

The University of Melbourne

Semester 2 Assessment, 2006

Department: Mathematics and Statistics

Subject Title: 620-374 Sampling and Forecasting

Exam Duration: Three hours

Reading Time: 15 minutes

This paper has ten (10) 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 120.

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. Parkville has 3,900 households and 8,300 residents (7,100 adults and 1,200 children). 100 households were surveyed and data on the following variables was collected:

X : the number of occupants;

Y : the number of adult occupants;

Z : the number of cars owned by occupants.

The following sample data will be needed for the remainder of the question.

$$\begin{aligned}\bar{x} &= 3.1, & \bar{y} &= 1.8, & \bar{z} &= 1.6 \\ s_X^2 &= 4.0, & s_Y^2 &= 1.0, & s_Z^2 &= 1.2 \\ s_{XY} &= 1.4, & s_{XZ} &= 1.3, & s_{YZ} &= 1.0\end{aligned}$$

- (a) Estimate the average number of cars per household using
- Just Z ;
 - X and Z using a ratio estimate;
 - Y and Z using a ratio estimate.
- (b) Why do we not use $(\sum_{i=1}^{100} z_i/x_i)/100$ to estimate μ_Z/μ_X , where μ_A is the population mean of variable A .
- (c) By estimating variances or otherwise, determine which estimate in part (a) is likely to be most accurate.
- (d) Let W_i be the number of cars registered to the i -th adult in the population. Explain how the sample described above can be viewed as a cluster-sample with respect to the variable W .

Using a cluster-sample type estimator, give a 95% confidence interval for the total number of cars in Parkville.

[3+2+6+8=19 marks]

2. (a) Suppose that Australia has 10,000,000 people of working age, made up of 5,000,000 men and 5,000,000 women. From a pilot sample of 100 men and 100 women the sample mean and variance of the variable “annual income” was calculated (in units of \$1,000).

$$\text{men: } \bar{x}_1 = 50, s_1^2 = 400; \text{ women: } \bar{x}_2 = 40, s_2^2 = 900.$$

You are given funds to survey 2,000 people and the cost of surveying men and women is the same. How would you split your sample between men and women to best estimate the average annual income across the working population of Australia?

- (b) Suppose that a population can be split into L strata of sizes N_1, \dots, N_L . A sample of size n is taken from the population but sample points are not pre-assigned to strata.

Let m_h be the number of sample points in stratum h .

(i) Let $N = \sum_h N_h$ and $W_h = N_h/N$, then prove that

$$\mathbb{E} \frac{1}{m_h} \approx \frac{1}{nW_h} + \frac{1 - W_h}{n^2 W_h^2}.$$

(ii) Let $\hat{\mu}_{st} = \sum_h W_h \bar{x}_h$, where \bar{x}_h is the sample average from stratum h , and let S_h^2 be the S^2 from stratum h . Prove that

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

You may use the formula for the variance of a SRS estimator without proof.

[3+10=13 marks]

3. Using a quick visual inspection from the street, the value of each house in Carlton is estimated, and the results recorded as X_1, \dots, X_{2000} . 100 houses are then chosen at random and a more thorough valuation conducted, giving values Y_1, \dots, Y_{100} . (For simplicity we assumed that the sampled houses are numbered 1 to 100.)

Given that (in units of \$1,000)

$$\mu_X = 600, \bar{x} = 620, \bar{y} = 630, s_X^2 = 22,500, s_{XY} = 14,400 \text{ and } s_Y^2 = 25,600$$

give a 95% CI for the mean value of a house in Carlton, using the regression method with slope $b = 1$.

How can you justify choosing $b = 1$ a-priori?

[8 marks]

4. You observe a random sample size n from an unknown population F ie $F \rightarrow \underline{x} = (x_1, x_2, \dots, x_n)$.

- Define the empirical distribution function \hat{F}
- Define the plug-in estimate $\hat{\theta}$ for a parameter of interest θ
- Derive the plug-in estimate for $\sigma_F^2 = \text{Var}_F(X)$ showing all your steps (where (x_1, x_2, \dots, x_n) are independent observations on X)
- Indicate how the effectiveness of plug-in estimators can be justified by reference to two distinct principles of statistical estimation.
- Define the ideal bootstrap estimate of $se_F(\hat{\theta}) =$ the standard error of a statistic $\hat{\theta}$ when sampling from F .
- Consider the following resampling program:

```
library(bootstrap)
treatmentgrp<-c(94,197,16,38,99,141,23)
results<-bootstrap(treatmentgrp,250,median)
brep<-results$thetastar
Estimate<-sd(brep)
```

- (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 $se_F(\hat{\theta})$. 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 25.

[1+1+3+2+1+8=16 marks]

5. You observe a random sample size n from an unknown population F ie $F \rightarrow \underline{x} = (x_1, x_2, \dots, x_n)$. Consider using the jackknife estimate of standard error \widehat{se}_{jack} to estimate $se_F(\hat{\theta})$ where $\hat{\theta} = s(\underline{x})$ is a statistic of interest.

- (a) Define the i th jackknife sample $x_{(i)}$
- (b) Define the i th jackknife replicate $\hat{\theta}_{(i)}$
- (c) Note that

$$\widehat{se}_{jack} = \left[\frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}(\cdot))^2 \right]^{1/2}$$

where $\hat{\theta}(\cdot) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}$.

Showing all steps prove that for $\hat{\theta} = \bar{x}$, $\widehat{se}_{jack} = \frac{s}{\sqrt{n}}$ where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

- (d) Consider the jackknife estimate of bias $\widehat{bias}_{jack} = (n-1)[\hat{\theta}(\cdot) - \hat{\theta}]$. Explain why a scaling factor of $(n-1)$ is used in its definition.

[1+1+5+2=9 marks]

6. You observe a random sample size n from an unknown population F ie $F \rightarrow \underline{x} = (x_1, x_2, \dots, x_n)$. You decide to estimate a parameter $\theta = t(F)$ using an estimator $\hat{\theta} = S(\underline{x})$.

- (a) State the definitions of $bias_F(\hat{\theta})$ and its ideal bootstrap estimate $bias_{\hat{F}}(\hat{\theta})$.
- (b) For the special case of $\theta = t(F) = \mu_F$ and $\hat{\theta} = \bar{x}$ show that $bias_F = bias_{\hat{F}}$ (carefully justifying each step).

[1+3=4 marks]

7. Briefly outline (preferably in point form and using appropriate formulae and/or examples) the basic idea behind the “importance sampling” method of variance reduction commonly used in simulations.

[5 marks]

8. You observe a random sample size n from an unknown population F ie $F \rightarrow \underline{x} = (x_1, x_2, \dots, x_n)$ and calculate the following bootstrap confidence intervals for a parameter θ : Percentile interval, Bootstrap t interval, BCa interval, ABC interval.

- (a) Under what circumstances would the percentile and BCa intervals be identical ?
- (b) The Percentile interval is first order accurate. What does this mean ?
- (c) State one disadvantage of the Bootstrap t interval.
- (d) Assume the Percentile interval for θ was [1.3, 2.4]. Give the endpoints of the Percentile interval for e^θ or state why you cannot from the available information.

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

9. (a) Describe the Holt-Winters method of exponential smoothing for a process $\{X_t\}_{t=0}^n$ with a linear trend and additive seasonality. Your description should specify how the level, trend and seasonal estimates are initialised and updated and how the m -step ahead forecast is calculated.
- (b) Let S_t be the estimated seasonal effect at time t . If the seasonal component of the process has period s , is it the case that $\sum_{k=1}^s S_{t-k} = 0$ for all t ? Justify your answer briefly.

[6+2=8 marks]

10. (a) Consider the following moving average (MA) estimate of the trend of a time-series X_t :

$$M_t = \sum_{k=-m}^m a_k X_{t+k}.$$

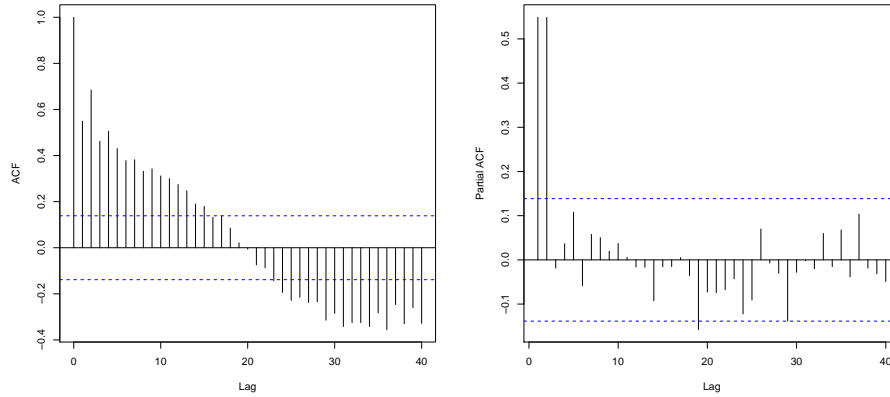
We take $m = 3$ and $a_{-3} = a_3 = 1/15$; $a_{-2} = a_2 = 2/15$ and $a_{-1} = a_0 = a_1 = 3/15$.

- (i) Is M_t equivalent to a 3×5 -MA estimate? Justify your answer.
 - (ii) Suppose that X_t has a seasonal component with period s . For what values of s does M_t remove the seasonal component of X_t ? Justify your answer.
- (b) What are the four main components of a time-series? Briefly describe each component.

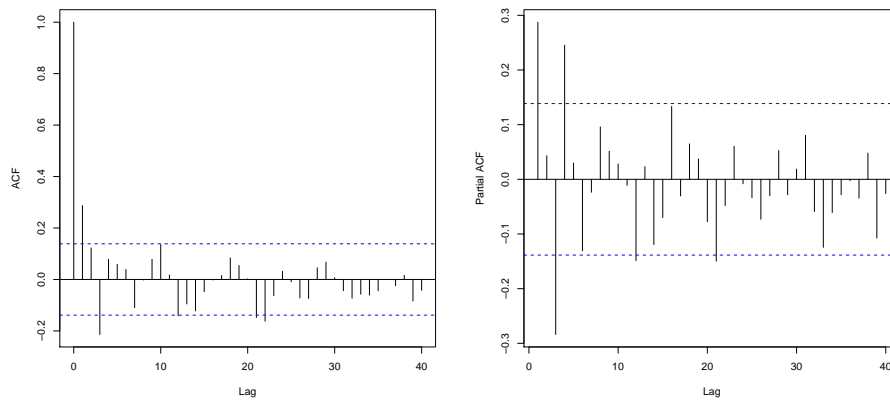
[6+4=10 marks]

11. The following are plots of the sample autocorrelation and partial autocorrelation functions of two stationary time-series. In each case indicate whether an $AR(p)$ or $MA(q)$ model would be most appropriate and give an estimate of the order (p or q) of the model.

Note that in these plots the acf starts at lag 0 but the pacf starts at lag 1.



Sample autocorrelation and partial autocorrelation A



Sample autocorrelation and partial autocorrelation B

[4 marks]

12. (a) (i) Define what it means for a stochastic process $\{X_t\}_{t=-\infty}^{\infty}$ to be *strictly stationary*.
- (ii) What does strict stationarity imply about the mean $\mu(t) = \mathbb{E}X_t$ and covariance $\gamma(s, t) = \mathbb{E}(X_s - \mu(s))(X_t - \mu(t))$?
- (iii) Define *second order* stationarity for the process $\{X_t\}_{t=-\infty}^{\infty}$.
- (b) Let $\{Z_t\}_{t=-\infty}^{\infty}$ be an i.i.d. sequence with mean 0 and variance σ^2 and let

$$X_t = \frac{1}{4}X_{t-2} + X_{t-4} - \frac{1}{4}X_{t-6} + Z_t + \frac{5}{2}Z_{t-1} + Z_{t-2}.$$

- (i) Express this model using the shift operator B .
- (ii) Give the general form of a Seasonal Integrated ARMA (SARIMA) process and then express X_t in this form, assuming a season of period 4.
- The order of a SARIMA process is written as $(p, d, q) \times (P, D, Q)$. What is the order of X_t ?

- (iii) Is X_t invertible (can it be written as an infinite order AR process)? Explain your answer.

[3+5=8 marks]

13. (a) Let $\{Z_t\}_{t=\dots,-1,0,1,\dots}$ be an i.i.d. sequence with mean 0 and variance σ^2 and let $\{X_t\}_{t=\dots,-1,0,1,\dots}$ be the AR(p) process given by

$$X_t = \sum_{j=1}^p \alpha_j X_{t-j} + Z_t.$$

Assuming that the process is stationary, show that $E X_t = 0$. Then derive the Yule-Walker equations for the autocorrelation (or equivalently autocovariance) of $\{X_t\}$.

- (b) Suppose that $X_t = -0.3X_{t-1} + 0.1X_{t-2} + Z_t$. What is the autocorrelation of $\{X_t\}$ in this case? Justify your answer.

[4+6=10 marks]

END of EXAM

Three sheets of formulas follow