Statistics Document 1

Thursday, August 24, 2023 6:39 PM

STATISTICS

90623

95123

#

Statistics Documen...

① MGF identification

i) Let M(t) be the Moment Generating Function (MGF) of a random variable Y. Given below are four MGFs written in terms of M(t) of four different random variables. Identify which one of the following is NOT a valid MGF.

A. M(t) * M(3t)

B. e^{-3t} * M(0.5t)

C. $\boxed{\frac{2}{3} * M(t)}$ → 2/3 = 1 $\frac{2}{3} \times 1 = \frac{2}{3} \neq 1$

D. $M(\frac{1}{3}t)$ M(0) = 1

for any R.V. @ t = 0 MGF = 1

Let X be a two-parameter exponential random variable such that X ~ Exp (λ, a). It has the following probability density function:

$$f(x) = \lambda e^{-\lambda(x-a)}, \quad x \geq a, \text{ where } \lambda, a > 0$$

$$M_X(t) = E(e^{tX}) = \int_a^\infty e^{tx} \lambda e^{-\lambda(x-a)} dx$$

$$= \int_a^\infty e^{\lambda a} \lambda e^{-x(\lambda-t)} dx$$

$$= e^{\lambda a} \lambda e^{-x(\lambda-t)} \Big|_a^\infty$$

$$= \frac{\lambda}{\lambda-t} e^{at} \quad \text{for } t < \lambda$$

$$= (1-t/\lambda)^{-1} e^{at}$$

Beta

Gamma

2 (ii) Show that the moment generating function of X is given by: $\left(1-\frac{t}{\lambda}\right)^{-1} * e^{at}$.

How $E(x) = M'_X(t)$

at 0

(iii) Using the result derived in part (ii), calculate E(X).

4) Which of the following is an expression representing the distribution function $F_x(X)$ of random variable X?

A. $e^{\lambda a} * (1 - e^{-\lambda x})$
B. $e^{\lambda a} * (e^{-\lambda a} - e^{-\lambda x})$ ✓
C. $e^{-\lambda x} * (1 - e^{-\lambda a})$
D. $e^{-\lambda a} * (e^{\lambda a} - e^{-\lambda x})$

Can be done by integrating the PDF for a to X.

If it results as

$$= e^{\lambda a} \times \left[e^{-\lambda a} - e^{-\lambda x} \right]$$

5

v) If $\lambda = 0.10$, $a = 0.05$, simulate two values from X using the distribution function as determined in part (iv) using values 0.226 and 0.304 from U (0, 1).

$$f_X(x) = u = e^{(\lambda a)}\left[e^{-\lambda a} - e^{-\lambda x}\right]$$

$$\frac{u}{e^{\lambda a}} - e^{-\lambda a} = -e^{-\lambda x}$$

$e^x = \exp(x)$

$$\ln\left[e^{-\lambda x}\right] = \ln\left[\left(e^{-\lambda a} - \frac{u}{e^{\lambda a}}\right)\right]$$

$$\Rightarrow -0.10.x = \ln\left[e^{-0.10\times0.05} - \frac{u}{e^{[0.10\times0.05]}}\right]$$

$$X = -10 \times \ln(1-u) \times e^{(-0.10.0.05)}$$

Ans

Put $u = 0.226$
$u = 0.304$ Ans

6

Q.6) A general insurance company is analyzing its portfolio of accidental insurance claims. The random variables X and Y represents the accidental loss amounts and allocated operating expenses respectively per policy.

Joint probability density function is given by:

$$f_{XY}(x, y) = 3/10^6\, e^{-(x/1000)}, \quad 0 < 3y < x < \infty$$
$$= 0 \quad \text{otherwise}$$

The industry has a practice of having ratio of 1:3(Y:X) between allocated operating expenses and accidental loss amounts. However, the Chief Financial Officer of the company feels that there is a comprehensive expenses management framework in the company and hence this ratio has fallen below 1:4.

i) Using the joint density function f_{XY} (x, y) given above, calculate the probability that the ratio has fallen below 1:4. You are given that current level of accidental loss amount per policy is 1,000.

$P\left(Y < \frac{X}{4}\right)$

$\Rightarrow P(Y < 250)$

$\Rightarrow P(3Y < X < \infty,\ 0 < Y < 250)$

$\int_0^{\infty}$

i) Using the joint density function $f_{XY}(x, y)$ given above, calculate the probability that the ratio has fallen below 1:4. You are given that current level of accidental loss amount per policy is 1,000.

$$= \int_0^{250} \left(\int_{3y}^{\infty} 3/10^6 \, e^{-\frac{x}{1000}} \, dx \right) dy$$

$$= \int_0^{250} \left(3/10^6 \, e^{-\frac{x}{1000}} \Big/ \left(-\frac{1}{1000}\right) \right)_{3y}^{\infty} dy$$

$$= \int_0^{250} \frac{3}{10^3} \left(e^{-\frac{3y}{1000}} \cdot e^{-\infty} \right) dy$$

$$= \frac{3}{10^3} \left(e^{-\frac{3y}{1000}} \Big/ \left(-\frac{3}{1000}\right) \right)_0^{250}$$

$$= e^0 - e^{-3(250/1000)}$$

$$= 1 - e^{-0.75} = 1 - 0.472$$

$$= 0.528$$

ii) Derive an expression for the marginal density function of random variable Y.

$$f(y) = \int_{x=3y}^{\infty} \frac{3}{10^6} e^{-\frac{x}{1000}} \, dx$$

90623
95123

$$= \frac{3}{10^6} \left(e^{-\frac{x}{1000}} \Big/ \left(-\frac{1}{1000}\right) \right)_{3y}^{\infty}$$

$$= \frac{3}{1000} e^{-3y/1000} \ldots$$

$\Rightarrow$ exponential Random Variable

Parameter is $-\frac{3}{1000}$

Random variable

Parameter is $-\frac{3}{1000}$

iii) Based on the expression obtained in part (ii) identify the correct distribution and parameters for random variable Y.

A. Y ~ Gamma(α=2, λ=1/1000) ✗
B. Y ~ Exp(λ=3/1000) ✓
C. Y ~ Gamma(α=2, λ=3/1000) ✗
D. Y ~ χ^2 with 3 degrees of freedom ✗

It is given that the marginal density function of random variable X i.e. $f_X(x)$ is given by the following expression:

$$f_X(x) = x/1000^2 e^{-x/1000} \quad 3y < x < \infty$$
$$= 0 \quad \text{otherwise}$$

iv) Justify why random variables X and Y cannot be independent based on $f_X(x)$ as given above and $f_Y(y)$ as determined in part (ii).

Rest

v) Identify the correct joint density function for which the random variables X and Y with marginal density functions as above are considered to be independent?

A. $g_{XY}(x, y) = 3x/10^9 \, e^{-(x+3y)/1000}$ $\quad 0 < 3y < x < \infty$
$\quad = 0$ $\quad$ otherwise

B. $g_{XY}(x, y) = 3x/10^6 \, e^{-(x+3y)/1000}$ $\quad 0 < 3y < x < \infty$
$\quad = 0$ $\quad$ otherwise

C. $g_{XY}(x, y) = 3xy/10^9 \, e^{-(x+3y)/1000}$ $\quad 0 < 3y < x < \infty$
$\quad = 0$ $\quad$ otherwise

D. $g_{XY}(x, y) = 3y/10^9 \, e^{(x-3y)/1000}$ $\quad 0 < 3y < x < \infty$
$\quad = 0$ $\quad$ otherwise

vi) *Suppose you decide to use the joint density function* $g_{XY}(x, y)$ as determined in part *(v), determine the conditional expectation* $E(Y \mid X > 950)$.

Q. 3)

i) Which of the following is a genuine difference between a "general" linear model and a "generalised" linear model?

Normal vs Non-Normal

~~Generalised~~ General   Generalised

A. The response variable is normally distributed in case of a "general" linear model whereas it can be non-normal in case of "generalised" linear models.

B. Link function can be used only for "generalised" linear models and cannot be used for a "general" linear model.

C. "Generalised" linear models can be non-linear in terms of the covariates whereas a "general" linear model has to be linear in terms of both parameters and covariates.

D. Interaction between various explanatory variables can be added in case of "generalised" linear models which is not possible in case of a "general" linear model.

Consider the discrete random variable Y with the following probability density function:

$f(y,\mu) = {}^{n}C_{ny} * \mu^{ny} * (1 - \mu)^{n-ny}$ where y = 0, 1/n, 2/n, 3/n, ………, 1

$$f(y) \Rightarrow \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

$$f(y) = {}^{n}C_{ny} \cdot \mu^{ny} \cdot (1-\mu)^{n-ny}$$

$$= \exp\left[n(y \ln\mu + (1-y)\ln(1-\mu)) + \ln({}^{n}C_{ny})\right]$$

$$= \exp\left[n(y\ln(\mu/(1-\mu)) + \ln(1-\mu)) + \ln({}^{n}C_{ny})\right]$$

$\theta = \ln(\mu/(1-\mu))$

$\phi = n$

$a(\phi) = 1/\phi$ $c(y,\phi) = \ln({}^{n}C_{ny})$

Using ~~ϕ~~ θ value $\mu = e^{\theta}/(1+e^{\theta})$

$b(\theta) = \ln(1+e^{\theta})$ …

$^{10}C_3$

$\boxed{4} \leftarrow$

$\boxed{10} \rightarrow \boxed{8}$

$\boxed{15}$

Negative Bin

ii) Show that the above discrete distribution belongs to the exponential family of distributions and specify different parameters of exponential family of distributions.

Let us define Z as a binomial variable such that Z = nY

$$E(Y) = \frac{e^{\theta}}{1+e^{\theta}} = \mu$$

$$E(y) = \frac{e^{\theta}}{1+e^{\theta}} = \mu$$

$$V(\mu) = \frac{e^{\theta}}{(1+e^{\theta})^2} = \mu(1-\mu)$$

$$V(y) = V(\mu) \times a(\phi) = \mu(1-\mu)/n$$

$$E(Z) = nE(y) = n \times \mu$$

$$V(Z) = n^2 V(y) \Rightarrow n\mu(1-\mu)$$

iii) Derive expressions for the mean and variance of the discrete distribution E(Z) and Var(Z), using your answer in part (ii). Also check whether the expressions match with mean and variance of Z ~ Bin(n,μ).

Bernoulli variable say → X wne μ → PD

$f(x) = \mu(1-\mu)$ for $x = 0, 1$

iv) Is the random variable Y a Bernoulli variable with parameter μ?

But PDF of Y is

$f(y, \mu) = {}^{n}C_{ny}\, \mu^{ny} (1-\mu)^{n-ny}$

$y = 0, 1/n, 2/n, 3/n \ldots$ - - - 1

Y is NOT Bernoulli

although mean, var of Y = var of Bernoulli

But this is not a sufficient condition ..

Rules of identification

(i) Single trial identi/prob

(ii) success remain const

logit link fns

v) Choose the canonical link function to be used in this case from the following options.

A. $g(\mu) = \log(\mu / (1 - \mu))$ ✓✓ establish connection

B. $g(\mu) = \mu$

C. $g(\mu) = \log \mu$

D. $g(\mu) = 1 / \mu$

mean → distr

Linear Combinations Predictor for variables..

Q. 4) A private tutor teaches actuarial subject CS2 in two batches – "Elite" and "Zenith". Those who have passed CS1 with more than 70 marks are a part of the "Elite" batch and others are a part of the "Zenith" batch.

He recently conducted a mock test for CS2 and is trying to analyse their performance by comparing the number of questions answered correctly. The mock question paper had a total of 7 questions and the data relating to number of correctly attempted questions is presented below:

Batch	"Elite"	"Zenith"
Total number of students	6	11
Total number of correct answers	33	37

He decides to use binomial distribution to model the number of questions correctly attempted by each candidate and defines two variables X_E and X_Z for "Elite" and "Zenith" batches respectively such that:

$$X_E \sim Bin(n=7, p_E)$$
$$X_Z \sim Bin(n=7, p_Z)$$

i) Which of the following is NOT a valid method used to determine a point estimate for the value of the unknown parameter using the information provided by above sample?

A. Method of Percentiles.
B. Non-Parametric Bootstrap Method.
C. Parametric Bootstrap Method.
D. None of the above.

ii) Determine the estimates $\hat{p}_E$ and $\hat{p}_Z$ and combined estimate $\hat{p}$ using the method of moments.

The tutor suggests an alternative model wherein he decides to use the method of maximum likelihood. He continues to use binomial distribution to model X_E and X_Z with $n = 7$, but for Elite Batch he uses the parameter 2Θ and for Zenith Batch he uses the parameter Θ, where $\Theta < 0.5$.

iii) Identify which one of the following corresponds to the log likelihood function of Θ given the observed data.

A. $\log L \propto 33 \text{ In } (2\Theta) + 9 \text{ In } (1\text{-}2\Theta) + 37 \text{ In } (\Theta) + 40 \text{ In } (1\text{-}\Theta)$
B. $\log L \propto 9 \text{ In } (2\Theta) + 33 \text{ In } (1\text{-}2\Theta) + 40 \text{ In } (\Theta) + 37 \text{ In } (1\text{-}\Theta)$
C. $\log L \propto 33 \text{ In } (\Theta) + 9 \text{ In } (1\text{-}\Theta) + 37 \text{ In } (2\Theta) + 40 \text{ In } (1\text{-}2\Theta)$
D. $\log L \propto 9 \text{ In } (\Theta) + 33 \text{ In } (1\text{-}\Theta) + 40 \text{ In } (2\Theta) + 37 \text{ In } (1\text{-}2\Theta)$

iv) Show using your answer in part (iii), that the maximum likelihood estimate for Θ is $\hat{\Theta}$ =0.412. You are NOT required to check that it is a maximum.

v) Without performing any formal test, by completing the below table, state which method gives a better estimate.

Total number of correct answers	**"Elite" Batch**	**"Zenith" Batch**
Method of Moments Estimate (using combined estimate $\hat{p}$)	_______	_______
Method of Maximum Likelihood Estimate	_______	_______
Observed Values	33	37

Distribution based ..

A telecommunications company operating more than 1,000 circles in the country is planning to perform Kaizen costing exercise to reduce its operational costs. Initially as a test run, the exercise is being done over 10 circles in the country. The operational costs data for these 10 circles under the traditional costing approach and Kaizen costing approach has been presented below:

↳ change for better

Costing Approach	Sample Size n	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} x_i^2$
Traditional	10	514.80	27804.64
Kaizen	10	401.40	18215.88

A statistical test is to be performed at the 5% level of significance, to determine whether Kaizen actually leads to cost reduction i.e. for:

H_0: There is no cost reduction versus H_1: There is cost reduction

i) Identify a suitable distribution for the test statistic.

A. t distribution
B. F distribution
C. Chi-square distribution
D. Normal distribution

$n < 10$ t

$n > 30$ large

determined by the Sample size ..

ii) Show that the 95% two-sided confidence interval for difference in mean operating costs is (-1.586, 24.266). What would you conclude in context of the null hypothesis?

$\bar{x}_A = 57.48$, $\bar{x}_B = 40.14$

$s_A^2 = \frac{1}{9}(27804.64 - 10 \cdot 57.48^2)$

$s_B^2 = \frac{1}{9}(18215.88 - 10 \cdot 40.14^2)$

Pooled var $s_p^2 = \frac{1}{18}(9 \times s_A^2 + 9 \times s_B^2)$

$95\%\ CI \Rightarrow (\bar{x}_A - \bar{x}_B) \pm t_{n_A+n_B-2} \times s_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$

df

=

If → (+, +) Suffn evidn to Rejct H_0

If → (−, +) Insuffn evidence to " H_0

→1 K

iii) Which one of the following is required to be assumed to perform the above mentioned test?

A. Normality of the sample and equal population variances
B. Normality of the sample and equal sample variances
C. Normality of the population and equal sample variances
D. Normality of the population and equal population variances

The 95% confidence interval as determined in part (ii) was presented to the CEO of the company and it was pitched to implement Kaizen costing over all the remaining circles in the country. The CEO was not convinced with the width of the confidence interval and suggested that the width should be reduced to below 20 in order to go ahead with full-blown implementation.

(9.30 − 10) (9.30 − 9.50)

suggested that the width should be reduced to below 20 in order to go ahead with full-blown implementation.

(9.30 - 10) (9.30 - 9.50)

(9.30 - 9.35)

iv) How many more circles shall be at least covered under the test run in order to reduce the width of the 95% confidence interval as determined in part (ii) as required by the CEO?

width of CI

$$2 \times t_{2.5\%, 2n-2} \sqrt{\frac{2}{n}} \sqrt{\frac{S_A^2(n-1) + S_B^2(n-1)}{n+n-1-1}}$$

$$= t_{2.5\%, 2n-2} \left(\quad \right)$$

Next Thursday

Economics fellow Public ..

v) At this reduced width, does your conclusion in part (ii) still hold true? Assume that the difference between the sample means remains the same.

Q. 6)

i) Generally, Kendal's Tau tends to be lower than Spearman's Rho. Consider n values for two rank variables R_x and R_y which have the following pairs:

(1, n), (2, n – 1), (3, n – 2) ………………………………… (n –2, 3), (n – 1, 2), (n, 1).

For the given values of the two rank variables, which of the following would be TRUE?

A. Both Kendal's Tau and Spearman's Rho would be equal to +1
B. Both Kendal's Tau and Spearman's Rho would be equal to –1
C. Both Kendal's Tau and Spearman's Rho would be positive but Kendal's Tau would be lower than Spearman's Rho
D. Both Kendal's Tau and Spearman's Rho would be negative but Kendal's Tau would be lower than Spearman's Rho

Inflation rates across ten different time periods for four developing economies were analysed and the sample Kendall's rank correlation coefficient between them was calculated as given in the below table:

Sample Kendall's Rank Correlation Coefficient Matrix

	Zubrowka	Freedonia	Genovia	Aldovia
Zubrowka	1.00	0.29	0.20	0.42
Freedonia	0.29	1.00	0.11	0.24
Genovia	0.20	0.11	1.00	0.16
Aldovia	0.42	0.24	0.16	1.00

ii) Using normal approximation, test whether sample Kendall's Rank correlation coefficient supports the hypothesis that the inflation rates for Freedonia and Genovia are positively correlated. Clearly state the null and alternate hypothesis, value of the statistic and conclusion at 0.05% level of significance.

Hint: *Variance = 2 (2n + 5) / 9n (n – 1)*

While calculating Kendall's Tau showing correlation between inflation rates for Freedonia and Genovia, the concordant and discordant pairs were calculated as follows:

Rank Freedonia	Rank Genovia	Number of Concordant Pairs	Number of Discordant Pairs
1	?	4	5
2	?	6	2
3	?	7	0
4	?	5	1
5	?	2	3
6	?	0	4
7	?	0	3
8	?	0	2
9	?	1	0
10	?	-	-
		25	**20**

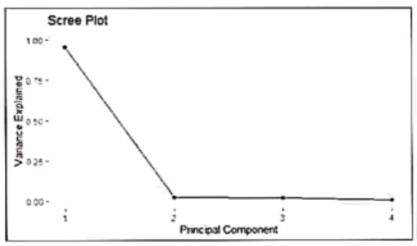
iii) Determine the values of "?" in the above table and hence calculate the sample rank correlation coefficient between inflation rates for Freedonia and Genovia using Spearman's method.

Principal Component Analysis was carried out to reduce the dimensionality of the inflation rates data-set in R using **prcomp** function using **scale = TRUE** and following extracts of the R output and scree-plot have been obtained:

```
Standard deviations (1, .., p=4):
[1] 1.9583973 0.3047250 0.2437022 0.1114980

Rotation (n x k) = (4 x 4):
                PC1         PC2        PC3         PC4
Zubrowka  -0.4993881  0.01113661 -0.8530755 -0.15083017
Freedonia -0.5015187 -0.48300910  0.3934253 -0.60033137
Genovia   -0.4932231  0.81303126  0.3066223 -0.04115716
Aldovia   -0.5057880 -0.32489745  0.1531718  0.78432047

Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.9584 0.30472 0.24370 0.11150
Proportion of Variance 0.9588 0.02321 0.01485 0.00311
Cumulative Proportion  0.9588 0.98204 0.99689 1.00000
```



iv) Determine the principal component(s) to be retained using the following criteria:

a) Retaining those components which represent at least 90% of the total variance;
b) Scree Test;
c) Kaiser Test.

v) The correlation coefficient between the PC1 and PC2 as determined above is -

A. -1
B. $+1$
C. 0
D. None of the above.

Q. 7) A Psephology Firm is conducting a survey to determine the proportion of people "p" who will vote for XJP political party in a particular municipal ward for the local body elections.

The Chief Psephologist is an Actuary and he is reviewing the responses collected by his team working on-field. A total of 200 responses have been collected. His prior beliefs about "p" based on historical vote share of XJP are given by a uniform distribution on the interval [0,1]. It turns out that after speaking to n_1 respondents, the n_1^{st} respondent happens to be a supporter of XJP. Others i.e. $(n_1 - 1)$ are non-supporters.

i) If n is a random variable that represents the number of respondents to be covered till one meets the first XJP supporter, then which of the following probability distributions would be suitable to model n?

A. Hyper-geometric Distribution
B. Geometric Distribution
C. Bernoulli Distribution
D. Binomial Distribution.

ii) Specify the posterior distribution of "p" after the actuary finds the first supporter of XJP political party.

iii) State the prior mean of "p" and calculate the posterior probability that "p" is greater than the prior mean using the posterior distribution determined in part (ii) and explain what it signifies. You may assume that $n_1 = 3$.

The second supporter is found after covering another n_2 respondents, third supporter after covering further n_3 respondents and so on until the 50^{th} supporter is found who happens to be the 200^{th} (last) respondent. In other words, $n_1 + n_2 + n_3 + \dots\dots\dots + n_{50} = 200$.

iv) *Show that after conducting the above survey, the posterior distribution of "p" is a Beta Distribution with parameters* $\alpha = 51$ and $\beta = 151$.

v) Determine the Bayesian estimate of "p" under "Squared Error Loss" and express it in the form of a credibility estimate. Also determine the value of the credibility factor Z.

vi) Based on the value of the credibility factor calculated in part (v), which of the following statements would be a reasonable inference for the survey?

A. XJP is slated to face defeat in the local body elections in the ward under consideration.

B. Under no circumstances, XJP would be able to achieve its historical vote share in the municipal ward in the coming local body elections.

C. There is a need to increase the sample size of the survey in order to take a credible view on the vote share of XJP in the upcoming elections in the ward.

D. It is virtually certain that the vote share of XJP will be reduced to almost a half of its historical vote share based on the survey results.

Q. 8) You are working as the Chief Statistical Officer in the Ministry of Health Care, Government of Actuaria. Your team has collected the following data relating to average heart rate (Y) and average systolic blood pressure (X_1) and average diastolic blood pressure (X_2) for 10 major cities in Actuaria.

City	Average Heart Rate beats per minute (Y)	Average Systolic Blood Pressure mm Hg (X_1)	Average Diastolic Blood Pressure mm Hg (X_2)
Orbit City	94	139	84
Emerald City	77	125	90
Shangri-La	81	126	81
Tomorrow Land	76	129	87
Kingsbury	63	112	94
Rivendell	76	118	78
Atlantis	91	153	92
Cloud City	73	104	70
Dark City	88	124	72
Thugs Mansion	84	134	79

The following multiple linear regression model was used to analyse the above data where Y was the response variable and X_1 and X_2 were the explanatory variables:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

The model was fitted in R and extracts from the R output for this model are given below:

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0127 -1.5872 -0.1759  1.7446  5.8372

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.56101   12.75027   3.730 0.007357 **
X1           0.69237    0.08818   7.852 0.000103 ***
X2          -0.66235    0.14896  -4.446 0.002985 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.322 on 7 degrees of freedom
Multiple R-squared:  0.9005, Adjusted R-squared: ????
F-statistic: 31.66 on 2 and 7 DF,  p-value: 0.0003113
```

i) Select the equation for the fitted multiple linear regression model using the extracts given above from the following four options:

A. $y = -4.0127 - 1.5872\ x_1 - 0.1759\ x_2 + e$
B. $y = 47.56101 + 0.69237\ x_1 - 0.66235\ x_2 + e$
C. $y = 12.75027 + 0.08818\ x_1 + 0.14896\ x_2 + e$
D. $y = 3.7300 + 7.8520\ x_1 - 4.4460\ x_2 + e$

ii) Calculate Adjusted R^2 *for the model.*

iii) Calculate the predicted average heart rate for Emerald City and Dark City. Also determine the value of residuals.

Some of the eminent officials from the ministry who are medical doctors by profession are of the opinion that instead of considering the levels of systolic and diastolic blood pressure, "pulse pressure" i.e. the difference between systolic and diastolic blood pressures should be considered as the explanatory variable for predicting the value of the average heart rate.

Let us define Average Pulse Pressure as Z where $Z = X_1 - X_2$.

$$\bar{y} = 80.30$$

$$\bar{z} = 43.70$$

$$\sum(y - \bar{y})^2 = 776.10\,;$$

$$\sum(z - \bar{z})^2 = 1472.10\,;$$

$$\sum(y - \bar{y})(z - \bar{z}) = 1013.90$$

iv) Using a bivariate linear regression model as given below, calculate the least square estimates of λ and μ and re-calculate the predicted average heart rate for Emerald City and Dark City clearly stating the value of residuals.

$$y = \lambda + \mu z + e$$

v) Calculate Adjusted R^2 for the bivariate regression model.

vi) What is the expected reduction in pulse pressure which will ensure a decrease in the heart rate by 2? Answer with reference to the bivariate regression model fitted in part (iv).

A. 3.51
B. 0.52
C. 1.38
D. 2.90

Your twenty-one year old son who is a freshly qualified statistics graduate has suggested that instead of using Z as the explanatory variable, using deviations from its mean defined as $W = (Z - \bar{Z})$ would provide a better fit.

The improvised bivariate linear regression model using W as the explanatory variable is given below:

$$y = \delta + £w + e$$

vii) Show that the least square estimators of the parameters of the improvised bivariate model are given by:

a) $\hat{\pounds} = \hat{\mu}$

b) $\hat{\delta} = \hat{\lambda} + \hat{\mu}\,\bar{z}$

where $\hat{\lambda}$ and $\hat{\mu}$ are the parameter estimators of the model stated in part (iv).

Hint: *Kindly note that* $\bar{w} = 0$.

Using the improvised bivariate linear regression model in part (vii), predicted average heart rate for Emerald City and Dark City and associated residuals were recalculated as follows:

City	$\hat{y}$	e
Emerald City	74.3079	2.6921
Dark City	86.0166	1.9834

Adjusted R^2 for the improvised bivariate model is 88.72%.

viii) Compare the three models based the predictions for Emerald City and Dark City and the values of Adjusted R^2 and briefly state your observations.