

STATISTICS AND PROBABILITY

Assignment 1: Descriptive Statistics and Relationships between Variables

In this assignment you are required to:

- collect single variable numerical data (obtain from a publication, or use data that you have generated yourself), and illustrate its statistical properties.
- collect real (numerical) bivariate data and calculate a correlation coefficient and regression equation, then make predictions. You should have reason to believe that there is an underlying linear relationship between the variables (almost certainly not a perfect straight line relationship).

* **NB: BEFORE CARRYING ANY STATISTICAL ANALYSIS ON YOUR BIVARIATE DATA, A SCATTERPLOT OF THE DATA MUST BE PLOTTED SO THAT YOU ENSURE THAT A STRAIGHT LINE OF BEST FIT CAN BE DRAWN THROUGH YOUR DATA. THE DATA SHOULD RELATE TO AN AREA OF RESOURCE/MANAGEMENT.**

FILE TECHNOLOGY

Present original data in tables.

DESCRIPTIVE STATISTICS

1. Using grouped or ungrouped data for a single variable, explain the **purpose** of your data and **how** the sample was obtained. In general, **why** is it necessary to carry out a sampling procedure?
2. Present your data graphically and/or pictorially in the most appropriate way.
3. Determine all possible measures of central tendency and spread and choose (with reasons) those that you think are most appropriate to your data. Results must be determined without using calculator or computer statistical functions, but you are encouraged to check your results using these functions.

RELATIONSHIPS BETWEEN VARIABLES

1. Plot the bivariate data that you have obtained on a scatter diagram and estimate trends and the degree of correlation between the variables by inspection.
2. Calculate, then check by re-calculation, the correlation coefficient using a calculator (Optional: check using the table method) and discuss the extent to which a linear relationship exists between the variables.
3. Determine the equation to the regression line for the data using a calculator, then accurately (using \hat{x} and/or \hat{y}) plot the regression line on the scatter diagram.
4. Use the regression equation to make two predictions of values for your data.
5. Describe the limitations of the use of regression lines for making predictions.

TUTORS REPORT:

- It would have been more beneficial to obtain real data from an area of Fire Technology for BOTH sections.

You must do calculations for spread of data, and comment on which measures of central tendency and spread you think are most appropriate. (Descriptive Statistics)
Check sense of your answer to Q5 on last page

→ you are required to address these points and resubmit.

TUTOR



MARK

CN

Statistics and Probability
Assignment 1
Descriptive Statistics

The following is the frequency distribution of a random sample of hourly earnings of employees obtained from a publication; total 509 employees

Purpose: To calculate the average hourly earnings and analyse the data from the average earnings.

Hourly earnings (\$)	Number of Employees
10	3
12	6
14	10
16	15
18	24
20	42
22	75
24	90
26	79
28	55
30	36
32	26
34	19
36	13
38	9
40	7

Name of publication? "
" " business"
Date?
i.e. need specific detail.
(This looks a bit like a textbook exercise!)

Hourly Earnings	Mid Value	No. of Employees	F*x	Step Deviation Method $u = (x-25)/2$	F*u
10-12	11	3	33	-7	-21
12-14	13	6	78	-6	-36
14-16	15	10	150	-5	-50
16-18	17	15	255	-4	-60
18-20	19	24	456	-3	-72
20-22	21	42	882	-2	-84
22-24	23	75	1725	-1	-75
24-26	25	90	2250	0	-398
26-28	27	79	2133	1	79
28-30	29	55	1595	2	110
30-32	31	36	1116	3	108
32-34	33	26	858	4	104
34-36	35	19	665	5	95
36-38	37	13	481	6	78
38-40	39	9	351	7	63
40-42	41	7	287	8	56
					693

$Ef = 509;$

$Efx = 13315;$

$Efu = 295$

$\text{MEAN } \bar{x} = \frac{\sum fx}{N} = \frac{13315}{509} = 26.16$

- L = Lower Limit of Median class
- N = Total Frequency
- h = Width of the Median class
- c = Cumulative frequency up to the class preceding the Median class

$\text{MEDIAN} = \frac{L + \frac{N}{2} - c}{f} \times h$

CI	Frequency	Cumulative Frequency
10-12	3	3
12-14	6	9
14-16	10	19
16-18	15	34
18-20	24	58
20-22	42	100
22-24	75	175
24-26	90	265
26-28	79	344
28-30	55	399
30-32	36	435
32-34	26	461
34-36	19	480
36-38	13	493
38-40	9	502
40-42	7	509

Median Class

$$= 24 + \frac{509 - 175}{2} \times 2$$

$$= 24 + \frac{334}{2} \times 2$$

$$= 24 + 167 \times 2$$

$$= 24 + 334$$

$$= 358$$

MEDIAN = 358

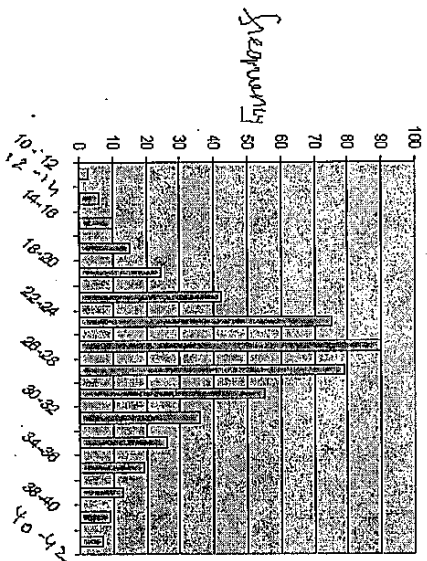
$$MODE = 24 + \frac{90 - 75}{90 - 75 + (90 - 79)} \times 2$$

$$= 24 + \frac{15}{26} \times 2 = 24 + \frac{30}{26}$$

$$= 24 + 1.154$$

$$= 25.154$$

Δ_1 = Excess of model frequency over frequency of preceding class
 Δ_2 = Excess of model frequency over frequency of preceding class
 h = Size of model class



Class interval

- Bar graph

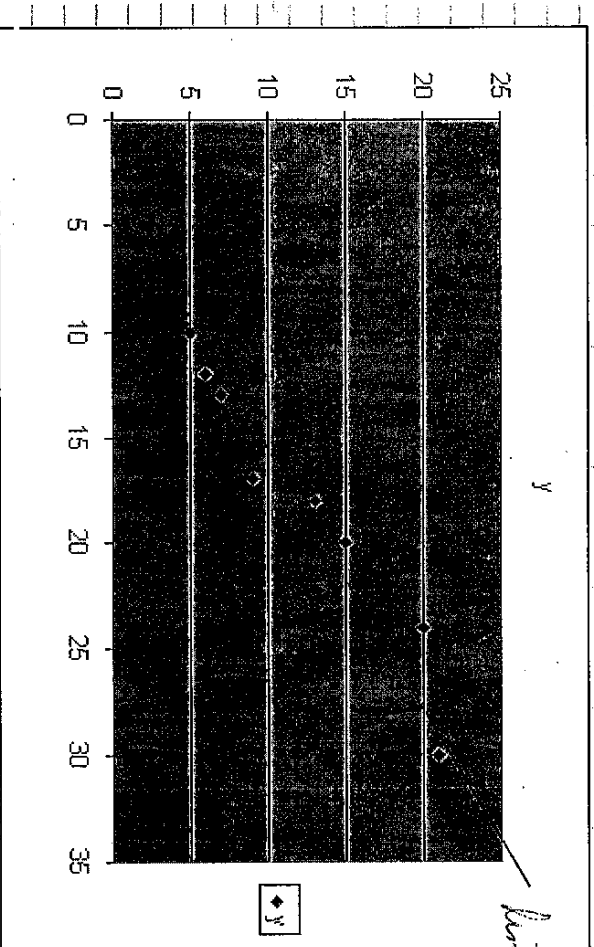
Relationships between Variables

1.

Scatter diagram

X	y
10	5
12	6
13	7
17	9
18	13
20	15
24	20
30	21

WHERE IS THIS DATA OBTAINED ?
Once again it looks like a textbook exercise.



From the scatter diagram we find that as x increases y also increases that is there is positive relation ship between x and y.

2.

The marks obtained by 8 students in Maths and science are given below;
 Maths = x Science = y

x	y	$x-18=x$	x^2	$y-12=y$	y^2
10	5	-8	64	-7	49
12	6	-6	36	-6	36
13	7	-5	25	-5	25
17	9	-1	1	-3	9
18	13	0	0	1	1
20	15	2	4	3	9
24	20	6	36	8	64
30	21	12	144	9	81

$$\Sigma x = 114 ; \Sigma xy = 282$$

$$\Sigma x^2 = 310$$

$$\Sigma y = 0$$

$$\Sigma y^2 = 274$$

Regression equation of y on x:

$$y - \hat{y} = b_{yx} (x - \hat{x})$$

$$b_{yx} = \frac{\Sigma yx}{\Sigma x^2} = \frac{282}{310} = 0.91$$

$$y - 12 = 0.91 (x - 18)$$

$$y - 12 = 0.19x - 16.38$$

$$y = -4.38 + 0.91x$$

Regression equation of x on y:

$$x - \hat{x} = b_{xy} (y - \hat{y})$$

$$x - 18 = b_{xy} (y - 12)$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{282}{274} = 1.029$$

P.T.O

$$x - 18 = 1.029(y - 12)$$

$$x = 1.029y + 5.652$$

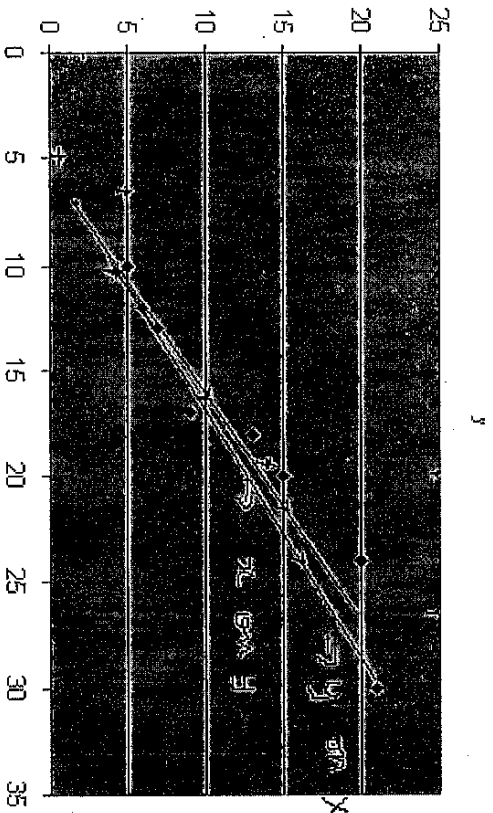
Coefficient of Correlation (r)

$$\begin{aligned} \therefore r &= 0.98 \\ \therefore r &= 0.98 \end{aligned}$$

$$0.97$$

This means marks in maths and science scored by students are closely related.

3.



4.

x on y

$$x = 1.029y + 5.652$$

$$\text{say } y = \boxed{12} \boxed{8}$$

$$x = 1.029(12) + 5.652 = \boxed{18}$$

and

$$x = 1.029(8) + 5.652 = \boxed{13.884}$$

$$y = -4.38 + 0.91x$$

$$x = 8, 12$$

$$y = -4.38 + 0.91(8) = 2.9$$

and

$$y = -4.38 + 0.91(12) = 6.54$$

5.

In making estimates from a regression it is important to remember that the assumption is being made that the relationship has not

Change since the regression was computed, another point worth remembering is that the relationship shown by the scatter diagram may not be the same if the equation is extended beyond the values in computing the equation, for example there may be a close relationship between the yield of a crop and the amount of fertilizers applied, with the yield increasing as the amount of fertilizer is increased, it would not be logical, however to extend this equation beyond the limits of the experiment for it is quite likely that if the amount of fertilizer were increased indefinitely, the yield would eventually decline as too much fertilizer was applied.