UNIVERSITI SAINS MALAYSIA

Second Semester Examination
2011/2012 Academic Session

June 2012

MSG 368 – Sample Survey and Sampling Technique
[Tinjauan Sampel dan Teknik Pensampelan]

Duration : 3 hours
[Masa : 3 jam]

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

[Sila pastikan bahawa kertas peperiksaan ini mengandungi DUA BELAS 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. Consider the problem of measuring the wingspan of bats in a certain region. In this region there are three species of bats : hoavy, silver-haired, and red. Let p_k = proportion of bats in the k^{th} subpopulation that have a wingspan greater than 5 inches. The investigator believes that approximately 30 bats live in the area. Of the 30 bats captured, 12 are from the first species under study and 3 of these have a wingspan greater than 5 inches.
   (a) Estimate p_1 .
   (b) Determine the bound on error of estimation.
   (c) Obtain the 90% confidence interval for p_1 .

[10 marks]

1. Pertimbangkan masalah mengukur lebar sayap kelawar di suatu kawasan tertentu. Di kawasan tersebut terdapat tiga jenis kelawar: hoavy, berbulu perak, dan merah. Biarkan p_k = kadaran kelawar dalam subpopulasi ke-k yang mempunyai lebar sayap yang lebih besar daripada 5 inci. Penyiasat percaya bahawa kira-kira 300 kelawar berada di kawasan itu. Daripada 30 kelawar yang ditangkap, 12 daripadanya adalah jenis pertama di bawah kajian dan 3 ini mempunyai lebar sayap yang lebih besar daripada 5 inci.
   (a) Anggarkan p_1.
   (b) Tentukan anggaran batas ralatnya.
   (c) Dapatkan selang keyakinan 95% bagi p_1.

[10 markah]

2. (a) State the methods of data collection. Select two of them and give two advantages and two disadvantages of each.

   (b) Define the following terms:
   (i) Target population
   (ii) Sampling frame
   (iii) Sampling Bias
   (iv) Sampling error

[20 marks]

2. (a) Apakah kaedah-kaedah pengumpulan data?. Pilih dua daripadanya dan berikan dua kebaikan dan dua keburukan bagi setiap satunya.

(b) Takrifkan sebutan-sebutan yang berikut:
   (i) Populasi sasaran
   (ii) Kerangka pensampelan
   (iii) Pensampelan berat sebelah
   (iv) Pensampelan ralat

[20 markah]

3. In simple random sampling without replacement where n units are selected from N units, show that

(a) the probability that the i^{th} unit is selected on the k^{th} draw is $1/N$.

(b) $Var(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ with $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{N}$

[25 marks]

3. Dalam pensampelan rawak ringkas tanpa penggantian dengan n unit dipilih daripada N unit, tunjukkan bahawa

(a) kebarangkalian bahawa unit i dipilih pada cabutan ke-k ialah $1/N$.

(b) $Var(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ *dengan* $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{N}$

[25 markah]

4. There is a relationship between the income of a household and the total floor space of the home of a household in a certain region of Sarawak. A simple random sample of size 4 has been selected without replacement. The table below contains the sample data.

Household	1	2	3	4
Floor space	116	81	73	99
Income	1200	950	650	1050

Given $\sum y^2 = 3867500$; $\sum x^2 = 35147$; $\sum xy = 367550$

The mean floor space of all houses in the population is 103.7 m^2.

(a) Compute the ratio estimator for the mean income.
(b) Compute the regression estimator for the mean income.
(c) Which of the two methods, in parts (a) or (b), is most appropriate in this case? Why?

[25 marks]

4. *Terdapat hubungan antara pendapatan isi rumah dan jumlah ruang lantai rumah bagi satu isi rumah di rantau yang tertentu di Sarawak. Satu sampel rawak ringkas yang bersaiz 4 telah dipilih tanpa penggantian. Jadual di bawah mengandungi data sampel.*

Isirumah	*1*	*2*	*3*	*4*
Ruang lantai	*116*	*81*	*73*	*99*
Pendapatan	*1200*	*950*	*650*	*1050*

Diberi $\sum y^2 = 3867500$; $\sum x^2 = 35147$; $\sum xy = 367550$

Min ruang lantai bagi semua rumah dalam populasi ialah 103.7 m^2.

(a) Kirakan penganggar nisbah bagi min pendapatan.

(b) Kirakan penganggar regresi bagi min pendapatan.

(c) Yang manakah antara kedua-dua kaedah, di bahagian (a) atau (b) paling sesuai dalam kes ini? Kenapa?

[25 markah]

5. A university has 807 faculty members. For each faculty member, the number of referred publications was recorded. The following data are from a sample of faculty, using the areas biological sciences, physical sciences, social sciences, and humanities as the strata.

Stratum	Number of Faculty Members in Stratum	Number of Faculty members in Sample
Biological Sciences	102	7
Physical Sciences	310	19
Social Sciences	217	13
Humanities	178	11
Total	807	50

The frequency table for number of publications in the strata is given below.

Number of Referred Publications	Number of Faculty Members			
	Biological	Physical	Social	Humanities
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0

(a) Estimate the total number of referred publications by faculty members in the university, and give the standard error.

(b) Estimate the proportion of faculty with no referred publications, and determine the bound on the error of estimation. Explain your results.

[25 marks]

5. *Sebuah universiti mempunyai 807 ahli fakulti. Bagi setiap ahli fakulti, bilangan penerbitan yang dirujuk telah direkodkan. Data berikut adalah daripada sampel fakulti, dengan menggunakan bidang sains biologi, sains fizikal, sains sosial, dan sains kemasyarakatan sebagai strata.*

Stratum	*Bilangan Ahli Fakulti dalam Stratum*	*Bilangan Ahli Fakulti dalam Sampel*
Sains Biologi	*102*	*7*
Sains Fizikal	*310*	*19*
Sains Sosial	*217*	*13*
Kemasyarakatan	*178*	*11*
Jumlah	*807*	*50*

Jadual frekuensi bagi bilangan penerbitan dalam strata adalah seperti di bawah.

Bilangan Penerbitan Yang Dirujuk	*Bilangan Ahli Fakulti*			
	Biologi	*Fizikal*	*Sosial*	*Kemasyarakatan*
0	*1*	*10*	*9*	*8*
1	*2*	*2*	*0*	*2*
2	*0*	*0*	*1*	*0*
3	*1*	*1*	*0*	*1*
4	*0*	*2*	*2*	*0*
5	*2*	*1*	*0*	*0*
6	*0*	*1*	*1*	*0*
7	*1*	*0*	*0*	*0*
8	*0*	*2*	*0*	*0*

(a) Anggarkan jumlah penerbitan yang dirujuk oleh ahli fakulti di universiti , dan berikan ralat piawai.

(b) Anggarkan kadaran fakulti yang tiada penerbitan yang dirujuk, dan tentukan batas ralat penganggaran. Jelaskan jawapan anda.

[25 markah]

6. The new candy Green Sweet is being test-marketed in the areas of Selangor and Johore. The market research firm decided to sample 6 cities from the 45 cities in the areas and then to sample supermarkets within cities, in order to know the number of cases of Green Sweet sold.

City	Number of Supermarkets	Number of cases sold
1	52	146, 180, 251, 152, 72, 181, 171, 361, 73, 186
2	19	99, 101, 52, 121
3	37	199, 179, 98, 63, 126, 87, 62
4	39	226, 129, 57, 46, 86, 43, 85, 165
5	8	12, 23
6	14	87, 43, 59

(a) State the type of sampling design used.
(b) State the primary sampling units and secondary sampling units.
(c) Estimate the total number of cases sold and place the bound on the error of estimation.
(d) Estimate the average number sold per supermarket. Determine the standard error of your estimates.

[35 markah]

6. *Gula-gula baru Green Sweet sedang dipasarkan di kawasan Selangor dan Johor. Firma penyelidikan pasaran telah memutuskan untuk mencuba 6 bandar raya daripada 45 bandar raya di kawasan-kawasan tersebut dan kemudian akan mensampelkan pasar raya dalam bandar raya terpilih supaya mengetahui bilangan kes Green Sweet yang telah dijual.*

Bandar raya	*Bilangan Pasar raya*	*Bilangan kes yang dijual*
1	*52*	*146, 180, 251, 152, 72, 181, 171, 361, 73, 186*
2	*19*	*99, 101, 52, 121*
3	*37*	*199, 179, 98, 63, 126, 87, 62*
4	*39*	*226, 129, 57, 46, 86, 43, 85, 165*
5	*8*	*12, 23*
6	*14*	*87, 43, 59*

(a) Nyatakan jenis reka bentuk pensampelan yang digunakan.
(b) Nyatakan unit pensampelan utama dan unit pensampelan kedua.
(c) Anggarkan jumlah kes-kes yang dijual dan tentukan batas ralat penganggaran .
(d) Anggarkan min bilangan yang dijual pada setiap pasar raya. Tentukan ralat piawai bagi penganggaran anda.

[35 markah]

7. A sampling design was conducted in a large city in China. The city comprises six districts. Within each district, a simple random sample of two neighborhood groups were selected, and all individuals within each neighborhood group were interviewed concerning their overall health status. The following table represents data from the survey on the number of individuals over 30 years of age who were edentulous (having lost teeth). District 1, 2, 4, and 6 each contain 200 neighborhood groups; district 3 contains 175 neighborhood groups; and district 5 contains 150 neighborhood groups. The data are as follows:

District	Neighborhood Group	Number of Persons > 30 years	Number of Edentulous Persons
1	1	28	7
	2	35	9
2	1	29	12
	2	43	26
3	1	61	19
	2	48	12
4	1	15	10
	2	39	28
5	1	21	9
	2	46	15
6	1	12	0
	2	25	4

(a) State the type of sampling design used.
(b) Estimate and construct 95% confidence interval for the total number of edentulous.

[15 marks]

7. *Satu reka bentuk pensampelan telah dijalankan di sebuah bandar besar di China. Bandar ini terdiri daripada enam daerah. Dalam setiap daerah, dua kumpulan kejiranan telah dipilih secara sampel rawak ringkas, dan kesemua individu dalam kumpulan setiap kejiranan telah ditemubual berkenaan status kesihatan secara keseluruhan. Jadual berikut mewakili data daripada kajian terhadap bilangan individu yang lebih daripada 30 tahun yang edentulous (kehilangan gigi). Daerah 1, 2, 4, dan 6 setiapnya mengandungi 200 kumpulan kejiranan; daerah 3 mengandungi 175 kumpulan kejiranan; dan daerah 5 mengandungi 150 kumpulan kejiranan. Data adalah seperti yang berukut:*

Daerah	***Kumpulan Kejiranan***	***Bilangan Individu > 30 tahun***	***Bilangan Individu yang Edentulous***
1	*1*	*28*	*7*
	2	*35*	*9*
2	*1*	*29*	*12*
	2	*43*	*26*
3	*1*	*61*	*19*
	2	*48*	*12*
4	*1*	*15*	*10*
	2	*39*	*28*
5	*1*	*21*	*9*
	2	*46*	*15*
6	*1*	*12*	*0*
	2	*25*	*4*

…8/-

(a) Nyatakan jenis reka bentuk pensampelan yang digunakan.

(b) Anggarkan dan binakan selang keyakinan 95% bagi jumlah bilangan edentulous.

[15 markah]

8. (a) Describe the three types of population in order to choose between systematic and simple random sampling.

(b) Crop yield for a large field of wheat is to be estimated by sampling small plots within the field while the grain is ripening. The field is on sloping land with higher fertility towards the lower side. Do you think simple random sampling or systematic random sampling will be effective in estimating τ, the total wheat yield? Provide a detailed explanation.

[15 marks]

8. (a) Huraikan tiga jenis populasi untuk memilih antara pensampelan sistematik dan rawak mudah.

(b) Hasil tanaman bagi sebuah ladang gandum adalah dianggarkan dengan mensampelkan plot-plot kecil dalam ladang sementara menunggu bijirin masak. Ladang tersebut berada pada tanah bercerun dengan kesuburan yang lebih tinggi ke arah bahagian bawah. Pada fikiran anda adakah pensampelan rawak mudah atau pensampelan rawak sistematik akan berkesan dalam menganggarkan τ , jumlah hasil gandum? Berikan penjelasan yang terperinci.

[15 markah]

9. If term $1/N_h$ is ignored relative to unity, show that

$$Var_{optimum}(\bar{y}_{st}) \leq Var_{proportional}(\bar{y}_{st})$$

[15 marks]

9. Jika sebutan $1/N_h$ diabaikan relatif kepada uniti, tunjukkan bahawa

$$Var_{optimum}(\bar{y}_{st}) \leq Var_{proportional}(\bar{y}_{st})$$

[15 markah]

10. A researcher would like to conduct a stratified random sampling from a population. If the cost survey function is in the form of

$$C = C_0 + \sum_{i=1}^{L} C_i n_i$$

where

C = RM10,000 and C_0 = RM 4,000 ,

determine the sample size to be taken from each stratum given the following information:

Stratum	C_i	S_i	N_i
1	4	13	120
2	5	16	250
3	2	14	80
4	3	8	300
5	6	12	500

[15 marks]

10. Seorang penyelidik ingin menjalankan pensampelan rawak berstrata daripada satu populasi . Jika fungsi kos penyelidikan adalah dalam bentuk

$$C = C_0 + \sum_{i=1}^{L} C_i n_i$$

dengan

C = *RM 10,000 and* C_0 = *RM 4,000* ,

tentukan saiz sampel yang diambil daripada setiap stratum dengan maklumat yang diberi seperti berikut:

Stratum	C_i	S_i	N_i
1	*4*	*13*	*120*
2	*5*	*16*	*250*
3	*2*	*14*	*80*
4	*3*	*8*	*300*
5	*6*	*12*	*500*

[15 markah]

Appendix

Sample	Sampel variance
$\sum_{i=1}^{n} \frac{y_i}{n}$	$\frac{s^2}{n}\left(\frac{N-n}{N}\right), s^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}$
$N\bar{y}$	$N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)$
$\frac{a}{n}$	$\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)$
$\frac{\sum_{i=1}^{n} N_i \bar{y}_i}{N}$	$\sum_{i=1}^{n} \frac{N_i^2}{N^2}\left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}$
$\sum_{i=1}^{n} \frac{N_i \hat{p}_i}{N}$	$\sum_{i=1}^{n} \frac{N_i^2}{N^2}\left(\frac{N_i - n_i}{N_i}\right)\frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}$
$\frac{\bar{y}}{\bar{x}}$	$\left(\frac{N-n}{nN}\right)\left(\frac{1}{\mu_x^2}\right)\left(\frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}\right)$
$\bar{y} + b(\mu_x - \bar{x})$, $b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$\left(\frac{N-n}{Nn}\right)\left(\frac{1}{n-2}\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$ $\approx \left(\frac{N-n}{Nn}\right)(S_y^2 - b^2 S_x^2)$
$\mu_x + \bar{d}$	$\left(\frac{N-n}{Nn}\right)\left(\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}\right)$
$\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}\right)$	$\left(\frac{N-n}{Nn\bar{M}^2}\right)\left(\frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1}\right)$

Sampel	Sample Variance
$\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}$	$\left(\frac{N-n}{Nn-\bar{M}^2}\right)\frac{\sum_{i=1}^{n}\left(a_i-\hat{p}m_i\right)^2}{n-1}$
$M\bar{y}$	$M^2\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}m_i\right)^2}{n-1}$
$\frac{N}{n}\sum_{i=1}^{n} y_i$	$N^2\left(\frac{N-n}{Nn}\right)S_t^2$ with $S_t^2=\sum_{i=1}^{n}\frac{(y_i-\bar{y}_t)^2}{n-1}$
$\hat{\mu}=\frac{1}{n\bar{M}}\sum_{i=1}^{n} M_i\bar{y}_i$	$\left(\frac{N-n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)S_b^2+\frac{1}{nN\bar{M}^2}\sum_{i=1}^{n}M_i^2\left(\frac{M_i-m_i}{M_i}\right)\left(\frac{S_i^2}{m_i}\right)$ with $S_b^2=\frac{\left(\sum_{i=1}^{n}M_i\bar{y}_i-\bar{M}\hat{\mu}\right)^2}{n-1}$ $S_i^2=\frac{\sum_{j=1}^{m_i}\left(y_{ij}-\bar{y}_i\right)^2}{m_i-1}$
$\hat{\mu}_r=\frac{\sum_{i=1}^{n}M_i\bar{y}_i}{\sum_{i=1}^{n}M_i}$	$\left(\frac{N-n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)S_r^2+\frac{1}{nN\bar{M}^2}\sum_{i=1}^{n}M_i^2\left(\frac{M_i-m_i}{M_i}\right)\left(\frac{S_i^2}{m_i}\right)$ with $S_r^2=\frac{\sum_{i=1}^{n_i}\left(M_i\bar{y}_i-\hat{\mu}_rM_i\right)^2}{n-1}$
$\hat{p}=\frac{\sum_{i=1}^{n}M_i\hat{p}_i}{\sum_{i=1}^{n}M_i}$	$\left(\frac{N-n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)S_r^2+\frac{1}{nN\bar{M}^2}\sum_{i=1}^{n}M_i^2\left(\frac{M_i-m_i}{M_i}\right)\left(\frac{\hat{p}_i\hat{q}_i}{m_i-1}\right)$ with $S_r^2=\frac{\sum_{i=1}^{n_i}\left(M_i\hat{p}_i-\hat{p}M_i\right)^2}{n-1}$

Sample Size
$n = \dfrac{N\sigma^2}{(N-1)D + \sigma^2}$; $D = \dfrac{B^2}{4}$; $D = \dfrac{B^2}{4N^2}$
$n = \dfrac{\sum_{i=1}^{L} \frac{N_i \sigma_i^{\,2}}{w_i}}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^{\,2}}$; $w_i = \dfrac{n_i}{n}$
$n = \dfrac{\left(\sum_{k=1}^{L} N_k \sigma_k / \sqrt{C_k}\right)\left(\sum_{i=1}^{L} N_i \sigma_i^{\,2} \sqrt{C_i}\right)}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^{\,2}}$, $n = \dfrac{(C - C_o)\sum_{i=1}^{L} N_i \sigma_i / \sqrt{C_i}}{\sum_{i=1}^{L} N_i \sigma_i \sqrt{C_i}}$ $n_i = \dfrac{n\, N_i \sigma_i / \sqrt{C_i}}{\sum_{i=1}^{L} N_i \sigma_i / \sqrt{C_i}}$ Optimal Allocation
$n = \dfrac{\left(\sum_{i=1}^{L} N_i \sigma_i\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^{\,2}}$; $n_i = n\left(\dfrac{N_i \sigma_i}{\sum_{i=1}^{L} N_i \sigma_i}\right)$ Neyman Allocation
$n = \dfrac{\sum_{i=1}^{L} N_i \sigma_i^{\,2}}{ND + \frac{1}{N}\sum_{i=1}^{L} N_i \sigma_i^{\,2}}$; $n_i = n\left(\dfrac{N_i}{\sum_{i=1}^{L} N_i}\right)$ Proportional Allocation
$n = \dfrac{\sum_{i=1}^{L} N^2{}_i\, p_i q_i / a_i}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i}$; $n_i = n\left(\dfrac{N_i \sqrt{p_i q_i / c_i}}{\sum_{i=1}^{L} N_i \sqrt{p_i q_i / c_i}}\right)$
$n = \dfrac{N\sigma^2}{ND + \sigma^2}$; $D = \dfrac{B^2 \mu x^2}{4}$; $D = \dfrac{B^2}{4}$; $D = \dfrac{B^2}{4N^2}$
$n = \dfrac{N\sigma_r^{\,2}}{ND + \sigma_r^{\,2}}$; $D = \dfrac{B^2 \left(\bar{M}\right)^2}{4}$; $D = \dfrac{B^2}{4N^2}$