# MANICALAND STATE UNIVERSITY OF APPLIED SCIENCES

## FACULTY OF ENGINEERING, APPLIED SCIENCES & TECHNOLOGY

### DEPARTMENT OF APPLIED STATISTICS

### MODULE: INTRODUCTION TO STATISTICS

### CODE: ASTA 101

### SESSIONAL EXAMINATIONS

### APRIL 2023

### DURATION: 3 HOURS

### EXAMINER: MR A CHAKAIPA

---

***INSTRUCTIONS***

1. *Answer **All** in Section A*
2. *Answer **three** questions in Section B.*
3. *Start a new question on a fresh page*
4. *Total marks 100*

***Additional material(s):*** Graph paper, *Non-programmable electronic scientific calculator, Statistical Tables.*

---

## SECTION A [40 MARKS]

*Answer **ALL** questions in this section*

A1

Define the following terms as used in Statistics

a.Confidence limits.

b.Coefficient of determination.

c.Two tailed test.

d.Critical values.

**[2, 2, 2, 2]**

A2

a) State any three assumptions of simple linear regression analysis.

b) A golfer observes that, when playing a particular hole at his local course, he hits a straight drive on 80 percent of the occasions when the weather is not windy but only on 30 percent of the occasions when the weather is windy. Local records suggest that the weather is windy on 55 percent of all days.
i.Show that the probability that, on a randomly chosen day, the golfer will hit a straight drive at the hole is 0.525
ii Given that he hits a straight drive at the hole, calculate the probability that the weather is windy.

**[3, 4, 5]**

**A3**

A company that manufactures wooden products (e.g garden furniture, ladders, benches) regularly maintains its lathe machines which are then used for cutting and shaping components. The manager would like to know whether the cost of machine maintenance is related is related to the age of the machines. For a

random sample of 12 lathe machines in the company's factory, the annual maintenance cost and age of each machine was recorded.

| Machine | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age(yrs) | 4 | 3 | 3 | 8 | 6 | 7 | 1 | 1 | 5 | 2 | 4 | 6 |
| Annual_cost($) | 45 | 20 | 38 | 65 | 58 | 50 | 16 | 22 | 38 | 26 | 30 | 35 |

i)Identify the independent variable and the dependent variable. Explain.

ii)Show the data graphically in a scatter plot. What relationship is observed?

iii)Compute the Pearson correlation coefficient between the age of lathe machines and their annual maintenance costs. Comment on the strength of the association.

iv)Use the method of least squares to find the best fitting line between the age of lathe machines and their annual maintenance costs.

v)What is the expected average maintenance cost of a lathe machine that is 5 years old? **[2, 2, 4, 4, 6, 2]**

## SECTION B [60 MARKS]

*Answer any **THREE** questions in this section*

**B4.**

In 1965, data on the connection between radioactive waste exposure and cancer mortality was published. The data was collected from 9 counties that were located near an Atomic Energy Commission facility in Hanford, Washington.

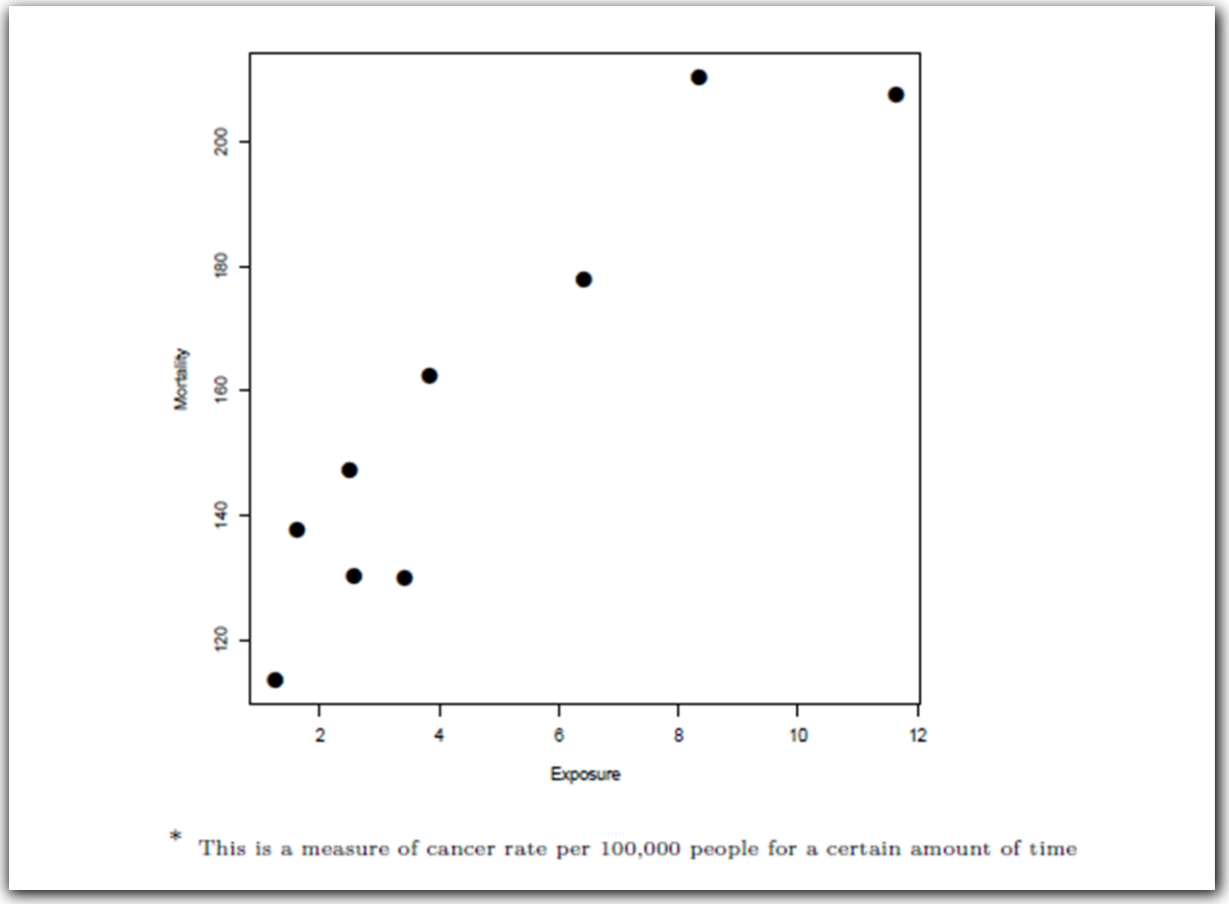
The data give the index of exposure and the cancer mortality rate during 1959-1964 for the nine counties affected. Higher index of exposure values represents higher levels of contamination.

| | | |
|---|---|---|
| *Variable Description* | *Country* | *Name of country* |
| | *Exposure* | *Index of exposure* |
| | *Mortality* | *Cancer mortality per 100,000 man-years* |

The data is as follows:

| | ***Country*** | ***Exposure*** | ***Mortality*** |
|---|---|---|---|
| *1* | *Umatilla* | *2.49* | *147.1* |
| *2* | *Morrow* | *2.57* | *130.1* |
| *3* | *Gilliam* | *3.41* | *129.9* |
| *4* | *Sherman* | *1.25* | *113.5* |
| *5* | *Wasco* | *1.62* | *137.5* |
| *6* | *HoodRiver* | *3.83* | *162.3* |
| *7* | *Portland* | *11.64* | *207.5* |
| *8* | *Columbia* | *6.41* | *177.9* |
| *9* | *Clatsop* | *8.34* | *210.3* |

The scatterplot:




Output from fitting the simple linear regression for predicting Mortality from Exposure is shown below:

```
> lm.out=lm(Mortality~Exposure)



> summary(lm.out)

Call:
lm(formula = Mortality ~ Exposure)

Residuals:
    Min      1Q  Median      3Q     Max
-16.295 -12.755   4.011   9.398  18.594

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  114.716      8.046  14.258 1.98e-06 ***
Exposure       9.231      1.419   6.507 0.000332 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.01 on 7 degrees of freedom
Multiple R-Squared: 0.8581,Adjusted R-squared: 0.8378
F-statistic: 42.34 on 1 and 7 DF,  p-value: 0.0003321



> anova(lm.out)
Analysis of Variance Table

Response: Mortality
          Df Sum Sq Mean Sq F value    Pr(>F)
Exposure   1 8309.6  8309.6  42.336 0.0003321 ***
Residuals  7 1373.9   196.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

**(a)** Write down the fitted regression line.

**(b)** What is the expected mortality rate for a country with an exposure index of 3.

**(c)** Is there a significant linear relationship between Mortality and Exposure? Provide a null hypothesis, alternative hypothesis, a test statistic, rejection criteria and conclusion.

**(d)** What is the estimated variance of the observations?

**(e)** Compute the Total Sums of Squares (SST) for this data?

**(f)** Compute the coefficient of correlation, $r$ and explain the nature and strength of the relationship between Mortality and Exposure.

**[3, 2, 5, 3, 3, 4]**

**B5**

**a)** Show that the least squares estimate $(\widehat{\beta_0}, \widehat{\beta_1})$ are unbiased estimates of $\beta_0$ and $\beta_1$ respectively.

**b)** Derive the variances of the estimates

Show that i) Var $(\widehat{\beta_1}) = \frac{\sigma^2}{S_{xx}}$

ii) Var $(\widehat{\beta_0}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$

where $\sigma^2$ is the Mean Square Error.

$S_{xx}$ is the variance of the independent variable X.

n is the number of data points.

**[4, 4, 6, 6]**

**B6**

a)Keffalos launched a new flavored ice cream. The marketing manager now wants to assess the product's success in the market place. If average sales per week were less than R5 650 per outlet, the product would be withdrawn. The results from a sample of 15 supermarket outlets countrywide showed that average sales per week were R5 280 with a sample standard deviation of R788.

Should the new pudding flavor be withdrawn? Advise the marketing manager by performing an appropriate statistical hypothesis test at the 10% level of significance. State an appropriate test and justify its use.

b) The quality control manager in a tyre manufacturing plant in Port Elizabeth wants to test a belief that the nature of defects found in the manufactured tyres depends upon the shift during which the defective tyre is produced. He therefore compiled the following two way table showing the number of defective tyres identified by shift and by nature of defect (i.e technical(operator induced), mechanical(machine fault), material(raw material quality)).

Is there evidence to substantiate a claim that there is a statistical association between the nature of defective tyres and the shift on which they are produced? Perform a statistical test at the 5% level of significance to establish if the two criteria are statistically independent or not.

| **Shift** | | **Nature of tyre defect** | | |
|---|---|---|---|---|
| | **technical** | **mechanical** | **material** | **Total** |
| **Morning** | **15** | **42** | **11** | **68** |
| **Afternoon** | **26** | **40** | **20** | **86** |
| **Night** | **29** | **25** | **14** | **68** |
| **Total** | **70** | **107** | **45** | **222** |

**[10, 10]**

**B7**

The distance travelled (in kms) by a courier service motorcycle on 50 trips was recorded by the driver.

| | | | | |
|---|---|---|---|---|
| 24 | 30 | 20 | 6 | 28 |
| 18 | 19 | 22 | 26 | 31 |
| 27 | 22 | 56 | 14 | 24 |
| 18 | 20 | 34 | 29 | 24 |
| 19 | 10 | 17 | 11 | 40 |
| 13 | 28 | 26 | 18 | 16 |
| 23 | 17 | 46 | 21 | 20 |
| 21 | 13 | 15 | 20 | 9 |
| 23 | 52 | 17 | 35 | 29 |
| 15 | 27 | 18 | 8 | 22 |

a) Construct a numerical frequency table of the distance travelled by a courier driver using the classes: 0-9, 10-19, 20-29, - - -

b) Draw a cumulative frequency curve(ogive).

c)Using the frequency distribution, calculate:

i)Mean

ii)Median

iii)Standard deviation

d)Calculate the coefficient of variation of the above distribution and interpret.

**[3, 4, 3, 4, 4, 2**]

**END OF QUESTION PAPER**