yang sesuai dan tafsirkan keputusan.

[16 markah]

FORMULAE

$G^2 = -2\sum\sum n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}$	$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$
$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$	$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{[\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})]^{1/2}}$
$R = \frac{\pi_1}{\pi_2}$	$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$
$\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$	$U = -\frac{\sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} \log\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right)}{\sum_{j=1}^{J} \pi_{+j} \log \pi_{+j}}$
$CMH = \frac{\left[\sum_k (n_{11k} - \hat{\mu}_{11k})\right]^2}{\sum_k \bar{\text{var}}(n_{11k})}$	$se(\log r) = \left[\frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}\right]^{1/2}$
$\hat{\mu}_{11k} = n_{1+k}n_{+1k}/n_{++k}$	$se(\log \hat{\theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2}$
$\widehat{Var}_{11k} = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k}-1)$	$\hat{\gamma} = \frac{C-D}{C+D}$
$P = \frac{(n_{11}+n_{12})!(n_{11}+n_{21})!(n_{21}+n_{22})!(n_{12}+n_{22})!}{n_{11}!n_{12}!n_{21}!n_{22}!n!}$	$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$
$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$	$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
$f_Y(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}$	$z = \hat{\beta}/ASE$

- ooo O ooo -