1. [10 marks] Suppose X and Y are two independent random variables such that $X \sim \mathsf{Poisson}(2)$ and $Y \sim \mathsf{Binomial}(5, 0.2)$. Define $W = X + 2Y$.

   (a) Compute $\mathbb{E}[W]$. [2 marks]

   Solution:

   $$\begin{aligned}\mathbb{E}[W] &= \mathbb{E}[X + 2Y] = \mathbb{E}[X] + 2\mathbb{E}[Y], \quad \text{[1 mark]} \\ &= 2 + 2 \times 5 \times 0.2 = 4\end{aligned}$$

   $\frac{1}{2}$ mark for each $\mathbb{E}[X]$ and $\mathbb{E}[Y]$

   (b) Compute $\mathrm{Var}(W)$. [2 marks]

   Solution:

   $$\begin{aligned}\mathrm{Var}(W) &= \mathrm{Var}(X + 2Y) = \mathrm{Var}(X) + \mathrm{Var}(2Y), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \mathrm{Var}(X) + 4\mathrm{Var}(Y), \quad [\tfrac{1}{2} \text{ mark}] \\ &= 2 + 4 \times 5 \times 0.2 \times 0.8 = 5.2\end{aligned}$$

   $\frac{1}{2}$ mark for each $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$

   (c) Compute $\mathrm{Cov}(X, W)$. [2 marks]

   Solution:

   $$\begin{aligned}\mathrm{Cov}(X, W) &= \mathrm{Cov}(X, X + 2Y) = \mathrm{Cov}(X, X) + \mathrm{Cov}(X, 2Y), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \mathrm{Cov}(X, X), \quad [\text{1 mark recognising} \mathrm{Cov}(X, 2Y) = 0] \\ &= \mathrm{Var}(X), \quad [\tfrac{1}{2} \text{ mark}] \\ &= 2\end{aligned}$$

   If the student lost $\frac{1}{2}$ mark in part (b) for gettting the variance of X wrong, don't penalise them again here if they have it wrong.

   (d) Compute $\mathbb{E}[W|X]$. [2 marks]

   Solution:

   $$\begin{aligned}\mathbb{E}[W|X] &= \mathbb{E}[X + 2Y|X] = \mathbb{E}[X|X] + 2\mathbb{E}[Y|X], \quad \text{[1 mark]} \\ &= X + 2\mathbb{E}[Y] = X + 2\end{aligned}$$

   $\frac{1}{2}$ mark for each $\mathbb{E}[X|X] = X$ and $\mathbb{E}[Y|X] = \mathbb{E}[Y]$. There is no additional mark for $\mathbb{E}[Y] = 1$. In particular, if they got $\mathbb{E}[Y]$ wrong in part (a), don't penalise them again here.

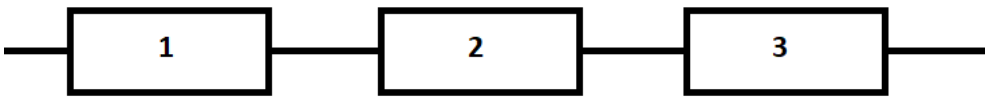
(e) Determine the moment generating function of W. [2 marks]

Solution:

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tW}] = \mathbb{E}[e^{t(X+2Y)}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{2tY}], \quad [\tfrac{1}{2} \text{ mark}] \\ &= M_X(t)M_Y(2t), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \exp(2(e^t - 1)) \times (1 - 0.2 + 0.2e^{2t})^5 \end{aligned}$$

$\frac{1}{2}$ mark for each of the moment generating functions $M_X(t)$ and $M_Y(2t)$.

2. [3 marks] Consider the system below comprised of three components. The system is working if there is a path from left to right through working components. The components fail independently of one another and the time to failure (in years) for each component has an Exponential(2) distribution. Determine the probability that the system is working at time t.



Solution: Let T_i be the time of failure for the i-th component. Then

$$\begin{aligned}\mathbb{P}(\text{Component } i \text{ working at time } t) &= \mathbb{P}(T_i > t) \\ &= \int_t^\infty 2e^{-2u}du = e^{-2t}, \quad \text{[1 mark]}\end{aligned}$$

It is sufficient to give this probability as e^{-2t} for the 1 mark. It is not necessary to show the additional working.

$$\begin{aligned}\mathbb{P}(\text{System working at time } t) &= \mathbb{P}(\text{all components working at time } t), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \mathbb{P}(T_1 > t, T_2 > t, T_3 > t), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \prod_{i=1}^{3} \mathbb{P}(T_i > t), \quad [\tfrac{1}{2} \text{ mark}] \\ &= e^{-6t} \quad [\tfrac{1}{2} \text{ mark}]\end{aligned}$$

3. [16 marks] A pair of random variables (X, Y) has a joint probability distribution in which Y has marginal probability density function

$$f_Y(y) = \begin{cases} 6y(1-y), & y \in (0,1) \\ 0, & \text{else} \end{cases}$$

and the conditional probability density function of X given $\{Y = y\}$ is uniform on the interval $(0, y)$.

(a) Write down the joint probability density function of (X, Y), clearly specifying the support of the distribution. [2 marks]

Solution: We have

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{y}, & x \in (0,y) \\ 0, & \text{else} \end{cases}$$

$\frac{1}{2}$ mark for writing the conditional pdf correctly. The joint probability density function is

$$\begin{aligned} f_{X,Y}(x,y) &= f_Y(y) f_{X|Y}(x|y), \quad \text{[1 mark]} \\ &= \begin{cases} 6(1-y), & y \in (0,1), x \in (0,y) \\ 0, & \text{else} \end{cases} \quad [\tfrac{1}{2} \text{ mark}] \end{aligned}$$

Don't penalise if students just write $6(1-y)$ for $y \in (0,1), x \in (0,y)$ and do not make it explicit that the pdf is zero outside this range. Same for the conditional pdf.

(b) Determine the marginal probability density function of X. [3 marks]

Solution:

$$\begin{aligned} f_X(x) &= \int f_{X,Y}(x,y)dy, \quad \text{[1 mark]} \\ &= \int_x^1 6(1-y)dy, \quad [\tfrac{1}{2} \text{ mark for the correct limits}] \\ &= \left[6y - 3y^2\right]_x^1 = 3 - 6x + 3x^2, \text{for } x \in (0,1) \end{aligned}$$

1 mark for the correct integration. $\frac{1}{2}$ mark for specifying the support as $x \in (0,1)$.

(c) Using the formula $\mathbb{E}[XY] = \mathbb{E}\left[Y\mathbb{E}\left[X \mid Y\right]\right]$ or otherwise, compute $\text{Cov}(X, Y)$. [5 marks]

Solution: First note

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \quad [\tfrac{1}{2} \text{ mark}]$$

To compute $\mathbb{E}[XY]$, we first determine the conditional expectation $\mathbb{E}[X|Y]$.

$$\mathbb{E}[X|Y] = \int x f_{X|Y}(x|y)dx = \int_0^y x \cdot \frac{1}{y} dx = \frac{y}{2}. \quad [1 \text{ mark}]$$

The explicit writing of the integral is not necessary to get the 1 mark since the mean of a uniform is given in the notes. Therefore,

$$\begin{aligned}\mathbb{E}[XY] &= \mathbb{E}[Y\mathbb{E}[X|Y]] = \mathbb{E}[\tfrac{1}{2}Y^2] \\ &= \int_0^1 \tfrac{1}{2}y^2 \cdot 6y(1-y)dy \\ &= \left[\tfrac{3}{4}y^4 - \tfrac{3}{5}y^5\right]_0^1 = 0.15 \quad [1 \text{ mark}]\end{aligned}$$

We also need $\mathbb{E}[Y]$ and $\mathbb{E}[X]$.

$$\mathbb{E}[Y] = \int y f_Y(y)dy = \int_0^1 y \cdot 6y(1-y)dy = \left[\frac{6}{3}y^3 - \frac{6}{4}y^4\right]_0^1 = 0.5, \quad [1 \text{ mark}]$$

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[\tfrac{1}{2}Y] = 0.25, \quad [1 \text{ mark}]$$

Putting this all together,

$$\text{Cov}(X, Y) = 0.15 - 0.5 \times 0.25 = 0.025, \quad [\tfrac{1}{2} \text{ mark}]$$

Other comments: If students make errors in those parts getting 1 mark and the error is just in the integration, give $\frac{1}{2}$.

(d) Suppose $Z = -\log Y$ (where log denotes the natural logarithm). Determine the probability density function of Z, clearly specifying the support of the distribution of Z. [4 marks]

Solution: As Y has support on (0,1), Z has support on $(0, \infty)$ [1 mark].

We first express the cdf of Z in terms of the cdf of Y. For $z > 0$,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leqslant z) = \mathbb{P}(-\log Y \leqslant z), \\ &= \mathbb{P}(\log Y \geqslant -z), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \mathbb{P}(Y \geqslant e^{-z}), \quad [\tfrac{1}{2} \text{ mark}] \\ &= 1 - F_Y(e^{-z}), \quad [\tfrac{1}{2} \text{ mark}] \end{aligned}$$

To get the pdf of Z we differentiate the cdf of Z:

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z), \quad [\tfrac{1}{2} \text{ mark}] \\ &= \frac{d}{dz}(1 - F_Y(e^{-z})) \\ &= f_Y(e^{-z})e^{-z} = 6e^{-2z}(1 - e^{-z}), \text{ for } z > 0. \end{aligned}$$

1 mark for correct differentiation. Don't penalise if they leave off the $z > 0$ part.

Note: After showing $F_Z(z) = \mathbb{P}(Y \geqslant e^{-z})$, students may explicitly determine $F_Z(z)$ as

$$F_Z(z) = \int_{e^{-z}}^{1} 6y(1-y)dy = \left[\frac{6}{2}y^2 - \frac{6}{3}y^3\right]_{e^{-z}}^{1} = 1 - 3e^{-2z} + 2e^{-3z}.$$

Give 1 mark if this is done correctly and 1 mark for getting the pdf by differentiating $F_Z(z)$.

Alternative solution: As Y has support on (0,1), Z has support on $(0, \infty)$ [1 mark].

Write $Z = g(Y)$, where $g(y) = -\log(y)$. Note $g'(y) = -y^{-1}$ [$\frac{1}{2}$ mark]. Now determine g^{-1}:

$$\begin{aligned} z = g(y) &= -\log(y) \\ e^{-z} = y &:= g^{-1}(z), \quad [1 \text{ mark}] \end{aligned}$$

Using the formula,

$$f_Z(z) = \left|\frac{1}{g'(g^{-1}(z))}\right| f_Y(g^{-1}(z)), \quad [\tfrac{1}{2} \text{ mark}]$$

we have

$$f_Z(z) = \left|\left(\frac{-1}{e^{-z}}\right)^{-1}\right| \cdot 6e^{-z}(1 - e^{-z}) = 6e^{-z}(1 - e^{-z}), \text{ for } z > 0.$$

[1 mark]

(e) The moment generating function of Z is

$$M_Z(s) = \frac{6}{6 - 5s + s^2}, \quad s < 2.$$

Using $M_Z(s)$ or otherwise, determine the expected value of Z. [2 marks]

Solution: We know $\mathbb{E}[Z] = M_Z'(0)$ [$\frac{1}{2}$ mark]. Then

$$M'(s) = \frac{-6 \times (-5 + 2s)}{(6 - 5s + s^2)^2}, \quad \text{[1 mark]}$$

and evaluating this at $s = 0$ gives $M'(0) = \frac{30}{36} = \frac{5}{6}$ [$\frac{1}{2}$ mark].

4. [8 marks] A study investigated the effect of playing computer games on heart rate. Twenty four individuals were recruited into the study and randomly assigned to play either an M-rated game or a G-rated game, with twelve participants in each group. Each participant's heart rate was measured before and after playing the video game for 20 minutes. The G-rated video game group had an average change (after − before) in heart rate of 3.2 beats per minute (bpm) and sample standard deviation 4.5 bpm. The M-rated video game group had an average change (after − before) in heart rate of 6.9 beats per minute (bpm) and sample standard deviation 5.7 bpm.

   (a) Do M-rated video games raise the heart rate more than G-rated video games? State the null and alternative hypotheses, and use an appropriate test statistic to determine the P-value. Based on the statistical test, what do you conclude? [5 marks]

   Solution: Let μ_M and μ_G denote the respective mean changes in heart rate after playing an M-rated/G-rated video game. We test

   $$H_0 : \mu_M = \mu_G, \quad \text{against} H_1 : \mu_M > \mu_G$$

   1 mark for correctly giving the null and alternative hypotheses. Give $\frac{1}{2}$ mark if the given alternative is two-sided.

   We need the pooled sample variance

   $$s_p^2 = \frac{(12-1)\times(4.5)^2 + (12-1)\times(5.7)^2}{12+12-2} = 26.37. \quad \text{[1 mark]}$$

   The test statistic is

   $$t = \frac{(\bar{x}_M - \bar{x}_G) - (\mu_M - \mu_G)}{s_p\sqrt{1/n_M + 1/n_G}} = \frac{(6.9-3.2)-0}{\sqrt{26.36}\sqrt{1/12+1/12}} = 1.7649. \quad \text{[1 mark]}$$

   The p-value is $\mathbb{P}(T_{22} > t) = 0.0457$. [1 mark]. Deduct $\frac{1}{2}$ mark if the degrees of freedom is incorrect or using the normal distribution as reference.

   This is moderate evidence against H_0, suggesting M-rated video games increase heart rate by more than G-rated video games. [1 mark]. Give $\frac{1}{2}$ mark if they only say 'moderate evidence against H_0' or 'reject H_0' without giving any context.

   Note: Student's may use the Welch approach, though this was not covered in class. The test statistic is

   $$t = \frac{(\bar{x}_M - \bar{x}_G) - (\mu_M - \mu_G)}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_G^2}{n_G}}} = \frac{(6.9-3.2)-0}{\sqrt{4.5^2/12 + 5.7^2/12}} = 1.7649$$

   (same value as before). The degrees of freed using the Welch approach is 20.876 and the p-value is 0.0461. If they get the degrees of freedom wrong using this approach deduct 1 mark.

(b) Construct a 95% confidence interval for the mean increase in heart rate after playing a G-rated video game for 20 minutes. [3 marks]

Solution: The confidence interval is

$$\text{estimate} \pm (\text{critical value}) \times s.e.(\text{estimate}). \quad \text{[1 mark]}$$

The critical value is $t_{0.975,11} = 2.200985$ [1 mark] Deduct $\frac{1}{2}$ mark if the degrees of freedom is wrong or using a normal distribution. Deduct $\frac{1}{2}$ mark if the quantile is wrong.

The standard error is

$$\frac{s}{\sqrt{n}} = \frac{4.5}{\sqrt{12}} \quad [\tfrac{1}{2} \text{ mark}]$$

The 95% confidence interval is 3.2 ± 2.859 (beats per minute), equivalently (0.3408,6.059) [$\frac{1}{2}$ mark]. Either form is acceptable.

5. [8 marks] A study was conducted in Australia on household use of information technology. As part of this survey, 400 adults aged between 25 and 54 were asked if they have ever experienced loss or damage due to a computer virus.

   (a) Of the 211 males surveyed, 28 had experienced loss or damage due to a computer virus. Of the 188 females surveyed, 17 had experienced loss or damage due to a computer virus. Construct a 99% confidence interval for the difference in the proportion of males and females affected. [3 marks]

   Let p_M be the proportion of males who have experienced loss due to a computer virus and let p_F be the proportion of females who have experienced loss due to a computer virus. We want a 99% confidence interval for $p_M - p_F$. The confidence interval is

   $$(\widehat{p}_M - \widehat{p}_F) \pm (\text{critical value}) \times s.e.(\widehat{p}_M - \widehat{p}_F).$$

   The estimates are

   $$\widehat{p}_M = \frac{28}{211} = 0.133, \qquad \widehat{p}_F = \frac{17}{188} = 0.09. \quad \text{[1 mark]}$$

   The standard error is

   $$s.e.(\widehat{p}_M - \widehat{p}_F) = \sqrt{\frac{\widehat{p}_M(1-\widehat{p}_M)}{n_M} + \frac{\widehat{p}_F(1-\widehat{p}_F)}{n_F}} = 0.0314 \quad \text{[1 mark]}$$

   The critical value is $z_{0.995} = 2.576$ [$\frac{1}{2}$ mark] The 99% confidence interval for $p_M - p_F$ is

   $$(0.133 - 0.09) \pm 2.576 \times 0.0314$$

   So 0.0423 ± 0.0808 equivalently (-0.0385,0.1230) [$\frac{1}{2}$ mark]. Either form is acceptable.

(b) The table below relates the respondent's age and whether they experienced loss or damage due to a computer virus.

Age	Yes	No
25 – 34	12	136
35 – 44	12	120
45 – 54	19	101

Based on this table, is there evidence of an association between age and experiencing loss or damage due to a computer virus?

[5 marks]

Solution: We test

- H_0: There is no association between age and 'experiencing loss or damage due to a computer virus'
- H_1: There is some association between age and 'experiencing loss or damage due to a computer virus'.

I don't think I can deduct a mark for not stating the null and alternative. If students have nothing else sensible in their answer to this question give them 1 mark for this.

The expected counts are obtained as: (row totals) $\times$ (column totals) / total. [1 mark] for the indication of how to compute the expected counts. Students will most likely be using MATLAB/R to compute these so I don't expect to see working. If students give the correct expected counts in some form, they get the mark. The table of expected counts is

Age	Yes	No
25 – 34	15.91	132.09
35 – 44	14.19	117.81
45 – 54	12.9	107.1

The test statistic is

$$X^2 = \sum_i \frac{(e_i - o_i)^2}{e_i} = 4.687 \quad \text{[1 mark]}$$

If the expected counts are wrong, don't penalise for the test statistic if they have the correct form. The degrees of freedom is (no. of rows -1) $\times$ (no. of columns -1) $= 2$ [1 mark]. The p-value is $\mathbb{P}(\chi^2_2 \geqslant 4.687) = 0.09598$ [1 mark]. There is weak evidence against the null hypothesis, suggesting some association between age and 'experiencing loss or damage due to a computer virus'. [1 mark] Give $\frac{1}{2}$ mark if they only say 'weak evidence against H_0' or 'reject H_0' without giving any context. Students get the mark if the conclusion is consistent with the reported p-value.

Other comments: Students doing this problem by hand may have rounding errors. Please don't penalise students for those.

6. [10 marks] Twenty nine university students were recruited into a study on target acquisition times in 3D gaming environments. Each student was assigned a different target acquisition difficulty level as measured by the Index of Difficulty (IoD) and their average time taken to acquire the target was recorded (TargetTime). The data are displayed in the figure below together with the fitted least squares line.

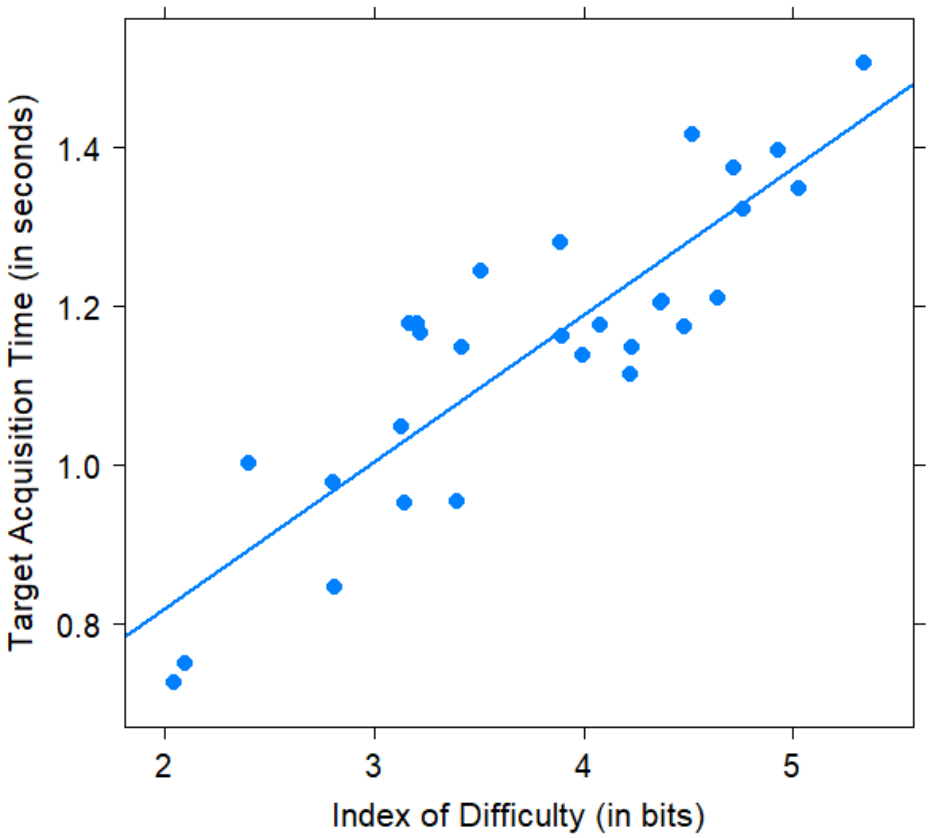


Figure 1: Plot of target acquisition times against index of difficulty.

The results of a linear regression fit for the relationship between TargetTime and IoD are given in the table below.

	Estimate	Std. Error
(Intercept)	0.42225	0.09725
IoD	0.19801	0.02526

(a) Briefly interpret the value 0.19801 in the regression output. [1 mark]

Solution: A 1 bit increase in IoD is associated with an increase of 0.198s in the mean time to acquire the target.

1 mark for this or something pretty similar. No marks for 'this is the slope' type answers. No part marks.

(b) Give a 95% confidence interval for the intercept term of the linear relationship between the time taken to acquire the target and IoD. [2 marks]

Solution: The confidence interval is

$$b_0 \pm (\text{critical value}) \times s.e.(b_0), \quad [\tfrac{1}{2} \text{ mark}]$$

where b_0 is the estimated intercept. The critical value is $t_{0.975,27} = 2.0518$ [1 mark] Deduct $\frac{1}{2}$ mark if the degrees of freedom is wrong or using a normal distribution. Deduct $\frac{1}{2}$ mark if the quantile is wrong.

So the 95% confidence interval is

$$0.42225 \pm 2.0518 \times 0.09725$$

which comes to 0.42225 ± 0.1995, equivalently $(0.2227, 0.6218)$ [$\frac{1}{2}$ mark]. Either form is acceptable..

Other comments: Maximum 1 mark if the student gives a confidence interval for the slope.

(c) Does the data provide evidence of a relationship between IoD and the time taken to acquire the target? State the null and alternative hypotheses, determine the appropriate test statistic and provide a bound on the P-value. Based on the statistical test, what do you conclude? [4 marks]

Solution: Let β_1 be the true slop in the linear model. We want to test

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0. \quad [1 \text{ mark}]$$

The test statistic is

$$t = \frac{b_1 - \beta_1}{s.e.(b_1)} = \frac{0.198 - 0}{0.0253} = 7.8389 \quad [1 \text{ mark}]$$

The p-value is $2 \times \mathbb{P}(T_{27} > 7.8389) < 2 \times 10^{-8}$. [1 mark] The p-value is very small. Don't worry if it is not precise. Deduct $\frac{1}{2}$ mark if wrong degrees of freedom or using the normal distribution.

There is strong evidence against the null hypothesis, suggesting a linear relationship between time to acquire target and index of difficulty. [1 mark] Give $\frac{1}{2}$ mark if they only say 'strong evidence against H_0' or 'reject H_0' without giving any context.

(d) The following figures were generated to check the assumptions underlying the linear regression. State the assumptions of the linear regression model and comment on their validity for this data with reference to Figures 1 and 2.
[3 marks]

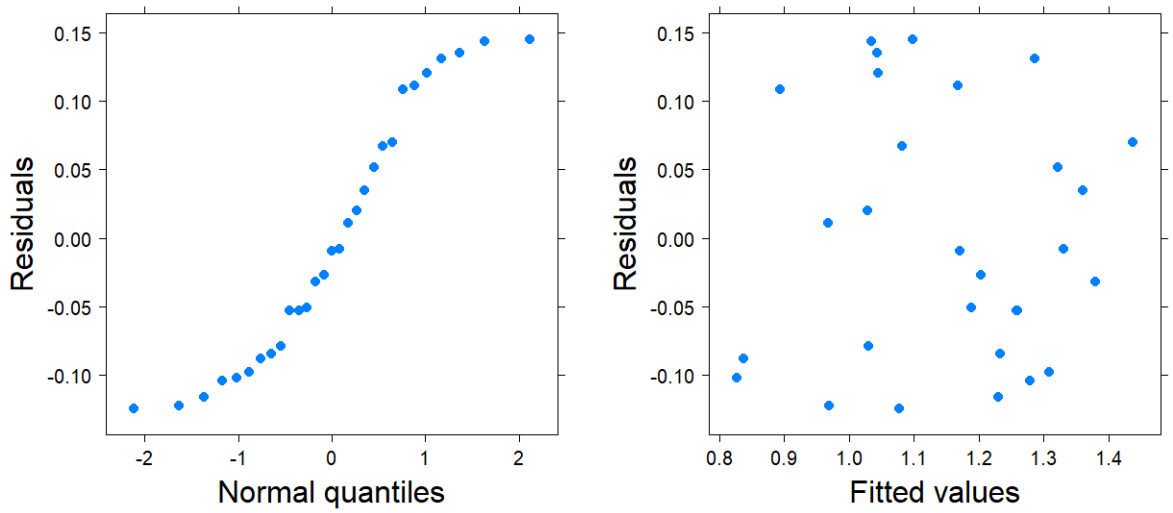


Figure 2: Left: Plot of residuals against quantiles of the standard normal distribution. Right: Plot of residuals against fitted values from the linear regression.

Solution: The assumptions of the linear regression model are:

- Linearity - The mean of the response is a linear function of the explanatory variable.
- Constant variance - The variance of the response is constant.
- Normality - The response has a normal distribution.
- Independence - The observations are independent.

[1 mark] for getting at least three of these. $\frac{1}{2}$ mark if only two are mentioned.

The residual vs normal quantiles plot shows a violation of the normality assumption. If the normality assumption held, then this plot would appear as roughly a straight line. The tails of the distribution appear lighter than we would expect from a normal distribution. [1 mark] Any statement that the plot is not straight so violates the normality assumption is fine.

The residual vs fitted values plot has no apparent trend which is consistent with the linearity assumption. The constant spread of points in this plot is consistent with the constant variance assumption. [1 mark] Any statement indicating that the linearity and constant variance assumptions are ok is fine. If only one of these assumptions is mentioned, then $\frac{1}{2}$ mark.

END OF EXAMINATION