Office Use Only

MONASH University

Semester Two 2016
Examination Period

Faculty of Business and Economics

EXAM CODES: **ETC2420 & ETC5242**

TITLE OF PAPER: **STATISTICAL METHODS FOR INSURANCE - Paper 1**

EXAM DURATION: 3 hours writing time

READING TIME: 10 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

☐ Berwick ☑ Clayton ☐ Malaysia ☐ Off Campus Learning ☐ Open Learning
☐ Caulfield ☐ Gippsland ☐ Peninsula ☐ Enhancement Studies ☐ Sth Africa
☐ Parkville ☐ Other (specify)

During an exam, you must not have in your possession, a book, notes, paper, electronic device/s, calculator, pencil case, mobile phone, smart watch/device or other material/item which has not been authorised for the exam or specifically permitted as noted below. Any material or item on your desk, chair or person will be deemed to be in your possession. You are reminded that possession of unauthorised materials, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations.

No exam paper or other exam materials are to be removed from the room.

AUTHORISED MATERIALS

OPEN BOOK	☑ **YES**	☐ **NO**
CALCULATORS only a HP 10bII+ calculator is permitted	☑ **YES**	☐ **NO**
SPECIFICALLY PERMITTED ITEMS	☑ **YES**	☐ **NO**

if yes, items permitted are:

- **Lecture notes**
- **Labs and solutions**
- **Quizzes and solutions**

Candidates must complete this section if required to write answers within this paper.

STUDENT ID: .... DESK NUMBER: ....

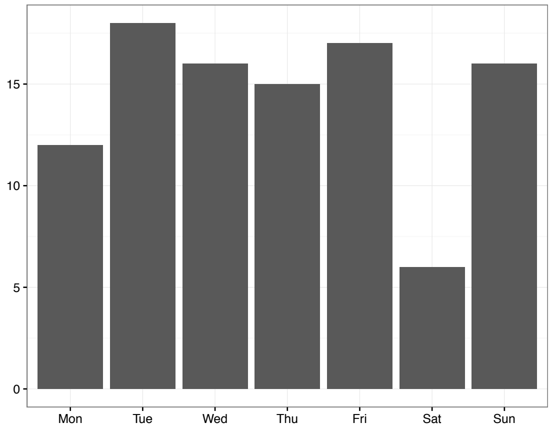
Instructions

There are 9 questions worth a total of 100 marks. You should attempt them all.

QUESTION 1

This question is about using random numbers to set up a computer experiment.

In a survey of CEOs of the top 100 global companies listed by Forbes magazine, the day of the week that they were born was recorded. Below is a bar chart of this data.



(a) Describe the distribution.

[2 marks]

Fairly uniform, except for Saturday. (Not bimodal, unimodal, multimodal, skewed.)

(b) If the probability that a CEO is born on any particular day is the same as any other day, what would you expect the bar chart to look like?

[1 marks]

In theory all the bars would be the same height. It would vary a little from this in samples.

(c) Describe how you could use simulation, to learn whether the count for Saturday is lower than we would expect if a births is equally likely on any day of the week.

[1 marks]

Sample 1:7, with replacement 100 times, and examine the count of the lowest day. Repeat this 1000 times. If the actual count for Saturday is lower than most of the low counts for the simulated data, then we would conclude that it is unlikely to have come from a population where every day was equally likely.

(d) Suppose the distribution of all baby births across days is not uniform, and follows this distribution:

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Prob	0.16	0.19	0.18	0.18	0.17	0.05	0.07

(a) How would you map random digits (0, 1, ... , 9) to days of the week, in order to set up a simulation to check in the data on CEOs is consistent with this probability distribution?

[2 marks]

Use the numbers in pairs: 00-15 would be mapped to 1, 16-38 would be mapped to 2, ...

(b) How would you use a sequence of random numbers to conduct the simulation? Write out the procedure.

[2 marks]

Simulate 100 pairs of digits, group them into the intervals defined above. Count the number in each interval. Keep the lowest count. Repeat 1000 times. Use the distribution of lowest counts to compare with the actual lowest count.

[Total: 8 marks]

— END OF QUESTION 1 —

QUESTION 2

This question is about using randomisation methods with data.

(a) You have the following data:

V1	V2
20	-5
12	3
31	-14
19	-8

and you have these two samples generated from the data:

A

V1	V2
20	-5
31	-14
31	-14
19	-8

B

V1	V2
20	-14
12	-5
31	3
19	-8

Label A and B as generated by either permutation or bootstrap randomisation methods.

[2 marks]

A is bootstrap, B is permutation.

(b) Compare and contrast bootstrap and permutation as methods for using randomisation in data analysis.

[3 marks]

Bootstrap generates samples similar to the current sample, and can be used to explore the variability of estimates. Permutation breaks association between variables, and can be used to do hypothesis testing.

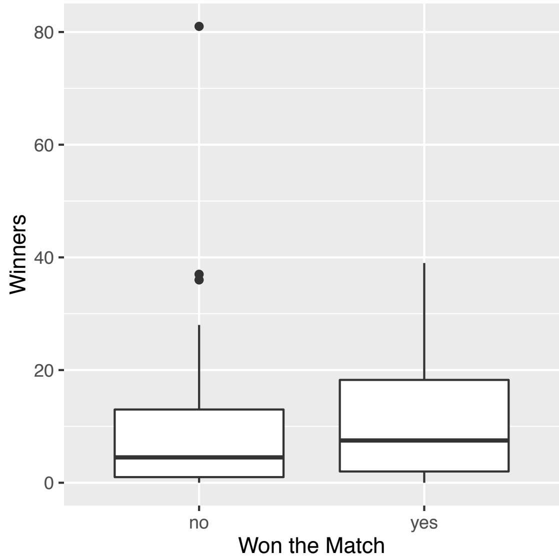
[Total: 5 marks]

— END OF QUESTION 2 —

QUESTION 3

This question is about decision theory.

As a coach, should you encourage your player to go for winners when playing a competitive game of tennis? Below is a side-by-side boxplot of the number of winners hit in the first round of the 2012 Australian Open women's matches against whether the player won or lost. And also the summary statistics for the two groups.



won	n	min	q1	median	q3	max	mean	sd
no	64	0	1.00	4.50	13.00	81	9.42	12.81
yes	64	0	2.00	7.50	18.25	39	11.47	11.30

(a) The classical test of two sample means is the t-test. Compute the t-statistic for this data.

[1 marks]

$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = 0.96$

(b) Write out the null and alternative hypotheses that corresponds to the t-test, which would help answer the coach's question.

[3 marks]

$H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 > \mu_2$ where group 1 is the player won the match.

(c) What assumptions does the classical t-test make? What concerns about satisfying these might you have after examining the side-by-side boxplot?

[2 marks]

Population of each sample needs to be close to a normal model. Samples need to be drawn independently. There are some outliers in the "No" group, and both distributions are skewed.

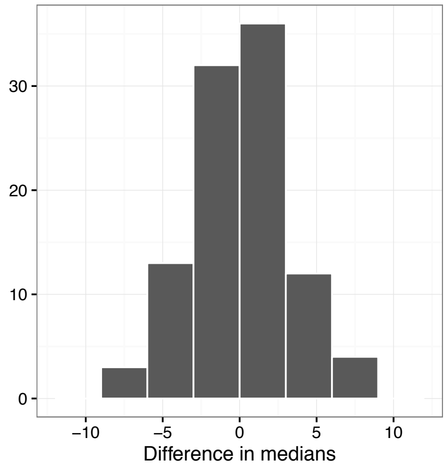
(d) Compute the difference in medians between the two groups (won, lost).

[1 marks]

3.00

(e) One hundred permutation samples are constructed. The median difference is computed for each. These are plotted below. Compute how many permutation samples have median differences larger than that computed on the data. 16, but it is ok to be off by one or two, 15 or 17 would be suitable. And the answer given here determines the corret answer in the next question, count/100.

[1 marks]



(f) Compute the permutation p-value based on the numbers in the previous question. 0.16

[1 marks]

(g) What null and alternative hypothesis pair is being tested with this permutation test?

[2 marks]

$H_o : median_1 - median_2 = 0$ vs $H_a : median_1 islargerthanmedian_2$

(h) Based on your p-value would you reject or fail to reject the null hypothesis?

[1 marks]

We would not reject H_o.

(i) Using your hypothesis test decision, what would your conclusion be? Should the coach advise their player to go for winners?

[2 marks]

Winners do not, on their own, improve a players chance of winning the match. (For women's Australian open matches.) Not on the basis of this data.

[Total: 14 marks]

— END OF QUESTION 3 —

QUESTION 4

This question is about statistical distributions.

(a) Using the Poisson density function, $P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x \in \{0, 1, 2, ...\}$, write down the likelihood function for n=2.

[2 marks]

$$l(\lambda|x_1, x_2) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \frac{\lambda^{x_2} e^{-\lambda}}{x_2!}$$

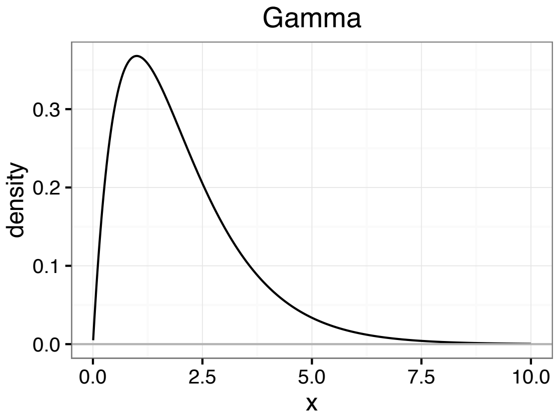
(b) For any probability density function, what is the total area under the curve? 1

[1 marks]

(c) Make a sketch of the Gamma(2,1), like given below, marking off the quantity that corresponds to $P(X > 5.0)$.

[2 marks]

A vertical line is drawn at x=5, and area under he curve to the right is shaded in.



(d) Which of the following most closely matches the value $P(X > 5.0)$? 0.04, 0.17, or 0.29? 0.04

[1 marks]

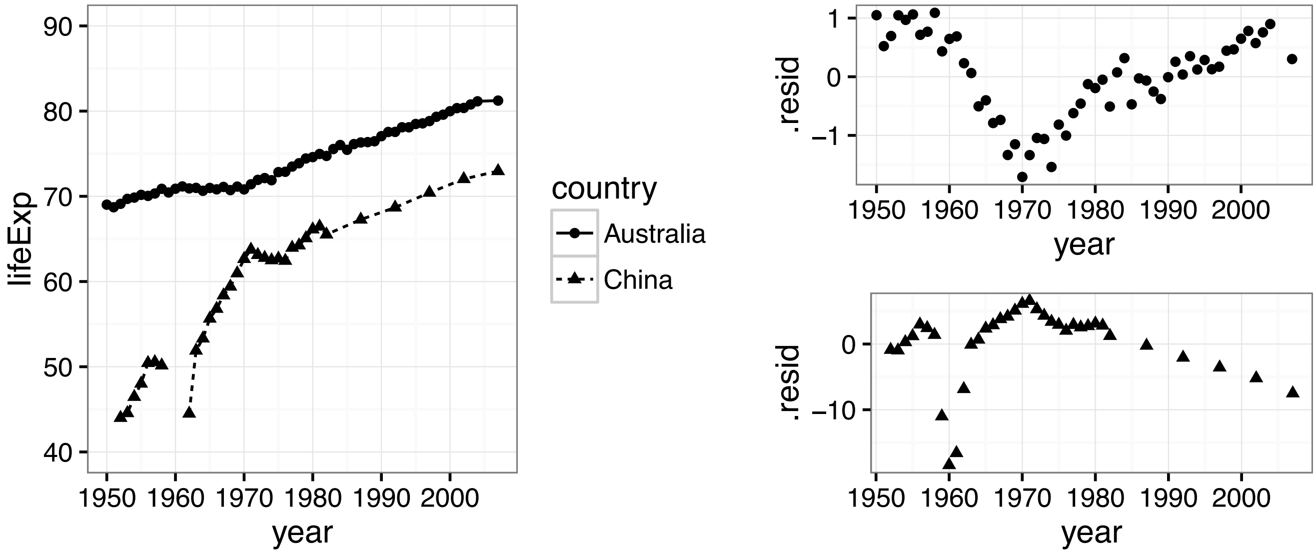
[Total: 6 marks]

— END OF QUESTION 4 —

QUESTION 5

This question is about linear models.

The life expectancy is modeled on year for data from the gapminder package in R, for Australia and China.



	Australia Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-375.50	11.96	-31.40	0.00
year	0.23	0.01	37.60	0.00

(a) Sketch the model fit for Australia.

[2 marks]

A line through (1950, 73) and (2010, 86.8)

(b) Explain what the intercept value of -375.5 means. Can we really have a negative life expectancy? What adjustment to the model would have lead to a more sensible intercept estimate?

[2 marks]

It means nothing useful! It is life expectancy in year 0. It would have been better to subtract 1950 from the year before modeling.

(c) Use the model to predict the life expectancy for 2000.

[1 marks]

-375.5+2000*0.23=84.5

(d) If the recorded life expectancy for 1970 is 70.81, and the predicted value is 77.6. Compute the residual.

[1 marks]

77.6-70.81=6.79

(e) The plot at top right shows the residuals for the model fit for Australia. Does this show that a linear model is a good fit? Explain your answer.

[2 marks]

It suggests that the model overfits life expectancy around 1960-1970. The relationship between year and life expectancy is not strictly linear.

(f) The remaining goodness of fit statistics for the fit are below. Compare the residual deviance with the null deviance and explain what this tells you about the goodness of fit.

[2 marks]

```
Degrees of Freedom: 55 Total (i.e. Null);  54 Residual
Null Deviance:     791.7
Residual Deviance: 29.12  AIC: 128.3
```

The residual deviance drops a huge amount from the null deviance: 791.7 to 29.12. That suggests that year explains a lot of the variation in life expectancy.

(g) We now switch the attention to China. A linear model has been fit for life expectancy on year. The hat values are calculated to assess leverage. The largest value is 0.222.

(i) Would this indicate that the point with this value has high leverage? ($n = 56$)
Yes. Values larger than $p/n = 1/56 = 0.0179$ would be considered to have high leverage.

(ii) Which year do you think the highest value corresponds to?
There is a temptation to say 1960, but leverage only measures deviation from mean of x, so it would have to be 2008.

[2 marks]

(h) The highest Cooks D value occurs at year 1960, and is 0.25. Would this indicate that (1960, 31.6) is an influential observation? Explain.

[2 marks]

There are two rules of thumb, $4/n = 0.07$ or 1 for looking at Cooks D. Based on the former 1960 would be considered to be an influential observation.

(i) Explain the difference between influence and leverage.

[2 marks]

Leverage just measures how far out on the predictor scale an observation is. Influence measures how much the model changes when the observation is used for model fitting and when it is left out.

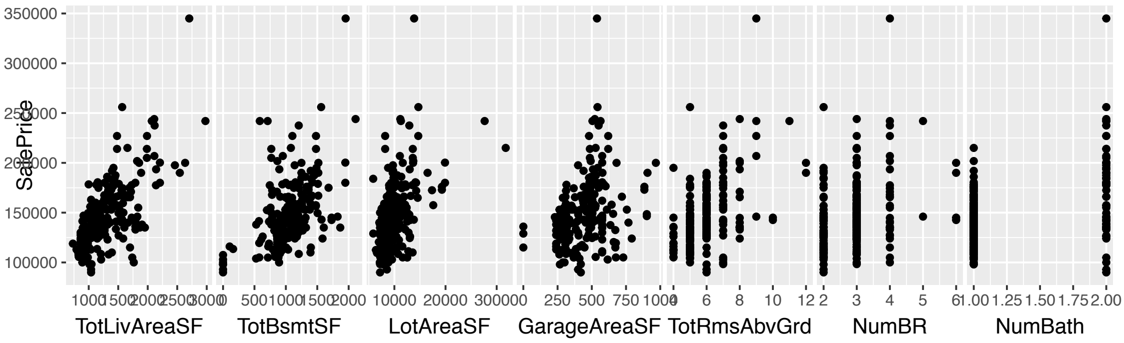
[Total: 16 marks]

— END OF QUESTION 5 —

QUESTION 6

This question is about multiple regression.

A linear model for Sales Price based on several house characteristics is fitted to a subset of the Ames Housing data. Two models are fitted, one with TotLivArea, TotBsmtSF, LotArea, GarageArea, TotRmsAbvGrd, NumBR, NumBath, and the second without the variable TotLivArea. This is a summary of the model fit. (Area is given in square feet, SF at the end of the variable name indicates this.)




```
> ah_glm <- glm(SalePrice~TotLivAreaSF+TotBsmtSF+LotAreaSF+
   GarageAreaSF+TotRmsAbvGrd+NumBR+NumBath, data=ah)
> ah_glm2 <- glm(SalePrice~TotBsmtSF+LotAreaSF+GarageAreaSF+
   TotRmsAbvGrd+NumBR+NumBath, data=ah)
```

	with TotLivAreaSF		without TotLivAreaSF	
	Estimate	Std.Error	Estimate	Std.Error
(Intercept)	48743.69	8163.77	31671.84	8590.28
TotLivAreaSF	48.29	6.88	-	-
TotBsmtSF	24.49	4.37	30.66	4.71
LotAreaSF	1.94	0.51	2.84	0.54
GarageAreaSF	23.27	9.58	22.77	10.56
TotRmsAbvGrd	1161.79	2190.47	9625.88	2015.33
NumBR	-9341.53	2923.26	-8783.01	3221.32
NumBath	1119.45	4091.19	11298.53	4216.78

(a) Predict the sales price for a house with 2000 SF of living space, 500 SF basement, on a 3000 SF lot, 500 SF garage, 6 rooms above ground, 3 bedrooms and 2 bathrooms.

[2 marks]

$48743.69 + 2000*48.29 + 500*24.49 + 3000*1.94 + 500*23.27 + 6*1161.79 - 3*9341.53 + 2*1119.45 = 156208.70$

(b) This is the model fitting code, and the variance inflation factors for each variable. How is variable inflation factor calculated? Which is the better model fit according to VIFs?

[2 marks]

$\frac{1}{1-R_j^2}$. The second model has lower VIFs, so is better. None of the VIFs are bigger than 10, but the sign of NumBR in the model is negative. It makes no sense for house price to go down when the number of bedrooms goes up, so this is an indicator of a multicollinearity problem.

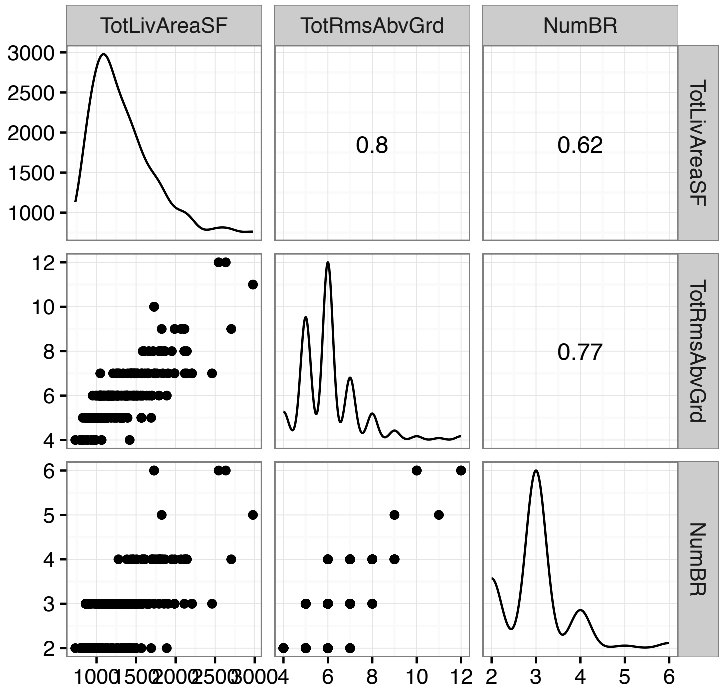
```
> vif(ah_glm)
TotLivAreaSF     TotBsmtSF     LotAreaSF GarageAreaSF TotRmsAbvGrd          NumBR
```

```
    3.664805     1.175415     1.244231     1.148607     4.230897     2.499457
     NumBath
    1.633853
> vif(ah_glm2)
   TotBsmtSF    LotAreaSF GarageAreaSF TotRmsAbvGrd        NumBR      NumBath
    1.127679     1.165199     1.148544     2.947134     2.497603     1.428313
```

(c) Below is a plot of the three of the predictors from the first model with highest VIFs. Describe the association between the three variables. In an ideal fit what would the pattern look like?

[2 marks]

Ideally the points should be scattered throughout the plot, that there is no association between these variables.



(d) For each increase in number of rooms above ground how much would you expect the sales price to increase? (Use the model containing `TotLivAreaSF`, and assume that all other predictor values remain constant.)

[1 marks]

$1161.79

(e) Assume that the living space is 2000 square feet. For each increase in number of bedrooms how much would you expect the sales price to increase? (Assume that all other predictor values remain constant.)

[1 marks]

-$9341.5

(f) Does the model containing `TotLivAreaSF` having a negative coefficient for NumBR indicate a problem? If so, does the second model fix the problem? If not, what would you do next?

[2 marks]

The sign of NumBR in the model is negative. It defies intuition for house price to go down when the number of bedrooms goes up. It is not so much an indicator of multicollinearity because the variances, of the estimates, are not inflated substantially, there is plenty of data to make the estimates accurately. But there is correlation between the predictors which makes interpretation of the coefficients a little more complicated. The negative sign for NumBRs (for both models) indicates, given fixed living space, the extra room negatively affects the house price. Nothing really needs to be done to the model. If we were concerned about the contradictory nature of the interpretation, though, we might first regress bedrooms on living space, and use the residuals as a predictor in the model. The coefficient for the new variable, number of bedrooms relative to living space, will (likely) be positive.

[Total: 10 marks]

— END OF QUESTION 6 —

QUESTION 7

This question is about modeling risk and loss, referring back to Lab 9 where we modeled risk and loss for locating a coffee shop at either Flinders St Underpass (Fl) or Melbourne Central (MC).

We can set this up as a decision theory problem, with Player A being the coffee shop is located at Fl and Player B being the coffee shop is located at MC. Suppose that Strategy 1 will be to have one employee, ... to strategy 4 is to have 4 employees, and strategy 5 will be the coffee shop is closed.

We will focus just on one day of the week, and one hour in that day to do calculations. Assume x_{Fl} is the number of pedestrians that pass by in that hour, and x_{MC} is the number of pedestrians. The proportion of pedestrians that will actually stop in to buy a coffee in that day (d) and time (t) is $p_{Fl}(d,t), p_{MC}(d,t)$ respectively. Assume each customer will spend \$4 when they come into the shop. To open the coffee shop costs \$100, and each employee adds an extra \$50 to costs. You need one employee for each 50 customers. If there are more than this, the additional customers will walk away rather than coming in to buy a coffee. At Flinders, the proportion of pedestrians passing by who will buy a coffee is 0.1 between 7-11am, 0.05 between 11-4, 0.01 between 4-8. At Melbourne Central, the proportion who will buy coffee is 0.08 between 7-11am, 0.06 between 11-4, 0.02 between 4-10pm. At all other times assume no purchases.

The goal is to earn the most money in the hour. Below is the payoff matrix (needs to be completed).

	MC				
Fl	1	2	3	4	5
1					
2					
3					
4					
5					

(a) The equation for measuring earnings at Fl (where e_{Fl} is the number of employees) is

$$\begin{aligned} y_{Fl} &= min(p_{Fl} \times x_{Fl}, e_{Fl} \times 50) \times 4 - 100 - (e_{Fl} - 1) \times 50 \;\; if \;\; e_{Fl} > 0, \\ &= 0 \;\; o.w. \end{aligned}$$

Write down the equation to measure the earnings at MC.

[2 marks]

$$\begin{aligned} y_{MC} &= min(p_{MC} \times x_{MC}, e_{Fl} \times 50) \times 4 - 100 - (e_{MC} - 1) \times 50 \;\; if \;\; e_{MC} > 0, \\ &= 0 \;\; o.w. \end{aligned}$$

(b) Write down the equation to measure the difference between earnings at Fl and MC.

[2 marks]

Assuming both shops are open

$$min(p_{Fl} \times x_{Fl}, e_{Fl} \times 50) \times 4 - (e_{Fl} - 1) \times 50 -$$
$$(min(p_{MC} \times x_{MC}, e_{MC} \times 50) \times 4 - (e_{MC} - 1) \times 50)$$

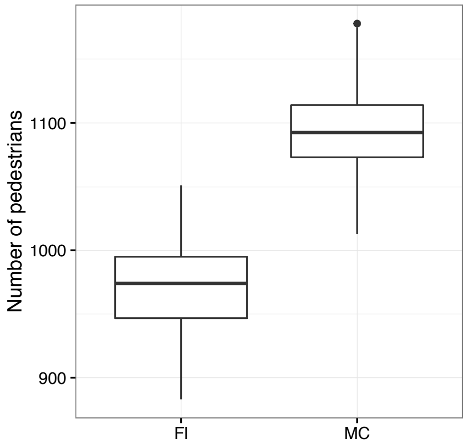
(c) The day is a Monday, and time is 10am. Complete the payoff matrix.

[3 marks]

$p_{Fl} = 0.1, p_{MC} = 0.06$, and then we would plug in e_{Fl}, e_{MC} for each column.

(d) We have built a generalised linear model of pedestrian counts based on 2015 data, and used this to simulate 100 predicted values for Wed 2pm at each location. A side-by-side boxplot is shown below of these values, and the summary statistics for each location are given in the table. At best, how many pedestrians can you expect at each location? And, at worst, how many? Fl ranges between 883 and 1051, MC ranges between 1013 and 1178.

[2 marks]



	min	q1	median	q3	max	mean	sd
Fl	883	946.75	974.00	995.00	1051	971.90	32.41
MC	1013	1073.00	1092.50	1114.00	1178	1092.68	33.58

(e) Use these values to compute the expected earnings difference under each strategy. And determine which is the best strategy for the player with the coffee shop at the Fl location.

[4 marks]

Worst case scenario would be if Fl has 883 pedestrians and MC has 1178 pedestrians walk by. This would correspond to a payoff matrix of

	MC			
Fl	1	2	3	4
1	0.00	-132.72	-182.72	-232.72
2	103.20	-29.52	-79.52	-129.52
3	53.20	-79.52	-129.52	-179.52
4	3.20	-129.52	-179.52	-229.52

which would suggest that Fl should have two employees on. And the only case where they would win is if MC only have one employee. This strategy, of having two employees for this time period provides the best odds of making more money than MC.

[Total: 13 marks]

— END OF QUESTION 7 —

QUESTION 8

This question is about Bayesian methods

(a) Of 30 music students, 20 can play the violin and 17 have had voice training. Furthermore, 15 have had voice training and can play the violin. One of the students chosen at random can play the violin, what is the probability that this student has had voice training? Explain.

[3 marks]

- $P(Violin) = \frac{20}{30}$
- $P(Training) = \frac{17}{30}$
- $P(Violin \cap Training = \frac{15}{30}$
- $P(Training|Violin) = \frac{15}{20} = \frac{3}{4} = 0.75$

(b) We are interested in estimating the probability p that it will rain tomorrow. Explain what is meant by a prior distribution, a posterior distribution and a conjugate prior distribution.

[3 marks]

- Before seeing any data, the prior distribution reflects our prior belief that it will rain tomorrow.
- After observing data, we can update our belief by computing a posterior distribution using the Bayes rule.
- Using a conjugate prior distribution is convenient since it allow us to obtain a closed-form expression for the posterior distribution. In particular, we do not have to compute the normalizing constant.

(c) We are interested in estimating the probability p that a coin will turn up heads. We will consider the maximum likelihood estimate and the optimal Bayes estimate under squared error risk. For the Bayes estimate, we use a uniform prior distribution, i.e. $\pi(p) = 1$.

- You toss the coin for the first time, and you see a tails. Compute both the MLE estimate and the Bayes estimate for p.

  [2 marks]

  - $\hat{p}_{MLE} = 0$
  - * $\pi(p) = 1$
    * $\pi(p|x_1 = T) = \frac{\pi(x_1=T|p)\pi(p)}{\int_0^1 \pi(x_1=T|p)\pi(p)dp} = \frac{(1-p)\cdot 1}{1/2} = 2(1-p)$
    * Under quadratic loss: $\hat{p}_{Bayes} = \mathbb{E}[p|x_1 = T] = \int_0^1 \pi(p|x_1 = T)p\ dp = \int_0^1 2(1-p)p\ dp = \frac{1}{3}$

    where

- You toss the coin for a second time, and you see a tails. Recompute both the MLE estimate and the Bayes estimate for p.

  [2 marks]

  - $\hat{p}_{MLE} = 0$
  - * $\pi(p) = 1$

* $\pi(p|x_1 = T, x_2 = T) = \frac{\pi(x_1=T,x_2=T|p)\pi(p)}{\int_0^1 \pi(x_1=T,x_2=T|p)\pi(p)dp} = \frac{(1-p)^2 \cdot 1}{1/3} = 3(1-p)^2$
* Under quadratic loss: $\hat{p}_{Bayes} = \mathbb{E}[p|x_1 = T] = \int_0^1 \pi(p|x_1 = T)p\ dp = \int_0^1 3(1-p)^2 p\ dp = \frac{1}{4}$

where

- Briefly discuss the results you obtain for the MLE and the Bayes estimate.

[2 marks]

As long as we do not see heads, the MLE estimate will be 0. While the Bayes estimate update his belief with every new observation.

(d) Briefly discuss why we often need to use numerical methods to compute the posterior distribution.

[2 marks]

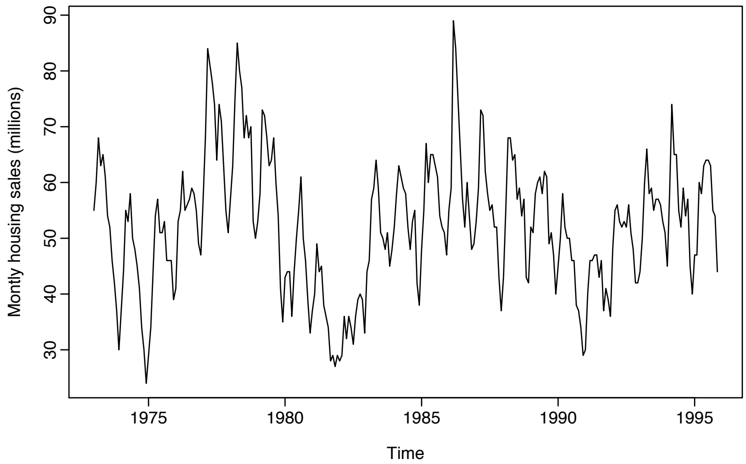
[Total: 14 marks]

— END OF QUESTION 8 —

QUESTION 9

This question is about time series methods.

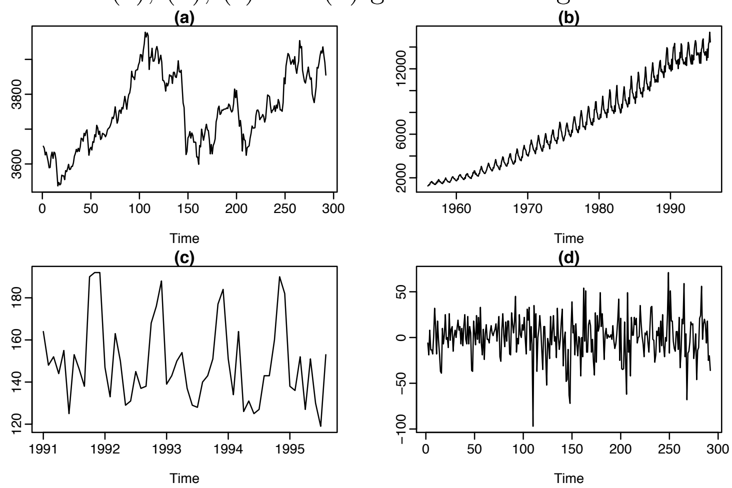
(a) Consider the sales of new one-family houses in the USA, Jan 1973 - Nov 1995 given in the Figure below.



- Describe this time series in terms of patterns you see (trend, cycle, seasonality, etc). [1 marks]

The time series contains a cycle, a yearly seasonality with a peak in mid-year

(b) Consider four time series (a), (b), (c) and (d) given in the Figure below.



- Which time series would be considered to be stationary? Explain. [2 marks]

  (a) and (c) have trend, and (c) has seasonality. So they are not stationary. (d) is stationary since the properties of the series do not depend on the time at which the series is observed.

- For each time series you think is not stationary, which transformation would you apply to make it stationary? [2 marks]

  Differencing for (a) and seasonal difference for (b) and (c).

(c) Suppose $\{\varepsilon_t\}$ is a i.i.d process with $\varepsilon_t \sim N(0, \sigma^2)$. We consider the autoregressive process $\{y_t\}$ where

$$y_t = \phi_1 y_{t-1} + \varepsilon_t.$$

- What is the formula to compute the autocorrelation function (ACF) $\gamma(k)$ for lag k of this process when $|\phi_1| < 1$?

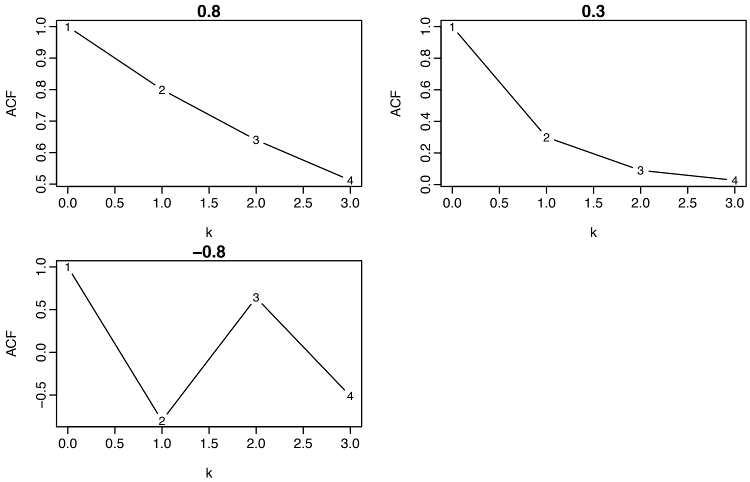
[1 marks]

$\gamma(k) = \phi_1^k$

- For $\phi_1 = \{0.8, 0.3, -0.8\}$, plot $(k, \gamma(k))$ for $k = 0, 1, 2, 3$. Explain what you observe.

[2 marks]

Solution:



We see the large decay in ACF for 0.8, and the slower decay for 0.3. For -0.8, we see how the ACF alternate when ϕ_1 is negative.

(d) Given T observations $y_1, \ldots, y_T$ from an AR(p) process.

- What is the formula to compute the sample ACF $\hat{\gamma}(k)$ where k is the lag.

[2 marks]

$\hat{\gamma}(k) = \frac{\sum_{t=k+1}^{T}(y_t-\bar{y})(y_{t-k}-\bar{y})}{\sum_{t=1}^{T}(y_t-\bar{y})^2}$ where $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$.

Describe the procedure to compute the standard error of $\hat{\gamma}(k)$ using the block bootstrap.

- Step 1, describe how do you generate bootstrap samples from the sample $y_1, \ldots, y_T$.

[2 marks]

Split the observed time series in blocks of size $> p$. Sample with replacement the blocks to get one bootstrapped time series $y_1^*, \ldots, y_T^*$. Repeat the procedure B times to get B time series $y_1^{(b)}, \ldots, y_T^{(b)}$ with $b = 1, \ldots, B$

- Step 2, how do you compute the bootstrapped standard errors?

[2 marks]

Compute $\hat{\gamma}_b(k)$ for each block $b = 1, \ldots, B$. Compute the standard deviation of the set $\{\hat{\gamma}_1(k), \ldots, \hat{\gamma}_B(k)\}$.

[Total: 14 marks]

— END OF QUESTION 9 —