

# The University of Melbourne

## Semester 2 Assessment, 2008

**Department:** Mathematics and Statistics

**Subject Title:** 620-374 Sampling and Forecasting

**Exam Duration:** Three hours

**Reading Time:** 15 minutes

**This paper has eight (8) pages** (including cover sheet and formula sheets).

**Authorised material:**

Hand-held electronic calculators may be used, provided all memories and programs are cleared.

**Instructions to Invigilators:**

Statistical tables will be provided.

**Instructions to Students:**

All questions may be attempted.

The number of marks for each question is indicated; this reflects the relative weighting of the questions.

The total number of marks available in this examination is 90.

Working and/or reasoning must be shown.

Formula for sample survey questions appear after the exam questions

**This paper is to be lodged with the Baillieu Library.**

1. The following data show the stratification of all the farms in a county by farm size, and the average acres of corn (maize) per farm in each stratum.

Farm Size (acres)	Number of Farms	Average Corn Acres	Standard Deviation		
	$N_h$	$\mu_Y$	$S_h$	$N_h S_h$	$N_h S_h^2$
0–40	394	5.4	8.3	3,270	27,100
41–80	461	16.3	13.3	6,130	81,500
81–120	391	24.3	15.1	5,900	89,200
121–160	334	34.5	19.8	6,610	130,900
161–200	169	42.1	24.5	4,140	101,400
201–240	113	50.1	26.0	2,940	76,400
241–	148	63.8	35.2	5,210	183,400

For a stratified sample of 100 farms, compute the sample sizes in each stratum under

- proportional allocation;
- optimum (Neyman) allocation.

By calculating variances, show that the Neyman allocation is better.

**[2 + 2 + 4 = 8 marks]**

2. (a) Explain the difference between stratified sampling and cluster sampling, specifying the desired qualities of strata and of clusters.
- (b) An industry is considering a revision of its retirement policy, and wants to estimate the proportion of employees that favour the new policy. The industry consists of 87 separate plants located throughout the country. Since results must be obtained quickly and with little cost, the industry decides to use cluster sampling, with each plant as a cluster. A simple random sample of 15 plants is selected, and the opinions of the employees in these plants are obtained by questionnaire. The results are given below.

Plant	Number of employees	Number in favour	Plant	Number of employees	Number in favour
	$X$	$Y$		$X$	$Y$
1	51	42	9	73	54
2	62	53	10	61	45
3	49	40	11	58	51
4	73	45	12	52	29
5	101	63	13	65	46
6	48	31	14	49	37
7	65	38	15	55	42
8	49	30			

For these data

$$\sum_{i=1}^{15} x_i = 911, \sum_{i=1}^{15} x_i^2 = 58,075, \sum_{i=1}^{15} y_i = 646,$$
$$\sum_{i=1}^{15} y_i^2 = 29,104 \text{ and } \sum_{i=1}^{15} x_i y_i = 40,730.$$

Estimate the proportion of employees in the industry who favour the new retirement policy, and estimate the standard deviation of the estimate.

**[3 + 5 = 8 marks]**

3. In 2007, a survey was conducted to estimate the average number of televisions per household in metropolitan Melbourne. At that time Melbourne had (approximately) 1,471,000 households and a population of 3,806,000. A sample of 500 households was selected at random, and the following variables measured

$X$  the number of occupants;

$Y$  the number of televisions.

From the sample data we have

$$\bar{x} = 2.8, \bar{y} = 2.4, s_x^2 = 1.9, s_y^2 = 2.6, s_{x,y} = 1.3$$

- (a) Estimate the average number of televisions per household using
- Just  $Y$ ;
  - $X$  and  $Y$  using a ratio estimate;
  - $X$  and  $Y$  using a regression estimate.
- (b) Estimate the variance of each estimator above, and thus decide which is the most accurate.
- (c) Let  $T_i$  be the number of televisions owned by person  $i$ . Explain how the sample described above can be viewed as a cluster-sample with respect to the variable  $T$ .

**[6 + 6 + 2 = 14 marks]**

4. The following estimates for the level  $L$ , slope  $B$  and seasonal effect  $S$  were obtained using the Holt-Winters method, for a time-series with a linear trend and additive seasonality of period 4.

time	$L$	$B$	$S$
$t - 4$	15.7	2.2	-0.9
$t - 3$	18.2	2.4	-0.5
$t - 2$	19.7	1.9	0.8
$t - 1$	20.7	1.4	0.6

The smoothing parameters used were  $\alpha = 0.3$  for the level,  $\beta = 0.5$  for the slope and  $\gamma = 0.8$  for the seasonal effect.

- (a) Update these values using the new observation  $X_t = 20.9$ .
- (b) Give forecasts for times  $t + 1$  and  $t + 2$ .
- (c) Describe how might you choose values for  $\alpha$ ,  $\beta$  and  $\gamma$  in practice.

**[6 + 3 + 2 = 11 marks]**

5. For each of the examples below, find the autocorrelation function of  $X_t$ .

Here  $\{Z_t\}_{t=-\infty}^{\infty}$  is an i.i.d. sequence with mean 0 and variance  $\sigma^2$ .

(a)

$$X_t = \mu + Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \beta_3 Z_{t-3} + \beta_4 Z_{t-4}.$$

(b)

$$X_t = X_{t-1} - \frac{3}{16} X_{t-2} + Z_t.$$

For the following example, find the autocorrelation at lag 1 (assuming  $|\alpha| < 1$ )

(c)

$$X_t = \alpha X_{t-1} + Z_t + \beta Z_{t-1}.$$

Hint: express the process as an MA( $\infty$ ) process.

**[2 + 4 + 4 = 10 marks]**

6. Let  $\{Z_t\}_{t=-\infty}^{\infty}$  be an i.i.d. sequence with mean 0 and variance  $\sigma^2$ , and define  $X_t$  by

$$X_t = X_{t-4} + \alpha(X_{t-1} - X_{t-5}) + Z_t + \beta Z_{t-4}.$$

(a) Express  $X_t$  using the backshift operator  $B$ .

Hence identify values of  $p$ ,  $d$ ,  $q$ ,  $P$ ,  $D$ ,  $Q$  and  $s$  such that  $X_t$  is a SARIMA( $p, d, q$ )( $P, D, Q$ ) $_s$  model.

Now suppose that you observe  $X_1, \dots, X_6$ .

(b) Express each the residuals  $R_1, \dots, R_6$  in terms of  $X_1, \dots, X_6$ .

(c) Express the one-step ahead forecast  $\hat{X}_7$  in terms of  $X_1, \dots, X_6$  and  $R_1, \dots, R_6$ .

(d) Express the two-step ahead forecast  $\hat{X}_8$  in terms of  $X_1, \dots, X_6$  only.

**[3 + 2 + 1 + 3 = 9 marks]**

7. Let  $X_1, \dots, X_n$  be an i.i.d. sample from the distribution  $F$ .

(a) Define the empirical distribution function  $\hat{F}$  as a sum of indicator functions.

- (b) Use the strong law of large numbers (SLLN) to show that  $\hat{F}(x)$  converges to  $F(x)$  as  $n \rightarrow \infty$ .
- (c) Use the central limit theorem (CLT) to obtain an approximate distribution for  $\hat{F}(x)$ , for large  $n$ .

Let  $X_1^*, \dots, X_n^*$  be an i.i.d. sample from  $\hat{F}$ .

- (d) Explain how to represent the  $X_i^*$  using a multinomial distribution. You may assume that the  $X_i$  are distinct.

[1 + 1 + 2 + 2 = 6 marks]

8. Let  $X$  be a bounded random variable, with continuous distribution function  $F$ . We define the *range*  $R$  of  $X$  to be largest possible value minus the smallest possible value, that is

$$R = \inf\{x : F(x) = 1\} - \sup\{x : F(x) = 0\}.$$

Let  $X_1, \dots, X_n$  be an i.i.d. sample from  $F$ , with observed values  $x_1, \dots, x_n$ .

- (a) What is  $\hat{R}$ , the plug-in estimator of  $R$ ?
- (b) Is  $\hat{R}$  biased, and why?
- (c) What is the bootstrap estimate of bias,  $\widehat{\text{bias}}_B \hat{R}$ ?
- (d) Describe two sources of error in  $\widehat{\text{bias}}_B \hat{R}$ , and explain how one of them can be measured.
- (e) Let  $\bar{R} = \hat{R} - \widehat{\text{bias}}_B \hat{R}$  be the bias-corrected estimate of  $R$ . Derive an estimate of the mean square error (MSE) of  $\bar{R}$ .
- (f) Show that for large bootstrap samples, we would estimate that the MSE of  $\bar{R}$  is smaller than the MSE of  $\hat{R}$  when

$$\widehat{\text{bias}}_B \hat{R} > \sqrt{3} \widehat{\text{sd}}_B \hat{R},$$

where  $\widehat{\text{sd}}_B \hat{R}$  is the usual bootstrap estimate of the standard deviation of  $\hat{R}$ .

- (g) What is the *controlled* estimate of bias,  $\overline{\text{bias}}_B \hat{R}$ ? In what situation would  $\overline{\text{bias}}_B \hat{R}$  and  $\widehat{\text{bias}}_B \hat{R}$  differ?  
(Do not derive the estimator, just state it.)

[2 + 2 + 1 + 3 + 4 + 3 + 4 = 19 marks]

9. Let  $\theta$  be a parameter of some distribution  $F$ , and  $\hat{\theta} = S(\mathbf{X})$  an estimator for  $\theta$ , using the i.i.d. sample  $\mathbf{X} = (X_1, \dots, X_n)$  taken from  $F$ .

Explain how to construct a bootstrap- $t$  confidence interval (CI) for  $\theta$ .

[5 marks]

# Sample Survey Formulae

## Simple Random Sampling (Without Replacement)

Population:  $\{Y_1, \dots, Y_N\}$ ,  $\mu_Y = \sum_{i=1}^N Y_i/N$ ,  $\tau_Y = \sum_{i=1}^N Y_i$ ,  $\sigma_Y^2 = \sum_{i=1}^N (Y_i - \mu_Y)^2/N$ ,  $S_Y^2 = N\sigma_Y^2/(N-1)$ .

Sample:  $\{y_1, \dots, y_n\}$ ,  $\bar{y} = \sum_{i=1}^n y_i/n$ ,  $s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$ .

### Estimating the mean

Estimator:  $\hat{\mu}_Y = \bar{y}$ .

Variance:  $\text{Var}(\hat{\mu}_Y) = S_Y^2(1-f)/n$  where  $f = n/N$ .

Variance estimator:  $\hat{\text{Var}}(\hat{\mu}_Y) = s_y^2(1-f)/n$ .

Sample size:  $n = N\sigma_Y^2/((N-1)D^2 + \sigma_Y^2)$ , where  $D^2 = \text{Var} \hat{\mu}_Y$ .

## Stratified Random Sampling

Population:  $N_h, \mu_h, \sigma_h^2$  and  $S_h^2$  are as above but for stratum  $h$ ;  $W_h = N_h/N$ .  $\mu_Y, N$  are as before and refer to the whole population.

Sample:  $n_h, \bar{y}_h, s_h^2$  and  $f_h = n_h/N_h$  are as above but for the subsample from stratum  $h$ .  $n$  is the whole sample size.

### Estimating the mean

Estimator:  $\hat{\mu}_{st} = \sum_{h=1}^L W_h \bar{y}_h$ .

Variance:  $\text{Var}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 S_h^2(1-f_h)/n_h$ .

Variance estimator:  $\hat{\text{Var}}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 s_h^2(1-f_h)/n_h$ .

Sample size:  $n = \sum_{h=1}^L (N_h^2 S_h^2 / w_h) / (N^2 V + \sum_{h=1}^L N_h S_h^2)$  where  $V = \text{Var}(\hat{\mu}_{st})$ ,  $w_h = n_h/n$ .

### Proportional allocation

Put  $n_h/n = N_h/N$  then

Variance:  $\text{Var}_{prop}(\hat{\mu}_{st}) = ((1-f)/n) \sum_{h=1}^L W_h S_h^2$

## Optimal allocation

For cost function  $C = c_0 + \sum c_h n_h$ , the cost  $C$  is minimised for a specified variance  $\text{Var}(\hat{\mu}_{st})$ , and the variance  $\text{Var}(\hat{\mu}_{st})$  is minimised for a fixed cost  $C$ , if

$$n_h \propto W_h S_h / \sqrt{c_h} \quad \text{that is} \quad \frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum (W_h S_h / \sqrt{c_h})}.$$

Thus,

$$n = \frac{(C - c_0) \sum (N_h S_h / \sqrt{c_h})}{\sum (N_h S_h \sqrt{c_h})}, \quad \text{if cost } C \text{ is fixed.}$$

$$n = \frac{(\sum W_h S_h / \sqrt{c_h}) (\sum W_h S_h \sqrt{c_h})}{V + (1/N) \sum W_h S_h^2}, \quad \text{if } V = \text{Var}(\hat{\mu}_{st}) \text{ is fixed.}$$

## Neyman allocation

Optimal allocation when  $c_h = c$  for all  $h$ .

$$\frac{n_h}{n} = \frac{W_h S_h}{\sum (W_h S_h)} = \frac{N_h S_h}{\sum (N_h S_h)}, \quad \text{Var}_{opt}(\hat{\mu}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}.$$

## Post-stratification

Let  $m_h$  be the number of units in stratum  $h$ .

$$\text{Var}_p(\hat{\mu}_{st} | m_h) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{m_h} (1 - f_h) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{m_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

$$\text{Var}_p(\hat{\mu}_{st}) \approx \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^L (1 - W_h) S_h^2.$$

## Cluster sampling

Population:  $\mu_Y$  = population mean per element,  $M$  = number of elements in population;

Clusters:  $N$  = number of clusters in population,  $m_h$  = size of cluster  $h$ ,  $\mu_h$  = mean of cluster  $h$ ,  $\tau_h$  = total of cluster  $h$ ,  $\sigma_h^2$  = variance of cluster  $h$ .  $\sigma_b^2$  = between cluster variance.

Sample:  $n$  = number of clusters in sample,  $\bar{y}_i$  = mean of cluster  $i$  in sample,  $t_i$  = total of cluster  $i$  in sample.

## One-stage cluster sampling with equal-sized clusters

Estimator:  $\hat{\mu}_{cl} = \sum_{h=1}^n \bar{y}_h / n$ .

Variance:  $\text{Var}(\hat{\mu}_{cl}) = S_b^2 (1 - f) / n$  where  $f = n/N$ ,  $S_b^2 = N \sigma_b^2 / (N - 1)$ .

Variance estimator:  $\hat{\text{Var}}(\hat{\mu}_{cl}) = s_b^2 (1 - f) / n$  where  $s_b^2$  is the sample variance of the selected cluster means

## One-stage cluster sampling with unequal-sized clusters

Estimator:  $\hat{\mu}_{clr} = \bar{t}/\bar{m}$  where  $\bar{t} = (\sum_i t_i)/n$ ,  $\bar{m} = (\sum_i m_i)/n$  (sample averages).

Variance:  $\text{Var}(\hat{\mu}_{clr}) \approx (N/M)^2((1-f)/n) \sum_{h=1}^N (\tau_h - \mu_Y m_h)^2 / (N-1)$ .

Variance estimator:  $\hat{\text{Var}}(\hat{\mu}_{clr}) = (1/\bar{m})^2((1-f)/n) \sum_{i=1}^n (t_i - \hat{\mu}_{clr} m_i)^2 / (n-1)$

## Ratio and regression estimators

Population:  $\mu_X$  is known,  $\mu_Y$  unknown,  $R = \mu_Y/\mu_X$ .

Sample: SRS with data  $\{(x_i, y_i); i = 1, 2, \dots, n\}$ ,  $r = \bar{y}/\bar{x}$ .

### Ratio estimator

Estimator:  $\hat{\mu}_{ratio} = r\mu_X$ .

Variance:  $\text{Var}(\hat{\mu}_{ratio}) \approx ((1-f)/n) \sum_{\ell=1}^N (Y_\ell - RX_\ell)^2 / (N-1)$

Variance estimator:  $\hat{\text{Var}}(\hat{\mu}_{ratio}) = ((1-f)/n) \sum_{i=1}^n (y_i - rx_i)^2 / (n-1)$   
 $= ((1-f)/n)(s_y^2 - 2r\hat{\rho}s_x s_y + r^2 s_x^2)$  where  $\hat{\rho} = s_{xy}/(s_x s_y)$

### Regression estimator

Estimator:  $\hat{\mu}_{lr} = \bar{y} + b(\mu_X - \bar{x})$ .

- When  $b = b_0$  is pre-assigned:

Variance:  $\text{Var}(\hat{\mu}_{lr}) = ((1-f)/n)(S_Y^2 - 2b_0 S_{XY} + b_0^2 S_X^2)$

Variance estimator:  $\hat{\text{Var}}(\hat{\mu}_{lr}) = ((1-f)/n)(s_y^2 - 2b_0 s_{xy} + b_0^2 s_x^2)$ .

The best value to assign  $b$  is  $\beta = S_{XY}/S_X^2$ , which minimises  $\text{Var}(\hat{\mu}_{lr})$ .

$$\min_b [\text{Var}(\hat{\mu}_{lr})] = \frac{(1-f)}{n} \left( S_Y^2 - \frac{S_{XY}^2}{S_X^2} \right) = \frac{(1-f)}{n} S_Y^2 (1 - \rho^2).$$

- When  $b$  is estimated from the sample:

Estimator for  $b$ :  $\hat{\beta} = s_{xy}/s_x^2$ .

Variance:  $\text{Var}[\hat{\mu}_{lr}(\hat{\beta})] \approx \text{Var}[\hat{\mu}_{lr}(\beta)] = ((1-f)/n) (S_Y^2 - S_{XY}^2/S_X^2)$

Variance estimator:  $\hat{\text{Var}}[\hat{\mu}_{lr}(\hat{\beta})] = ((1-f)/n) ((n-1)/(n-2)) (s_y^2 - s_{xy}^2/s_x^2)$   
 $= ((1-f)/n) ((n-1)/(n-2)) s_y^2 (1 - \hat{\rho}^2)$  where  $\hat{\rho} = s_{xy}/(s_x s_y)$ .