CORRELATION

Dr. Asheesh Srivastava
Professor, Head & Dean
Department of Educational Studies
School of Education,
Mahatma Gandhi Central University,
Motihari, East Champaran, Bihar-845401
profasheesh@mgcub.ac.in

❖ The idea of correlation

X and Y work for a company. X drives a Toyota, costing Rs. 8,00,000/- and Y drives an Maruti 800, costing 3,00,000/-. Which man has the greater salary?

In this case, we can reasonably assume that it must be X who earns more, as he drives the more expensive car. As he earns a larger salary, the chances are that he can *afford* a more expensive car.

We can't be absolutely certain, however. It could be that X's Toyota was a gift from a friend or he could have stolen it! However, most of the time, an expensive car means a larger salary.

In this case, we say that there is a (relationship) *correlation* between someone's salary and the cost of the car that he/she drives. This means that as one figure changes, we can expect the other to change in a fairly regular way.

- Another Question

➢ **Both X and Y are married. Which man has more children?**

Ah! This is a different matter. How much a man earns does not influence how many children he has (as far as I feel). In this case, we say that there is no correlation between a person's salary and how many children he/she has. As one figure changes, we cannot say how the other figure will change.

Scatter diagrams

Assumptions

1. The sample of paired data (x, y) is a random sample.

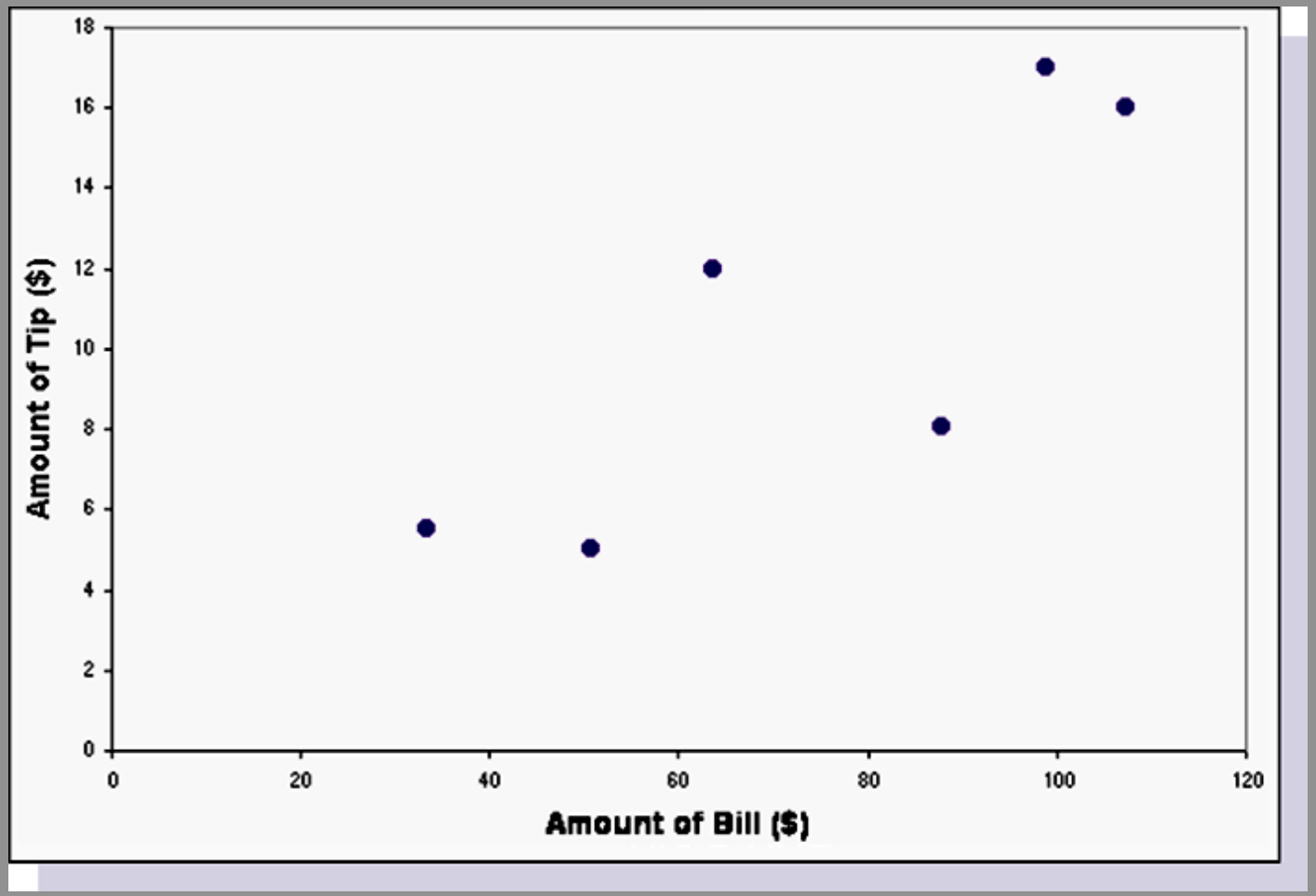
2. The pairs of (x, y) data have a bivariate normal distribution.

Definition

- **Scatter plot (or scatter diagram)**

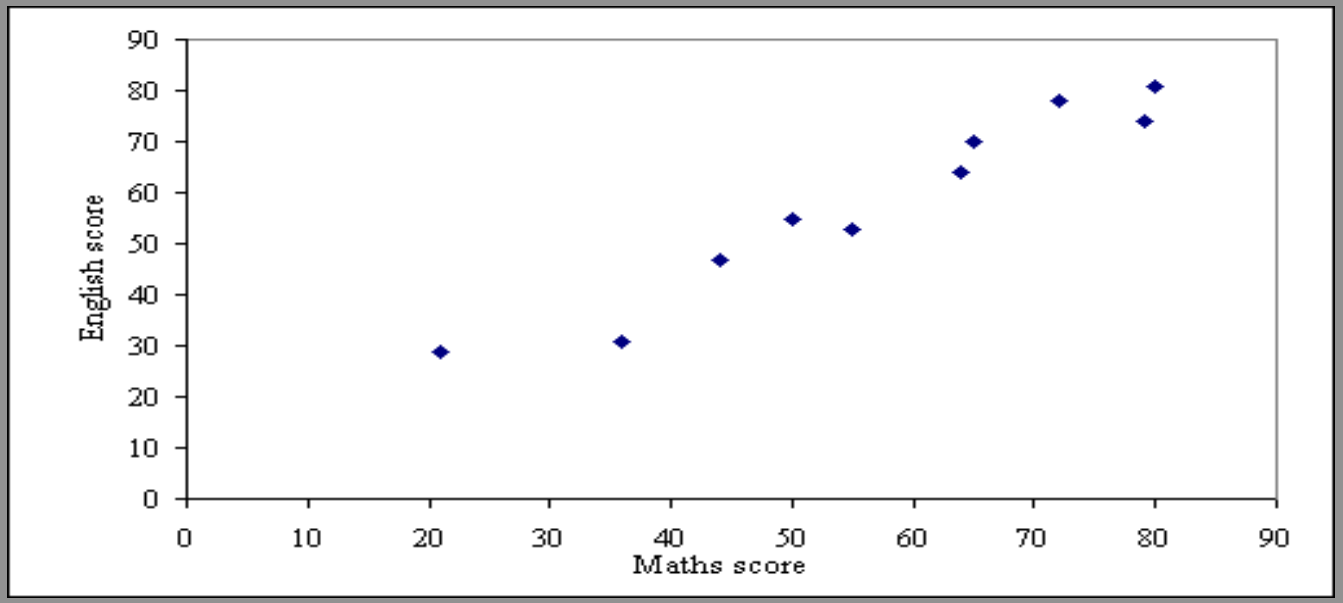
  is a graph in which the paired (x_i, y_i) sample data are plotted with a horizontal x axis and a vertical y axis. Each individual (x_i, y_i) pair is plotted as a single point.

Scatter Diagram of Paired Data



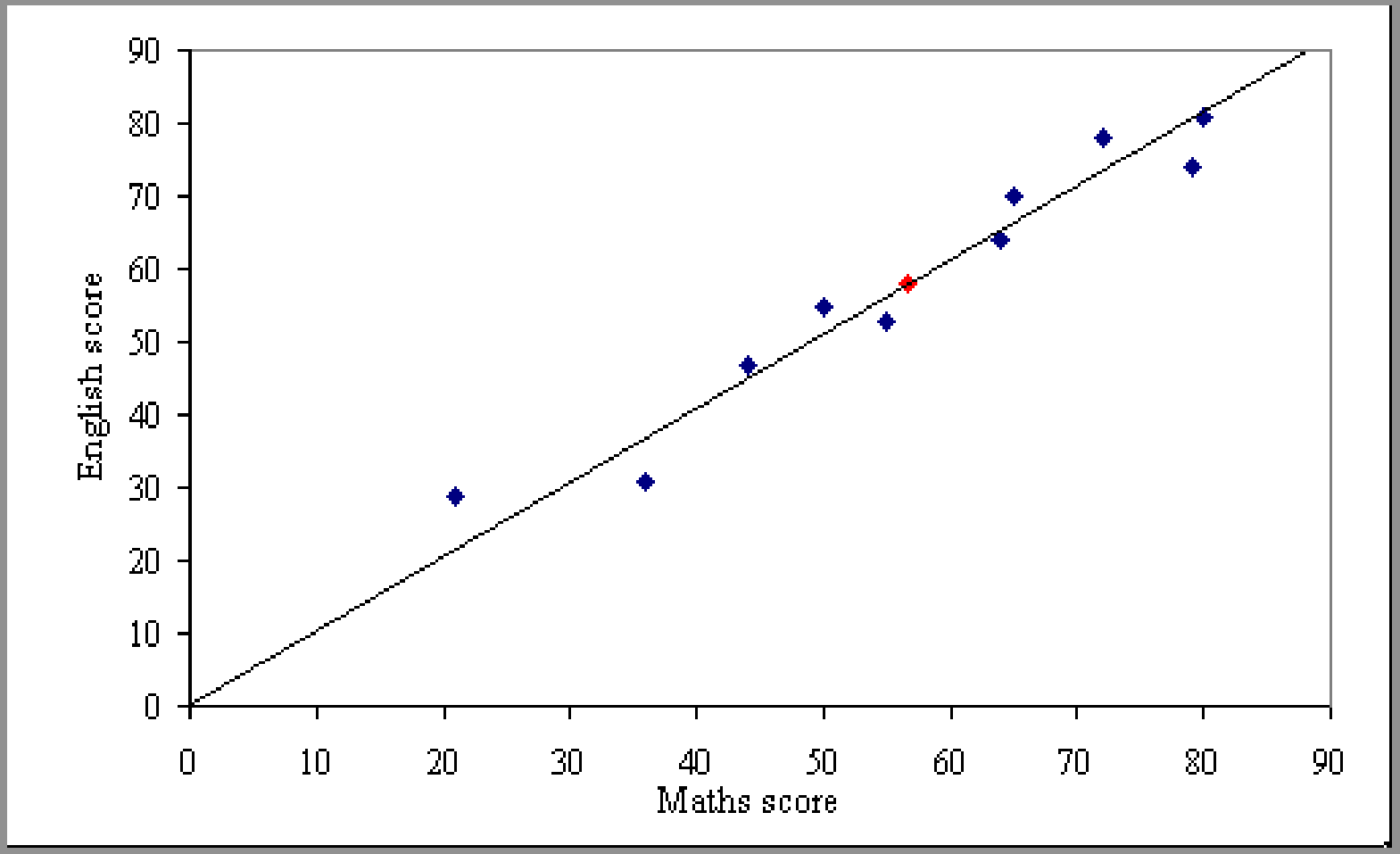
- Consider the results of two examinations

Maths	72	65	80	36	50	21	79	64	44	55
English	78	70	81	31	55	29	74	64	47	53



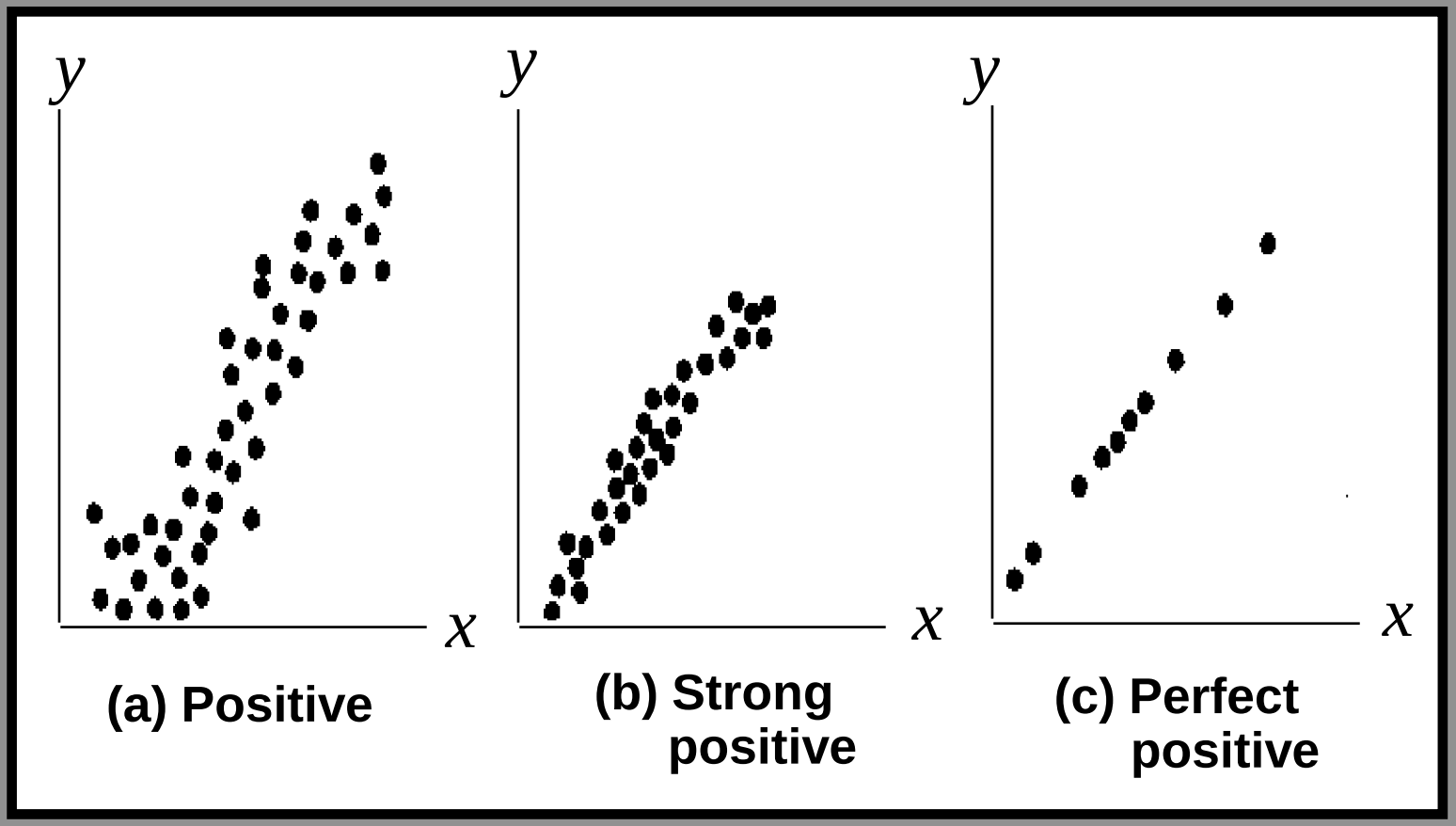
- You can see that the points follow a fairly strong pattern. People who are good in Maths tend to be good in English as well. The marks lie fairly close to an imaginary straight line that we can draw on the graph.

- The diagram below shows a straight line and also shows a point (in red) which I shall explain later:


English score
Maths score
0
10
20
30
40
50
60
70
80
90

- The fact that the points lie close to the straight line is called a **strong correlation**. The fact that this line points upwards to right - indicating that the English mark tends to increase as the math mark increases - is called a **positive correlation**.

Positive (Linear) Correlation



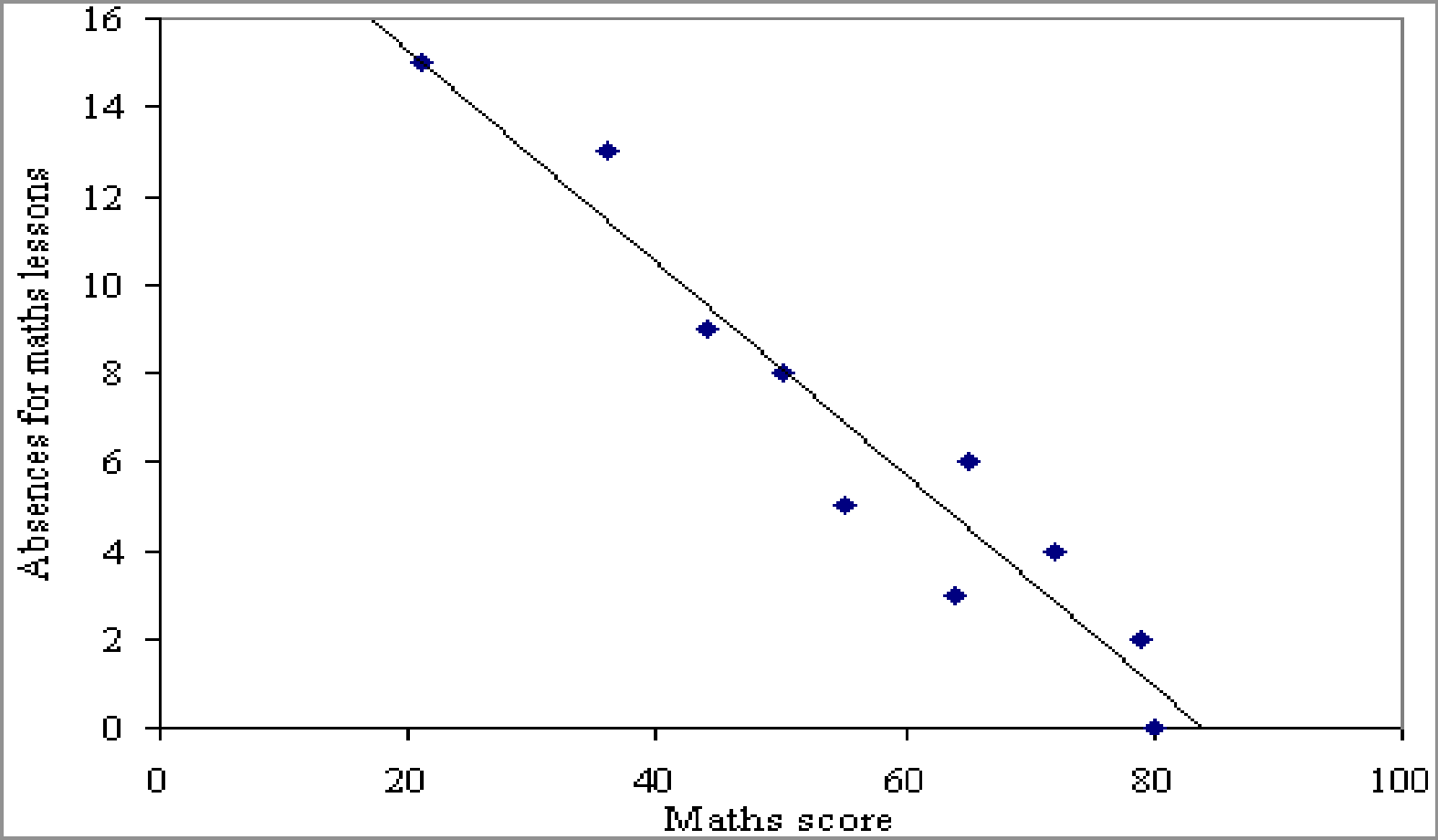
Scatter Plots

Negative Correlation

In the following table, We have duplicated the math marks for the ten students and this time added the number of absences from math lessons for each student:

Maths	72	65	80	36	50	21	79	64	44	55
Absences	4	6	0	13	8	15	2	3	9	5

In this case, the scatter diagram looks like this. The regression line has also been added.



- Again, there is a good correlation between the math scores and the absences from math lessons, except that as the number of absences increases, the math score goes down. This is referred to as *negative correlation.*

We can use the line of best fit to make predictions.

For example, what score would a student have received if he had been absent 10 times. According to the graph, it would have been about 41.

Similarly, if a student received a 30 marks, how many times would you expect him to have been absent? From the graph, it seems to be about 13 times.

- **However, this graph shows well the limitations of making predictions.**

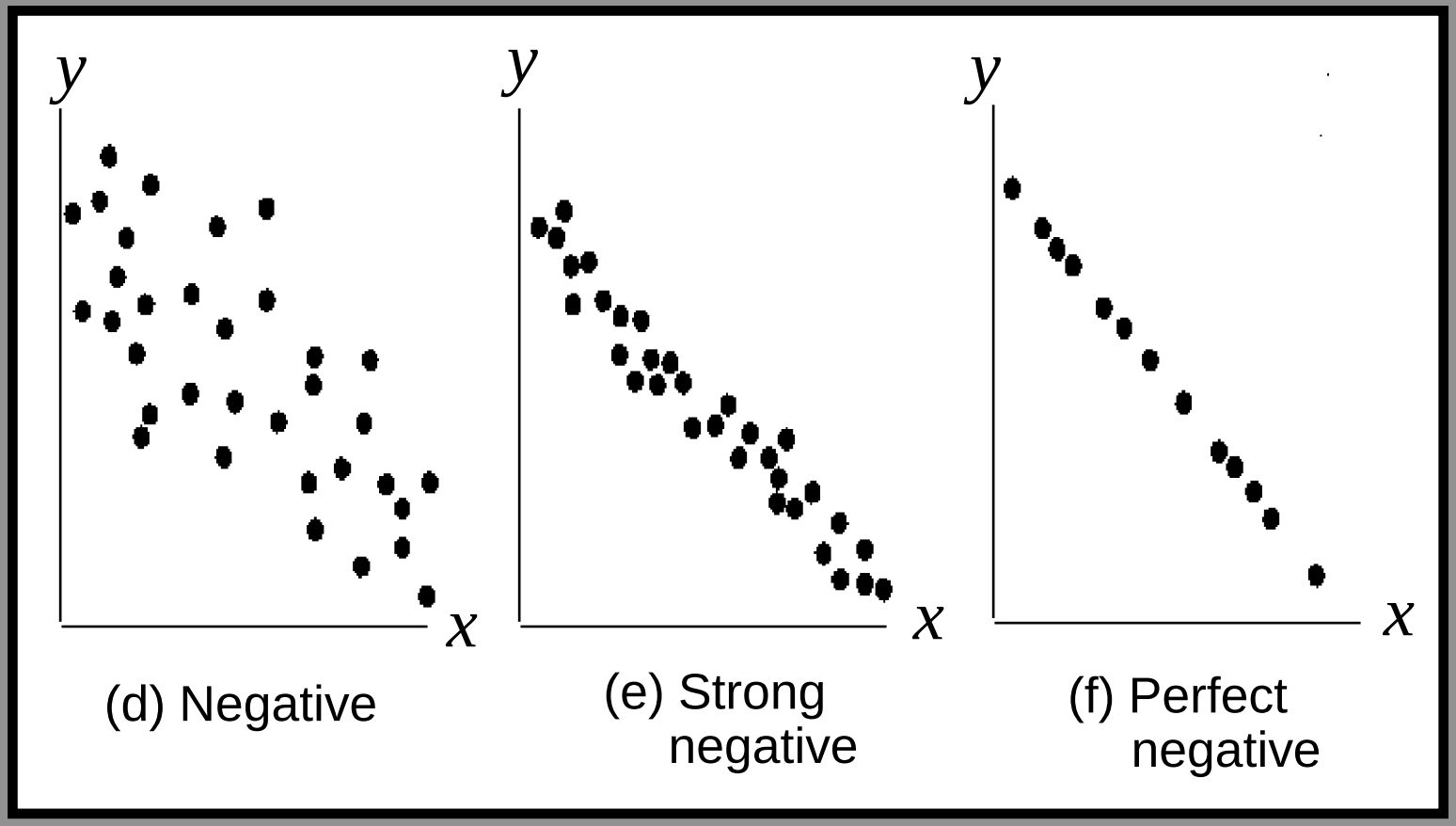
What score would someone have received if they had been absent for all 30 math lessons? According to the graph, the score would be less than zero!

Similarly, how many times would a student have had to be absent in order to gain a score of 90? Well, the line hits the horizontal axis when the score is just over 80, so in order to get a score of 90, a student would have to be absent a negative number of times.

Clearly, these conclusions are irrelevant, and they lead us to another general principle:

You can only use linear regression to draw conclusions about values within the range of the data point themselves. You might just be able to get away with drawing conclusions about values just outside that range, but the further away from the data range you move, the less reliable the conclusions become.

Negative Linear Correlation



(d) Negative

(e) Strong negative

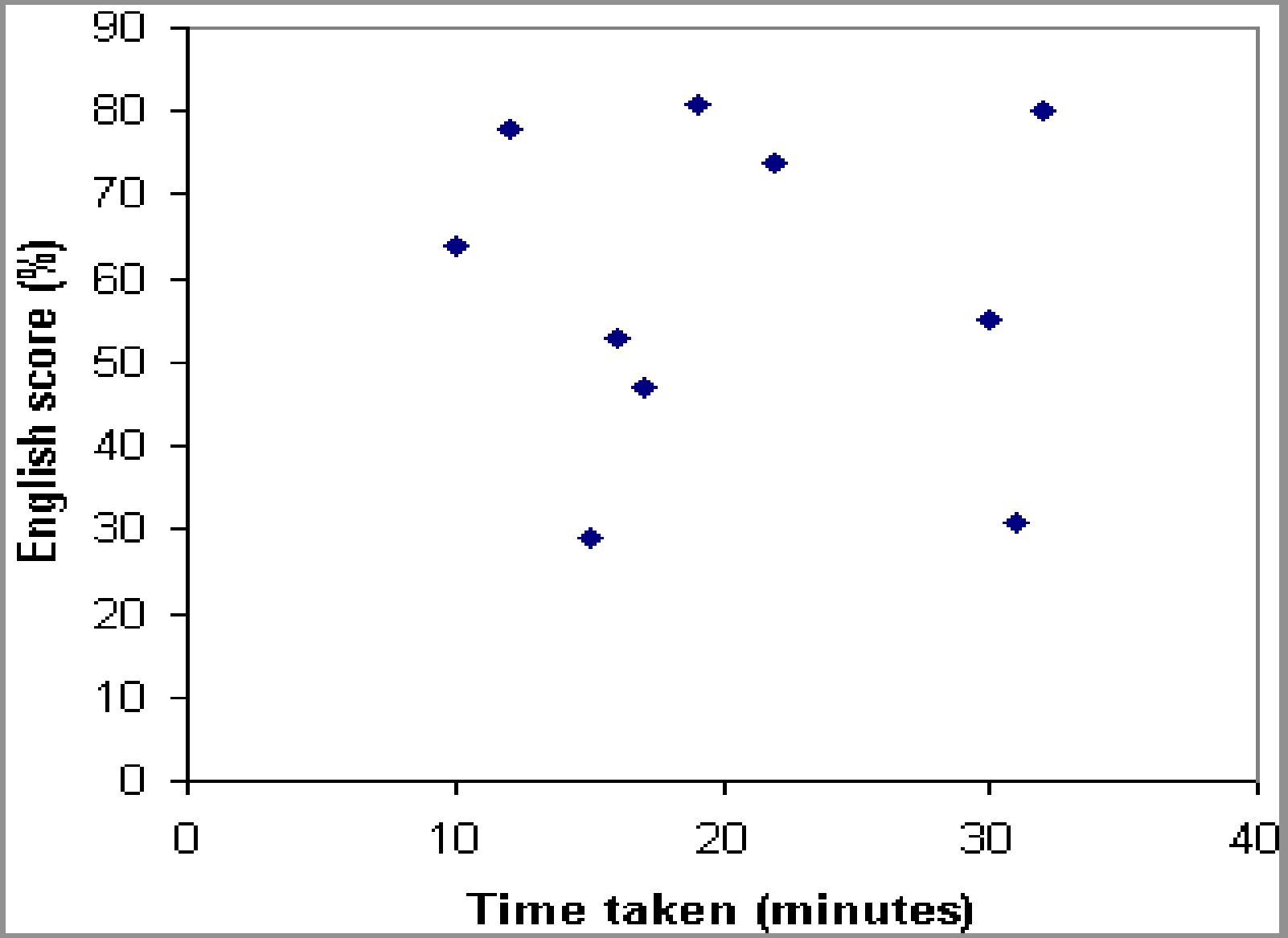
(f) Perfect negative

Scatter Plots

No correlation

- Finally, one more table, this time showing the English marks compared with the average length of time the students spend traveling to college each morning, recorded in minutes.

English score	78	70	81	31	55	29	74	64	47	53
Time	12	32	19	31	30	15	22	10	17	16

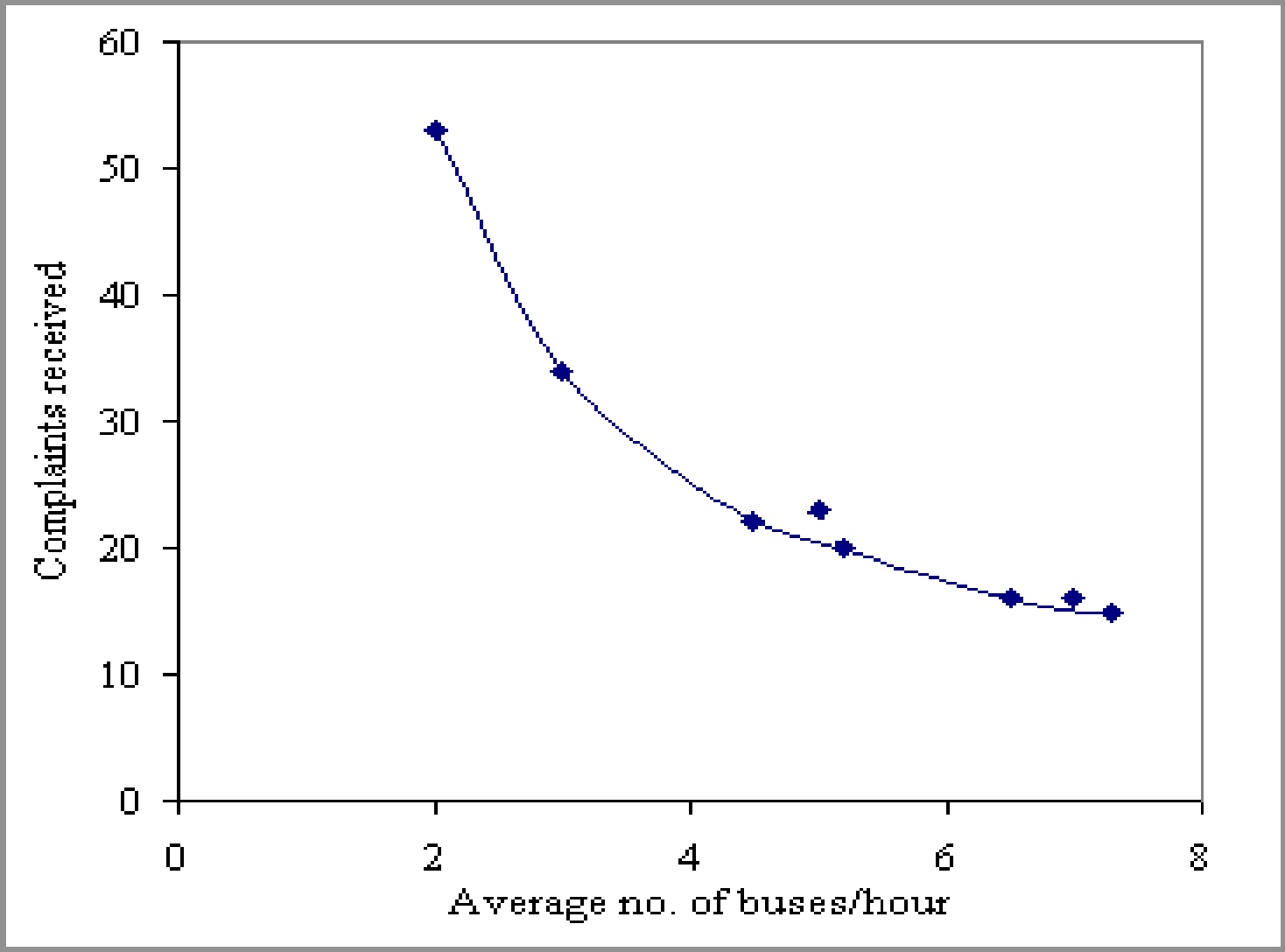

English score (%)
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
Time taken (minutes)

In this case, the scatter diagram shows no particular pattern. It is clear that we can't draw a straight line anywhere near the data points, and we say that there is *no correlation* between the length of time taken to travel to college and the final English mark that a student gets.

We cannot predict the English mark of any student based on how long it takes him to get to college. Nor can we predict how long it takes a student to get to college given that student's English mark.

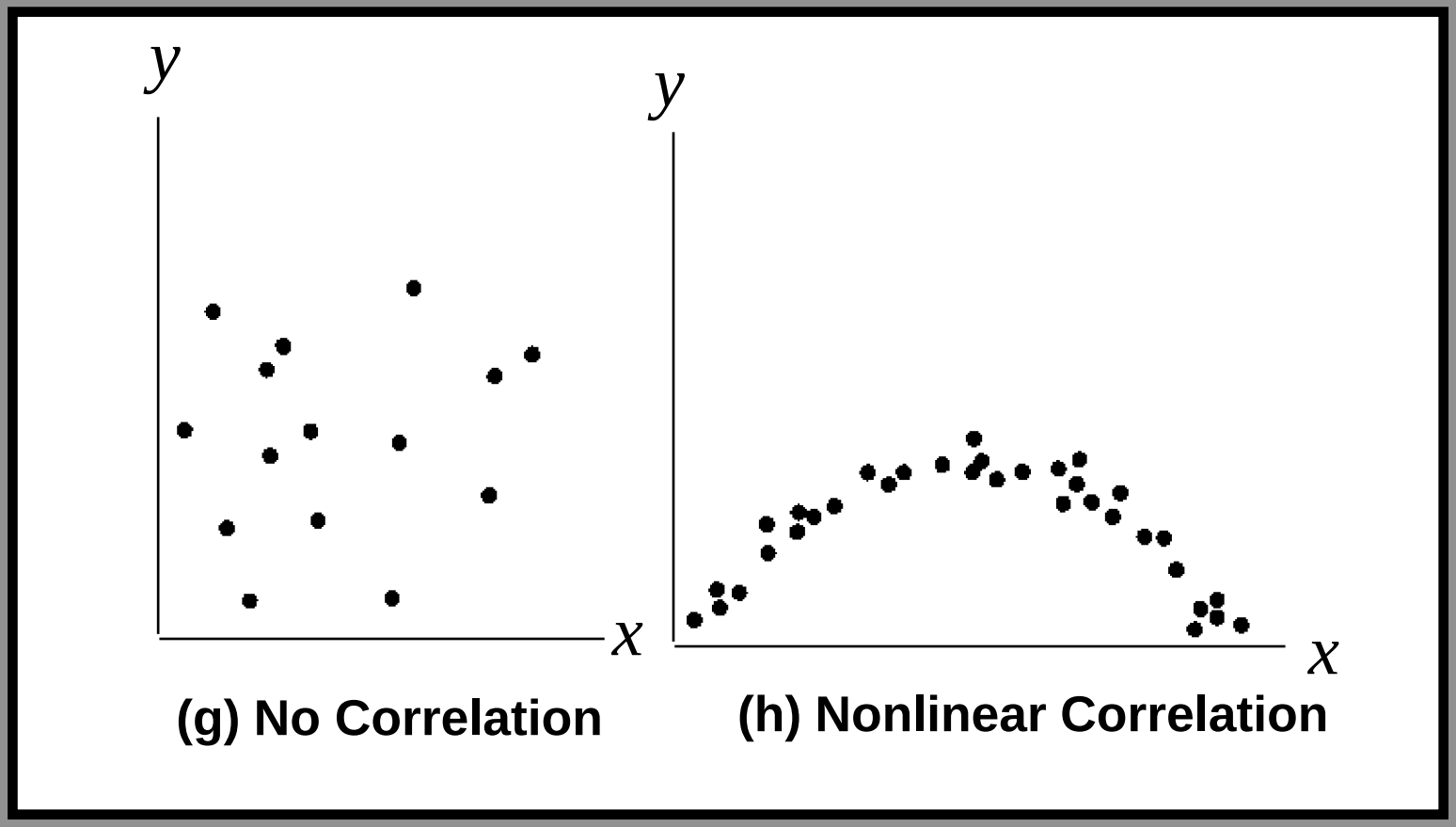
Non-linear correlations

- A bus company wanted to discover if there was any relationship between the number of buses it ran and the number of complaints it received. It carried out a survey testing the average number of buses per hour for different days, and the number of complaints that it received on those days. Here are the results:


60
50
40
30
20
10
0
Complaints received
0
2
4
6
8
Average no. of buses/hour

- We can see, there is a negative correlation between the number of buses per hour and the number of complaints, but in this case, a *curved* line fits the data better than a straight line. We are about to investigate the rule that lets you fit a straight line to the data points - it is enough to say at this point that similar rules exist which let you fit various curved lines to the data points as well.

No Linear Correlation



Scatter Plots

Correlation and Cause

Just because two variables are correlated, does not mean that one of the variables is the cause of the other. It could be the case, but it does not necessarily follow:

There is a strong positive correlation between the number of cigarettes that one smokes a day and one's chances of contracting lung cancer.

The percentage of heavy smokers who contract lung cancer is higher than the percentage of light smokers who develop the disease, and both figures are higher than the percentage of non-smokers who get lung cancer.

In this case, the cigarettes are definitely causing the cancer.

Although a correlation between two variables doesn't mean that one of them causes the other, it can suggest a way of finding out what the true cause might be.

There may be some underlying variable that is causing both of them. For instance, if a survey found that there is a correlation between the time that people spend watching television and the amount of crime that people commit, it could be because unemployed people tend to sit around watching the television, and that unemployed people are more likely to commit crime. If that were the case, then unemployment would be the true cause!

Definition

❖ Correlation Coefficient *r*

measures strength of the linear relationship between variable *x* and *y* values in a sample

$$r = \frac{Cov(x, y)}{sd_x \, sd_y}$$

sd_x is the standard deviation of the variable x and sd_y is the standard deviation of the variable y.

Properties of the Correlation Coefficient *r*

1. **Range of correlation coefficient is $-1 \leq r \leq 1$**
2. **Value of *r* does not change if all values of either variable are converted to a different scale.**
3. **The r is not affected by the choice of x and y. Interchange x and y and the value of r will not change.**

Common Errors Involving Correlation

1. **Causation: It is wrong to conclude that correlation implies causality.**

2. **Linearity: There may be some relationship between x and y even when there is no significant linear correlation.**

Rank Correlation

$$\rho = 1 - [6 (\sum d_i^2) / n (n^2-1)]$$

Correlation Calculations

Rank Correlation - ρ

Pearson's - r

Rank Correlation, cont

$$\rho = 1 - [6 (\sum d_i^2) / n (n^2-1)]$$

Hits	Rank	HR	Rank	d_i	d_i^2
1	10	3	8	2	4
2	9	4	7	2	4
3	8	5	6	2	4
4	7	1	10	-3	9
5	6	7	4	2	4
6	5	6	5	0	0
7	4	2	9	-5	25
8	3	10	1	2	4
9	2	9	2	0	0
10	1	8	3	2	4

$\rho = 1- [6(58)/10(10^2-1)]$

$\rho = 1- [348 / 10 (100 -1)]$

$\rho = 1- [348 / 990]$

$\rho = 1- 0.352$

$\rho = 0.648$

$(\sum d_i^2 = 58)$

n=10

Pearson's r

Hits (x)	HR (y)	xy
1	3	3
2	4	8
3	5	15
4	1	4
5	7	35
6	6	36
7	2	14
8	10	80
9	9	81
10	8	80
$\Sigma x_i/n = 5.5$	$\Sigma x_i/n = 5.5$	$\Sigma x_i y_i/n = 32.86$

$$r = \frac{\mathrm{Cov}(x, y)}{sd_x\, sd_y}$$

r = 32.86 - (5.5) (5.5)/(3.03) (3.03)

r = 35.86 - 30.25 / 9.09

r = 5.61 / 9.09

r = 0.6172

Is there a significant linear correlation?

Data from the Garbage Project

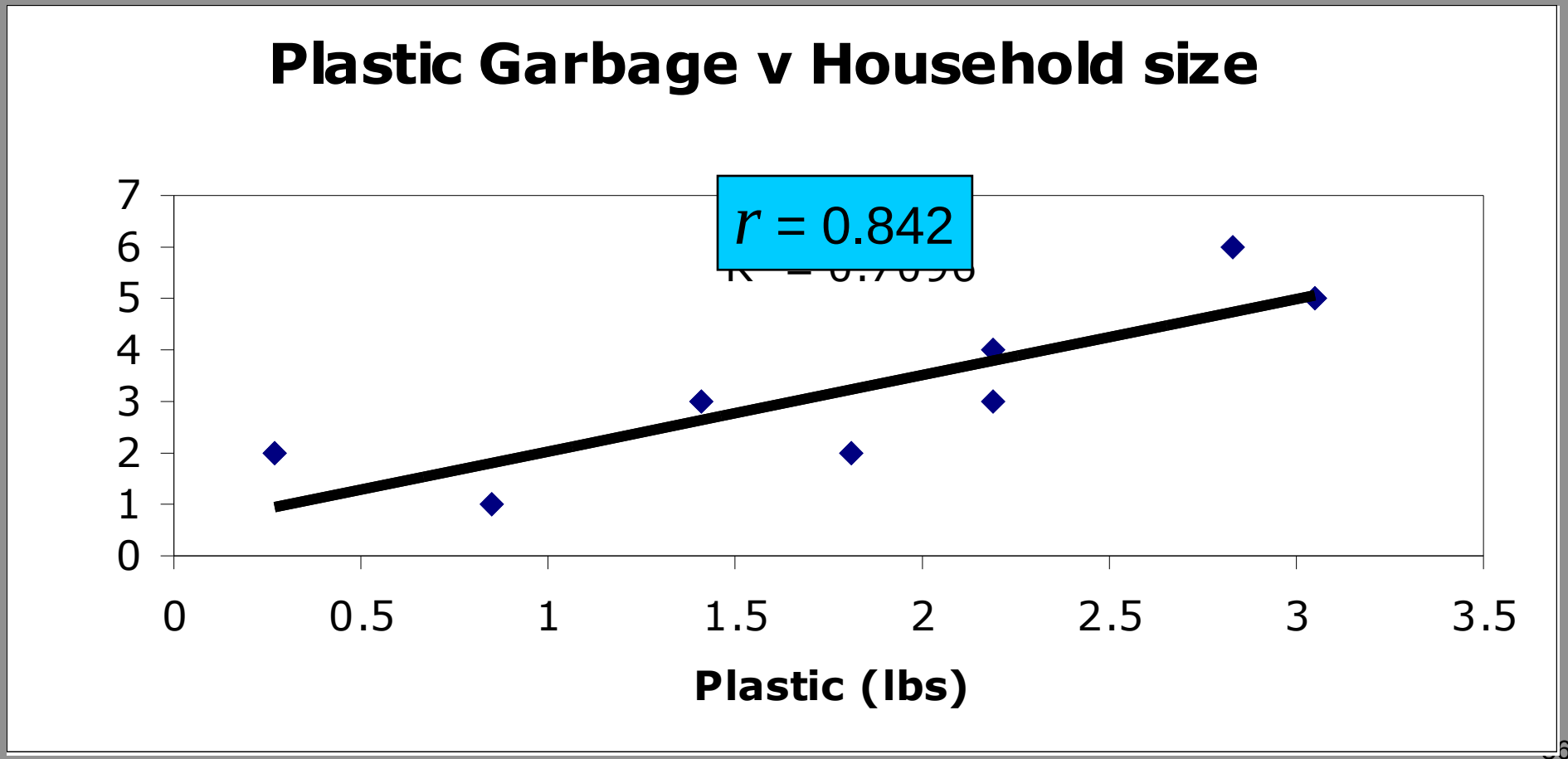
x	Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y	Household	2	3	3	6	4	2	1	5

Plastic	Household
0.27	2
1.41	3
2.19	3
2.83	6
2.19	4
1.81	2
0.85	1
3.05	5

Is there a significant linear correlation?

Data from the Garbage Project

x Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y Household	2	3	3	6	4	2	1	5



Is there a significant linear correlation?

$n = 8$ $\alpha = 0.05$ $H_0: \rho = 0$

$H_1: \rho \neq 0$

Test statistic is $r = 0.842$

Critical values are $r = -0.707$ and 0.707
(Table R with $n = 8$ and $\alpha = 0.05$)

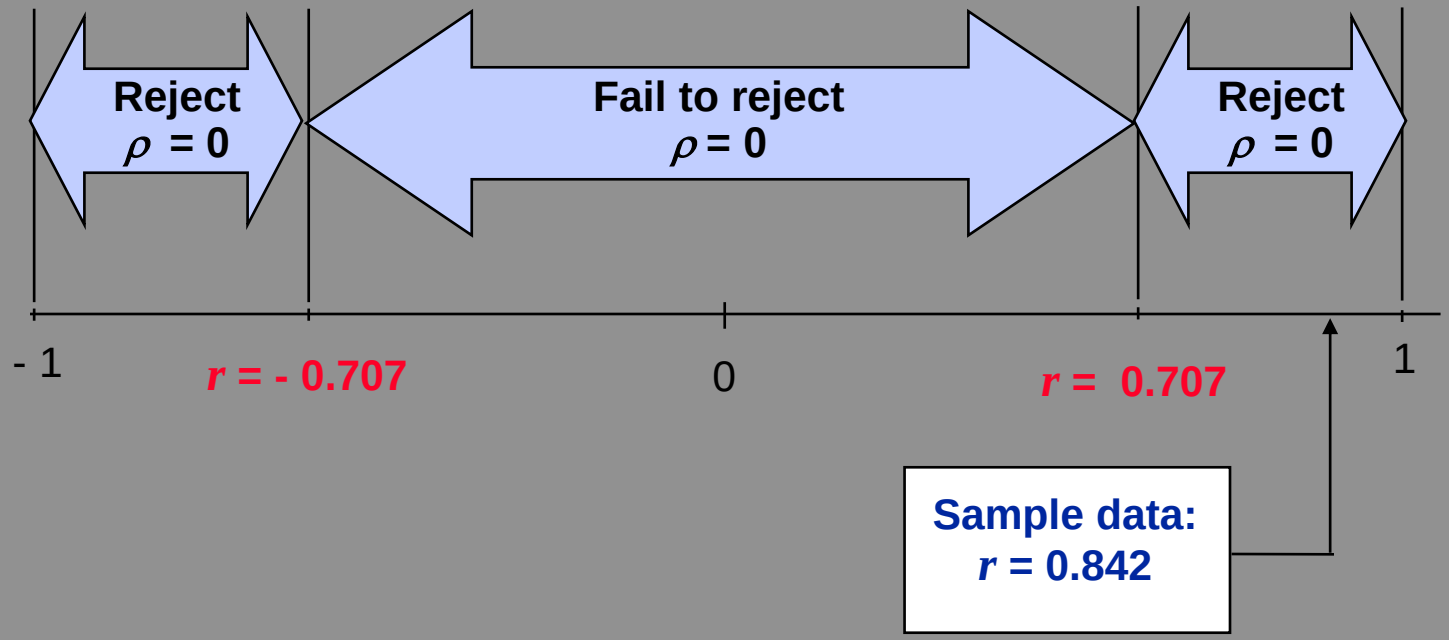
n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

TABLE R Critical Values of the Pearson Correlation Coefficient r

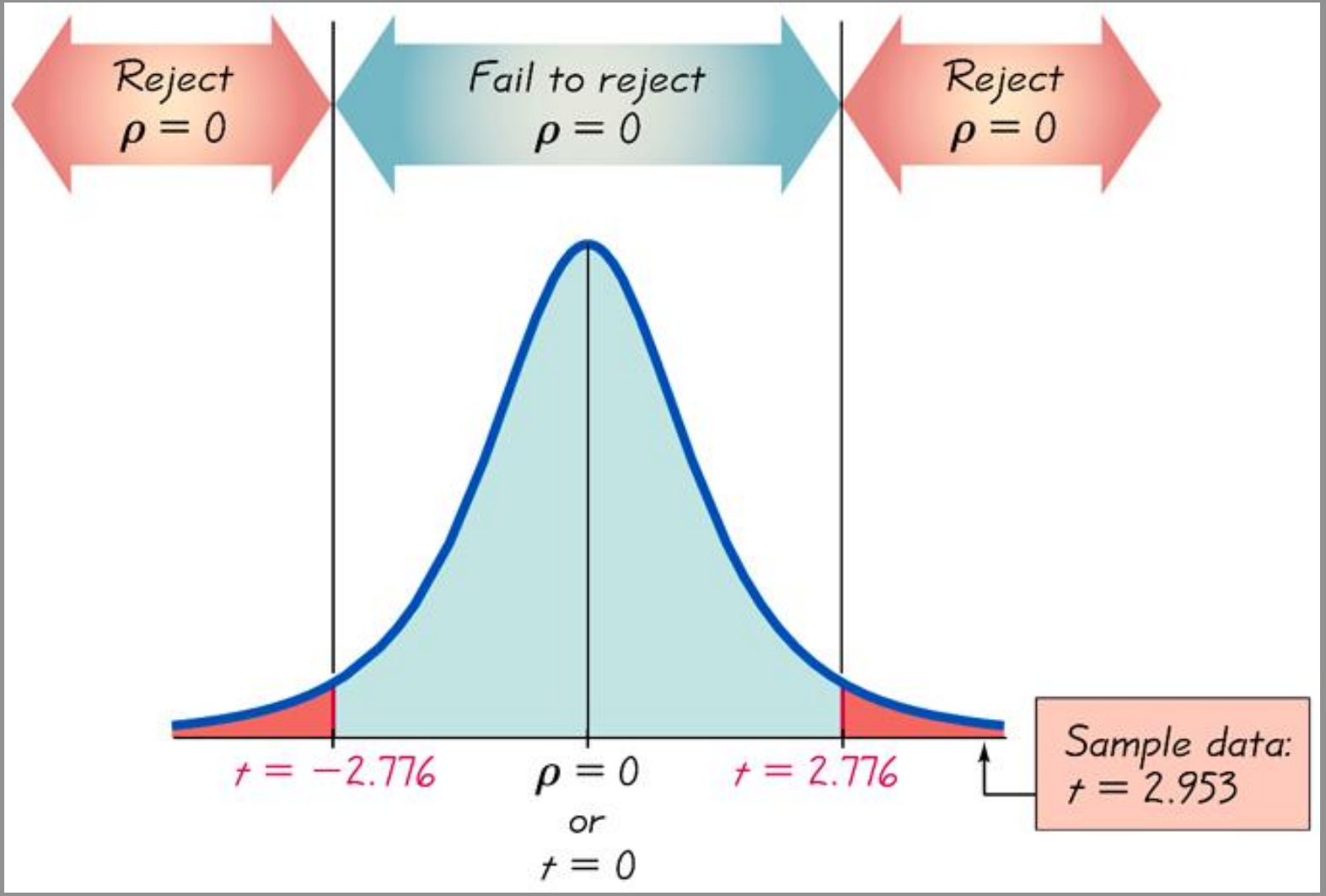
Is there a significant linear correlation?

0.842 > 0.707, That is the test statistic does fall within the critical region.

Therefore, we REJECT H_0: $\rho = 0$ (no correlation) and conclude there is a significant linear correlation between the weights of discarded plastic and household size.



Test Statistic is *t*



Formal Hypothesis Test

- To determine whether there is a significant linear correlation between two variables
- let H_0: $\rho = 0$

  H_1: $\rho \neq 0$

Test Statistic is t

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Critical values:

use Table t with
degrees of freedom = n - 2

Interpreting the Correlation Coefficient

- If the absolute value of *t* exceeds the value in t-Table, conclude that there is a significant linear correlation.
- Otherwise, there is not sufficient evidence to support the conclusion of significant linear correlation.
- Remember to use n-2

Thank you