

The University of Melbourne

Semester 2 Assessment, 2005

Department: Mathematics and Statistics

Subject Title: 620-374 Sampling and Forecasting

Exam Duration: Three hours

Reading Time: 15 minutes

This paper has eleven (11) 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.

- At the University of Melbourne there are approximately 2,000 general staff, 3,000 academic staff and 40,000 students.

A survey was conducted to measure the average number of cappuccinos drunk per week by a member of the university community, with the following results

Sample no.	Type	Response
1	General	4
2	General	6
3	Academic	0
4	Academic	10
5	Student	2
6	Student	12
7	Student	18
8	Student	0

For each of the three cases below, estimate the mean and standard deviation of the number of cappuccinos drunk per person per week. In each case state clearly what type of sample survey is being used.

You may find the following quantities useful

$$s^2 = 40.86, \quad s_{Gen}^2 = 2, \quad s_{Acad}^2 = 50, \quad s_{Stud}^2 = 72.$$

It is sufficient to give answers correct to 2 decimal places.

- Supposing that the samples were selected at random from the whole population, but you *do not* know the type of each sample point.
- Supposing that the samples were selected at random from the whole population, and you *do* know the type of each sample point.
- Supposing that the number of samples of each type was chosen *before* the sample took place.

[12 marks]

- In Carlton North there are approximately 3,900 households.

Households were selected at random and the number of university students living in each household recorded, together with the size of the household.

House	No. Occupants h_i	No. Uni Students u_i
1	3	0
2	2	0
3	4	4
4	4	1
5	3	1
6	2	0

Note that, to 2 decimal places,

$$\sum_{i=1}^6 (u_i - \bar{u})^2 = 14.4 \text{ and } \sum_{i=1}^6 (u_i - h_i \bar{u} / \bar{h})^2 = 9.11$$

- (a) (i) Give an estimate and 95% confidence interval for the total number of university students in Carlton North.
- (ii) Now suppose that you are told that Carlton North has 8,300 residents. Using this extra information, give a different estimate and 95% confidence interval for the total number of university students.
- (iii) It is noted that most students in Carlton North live in share homes, so that the number of students in a home is often either 0 or the size of the household. Given this, do you think your estimates at (i) and (ii) would be better or worse than an estimate using the same number of people, sampled at random from the whole population of Carlton North? Explain your answer.

[12 marks]

- (b) Suppose that you are estimating the population mean μ_Y using two methods: simple random sampling and cluster sampling with clusters of equal size. The total number of sample points is the same in each case.

Prove that the cluster sample is more precise than the simple random sample if and only if $\bar{S}^2 > S^2$ where \bar{S}^2 is the average cluster variance and S^2 is the population variance. (You may assume any of the results in the formula sheet without proof.)

[6 marks]

3. A pilot survey of 20 small businesses (less than 20 employees) produced the following data for the number of employees x and number employed under enterprise bargaining agreements y

x	18	10	8	12	5	13	7	19	8	14	18	12
y	18	7	4	10	0	9	2	11	5	7	9	11

For these data the mean vector and covariance matrix are

$$\begin{pmatrix} 12 & 7.75 \end{pmatrix}'$$

and

$$\begin{pmatrix} 19.66666667 & 16.83333333 \\ 16.83333333 & 20.85416667 \end{pmatrix}$$

There are (roughly) 1,140,000 small business in Australia employing 2,550,000 people. We wish to estimate the total number of people employed by small businesses under enterprise bargaining agreements. Determine whether it is better to use a ratio estimate or the mean per unit estimator (treating the y 's as if they each come from a SRS with no auxiliary information). [10 marks]

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

- (a) Define the empirical distribution function \widehat{F}
- (b) Define the plug-in estimate $\widehat{\theta}$ for a parameter of interest θ
- (c) Derive the plug-in estimate for $\theta = E_F(X^2)$ showing all your steps (where (x_1, x_2, \dots, x_n) are independent observations on X)
- (d) Indicate how the effectiveness of plug-in estimators can be justified by reference to two distinct principles of statistical estimation.
- (e) 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) List all possible values appearing in the vector results\$thetastar.
- (iii) Derive a theoretical expression for the probability that a bootstrap replicate generated by the program is less than 20.

[11 Marks]

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

- (a) Define the i th jackknife sample $x_{(i)}$
- (b) Define the i th jackknife replicate $\widehat{\theta}_{(i)}$
- (c) For the statistic $\widehat{\theta} = \bar{x}$ state the value of \widehat{se}_{jack} . (Marks only for answer: no need to give a derivation).
- (d) Consider the jackknife estimate of bias: $\widehat{bias}_{jack} = (n-1)[\frac{1}{n} \sum_{i=1}^n \widehat{\theta}_{(i)} - \widehat{\theta}]$. Explain why a scaling factor of $(n-1)$ is used in its definition.

[5 Marks]

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

- (a) State the definitions of $bias_F(\widehat{\theta})$ and its ideal bootstrap estimate $bias_{\widehat{F}}(\widehat{\theta})$.

- (b) For the special case of $\theta = t(F) = \mu_F$ and $\hat{\theta} = \bar{x}$ show that $\text{bias}_F = \text{bias}_{\hat{F}}$ (carefully justifying each step).
- (c) Now for general $\hat{\theta}$ and $S(\underline{x})$, write down an algorithm for calculating $\widehat{\text{bias}}_B$, the bootstrap estimate of bias.
- (d) State, with reasons, an approximate distribution for $\widehat{\text{bias}}_B$.
- (e) Write down the definition of a resampling vector. Give an example of a bootstrap sample \underline{x}^{*1} from \underline{x} and write down the corresponding resampling vector \underline{P}^{*1} .
- (f) If $\hat{\theta} = S(\underline{x}^*) = T(\underline{P}^*)$ (where \underline{P}^* is the resampling vector corresponding to \underline{x}^*), consider the following two estimates of bias calculated from B bootstrap samples:

$$E_1 = \frac{1}{B} \sum_{b=1}^B T(\underline{P}^{*b}) - T(\underline{P}^0)$$

$$E_2 = \frac{1}{B} \sum_{b=1}^B T(\underline{P}^{*b}) - T(\bar{\underline{P}}^*).$$

- (i) Write down formulae defining \underline{P}^0 and $\bar{\underline{P}}^*$.
- (ii) State which estimator has the smaller variance and name the variance reduction method applied to derive it.
- (iii) We can think of the variance reduced estimator as ‘correcting’ the other estimator. Give an intuitive explanation of the basis for this correction.

[15 Marks]

7. Briefly outline (preferably in point form and using appropriate formulae and/or examples) the basic idea behind the “control variate” method of variance reduction commonly used in simulations. [5 marks]
8. You observe a random sample size n from an unknown population F i.e. $F \rightarrow \underline{x} = (x_1, x_2, \dots, x_n)$ and calculate the following bootstrap confidence intervals for a parameter θ : Percentile interval, Bootstrap t interval, BCa interval, ABC interval.

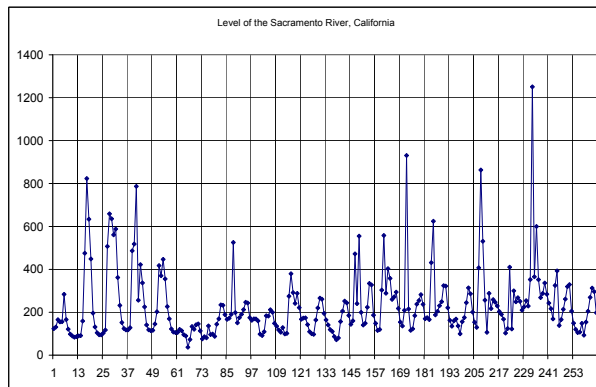
- (a) Under what circumstances would the percentile and BCa intervals be identical?
- (b) The BCa interval is second order accurate. What does this mean?
- (c) State one disadvantage of the Bootstrap t interval.

[4 marks]

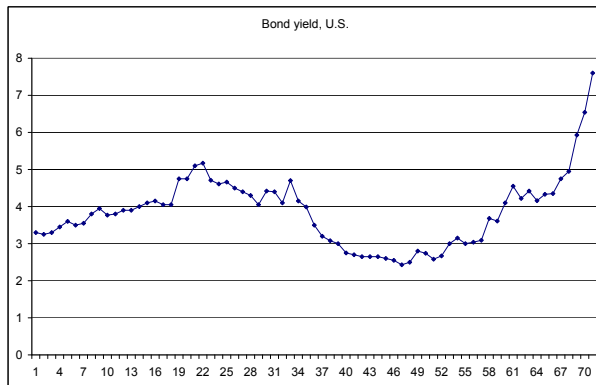
9. For each of the following time-series indicate whether or not the following features are present

- (i) Trend;
- (ii) Seasonality (multiplicative or additive);
- (iii) Cyclic behaviour.

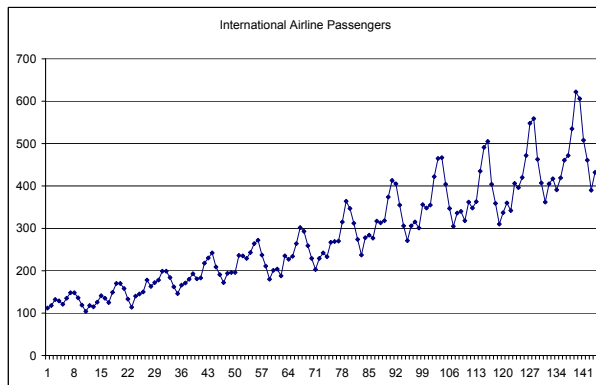
In addition explain briefly (in one or two sentences) the difference between seasonality and cyclic behaviour. [5 marks]



Time series A

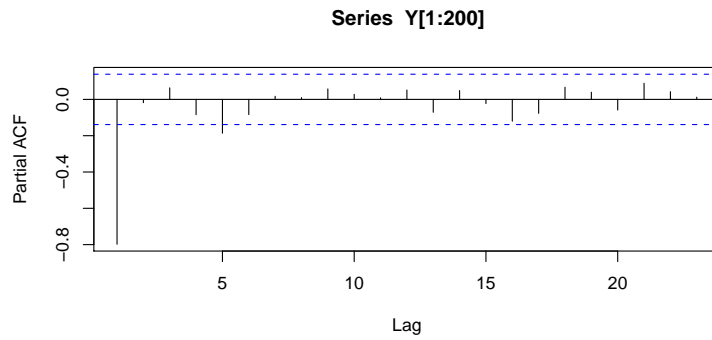
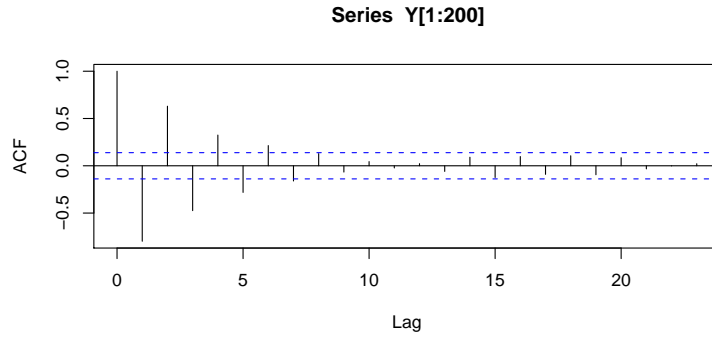


Time series B

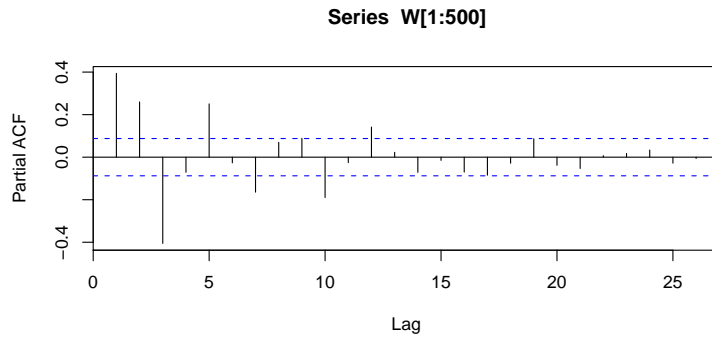
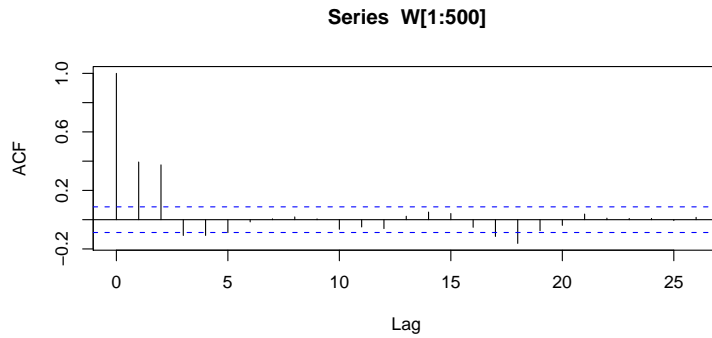


Time series C

10. The following are plots of the sample autocorrelation and partial autocorrelation functions for 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. Explain your choices. [6 marks]



Time Series A



Time Series B

11. 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
$t - 4$	2.8467	0.0129	-0.0549
$t - 3$	2.8434	0.0080	-0.0198
$t - 2$	2.8502	0.0076	0.0122
$t - 1$	2.8631	0.0092	0.0598

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

- (i) Update these values using the new observation $X_t = 2.8202$.
- (ii) Give forecasts for times $t + 1$ and $t + 2$.
- (iii) Briefly explain how the behaviour of the Holt-Winters' smoother changes as α , β and γ are increased/decreased.

Describe how might you choose values for α , β and γ in practice.

[9 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 $\{X_t\}_{t=-\infty}^{\infty}$.

[4 marks]

- (b) Let $\{Z_t\}_{t=-\infty}^{\infty}$ be an i.i.d. sequence with mean 0 and variance σ^2 and let

$$X_t = X_{t-1} + X_{t-4} - X_{t-5} + Z_t + \frac{1}{6}Z_{t-1} - \frac{1}{3}Z_{t-2}.$$

- (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 stationary? Explain your answer.

[8 marks]

- (c) Let $\{Z_t\}_{t=-\infty}^{\infty}$ be an i.i.d. sequence with mean 0 and variance σ^2 and let

$$X_t = \frac{1}{12}X_{t-1} + \frac{1}{12}X_{t-2} + Z_t.$$

Find the autocorrelation function of X_t .

[8 marks]