ETC2420/5242

Practice exam 2016

Instructions

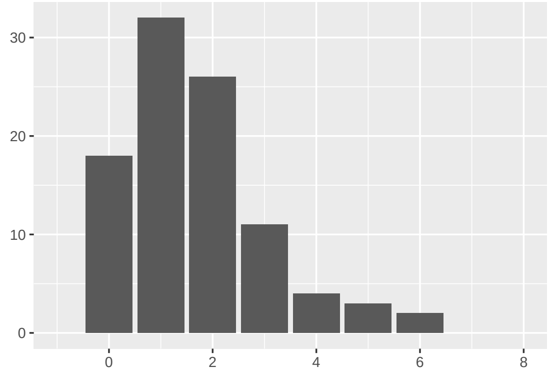
- There are 9 questions worth a total of 100 marks. You should attempt them all.
- Open book
- You can use the approved calculator

NOTE: The questions on this exam are to give you a sense of questions that will be on the exam, and of the length of the exam. They are not the same as the actual exam questions.

QUESTION 1

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

We are going to model the number of customers in the coffee shop for every 5 minutes of a 8 hour working day. Below is a bar chart of the data for one day.



(a) Describe the distribution.

[2 marks]

Right-skewed, unimodal, discrete.

(b) Which of these probability density functions is most likely the best match for modeling the distribution?

[2 marks]

A: $P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x \in \{0, 1, 2, ...\}$

B: $P(X = x \mid n, p) = \left(\begin{array}{c} n \\ p \end{array} \right) p^x (1-p)^{n-x} \quad x \in \{0, 1, 2, ..., n\}$

A, it is a Poisson-type model. Both of these are discrete, but a binomial model derives from a success/failure problem rather than a counting problem.

(c) There are $n = 96$ 5 minute intervals in one day. How could you use maximum likelihood estimation to find a suitable value for the population parameter? (Write down the likelihood function.)

[2 marks]

$\frac{\lambda_1^x e^{-\lambda}}{x_1!} \times ... \times \frac{\lambda_{96}^x e^{-\lambda}}{x_{96}!}$. Find the value of λ which maximises this function.

(d) Describe how you could use simulation, to learn whether the count (number of 5 minute intervals) when there are more than 4 customers in the coffee shop is higher than expected.

[2 marks]

Using the maximum likelihood estimates obtained we could sample values from the distribution, and compare these with the observed counts.

(e) Which of these sequences of heads and tails is more likely to have been generated by actually doing a coin flip? Explain your choice.

[2 marks]

A: HTHHTHTTHH

B: THTHHHHTTH

B, because it has the longest run, but still quite likely to see one of length 4. People are more likely to write down a “balanced” sequence with an even number of H’s and T’s, and not a run of 4.

[Total: 10 marks]

— END OF QUESTION 1 —

QUESTION 2

This question is about using randomisation methods with data.

(a) You have the following sequence of random numbers:

```
3 6 6 2 9 4 6 4 0 6 0 8 4 0 5 6 6 6 5 9 6 4 3 2 1 5 0 6 0
9 4 6 9 1 5 0 3 2 6 0 0 4 5 6 7 2 5 4 5 1 0 1 9 1 8 9 7 4
0 9 7 3 9 7 9 5 2 7 3 5 7 0 2 6 7 2 0 9 2 3 7 3 1 6 9 5 3
7 3 9 0 4 2 0 7 2 5 3 9 9
```

You have 8 people to choose from, who are willing to answer questions about their internship experiences:

`Han, Gabrielle, Emma, Carson, Sam, Tina, Wei, Viv`

You can't interview all of them, only 5 will be interviewed.

(a) Map the possible digits $\{0, ..., 9\}$ to people's names.

[2 marks]

0=Han, 1=Gabrielle, 2=Emma, 3=Carson, 4=Sam, 5=Tina, 6=Wei, 7=Viv, ignore 8, 9

(b) Use the numbers to select 5 people to be interviewed.

[1 marks]

Carson, Wei, Emma, Sam, Han

(c) If you were to do a bootstrap sample instead (doesn't make sense here, but just imagine), how would the sample of people differ?

[2 marks]

Carson, Wei, Wei, Emma, Sam

[Total: 5 marks]

— END OF QUESTION 2 —

QUESTION 3

This question is about decision theory.

(a) For this pair of hypotheses, which is the alternative hypothesis?

[2 marks]

$H_o: \mu_1 = \mu_2 \quad vs \quad H_a: \mu_1 \neq \mu_2$

$H_a: \mu_1 \neq \mu_2$

(b) If the p-value is large, e.g. 0.23, you would typically fail to reject the null hypothesis. True or False

[2 marks]

True

(c) We have the following sample of errors from two different prediction models, based on taking slightly different training and test samples:

	n	min	q1	median	q3	max	mean	sd
Model A	20	0.060	0.065	0.070	0.085	0.10	0.072	0.02
Model B	15	0.050	0.065	0.075	0.080	0.11	0.077	0.015

(a) We want to know whether model A or B has better predictive power, so we are testing $H_o: \mu_A = \mu_B$ vs $H_a: \mu_A \neq \mu_B$
Compute the classical t-statistic value for this test.

[2 marks]

$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}} = \frac{0.072-0.077}{\sqrt{0.02^2/20+0.015^2/15}}$

(b) If the distribution of test errors was bimodal for either model, would you still be able to use the t-test? What concerns might you have about this?

[2 marks]

Population of each sample needs to be close to a normal model. Bimodality violates this assumption. Samples need to be drawn independently.

(c) Instead of a classical t-test, you conduct a permutation test, of the equality of the errors of the two models. Of 1000 permutations, you observe 30 permutations that have a mean difference larger than $\bar{x}_A - \bar{x}_B$ and 90 that have mean difference larger than $\bar{x}_B - \bar{x}_A$. Compute the p-value for the test.

[2 marks]

120/1000=0.12

(d) Based on your p-value calculation would you reject, or fail to reject H_o?

[2 marks]

Fail to reject

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

[2 marks]

Models A and B provide effectively the same predictive error

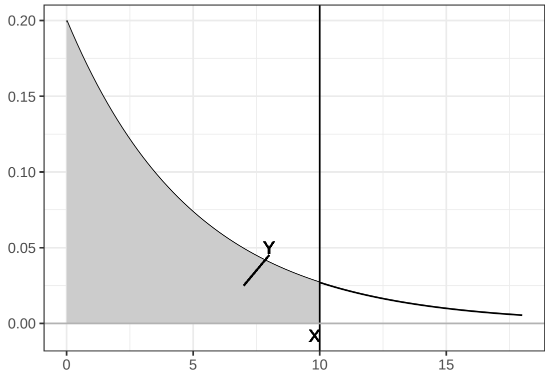
[Total: 14 marks]

— **END OF QUESTION 3** —

QUESTION 4

This question is about statistical distributions.

For the exponential density function, $f(x \mid \lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{(-x/\lambda)^k}, \quad x \geq 0$, shown below,



(a) Which letter represents a probability? X or Y

[1 marks]

Y

(b) What possible values can the quantiles for an exponential distribution take?

[1 marks]

$0 - \infty$

(c) What is the parameter of the exponential distribution? X, f, λ, k

[1 marks]

λ, k

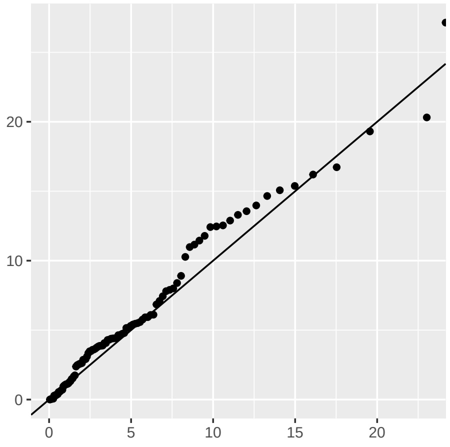
(d) Which of these would be bigger? $P(X > 5)$ or $P(X > 7)$

[1 marks]

$P(X > 5)$

(e) The following plot is a QQ-plot of a sample, measured against the quantiles of the $exp(0.2)$ distribution. Does it indicate that the sample might be from this distribution?

[2 marks]



Looks ok, close to the line, just a little wobble

[Total: 6 marks]

— END OF QUESTION 4 —

QUESTION 5

This question is about linear models.

We want to estimate the number of medals a country will earn in the 2008 olympics based on the standardised GDP of the country. The table below summarises the model fit:

```
Residuals:
   Min     1Q Median     3Q    Max
-27.78  -4.73  -2.76   1.34  51.91

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.00       1.18    9.32  1.5e-14 ***
GDP_std        16.08       1.19   13.55  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11 on 83 degrees of freedom
Multiple R-squared:  0.689,Adjusted R-squared:  0.685
F-statistic:  184 on 1 and 83 DF,  p-value: <2e-16
```

(a) Write down the equation for the fitted model.

[2 marks]

$M\hat{2}008 = 11.00 + 16.08 \times GDP_{std}$

(b) Explain what the intercept value of 11.00 means. Does it make sense in the context of this data?

[2 marks]

For a country with standardized GDP score of 0, they are expected to score 11 medals. Yes, it makes sense, because the GDP values are standardised so 0 is in the range of the data.

(c) Use the model to predict the 2008 medals for China, who has a standardised GDP of 3.18.

[2 marks]

$11.00 + 16.08 * 3.18 = 62$

(d) If China scored 100 medals, compute the residual.

[2 marks]

100-62=38

(e) How well did the model predict China's medal count for 2008? Use the summary of the distribution of residuals to help you answer this.

[2 marks]

Under-estimating the count by 38 medals is not very good. But the residuals are as high as 51.91, and as low as -27.78, so it is not the worst of the predicted counts.

(f) How much variation in medal count does the standardised GDP explain?

[2 marks]

68.9%

(g) Compute a classical 95% confidence interval for the slope, given the $t_{83,0.05} = 2$. Would this suggest that the true slope is significant?

[2 marks]

$16.08 \pm 2 \times 1.19 = (14, 18)$. The interval is a lot different from 0, so the slope is significantly different from 0.

(h) China has a leverage value of 0.13. What does this tell us about how influential the observation is on the model fit?

[2 marks]

0.13 is a large leverage value. However it only tells us about the potential of the observation to affect the model, and we would need to look at Cooks D to check influence.

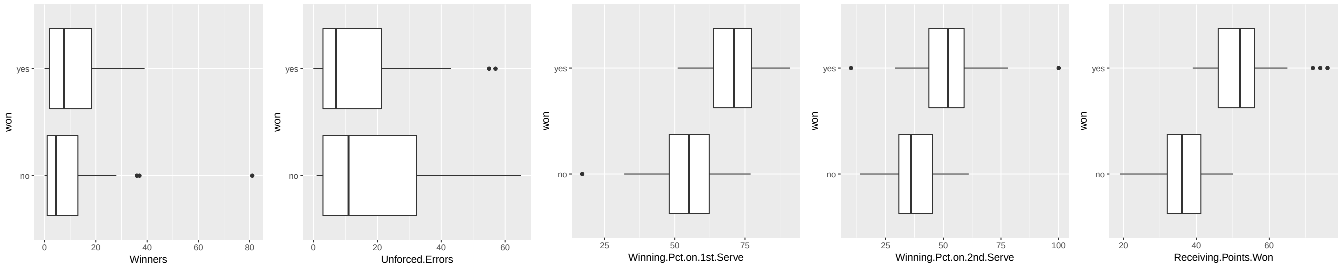
[Total: 16 marks]

— END OF QUESTION 5 —

QUESTION 6

This question is about multiple regression, and abut compiling data.

We have collected data from the 2012 Australian Open women's tennis round 1 matches. The data has been scraped off the web site, and contains player statistics including winners, unforced errors, serving points win %, and receiving points won. We want to build a model for whether the player wins the game or not.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.53994	0.16	-9.79	0.00
Winners	0.00552	0.00	1.38	0.17
Unforced.Errors	-0.00578	0.00	-1.90	0.06
Winning.Pct.on.1st.Serve	0.01166	0.00	5.28	0.00
Winning.Pct.on.2nd.Serve	0.00683	0.00	3.19	0.00
Receiving.Points.Won	0.02341	0.00	8.26	0.00

(a) For a player making 40 winners, 30 unforced errors, 70% first serve winning, 50% second serve winning, and 40 receiving points won, predict whether they would win the match or not.

[2 marks]

$-1.53994 + 0.00552 * 40 - 0.00578 * 30 + 0.01166 * 70 + 0.00683 * 50 + 0.02341 * 40 = 0.6$. This would suggest they have a good chance of winning.

(b) What type of response variable is it? (Categorical, quantitative, ...) Is the linear model that was fit the best for this type of response? If not, what would be better?

[3 marks]

Categorical response. A logistic regression model would be better. There is a chance that this model would make predictions above 1 and below 0, which don't make sense in the context of winning or losing.

(c) Are all of the variables included in the model important for predicting winning?

[1 marks]

Winners would appear to be not important based on the p-value.

(d) Below are the vifs for the variables included in the model. What does vif measure? Do these values indicate any problems with the model?

[2 marks]

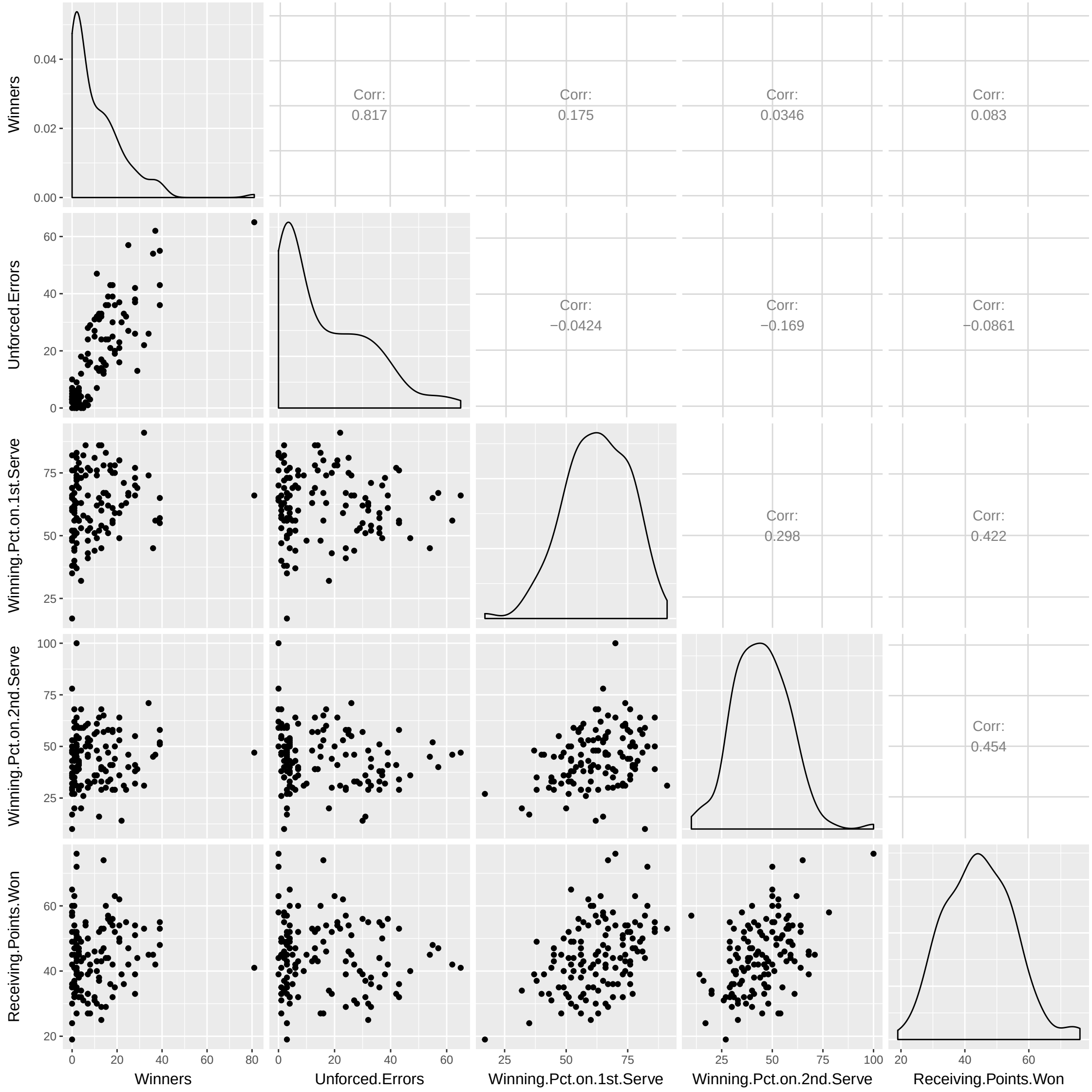
```
> vif(aus_lm)
                 Winners          Unforced.Errors Winning.Pct.on.1st.Serve
                3.661416                 3.652041                 1.334914
Winning.Pct.on.2nd.Serve     Receiving.Points.Won
                1.356313                 1.425647
```

vif is variance inflation factor, and measures multicollinearity between predictors. None of these values is very high. There doesn't look to be a problem with the model.

(e) Below is a plot of the predictors. Describe the association between the five variables. In an ideal fit what would the pattern look like?

[2 marks]

There is an outlier on winners. Other than this, there is some positive association between several variables but it is not very strong. Ideally the points should be scattered throughout the plot, that there is no association between these variables.



[Total: 10 marks]

— END OF QUESTION 6 —

QUESTION 7

This question is about modeling risk and loss.

(a) Monty Hall game: Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

(a) What is the probability that door one is chosen by the player?

[1 marks]

1/3

(b) Given that the player chose door 1, and the car is behind door 2, what is the probability that host shows you door 2?

[1 marks]

0

(c) Given that the player chose door 1, and the car is behind door 1, what is the probability that host shows you door 2?

[2 marks]

1/2

(d) What are the possible decisions that the player has? Write down the decision table, with possible strategies and the outcomes that can result, and the probabilities of each occurring.

[3 marks]

Possible strategies are to stay with their original choice (A), or switch their choice (B).

	Outcomes	
Strategies	Win car	Don't win car
A	1/3	2/3
B	2/3	1/3

(e) Suppose the game has a cost involved in playing. The value of the car is $20,000, but you have to pay $5,000 to play the game. Fill in the payoff matrix with the gains obtained with each strategy.

[2 marks]

	Outcomes	
Strategies	Win car	Don't win car
A	$15,000	-$5,000
B	$15,000	-$5,000

(f) Compute the expected gain under each strategy?

[2 marks]

A: 15000/3-5000*2/3=1666.7; B: 15000*2/3-5000/3=8333.3

(g) What is the minimax principle? How would it apply to the Monty Hall game?

[2 marks]

The minimax approach means we are focused on trying to avoid the worst case, and minimise the losses in this situation. The worst case in Monty Hall is that they player does not win the car. To minimise this loss, we would use strategy B, because the chance of not winning is lower.

(h) A version of the Monty Hall game was played on TV in 1963. During this time the number of people who chose to switch was 179, out of 326 contestants. What does this tell you about their use of the best strategy?

[2 marks]

The proportion of contestants who switched is 0.5491, which is very close to 0.5. People were not generally using the best winning strategy.

(i) In the contemporary game Deal or No Deal a similar type of game to Monty Hall is played, with one significant difference, that the host does not know where the best prizes are hiding. Would the strategy of switch your choice work better than keep your choice in this situation? Explain.

[2 marks]

In this game, you obtain no advantage form the host's action, so you have the same chance of winning whether to switch or stay.

[Total: 17 marks]

— END OF QUESTION 7 —

QUESTION 8

This question is about Bayesian methods

(a) A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase “free money” is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention “free money”. What is the probability that it is spam?

[4 marks]

- $P(S) = 0.8$
- $P(F|S) = 0.10$
- $P(F|\neg S) = 0.01$

$$\begin{aligned} P(S|F) &= \frac{P(F|S)P(S)}{P(F)} \qquad (1) \\ &= \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|\neg S)P(notS)} \qquad (2) \\ &= \frac{0.10 \times 0.8}{0.10 \times 0.8 + 0.01 \times 0.2} \qquad (3) \\ &= 0.9756 \qquad (4) \end{aligned}$$

(b) When maximizing the posterior probability is equivalent to maximizing the likelihood? Proof it using the bayes rule.

[4 marks]

They are equivalent under an uninformative prior, i.e. $\pi(\theta) = 1$

$$\begin{aligned} \pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{f(x)} \\ &= f(x|\theta)k(x) \\ &\propto f(x|\theta) \end{aligned}$$

$\implies \text{Maximize}_\theta\ \pi(\theta|x) \equiv \text{Maximize}_\theta\ f(x|\theta)$

(c) Briefly explain when do we need Monte Carlo sampling methods to compute the posterior distribution.

[3 marks]

If we are using non-conjugate prior distributions, then the normalizing constant can be hard to compute, especially in high dimensions. In order to compute the posterior distribution, we need to use Monte Carlo sampling methods to estimate this integral.

[Total: 11 marks]

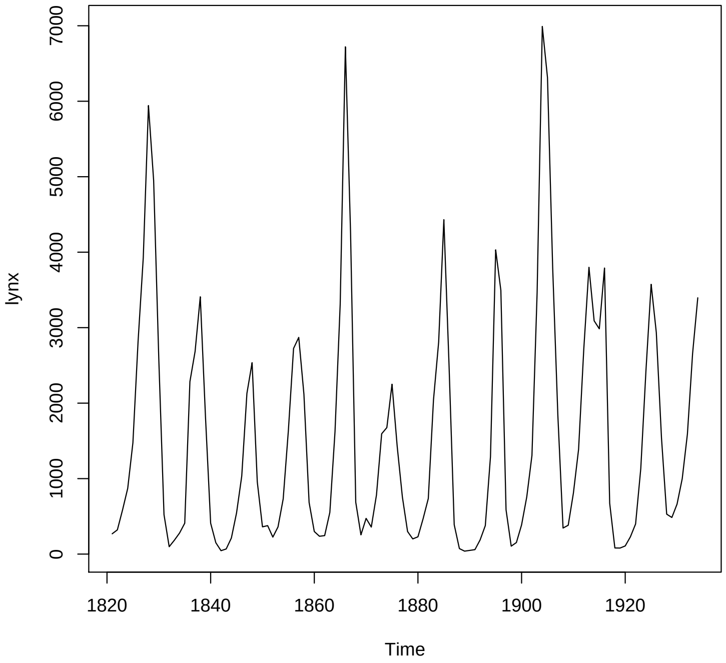
— END OF QUESTION 8 —

QUESTION 9

This question is about time series methods.

(a) Consider the annual numbers of lynx trappings for 1821-1934 in Canada given in the Figure below.

[3 marks]



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

The time series contains (aperiodic) cycles, but no obvious overall trend.

(b) Consider the following autoregressive process

$$Y_t = Y_{t-1} + Z_t$$

where $Y_1 = Z_1$ and $\mathbb{E}[Z_t] = \mu_Z$ and $\text{Var}(Z_t) = \sigma_Z^2$.

[4 marks]

Compute the mean function and the variance function.

$$\mathbb{E}[Y_t] = t\mu_Z \tag{5}$$

$$\text{Var}(Y_t) = t\sigma_Z^2 \tag{6}$$

Is this process strictly stationary? weakly stationary? why?

The process is neither strictly or weakly stationary since both the mean and variance changes with time.

(c) Consider the following non-stationary autoregressive process

$$Y_t = 2Y_{t-1} - Y_{t-2} + Z_t$$

where Z_t is a white noise process with $\mathbb{E}[Z_t] = 0$ and $\text{Var}(Z_t) = \sigma_Z^2$.

What transformation would make this process stationary? Show that after applying the transformation, the new process is indeed stationary.

[4 marks]

Second differencing would make it stationary.

We have

$$\begin{aligned} Y_t'' &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \quad (7) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \quad (8) \\ &= Z_t \quad (9) \end{aligned}$$

which is stationary.

[Total: 11 marks]

— END OF QUESTION 9 —