Bivariate measurement data

Let to consider the description of data sets concerning **two** variables, *x* and *y*, each measured on a numerical scale.

Thus, two numerical observations (x,y) are recorded for each sampling unit.

These observations are **paired** in the sense that an (*x*,*y*) pair rises from the same sampling unit. For *n* sampling units, we can write the measurement pairs as:

$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$$

univariate distribution or
marginal distribution

A major purpose of collecting bivariate data is to answer such such question as:

✓ Are the variables related?

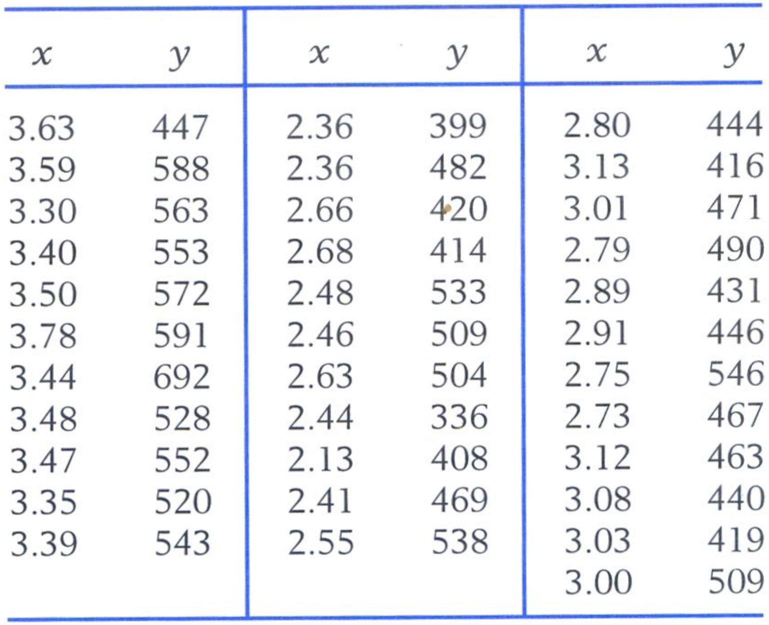
✓ What form of relationship is indicated by the data?

✓ Can we quantify the strength of their relationship?

✓ Can we predict one variable from the other?

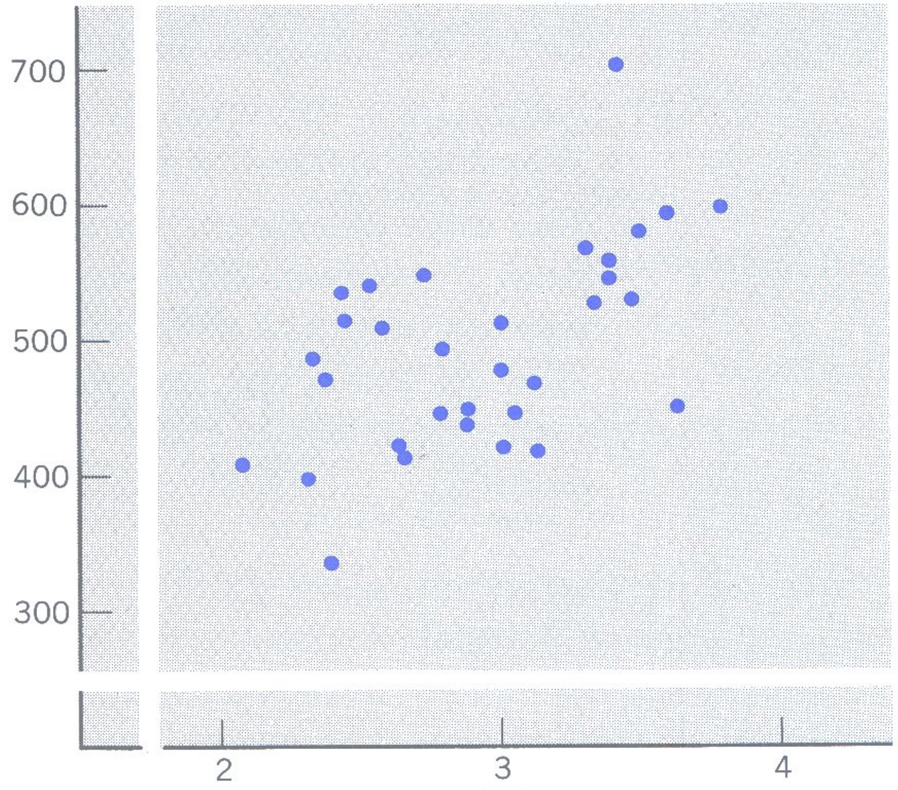
An important first step in studying the relationship between two variables is to graph the data. The resulting diagram is called a scatter diagram.

Table of bivariate data



x	y	x	y	x	y
3.63	447	2.36	399	2.80	444
3.59	588	2.36	482	3.13	416
3.30	563	2.66	420	3.01	471
3.40	553	2.68	414	2.79	490
3.50	572	2.48	533	2.89	431
3.78	591	2.46	509	2.91	446
3.44	692	2.63	504	2.75	546
3.48	528	2.44	336	2.73	467
3.47	552	2.13	408	3.12	463
3.35	520	2.41	469	3.08	440
3.39	543	2.55	538	3.03	419
				3.00	509

Scatter diagram



example

(Multiple Scatter Diagram)

Concern was raised by environmentalists that spills of contaminants were affecting wildlife in and around an adjacent lake. Estrogenic contaminants in the environment can have grave consequences on the ability of living things to reproduce. Researchers examined the reproductive development of young male alligators hatched from eggs taken from around (1) Lake Apopka, the lake which was contaminated, and (2) Lake Woodruff, which acted as a control. The contaminants were thought to influence sex steroid concentrations. The concentrations of two steroids, estradiol and testosterone, were determined by radioimmunoassay.

Lake Apopka

Estradiol	38	23	53	37	30
Testosterone	22	24	8	6	7

Lake Woodruff

Estradiol	29	64	19	36	27	16	15	72	85
Testosterone	47	20	60	75	12	54	33	53	100

(a) Make a scatter diagram of the two concentrations for the Lake Apopka alligators.

(b) Create a multiple scatter diagram by adding to the same plot the pairs of concentrations for the Lake Woodruff male alligators. Use a different symbol for the two lakes.

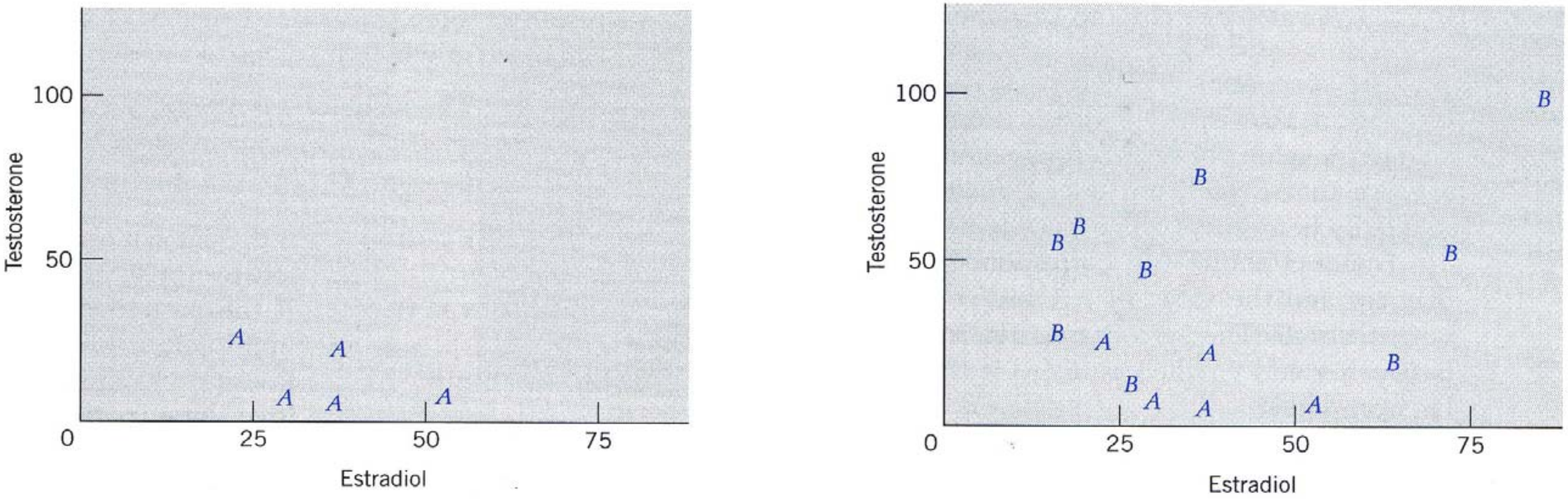
(c) Comment on any major differences between the two groups of male alligators.

SOLUTION

(a) Figure 2*a* gives the scatter diagram for the Lake Apopka alligators.

(b) Figure 2*b* is the multiple scatter diagram with the points for Lake Woodruff marked as *B*.

(c) The most prominent feature of the data is that the male alligators from the contaminated lake have, generally, much lower levels of testosterone than those from the nearly pollution-free control lake (The *A*s are at the bottom third of the scatter diagram.) Low testosterone levels in males has grave consequences regarding reproduction.

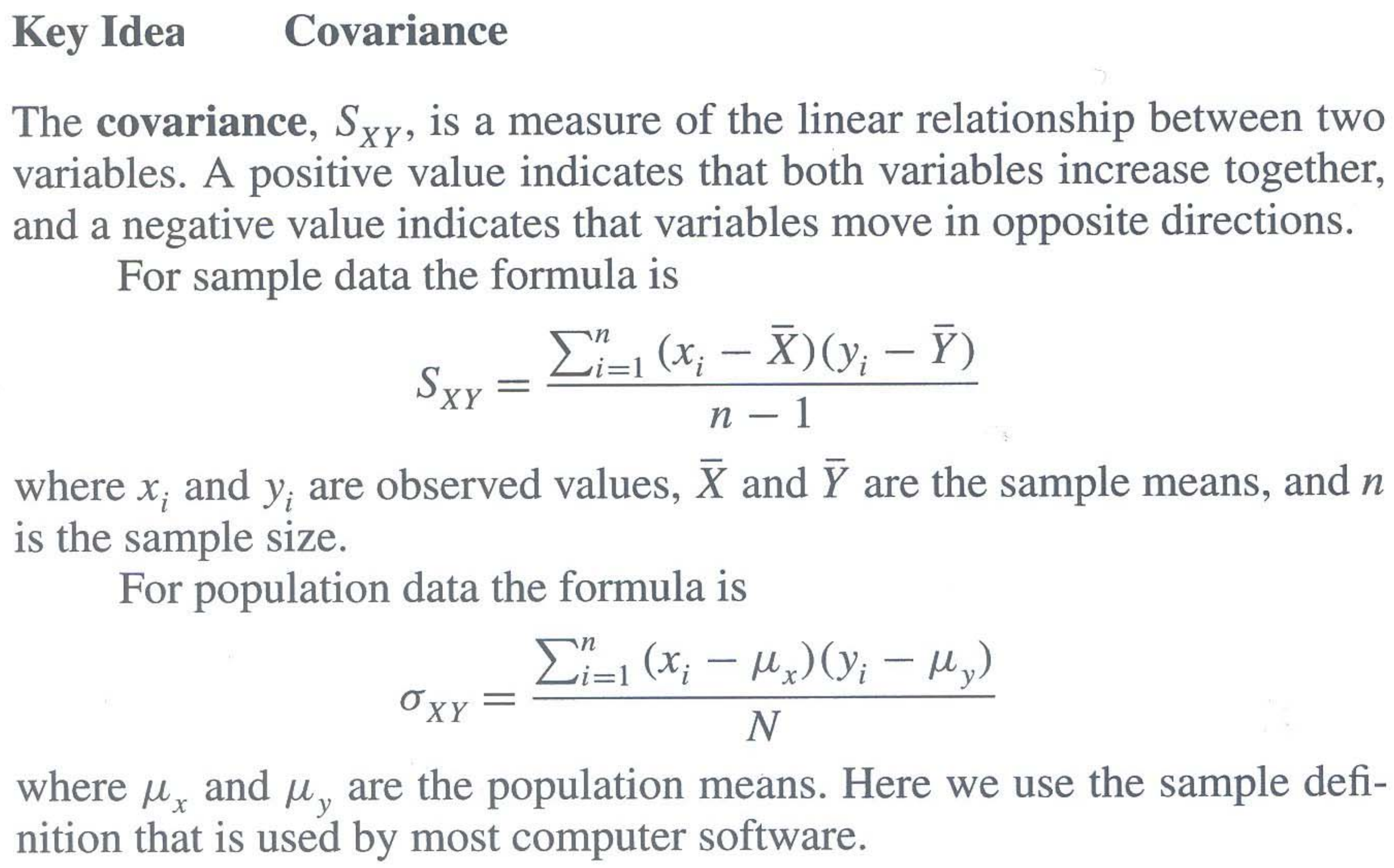


(*a*) Scatter diagram for Lake Apopka

(*b*) Multiple scatter diagram

Scatter diagrams. *A* = Lake Apopka. *B* = Lake Woodruff.

Covariance: the strength of a bivariate relationship



Key Idea **Covariance**

The **covariance**, S_{XY}, is a measure of the linear relationship between two variables. A positive value indicates that both variables increase together, and a negative value indicates that variables move in opposite directions.

For sample data the formula is

$$S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n - 1}$$

where x_i and y_i are observed values, $\bar{X}$ and $\bar{Y}$ are the sample means, and n is the sample size.

For population data the formula is

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{N}$$

where μ_x and μ_y are the population means. Here we use the sample definition that is used by most computer software.

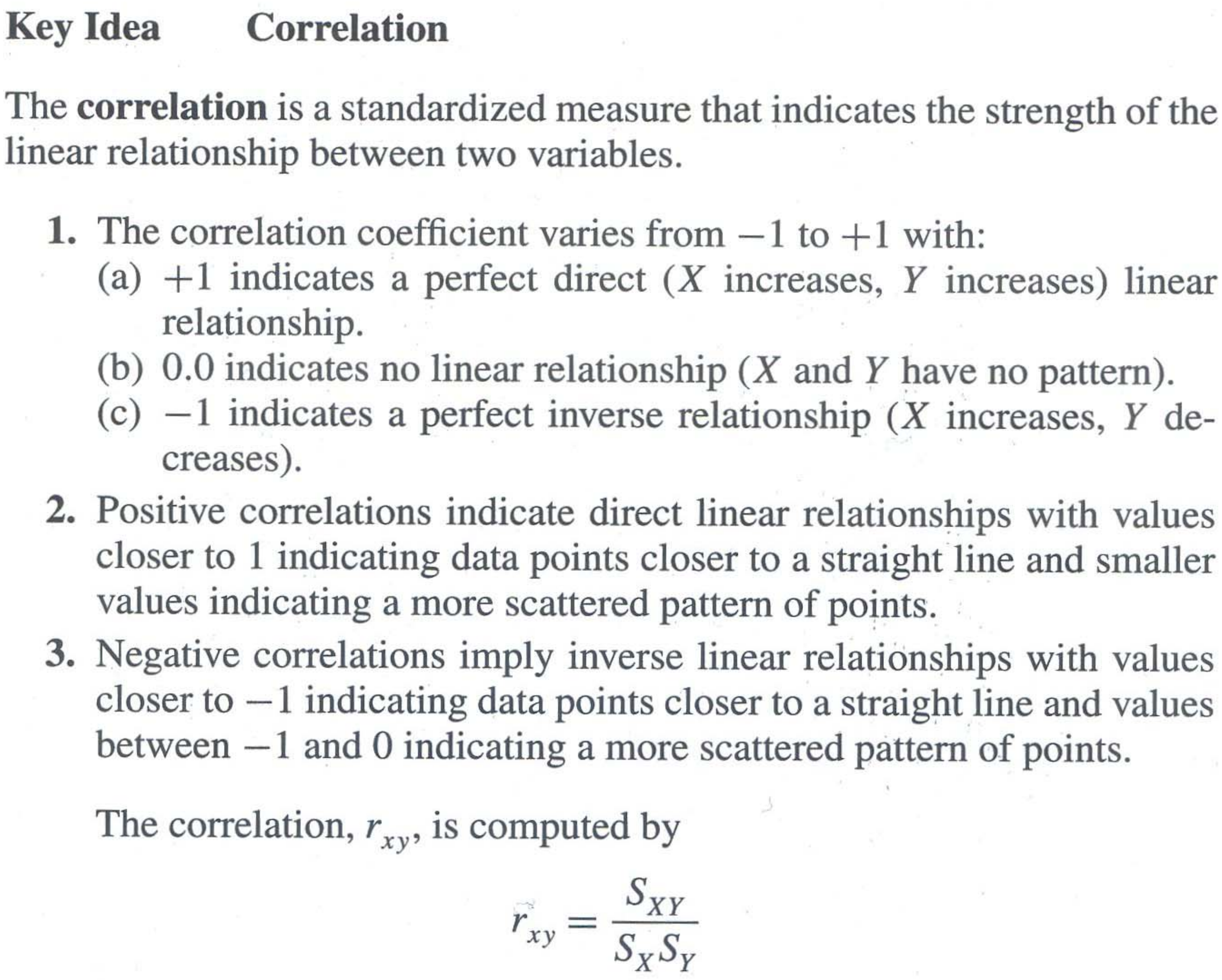
Let to have x = [6,7,8,9,10] and y =[80,60,70,40,0], first we have to calculate the mean of the univariate (marginal distributions):

$$
\begin{aligned}
\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{6+7+8+9+10}{5} \\
&= 8.0 \\
\bar{Y} &= \frac{\sum_{i=1}^{n} y_i}{n} \\
&= \frac{80+60+70+40+0}{5} \\
&= 50
\end{aligned}
$$

Next we can compute the **covariance**:

$$
\begin{aligned}
S_{XY} &= \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n-1} \\
&= \frac{(6-8)(80-50) + (7-8)(60-50) + \cdots + (10-8)(0-50)}{5-1} \\
&= -45
\end{aligned}
$$

Standardising covariance: the correlation coefficient



Key Idea **Correlation**

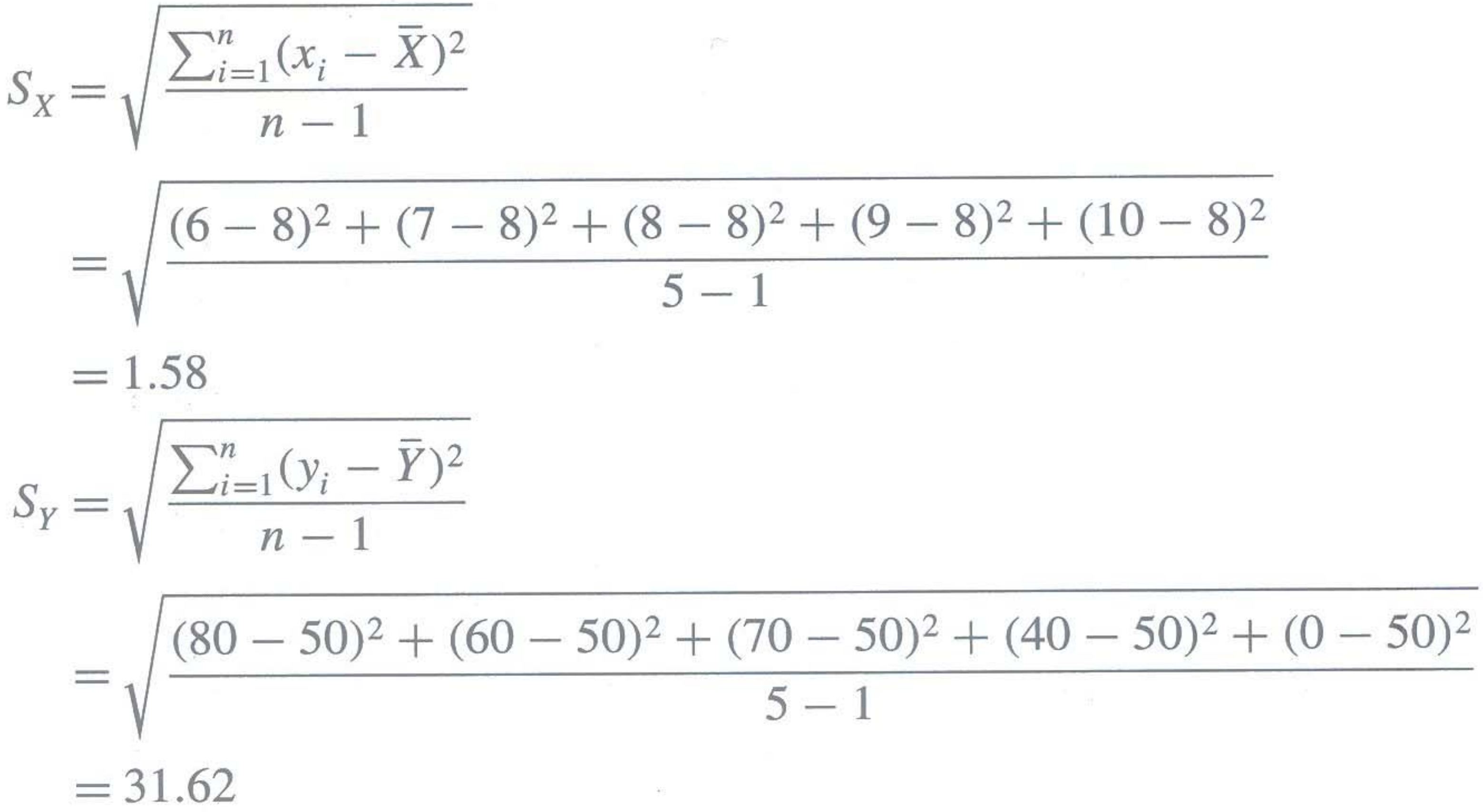
The **correlation** is a standardized measure that indicates the strength of the linear relationship between two variables.

1. The correlation coefficient varies from -1 to $+1$ with:
   (a) $+1$ indicates a perfect direct (X increases, Y increases) linear relationship.
   (b) 0.0 indicates no linear relationship (X and Y have no pattern).
   (c) -1 indicates a perfect inverse relationship (X increases, Y decreases).
2. Positive correlations indicate direct linear relationships with values closer to 1 indicating data points closer to a straight line and smaller values indicating a more scattered pattern of points.
3. Negative correlations imply inverse linear relationships with values closer to -1 indicating data points closer to a straight line and values between -1 and 0 indicating a more scattered pattern of points.

The correlation, r_{xy}, is computed by

$$r_{xy} = \frac{S_{XY}}{S_X S_Y}$$

Let to consider again the vectors x = [6,7,8,9,10] and y =[80,60,70,40,0], first we have to calculate the standard deviation of the univariate (marginal distributions):

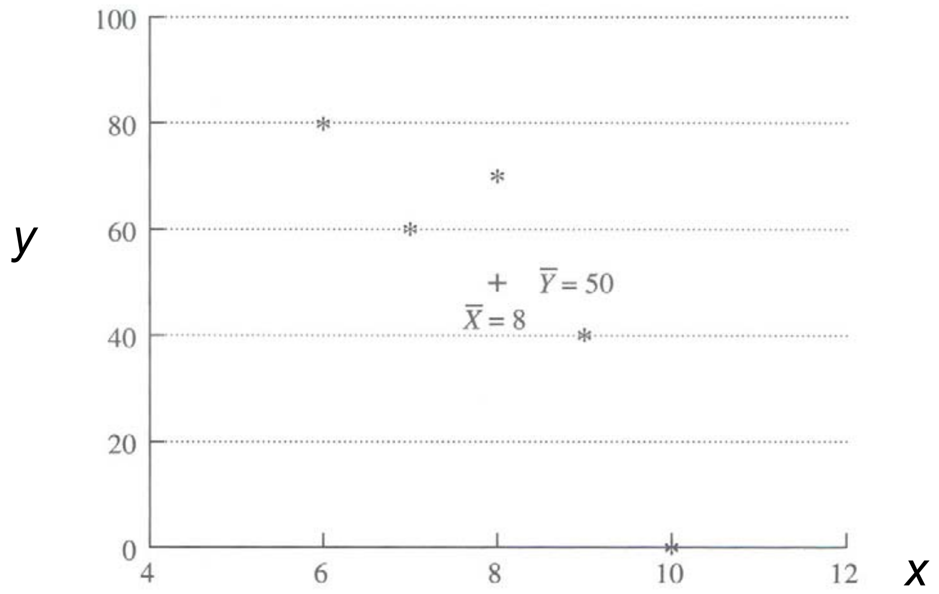


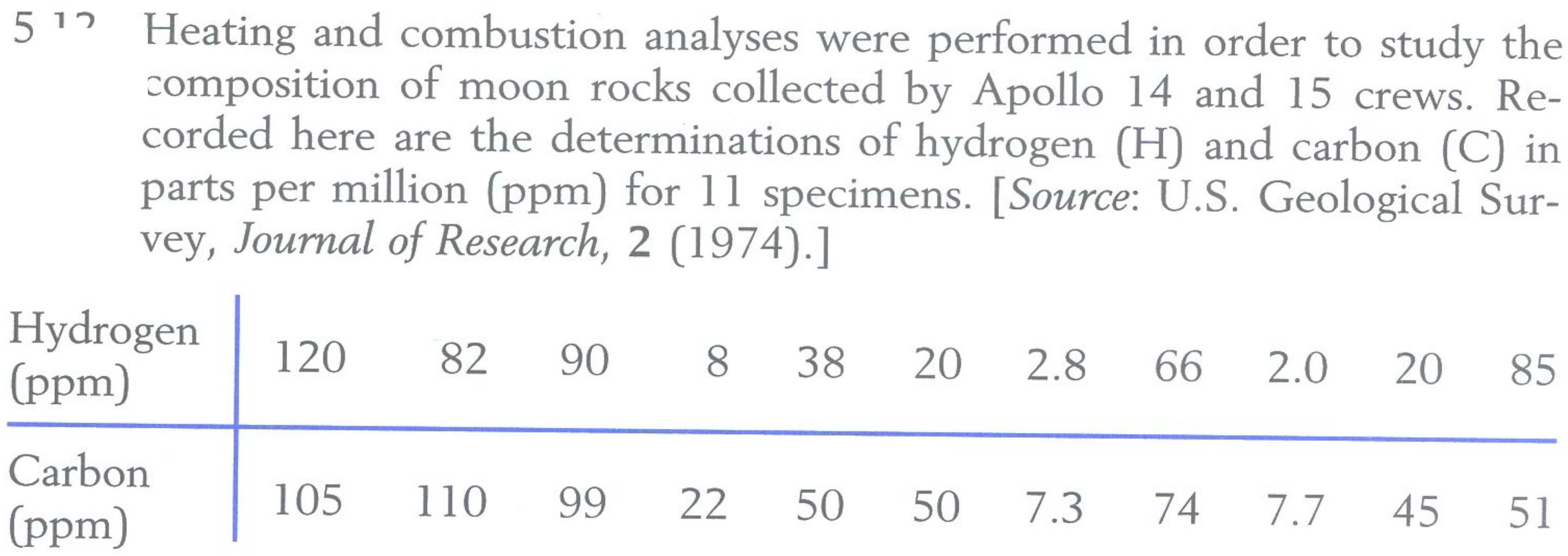
$$
\begin{aligned}
S_X &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}} \\
&= \sqrt{\frac{(6-8)^2 + (7-8)^2 + (8-8)^2 + (9-8)^2 + (10-8)^2}{5-1}} \\
&= 1.58 \\
S_Y &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n-1}} \\
&= \sqrt{\frac{(80-50)^2 + (60-50)^2 + (70-50)^2 + (40-50)^2 + (0-50)^2}{5-1}} \\
&= 31.62
\end{aligned}
$$

Using these results with the covariance computed previously, we found that:

$$
\begin{aligned}
r_{xy} &= \frac{S_{XY}}{S_X S_Y} \\
&= \frac{-45}{(1.58)(31.62)} \\
&= -0.90
\end{aligned}
$$

The correlation of -0.90 indicates that the relationship is downward sloping and that it is close to a straight line. The figure will provide an image of this strong negative correlation!

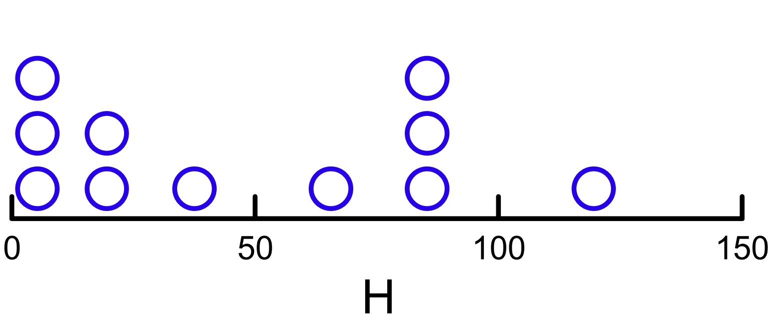


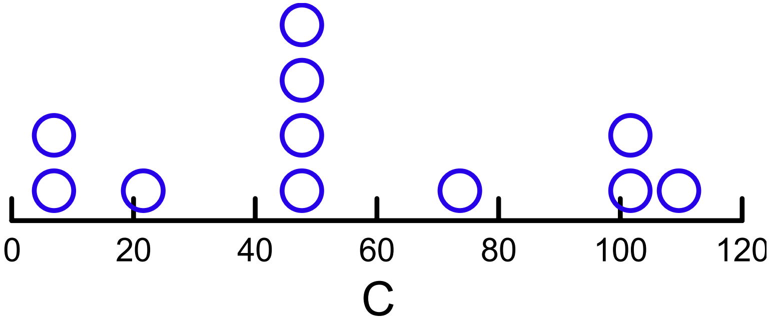


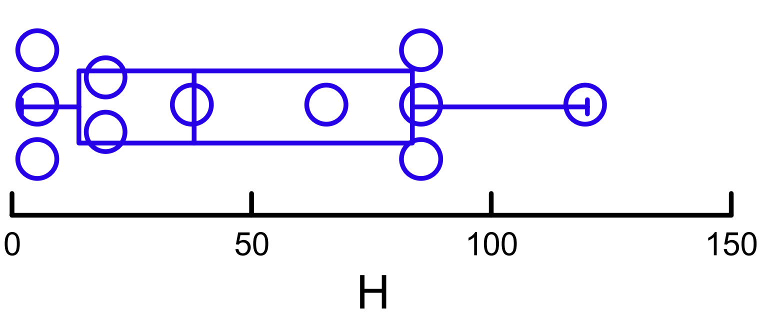
5.12 Heating and combustion analyses were performed in order to study the composition of moon rocks collected by Apollo 14 and 15 crews. Recorded here are the determinations of hydrogen (H) and carbon (C) in parts per million (ppm) for 11 specimens. [*Source*: U.S. Geological Survey, *Journal of Research*, **2** (1974).]

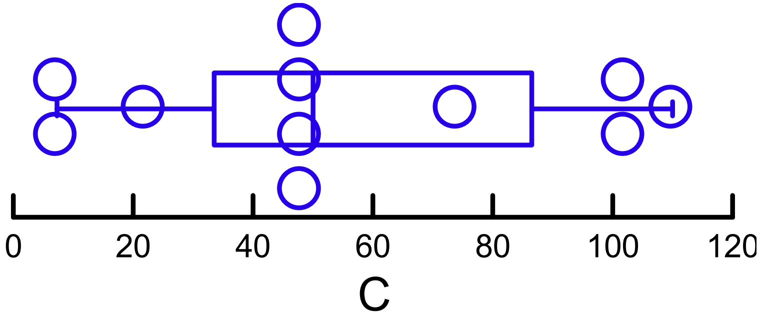
Hydrogen (ppm)	120	82	90	8	38	20	2.8	66	2.0	20	85
Carbon (ppm)	105	110	99	22	50	50	7.3	74	7.7	45	51

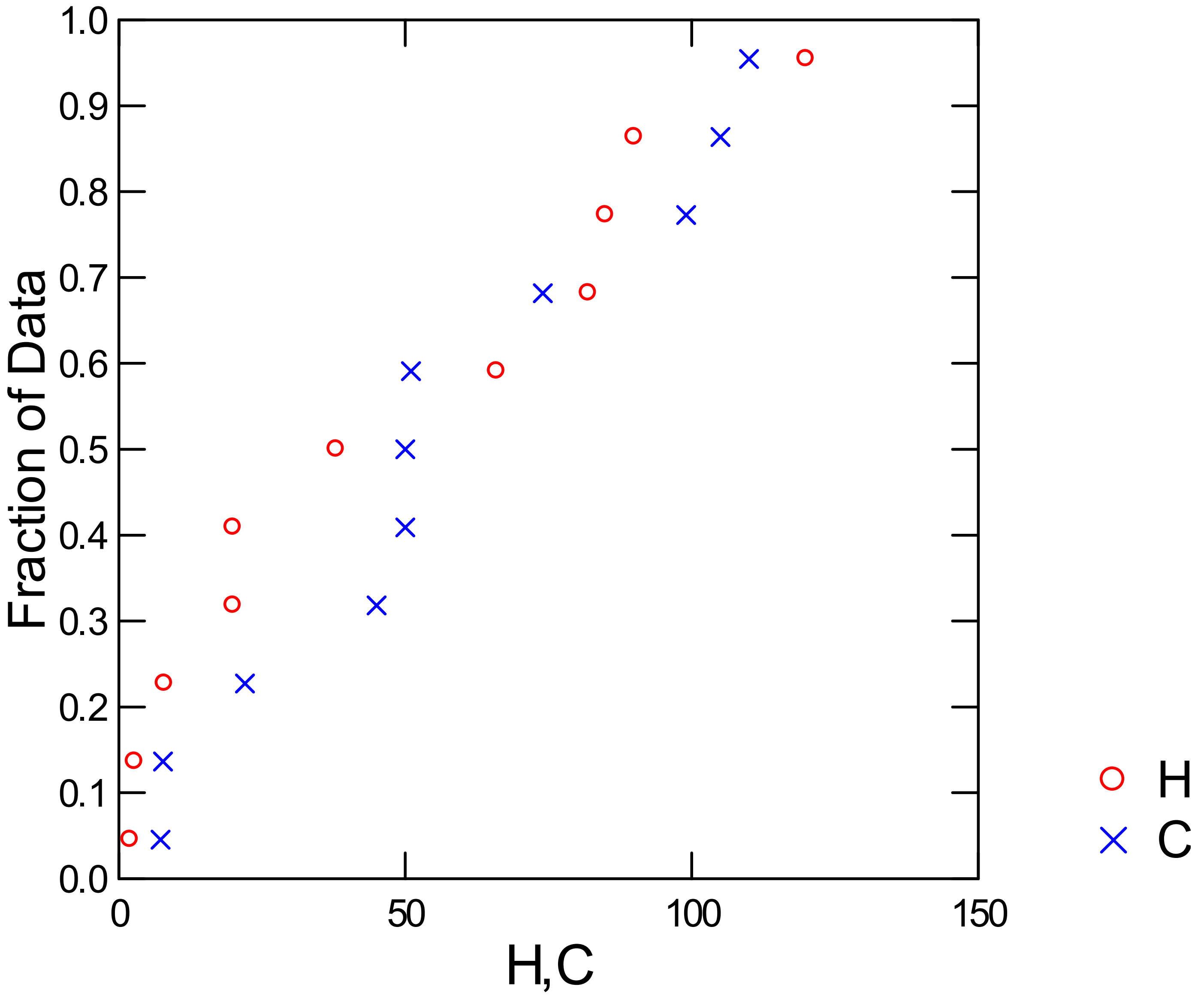
- ❑ Analyse the univariate behaviour of H and C by using box plot and/or dot plot;
- ❑ Summarise the univariate statistical data;
- ❑ Calculate covariance and correlation and plot the scatter diagram

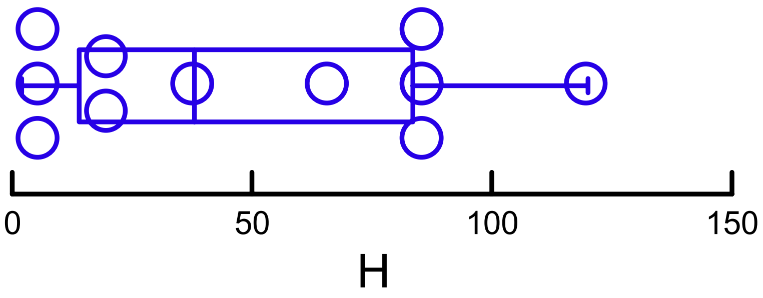

0
50
100
150
H

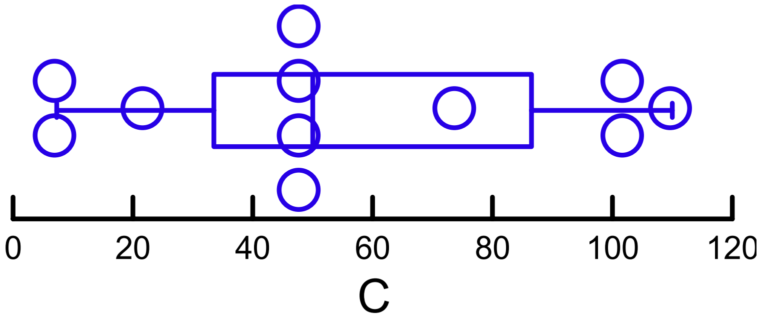

0
20
40
60
80
100
120
C


0
50
100
150
H

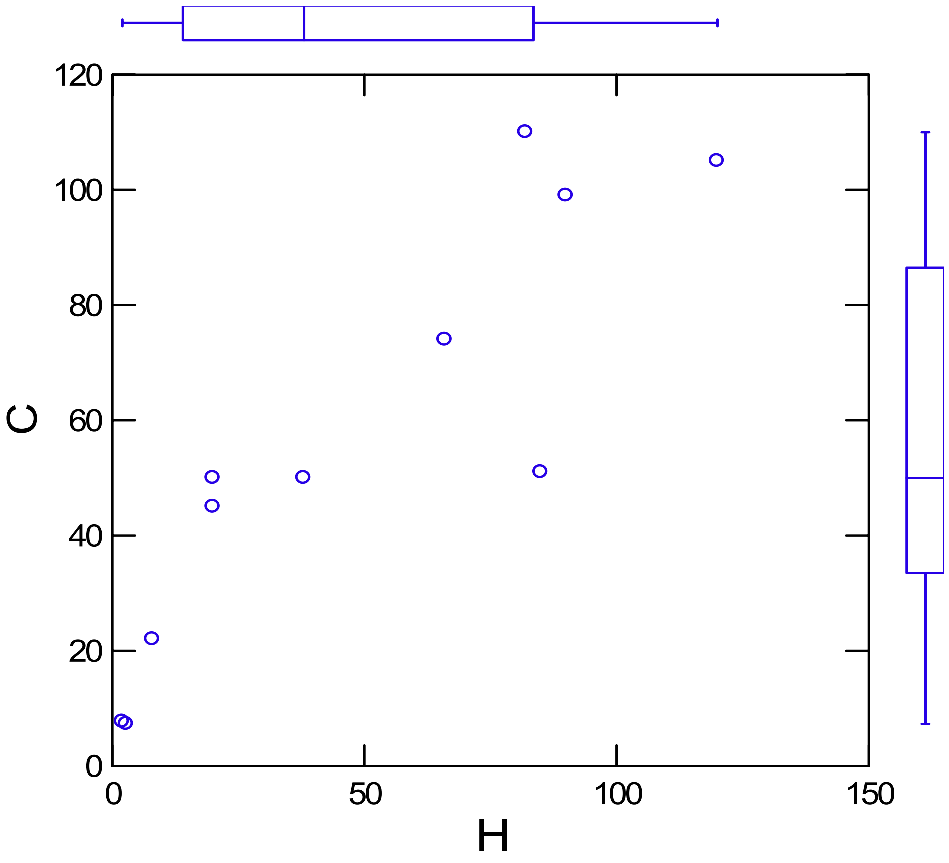

0
20
40
60
80
100
120
C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Fraction of Data
0
50
100
150
H,C
○ H
× C





	H	**C**
N of Cases	11	11
Minimum	2.000	7.30
Maximum	120	110
Range	118	103
Interquartile Range	73.25	65.00
Median	38.00	50.00
Arithmetic Mean	48.53	56.46
Geometric Mean	25.81	41.33
Standard Deviation	50	36.83
Variance	1,722	1,357
Coefficient of Variation	0.86	0.65

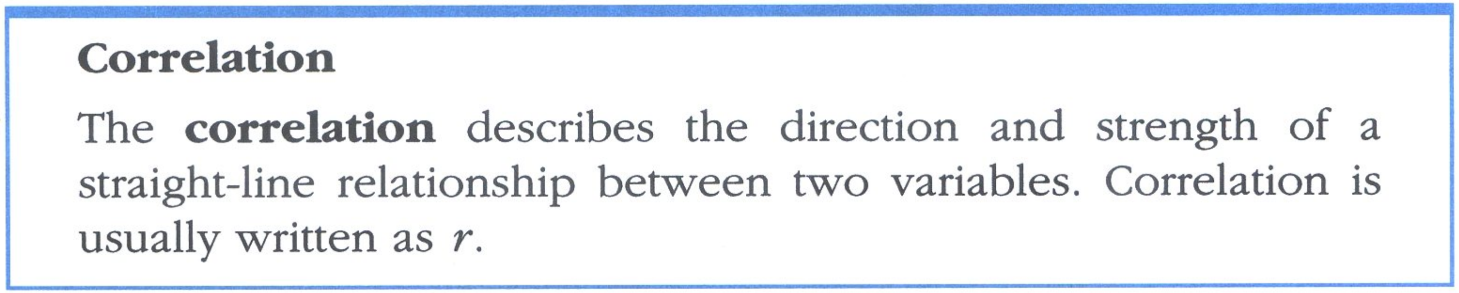


Correlations

		H	C
H	Pearson Correlation	1.000	.891**
	Sig. (2-tailed)	.	.000
	Sum of Squares and Cross-products	17220.98	13625.40
	Covariance	1722.098	1362.540
	N	11	11
C	Pearson Correlation	.891**	1.000
	Sig. (2-tailed)	.000	.
	Sum of Squares and Cross-products	13625.40	13566.31
	Covariance	1362.540	1356.631
	N	11	11

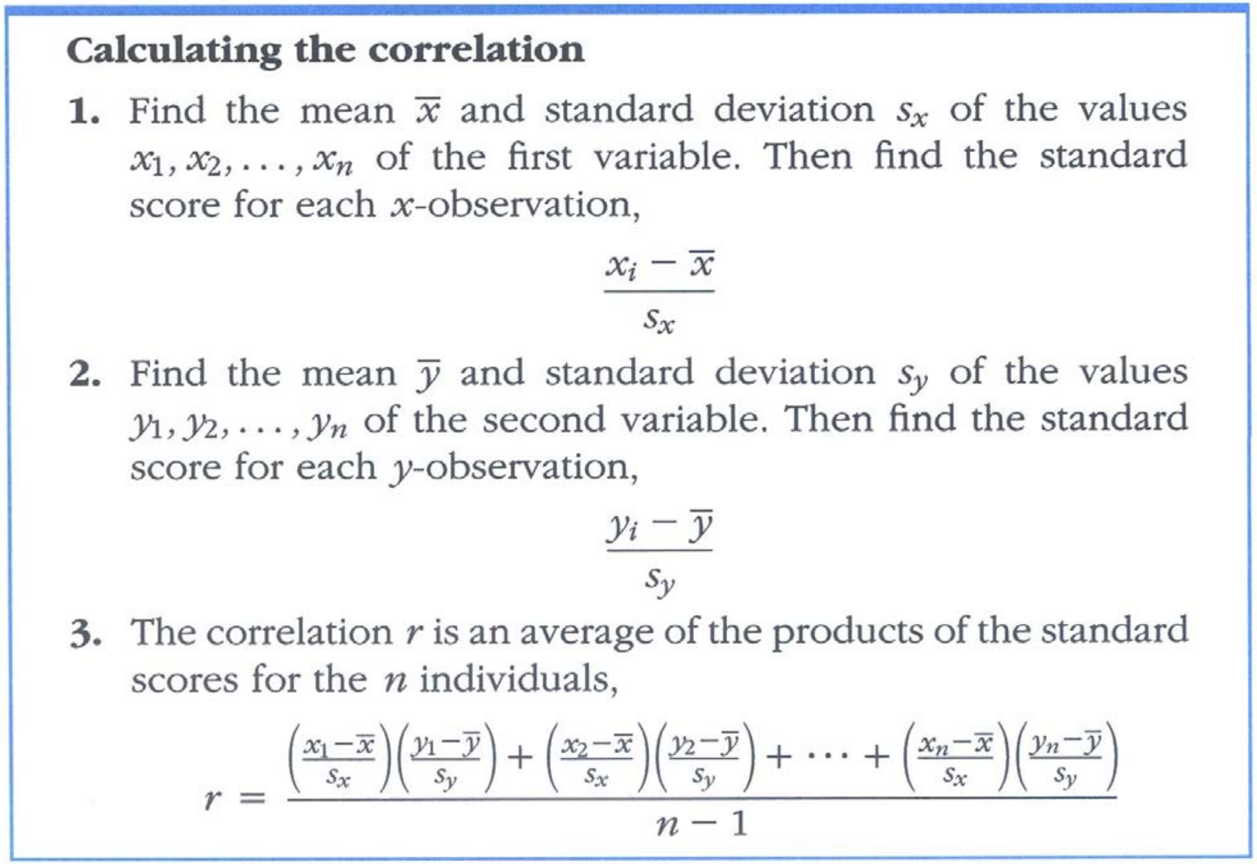
**. Correlation is significant at the 0.01 level (2-tailed).

Summarising



Correlation

The **correlation** describes the direction and strength of a straight-line relationship between two variables. Correlation is usually written as r.



Calculating the correlation

1. Find the mean $\overline{x}$ and standard deviation s_x of the values $x_1, x_2, \ldots, x_n$ of the first variable. Then find the standard score for each x-observation,

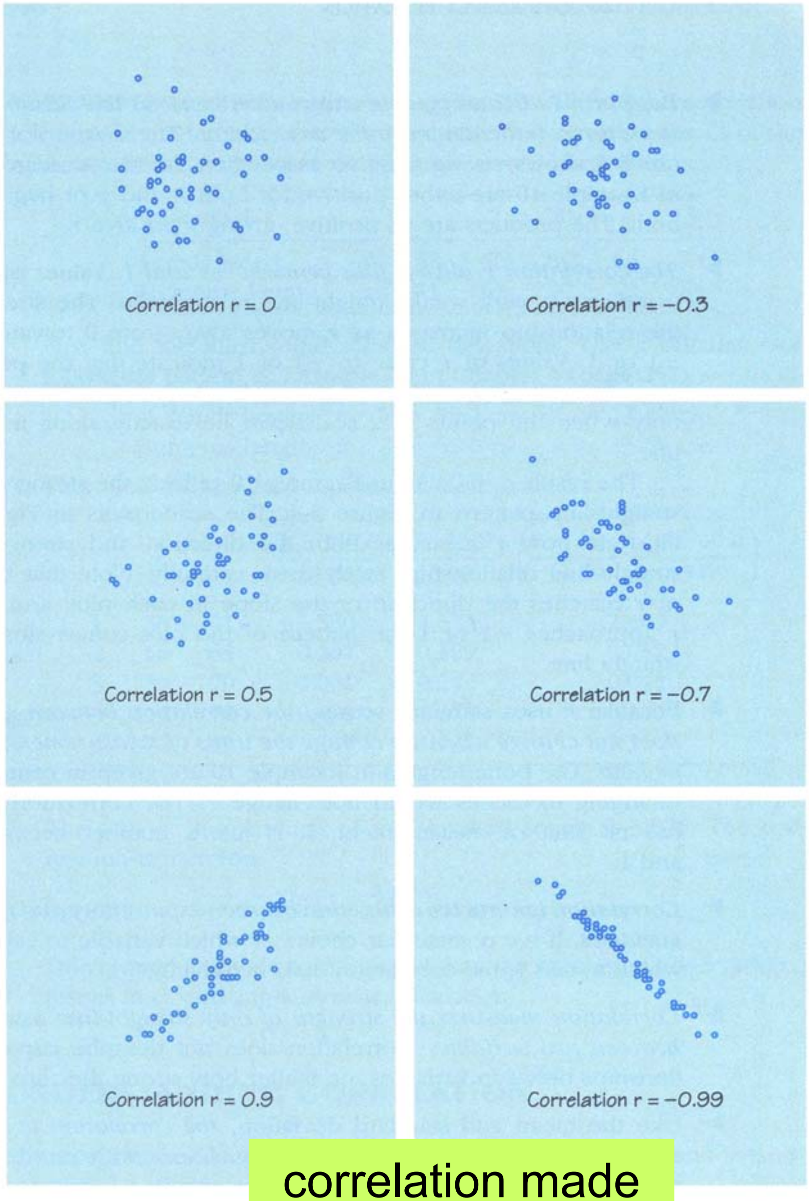
$$\frac{x_i - \overline{x}}{s_x}$$

2. Find the mean $\overline{y}$ and standard deviation s_y of the values $y_1, y_2, \ldots, y_n$ of the second variable. Then find the standard score for each y-observation,

$$\frac{y_i - \overline{y}}{s_y}$$

3. The correlation r is an average of the products of the standard scores for the n individuals,

$$r = \frac{\left(\frac{x_1-\overline{x}}{s_x}\right)\left(\frac{y_1-\overline{y}}{s_y}\right) + \left(\frac{x_2-\overline{x}}{s_x}\right)\left(\frac{y_2-\overline{y}}{s_y}\right) + \cdots + \left(\frac{x_n-\overline{x}}{s_x}\right)\left(\frac{y_n-\overline{y}}{s_y}\right)}{n-1}$$



correlation made visible

Positive *r* indicates positive association between the variables and negative r indicates negative association.

The correlation *r* always falls between -1 and 1.

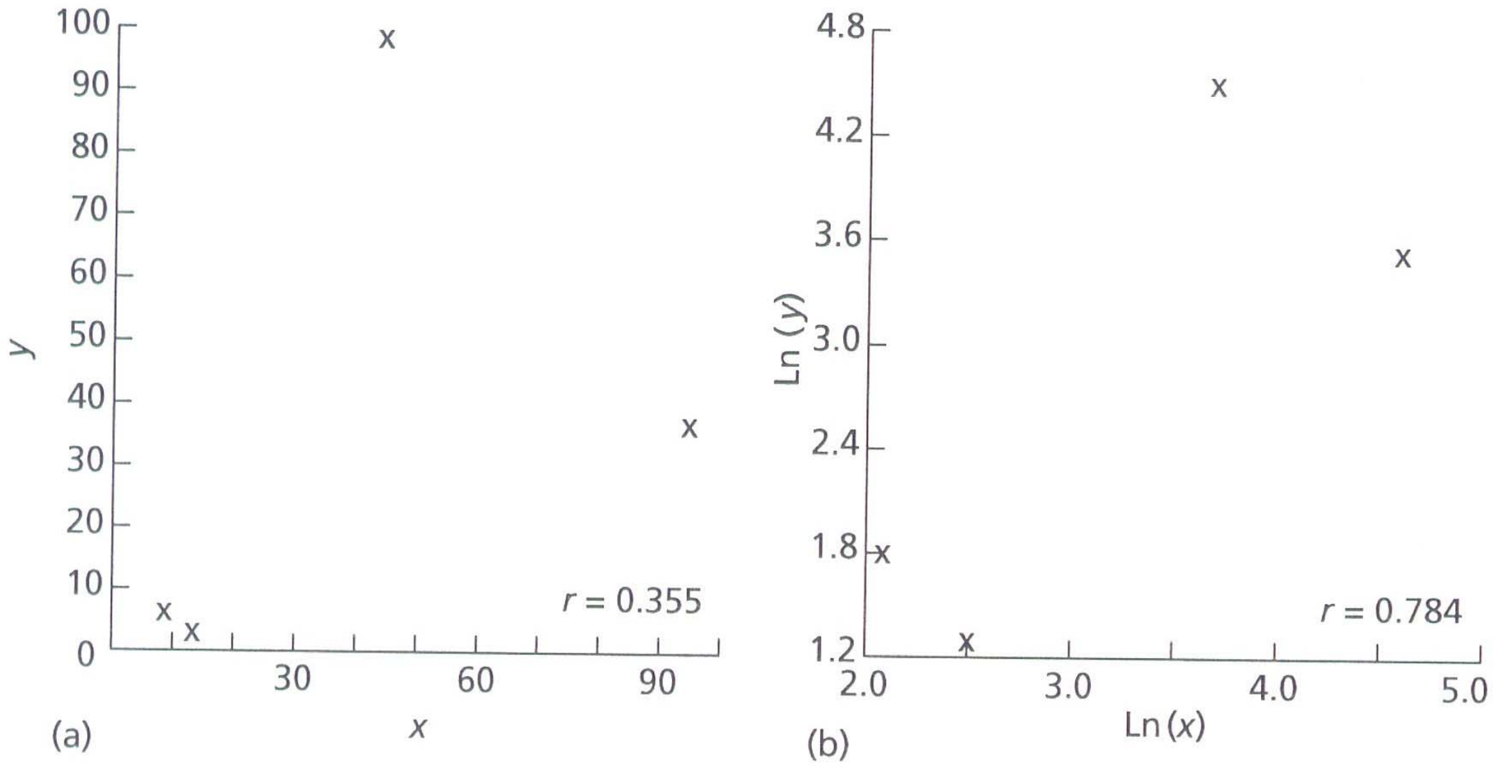
Because *r* uses standard scores, the correlation between x and y does not change when we change the units of the variables.

Correlation ignores the distinction between explanatory and response variables.

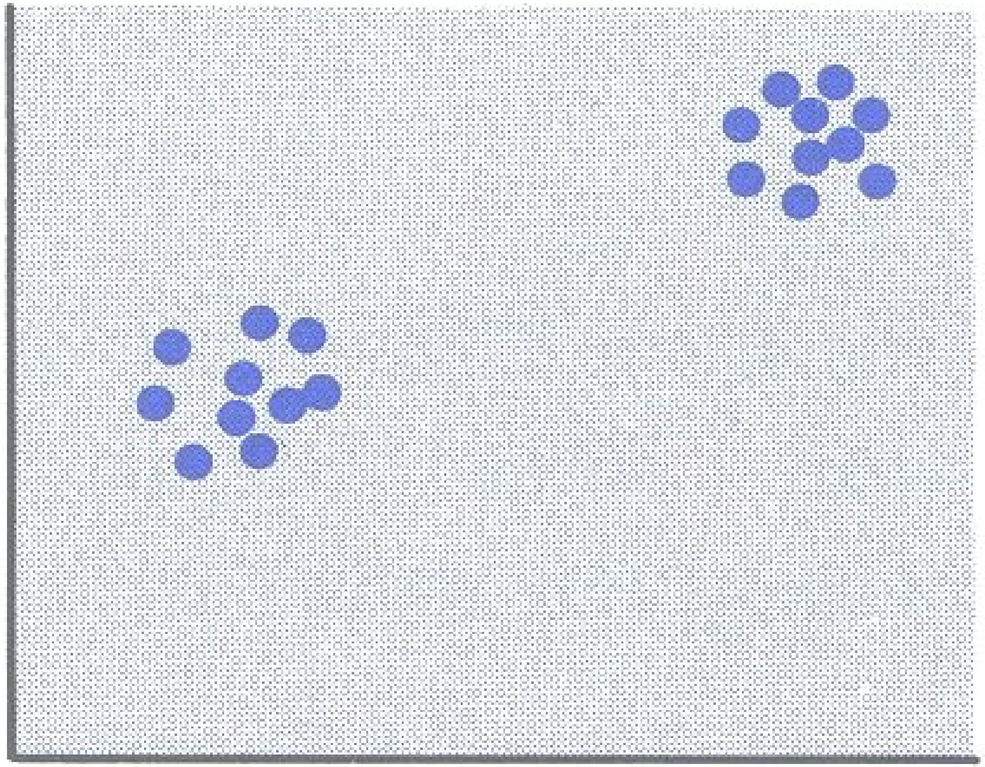
Correlation measures the strength of only straight line association between two variables

Like the mean and the standard deviation the correlation is strongly affected by a few **outlying** observations.

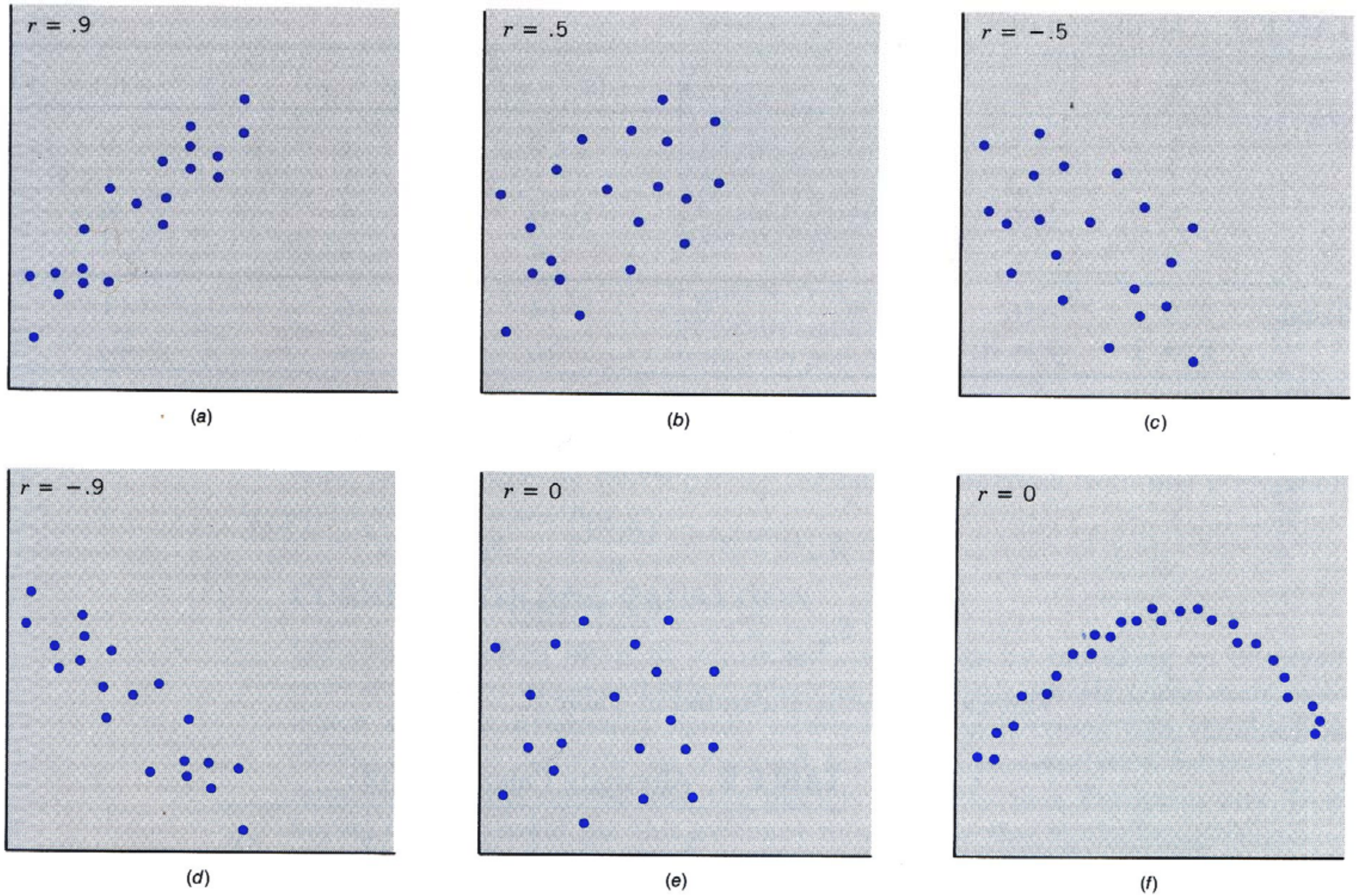
Use *r* with caution when **outliers** appear in the scatterplot!



Moreover r is not appropriate when data are clustered!



Correspondence between the values of *r* and the amount of **scatter**



(*a*) (*b*) (*c*)

(*d*) (*e*) (*f*)