

# The University of Melbourne

## Semester 2 Assessment, 2007

**Department:** Mathematics and Statistics

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

**Exam Duration:** Three hours

**Reading Time:** 15 minutes

**This paper has nine (9) 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. A finite population is divided into strata, of size  $N_h$ ,  $h = 1, \dots, L$ . We wish to estimate the population mean  $\mu$  of some variable of interest  $Y$ . A sample of size  $n_h$  is taken from stratum  $h$ , for each  $h$ .

(a) Let  $\bar{y}_h$  be the sample mean from stratum  $h$ . Let  $I_{h,i} = 1$  if item  $i$  from stratum  $h$  is included in the sample, and 0 otherwise. Express  $\bar{y}_h$  in terms of the  $I_{h,i}$ , and hence show that  $\mathbb{E}\bar{y}_h = \mu_h$ , the mean of  $Y$  over stratum  $h$ .

(b) Let  $N = \sum_{h=1}^L N_h$  and  $n = \sum_{h=1}^L n_h$ . Which of the following two estimators do you prefer, and why? Prove any claims you make about these estimators.

(i)  $\hat{\mu}_1 = \sum_{h=1}^L (N_h/N) \bar{y}_h$

(ii)  $\hat{\mu}_2 = \sum_{h=1}^L (n_h/n) \bar{y}_h$ .

(c) Suppose that the cost of taking a sample of size  $n_h$  from stratum  $h$  is  $c_h n_h^2$ , for some positive constant  $c_h$ . Show that to minimise the variance of  $\hat{\mu}_1$ , subject to a fixed total cost, the sample sizes should satisfy

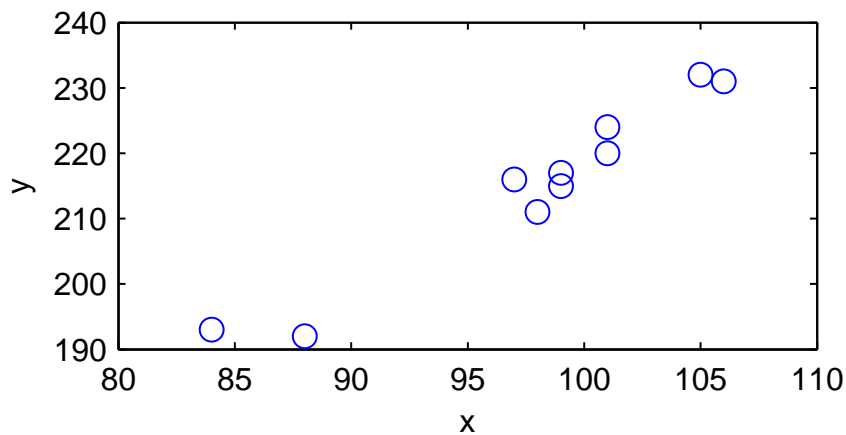
$$n_h \propto \left( \frac{(N_h/N)^2 S_h^2}{c_h} \right)^{1/3}$$

where  $S_h^2 = N_h \sigma_h^2 / (N_h - 1)$  and  $\sigma_h^2$  is the variance of stratum  $h$ .

[3 + 3 + 6 = 12 marks]

2. A freight loading facility deals with 100 trains in a year. The length  $X_i$  (total number of freight cars) is known for each train, with population mean  $\mu_X = 100$ , but the total weight of each train  $Y_i$  is only known for a sample of 10 trains:

$x_j$	101	106	99	84	88	98	101	99	105	97
$y_j$	220	231	215	193	192	211	224	217	232	216



Give an estimate of  $\tau_Y$ , the total weight of trains over the year, together with a 95% confidence interval. You may find the following sample statistics useful (all given to one decimal place accuracy):

$$\bar{x} = 97.8, \bar{y} = 215.1, s_x^2 = 47.7, s_y^2 = 187.2, s_{xy} = 92.0.$$

Your choice of estimator is important: briefly justify the choice you make.

[6 marks]

3. A population is made up of 20 separate clusters of size 5. We wish to estimate the mean  $\mu_Y$  of some variable of interest  $Y$ . It costs \$2 to sample a whole cluster, or \$1 to sample an individual, and you have a budget of \$20.

Let  $S_h^2 = N_h \sigma_h^2 / (N_h - 1)$  where  $\sigma_h^2$  is the variance of  $Y$  over cluster  $h$ ,  $\bar{S}^2 = \sum_{h=1}^{20} S_h^2 / 20$ , and let  $S^2 = N \sigma^2 / (N - 1)$  where  $\sigma^2$  is the population variance of  $Y$ . For what values of  $\bar{S}^2$  and  $S^2$  would you prefer a cluster sample to a simple random sample?

[8 marks]

4. A population is made up of 10 clusters each of size 100. We wish to estimate the mean  $\mu$  of some variable of interest  $Y$ .

Consider the following two-stage sampling procedure: choose 5 clusters at random, then from each chosen cluster take a simple random sample of size 10. Let  $k(i) \in \{1, \dots, 10\}$  be the index of the  $i$ -th sample cluster and let  $\hat{\mu}_j$  be the sample mean from the  $j$ -th cluster, then put

$$\hat{\mu} = \sum_{i=1}^5 \hat{\mu}_{k(i)} / 5.$$

- (a) Prove that  $\hat{\mu}$  is unbiased.  
 (b) Explain how you would use bootstrapping to estimate  $\text{Var } \hat{\mu}$ . Take care to explain how you resample.

[4 + 4 = 8 marks]

5. Each of the following is a SARIMA( $p, d, q$ )( $P, D, Q$ ) $_s$  process (after appropriate differencing). Identify  $p, d, q, P, D, Q$  and  $s$  in each case.

Here  $Z_t$  is a sequence of i.i.d. random variables with mean zero and finite variance and  $B$  is the shift operator.

(i)

$$(1 - \alpha B)(1 - B^{12})X_t = (1 - \beta B^{12})Z_t.$$

(ii)

$$X_t = X_{t-1} + X_{t-4} - X_{t-5} + Z_t + \beta_1 Z_{t-1} + \beta_4 Z_{t-4} + \beta_1 \beta_4 Z_{t-5}.$$

(iii)

$$X_t = (\mu + at)S_t + Z_t, \text{ where } S_t \text{ is periodic with period 12.}$$

[6 marks]

6. Which stationary invertible ARMA processes  $\phi(B)X_t = \theta(B)Z_t$  have the following autocovariances? Specify  $\phi$  and  $\theta$  in each case.

(a)

$$\gamma(k) \begin{array}{l|l} k & 0 \quad 1 \quad \geq 2 \\ \hline & 1.25 \quad 0.5 \quad 0 \end{array}$$

(b)

$$\gamma(k) \begin{array}{l|l} k & 0 \quad 1 \quad \geq 2 \\ \hline & 8 \quad 2 \quad (5/6)^k + 7(1/6)^k \end{array}$$

[7 + 5 = 12 marks]

7. Complete the table of residuals and forecasts below, for the process

$$X_t = 0.5X_{t-1} + Z_t - Z_{t-1}.$$

Here  $X_t$  is the observed process (observed at times  $t = 1, \dots, 4$ ), and  $Z_t$  a sequence of i.i.d. mean zero finite variance random variables. The residual at time  $t$  is denoted  $R_t$ .

$t$	$X_t$	$R_t$
0	0	0
1	8	?
2	-12	?
3	-6	?
4	21	?
5	?	0

[5 marks]

8. Consider the following time-series:

$t$	1	2	3	4	5	6	7
$X_t$	1.4	1.3	1.3	1.6	2.9	3.2	3.2

(a) Estimate  $X_8$  using a moving average of length  $k = 3$ .

(b) Estimate  $X_8$  using a simple exponential smoother with parameter  $\alpha = 0.7$ .

- (c) Use the mean square error (MSE) to compare the accuracy of your estimates from (a) and (b). Which do you prefer?

[1 + 3 + 3 = 7 marks]

9. Let  $\underline{x} = (x_1, x_2, \dots, x_n)^T$  be a random sample size  $n$  from an unknown distribution  $F$ .

- (a) Define the empirical distribution function  $\widehat{F}$ .  
 (b) Define the plug-in estimate  $\widehat{\theta}$  for a parameter of interest  $\theta$ .  
 (c) Suppose that the distribution  $F$  is bounded. Derive the plug-in estimate for  $\max_F X$ , showing all your steps.  
 (d) Give two justifications for the use of plug-in estimators.

[1 + 1 + 3 + 3 = 8 marks]

10. Let  $\underline{x} = (x_1, x_2, \dots, x_n)^T$  be a random sample size  $n$  from an unknown distribution  $F$ . Let  $\widehat{\theta} = S(\underline{x})$  be an estimate of some parameter of interest  $\theta$ .

- (a) Define the ideal bootstrap estimate of  $se_F(\widehat{\theta})$ , the standard error of  $\widehat{\theta}$  when sampling from  $F$ .  
 (b) Consider the following resampling program:

```
library(bootstrap)
original.sample <- c(10.1, 12.3, 9.8, 7.4, 9.9, 12.1, 11.7, 5.3, 9.9, 7.8)
results <- bootstrap(original.sample, 250, max)
bootstrap.replicates <- results$thetastar
estimate <- sd(bootstrap.replicates)
```

- (i) Write down the algorithm the program applies to generate `estimate`.  
 (ii) Name and concisely explain the two sources of error which would arise in using `estimate` to estimate  $se_F(\widehat{\theta})$ . State how you might reduce each type of error.  
 (iii) Derive a theoretical expression for the probability that a bootstrap replicate generated by the program is less than 10.

[1 + 7 = 8 marks]

11. Let  $\underline{x} = (x_1, x_2, \dots, x_n)^T$  be a random sample size  $n$  from an unknown distribution  $F$ , and let  $\widehat{\theta}$  be the plug-in estimate of some parameter  $\theta = t(F)$ .

- (a) Define the bootstrap estimate of bias,  $\widehat{bias}_B \widehat{\theta}$ , and use it to define a bias corrected estimate  $\bar{\theta}$  of  $\widehat{\theta}$ .  
 (b) Derive an estimator of  $\text{Var} \bar{\theta}$  and thus an estimator of the mean square error (MSE) of  $\theta$

- (c) Show that for large bootstrap samples, we would estimate that the MSE of  $\bar{\theta}$  is less than the MSE of  $\hat{\theta}$  when

$$\widehat{bias}_B \hat{\theta} > \sqrt{3} \widehat{se}_B \hat{\theta},$$

where  $\widehat{se}_B \hat{\theta}$  is the usual bootstrap estimate of the standard error of  $\hat{\theta}$ .

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

END of EXAM

Three sheets of formulas follow

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

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