UNIVERSITI SAINS MALAYSIA

Second Semester Examination
2012/2013 Academic Session

June 2013

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 FOURTEEN pages of printed material before you begin the examination.

[Sila pastikan bahawa kertas peperiksaan ini mengandungi EMPAT BELAS 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. (a) Define the following terms:
   (i) Target population
   (ii) Variable of study
   (iii) Elementary unit
   (iv) Sampling frame
   (v) Enumeration unit.

   (b) For the following problem indicate how you would carry out a sample and specify the entities of part (a):

   Suppose we wish to do a nutritional survey in order to estimate the average amount of fiber consumed by individuals in a certain city. Suppose there is no list of families available for the city but there is a map of the city showing each block in detail.

   [10 marks]

1. (a) Takrifkan istilah-istilah berikut:
   (i) Populasi sasaran
   (ii) Pembolehubah kajian
   (iii) Unit asas
   (iv) Rangka pensampelan
   (v) Unit pensampelan

   (b) Bagi masalah yang berikut, nyatakan bagaimana anda akan menjalankan sampel dan nyatakan entiti dari bahagian (a):

   Katakan kita ingin menjalankan suatu kajian pemakanan dalam usaha untuk menganggar amaun min serat yang diperlukan oleh individu di sebuah bandar tertentu. Andaikan tiada senarai keluarga yang disediakan untuk bandar tetapi terdapat peta bandar yang menunjukkan setiap blok secara terperinci.

   [10 markah]

2. Briefly describe two advantages and two disadvantages of using interviewers for survey data collection

   [8 marks]

2. Terangkan secara ringkas dua kelebihan dan dua kelemahan apabila menggunakan penemuramah untuk pengumpulan data kajian.

   [8 markah]

3. A simple random sample of 10 farms is selected from a population of 50 farms in a particular district. The numbers of cows in the sample farm are 23, 14, 38, 11, 7, 31, 9, 18, 12, 25.

(a) Estimate the mean number of cows per farm and determine the bound on the error of estimation.

(b) Estimate the total number of cows in this population and estimate the variance of the estimator.

(c) (i) What is the probability for selecting any particular farm in the population.

(ii) How many possible samples are there?

(iii) What is the probability of selecting the particular sample in this problem?

(d) Find the approximate 95% confidence interval for the population mean and population total.

[20 marks]

3. Satu sampel rawak ringkas sebanyak 10 ladang telah dipilih daripada populasi yang terdiri daripada 50 ladang di suatu daerah tertentu. Bilangan lembu di ladang yang di sampel adalah 23, 14, 38, 11, 7, 31, 9, 18, 12, 25.

(a) Anggarkan min bilangan lembu bagi setiap ladang dan tentukan batas ralat penganggaran.

(b) Anggarkan jumlah bilangan lembu dalam populasi dan anggarkan variansnya.

(c) (i) Apakah kebarangkalian untuk memilih mana-mana ladang tertentu dari populasi?

(ii) Berapa banyakkah sampel yang mungkin?

(iii) Apakah kebarangkalian memilih sampel yang tertentu dalam masalah ini?

(d) Cari anggaran selang keyakinan 95% bagi min populasi dan jumlah populasi.

[20 markah]

4. A wildland manager would like to estimate the total number of caribou in south-central Alaska. The sample unit is a 4-square mile area. A count of caribou is made on each unit selected. Based on a preliminary survey, the area utilized by the herd is divided into locations, and the rough estimates of standard deviations and costs for surveying sample units in each stratum are given in the following table:

Location *(h)*	N_h	σ_h	c_h
1	400	75	6
2	30	60	6
3	61	600	6
4	18	150	8
5	70	350	8
6	120	100	10

(a) By using the optimal allocation,
   (i) determine the allocation to each location assuming that it is desired to estimate the total number of caribou to be within 5000 caribou with 95% probability.
   (ii) calculate the total cost of the survey (assuming no fixed overhead cost).

(b) By using the proportional allocation,
   (i) determine the allocation to each location assuming that it is desired to estimate the total number of caribou to be within 5000 caribou with 95% probability.
   (ii) calculate the total cost of the survey.

[20 marks]

4. *Seorang pengurus tanah liar ingin menganggar jumlah caribou di selatan tengah Alaska. Unit sampel adalah sebuah kawasan 4 batu persegi. Pengiraan caribou dibuat pada setiap unit yang dipilih. Berdasarkan kajian awal, kawasan yang digunakan oleh binatang tersebut dibahagikan kepada beberapa lokasi dan anggaran kasar sisihan piawai dan kos untuk peninjauan unit sampel diberi dalam jadual yang berikut:*

Lokasi (h)	N_h	σ_h	c_h
1	400	75	6
2	30	60	6
3	61	600	6
4	18	150	8
5	70	350	8
6	120	100	10

(a) Dengan menggunakan peruntukan optimum,

(i) tentukan peruntukan bagi setiap lokasi dengan anggapan bahawa adalah dikehendaki untuk menganggarkan jumlah caribou agar berada dalam lingkungan 5000 caribou dengan kebarangkalian 95%.

(ii) hitung jumlah kos kaji selidik (andaian tiada kos "overhead" tetap).

(b) Dengan menggunakan peruntukan berkadaran,

(i) tentukan peruntukan bagi setiap lokasi dengan anggapan bahawa adalah dikehendaki untuk menganggarkan jumlah caribou dengan kebarangkalian 95%.

(ii) hitung jumlah kos bagi kajian ini.

[20 markah]

5. A researcher from USM has developed a test to measure the degree of awareness of current events. She would like to estimate the average score that would be achieved on this test by all students at Penang High School. The principal will not let the researcher to randomly select students out of classes in a session, but he will allow her to interrupt a small number of classes for the purpose of giving the test to every student of the class. Thus, the researcher selects 5 classes at random from the 40 classes in a particular session. The test is given to each student of the sampled classes, and the result are shown below. It is also known that there are 850 students enrolled in Penang High School.

Class i	Number of students in class i	Average score for class i
1	21	13
2	18	14
3	22	17
4	32	11
5	17	15

(a) State the type of sampling design used.
(b) State the primary sampling unit.
(c) Estimate the average score that would be achieved and place a bound on the error of estimation.
(d) How many classes should the researcher select for her sample if she wants the bound on the error of estimation to be less than 2?

[12 marks]

5. Seorang penyelidik dari USM telah membangunkan ujian untuk mengukur tahap kesedaran mengenai peristiwa-peristiwa semasa. Beliau ingin menganggar nilai min yang akan dicapai dalam ujian tersebut oleh semua pelajar di Sekolah Tinggi Pulau Pinang. Pengetua tidak akan membenarkan penyelidik memilih pelajar-pelajar secara rawak dari kelas dalam sesi tersebut, tetapi beliau akan membenarkan penyelidik mengganggu sebilangan kecil dari kelas-kelas yang bertujuan untuk memberi ujian kepada setiap pelajar dari kelas yang terpilih. Penyelidik memilih 5 kelas secara rawak daripada 40 kelas dalam sesi tertentu. Ujian itu diberikan kepada setiap pelajar dari kelas yang disampel, dan hasilnya adalah seperti di bawah. Juga diketahui terdapat seramai 850 pelajar yang mendaftar di Sekolah Tinggi Pulau Pinang.

Kelas i	*Bilangan pelajar dalam kelas i*	*Min skor untuk kelas i*
1	*21*	*13*
2	*18*	*14*
3	*22*	*17*
4	*32*	*11*
5	*17*	*15*

(a) Nyatakan jenis reka bentuk pensampelan yang digunakan.
(b) Nyatakan unit pensampelan utama.
(c) Anggarkan min skor yang akan dicapai dan tentukan batas ralat penganggaran.
(d) Berapa banyakkah kelas penyelidik perlu memilih untuk sampel beliau jika dia ingin batas ralat penganggaran kurang daripada 2?

[12 markah]

6. Ecologists wish to estimate the number of species of breeding birds in a population of 38 recreational area that had been developed with as little disturbance as possible to natural vegetation. They took a random sample of 15 recreational areas with a sampling of 2-hectare plot. On each plot they recorded the percentage of canopy diversity and the number of species of breeding birds. The following data are obtained:

Canopy Diversity	8	39	45	10	18	21	0	37	24	28	29	12	20	44	32
Species of Breeding Birds	9	13	10	8	9	10	4	10	11	12	8	7	8	14	11

$\sum x^2 = 11549; \qquad \sum y^2 = 1470; \qquad \sum xy = 3895$

The mean canopy diversity on all 38 areas was 26 percent. Assume the 2-hectare plot captures all species of breeding birds in each area.

(a) Identify the sampling unit, the variable of interest and any auxillary information associated with the unit.
(b) Estimate the mean number of species per area in three different methods:
  (i) simple mean
  (ii) ratio estimate
  (iii) regression estimate
(c) Calculate the standard error for each method and compare.
(d) Based from the answer (c), which method appears to be the most appropriate?

[25 marks]

6. Para ahli ekologi ingin menganggar bilangan spesis burung pembiakan dalam populasi dari 38 kawasan rekreasi yang telah dibangunkan dengan sedikit gangguan yang mungkin untuk tumbuh-tumbuhan semula jadi. Mereka mengambil sampel rawak sebanyak 15 kawasan rekreasi dengan pensampelan plot seluas 2 hektar. Pada setiap plot mereka mencatatkan peratusan kepelbagaian kanopi dan bilangan spesis burung pembiakan. Data yang berikut diperolehi:

Kepelbagaian kanopi	8	39	45	10	18	21	0	37	24	28	29	12	20	44	32
Spesis burung pembiakan	9	13	10	8	9	10	4	10	11	12	8	7	8	14	11

$\sum x^2 = 11549; \qquad \sum y^2 = 1470; \qquad \sum xy = 3895$

Min kepelbagaian kanopi pada semua 38 kawasan adalah 26 peratus. Andaikan plot 2 hektar dapat merangkupi semua spesis burung pembiakan di setiap kawasan.

(a) Camkan unit pensampelan, pembolehubah kepentingan, dan apa-apa maklumat tambahan berkaitan dengan unit.

(b) Anggarkan min bilangan spesis bagi setiap kawasan dengan tiga kaedah yang berbeza:
- *(i) min mudah*
- *(ii) anggaran nisbah*
- *(iii) anggaran regresi*

(c) Kirakan ralat piawai bagi setiap kaedah dan bandingkan.

(d) Berdasarkan jawapan dari (c), kaedah manakah yang paling sesuai?

[25 markah]

7. Consider a small population with periodicity with N = 12 with the values $y_1 = 1, y_2 = 2, y_3 = 3, y_4 = 1, y_5 = 2, y_6 = 3, y_7 = 1, y_8 = 2, y_9 = 3, y_{10} = 1, y_{11} = 2, y_{12} = 3$. Consider all systematic samples of size n = 4.

(a) List all possible samples.

(b) Show that $\bar{y}_{sys}$ is unbiased

(c) Determine the intra class correlation coefficient, ρ_w.

(d) Find $Var(\bar{y}_{sys})$

(e) Verify that it would be much better to use a simple random sample of size 4.

[20 marks]

7. Pertimbangkan satu populasi kecil dengan kekalaan N = 12 dengan nilai $y_1 = 1, y_2 = 2, y_3 = 3, y_4 = 1, y_5 = 2, y_6 = 3, y_7 = 1, y_8 = 2, y_9 = 3, y_{10} = 1, y_{11} = 2, y_{12} = 3$. *Pertimbangkan semua sampel sistematik bersaiz n = 4.*

(a) Senaraikan semua sampel yang mungkin.

(b) Tunjukkan bahawa $\bar{y}_{sys}$ *adalah saksama.*

(c) Tentukan pekali korelasi kelas intra, ρ_w.

(d) Cari $Var(\bar{y}_{sys})$

(e) Tentusahkan bahawa adalah lebih baik untuk menggunakan sampel rawak ringkas bersaiz 4.

[20 markah]

8. (a) Suppose in a two-stage cluster sample that all population cluster sizes are equal ($M_i = M$ for all *i*) and that all sample sizes for the cluster are equal ($m_i = m$ for all *i*). Show that $\hat{\tau}_{unb} = \hat{\tau}_r$, and, hence, $\hat{\mu}_{unb} = \hat{\mu}_r$.

(b) Show that if Neyman allocation is used, the estimated variance of the population average, $\bar{y}_{st}$ is:

$$\hat{V}_{opt}(\bar{y}_{st}) = \frac{\sum(W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}$$

[15 marks]

8. *(a) Katakan dalam sampel berkelompok 2-tahap, semua saiz populasi kelompok adalah sama* ($M_i = M$ *untuk semua i*) *dan semua saiz sampel untuk kelompok adalah sama* ($m_i = m$ *untuk semua i). Tunjukkan bahawa* $\hat{\tau}_{unb} = \hat{\tau}_r$ *, dan , seterusnya,* $\hat{\mu}_{unb} = \hat{\mu}_r$.

(b) Tunjukkan jika peruntukkan Neyman digunakan, anggaran varians bagi min populasi, $\bar{y}_{st}$ *adalah:*

$$\hat{V}_{opt}(\bar{y}_{st}) = \frac{\sum(W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}$$

[15 markah]

9. An inspector samples cans from a truckload of canned creamed corn to estimate the total number of worm fragments in the truckload. The truck has 580 boxes; each box contains 24 cans. The inspector samples 12 boxes at random, and subsamples 3 cans randomly from each selected box.

Box	1	2	3	4	5	6	7	8	9	10	11	12
Can 1	1	4	0	3	4	0	5	3	7	3	4	0
Can 2	5	2	1	6	9	7	5	0	3	1	7	0
Can 3	7	4	2	6	8	3	1	2	5	4	9	0

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Factor	149.639	11	13.604	3.004	.012
Error	108.667	24	4.528		
Total	258.306	35			

(a) State the type of sampling design used.
(b) State the primary sampling unit and secondary sampling unit.
(c) Estimate the total number of worm fragments and place a bound on the error of estimation. Explain your results.

[20 marks]

9. *Seorang inspector mensampelkan tin daripada lori yang dimuatkan dengan tin jagung berkrim untuk menganggarkan jumlah serpihan cacing dalam lori. Lori tersebut mempunyai 580 kotak; setiap kotak mengandungi 24 tin. Inspektor mensampelkan 12 kotak secara rawak, dan subsampel 3 tin secara rawak daripada setiap kotak yang terpilih.*

Kotak	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Tin 1	*1*	*4*	*0*	*3*	*4*	*0*	*5*	*3*	*7*	*3*	*4*	*0*
Tin 2	*5*	*2*	*1*	*6*	*9*	*7*	*5*	*0*	*3*	*1*	*7*	*0*
Tin 3	*7*	*4*	*2*	*6*	*8*	*3*	*1*	*2*	*5*	*4*	*9*	*0*

ANOVA

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Factor	*149.639*	*11*	*13.604*	*3.004*	*.012*
Error	*108.667*	*24*	*4.528*		
Total	*258.306*	*35*			

(a) Nyatakan jenis reka bentuk pensampelam yang digunakan.
(b) Nyatakan unit pensampelan utama dan unit pensampelan kedua.
(c) Anggarkan jumlah serpihan cacing dan tentukan batas ralat penganggaran. Terangkan jawapan anda.

[20 markah]

Appendix

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

…12/-

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

…13/-

Sample	Sample variance
$\hat{p} = \dfrac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}$	$\left(\dfrac{N-n}{N}\right)\left(\dfrac{1}{n\bar{M}^2}\right)S_r^2 + \dfrac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2\left(\dfrac{M_i - m_i}{M_i}\right)\left(\dfrac{\hat{p}_i\hat{q}_i}{m_i - 1}\right)$ with $S_r^2 = \dfrac{\sum_{i=1}^{n_i}\left(M_i\hat{p}_i - \hat{p}M_i\right)^2}{n-1}$

Sample Size
$n = \dfrac{N\sigma^2}{(N-1)D + \sigma^2}$; $D = \dfrac{B^2}{4}$; $D = \dfrac{B^2}{4N^2}$
$n = \dfrac{\sum_{i=1}^{L} \dfrac{N_i\sigma_i^2}{w_i}}{N^2 D + \sum_{i=1}^{L} N_i\sigma_i^2}$; $w_i = \dfrac{n_i}{n}$
$n = \dfrac{\left(\sum_{k=1}^{L} N_k\sigma_k / \sqrt{C_k}\right)\left(\sum_{i=1}^{L} N_i\sigma_i\sqrt{C_i}\right)}{N^2 D + \sum_{i=1}^{L} N_i\sigma_i^2}$, $n = \dfrac{(C - C_o)\sum_{i=1}^{L} N_i\sigma_i / \sqrt{C_i}}{\sum_{i=1}^{L} N_i\sigma_i\sqrt{C_i}}$ $n_i = \dfrac{n\, N_i\sigma_i / \sqrt{C_i}}{\sum_{i=1}^{L} N_i\sigma_i / \sqrt{C_i}}$ 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

…14/-

Sample Size

$$n = \frac{\sum_{i=1}^{L} N^2{}_i p_i q_i / a_i}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i} \quad ; \quad n_i = n \left(\frac{N_i \sqrt{p_i q_i / c_i}}{\sum_{i=1}^{L} N_i \sqrt{p_i q_i / c_i}} \right)$$

$$n = \frac{N\sigma^2}{ND + \sigma^2} \quad ; \quad D = \frac{B^2 \mu x^2}{4} \quad ; \quad D = \frac{B^2}{4} \quad ; \quad D = \frac{B^2}{4N^2}$$

$$n = \frac{N\sigma_r{}^2}{ND + \sigma_r{}^2} \quad ; \quad D = \frac{B^2 \left(\bar{M}\right)^2}{4} \quad ; \quad D = \frac{B^2}{4N^2}$$

Intra class correlation coefficient

$$\rho_w = \frac{2}{\sigma^2 nk(n-1)} \sum_{i=1}^{k} \sum_{j<u}^{n} \left(y_{ij} - \mu\right)\left(y_{iu} - \mu\right)$$

-ooo000ooo-