Correlation

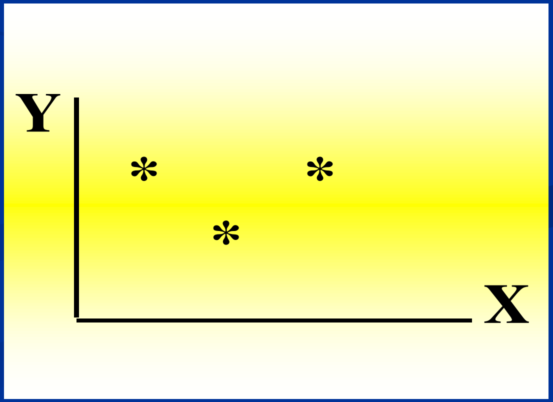
Correlation

Finding the relationship between two quantitative variables without being able to infer causal relationships

Correlation is a statistical technique used to determine the degree to which two variables are related

Scatter diagram

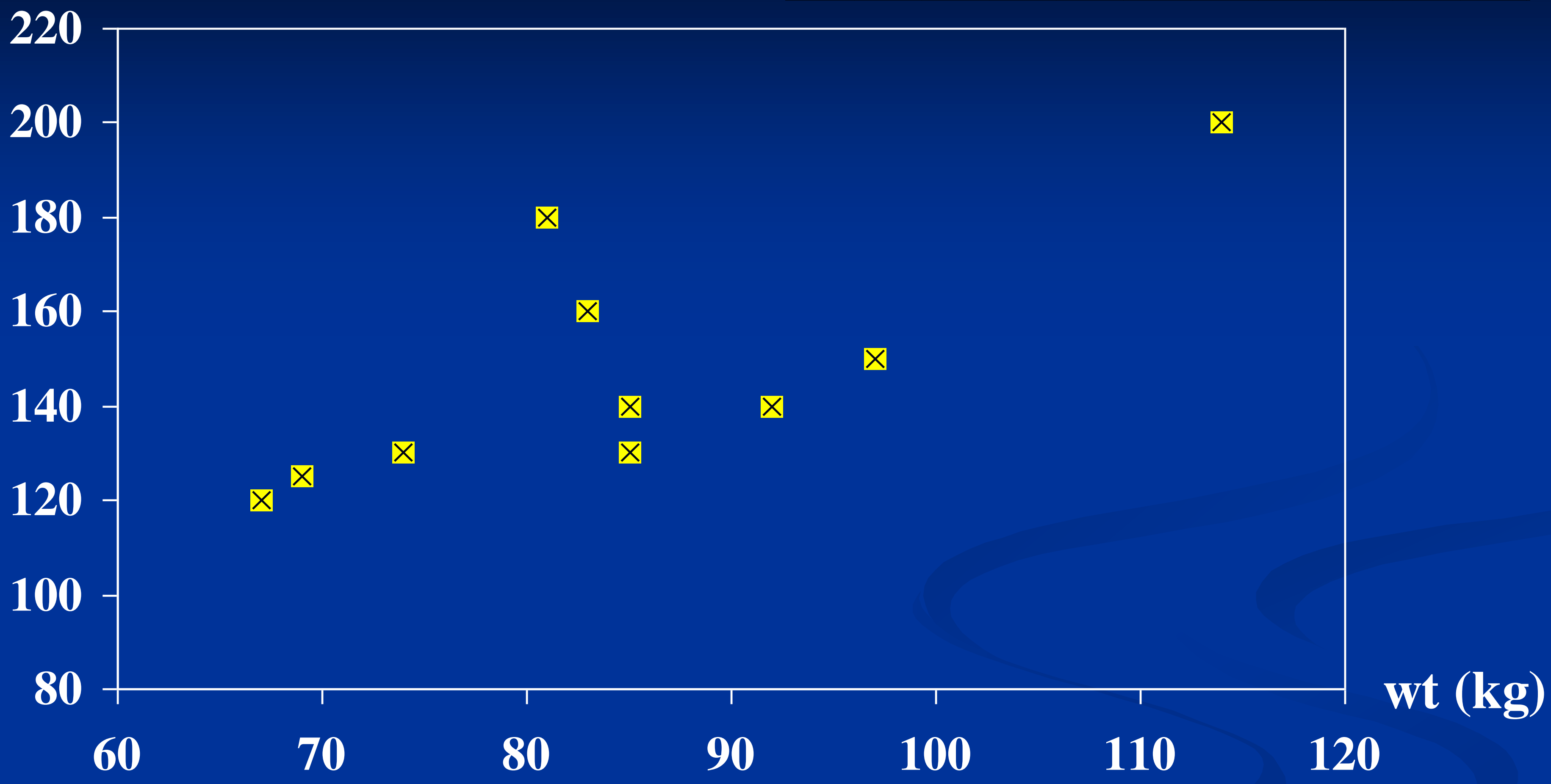
- Rectangular coordinate
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined
- No frequency table



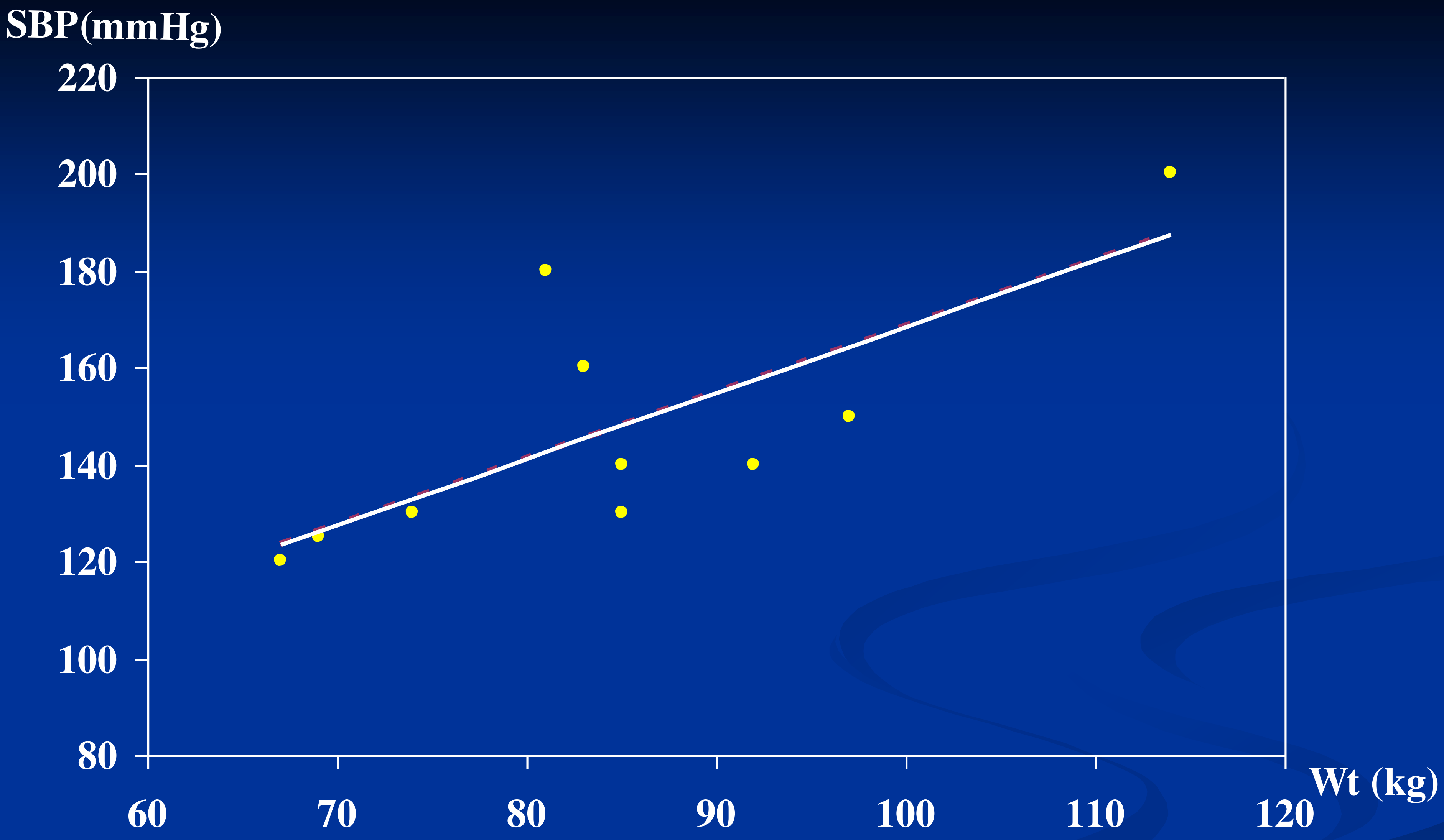
Example

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP mHg)	120	125	140	160	130	180	150	140	200	130

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP mHg)	120	125	140	160	130	180	150	140	200	130



Scatter diagram of weight and systolic blood pressure



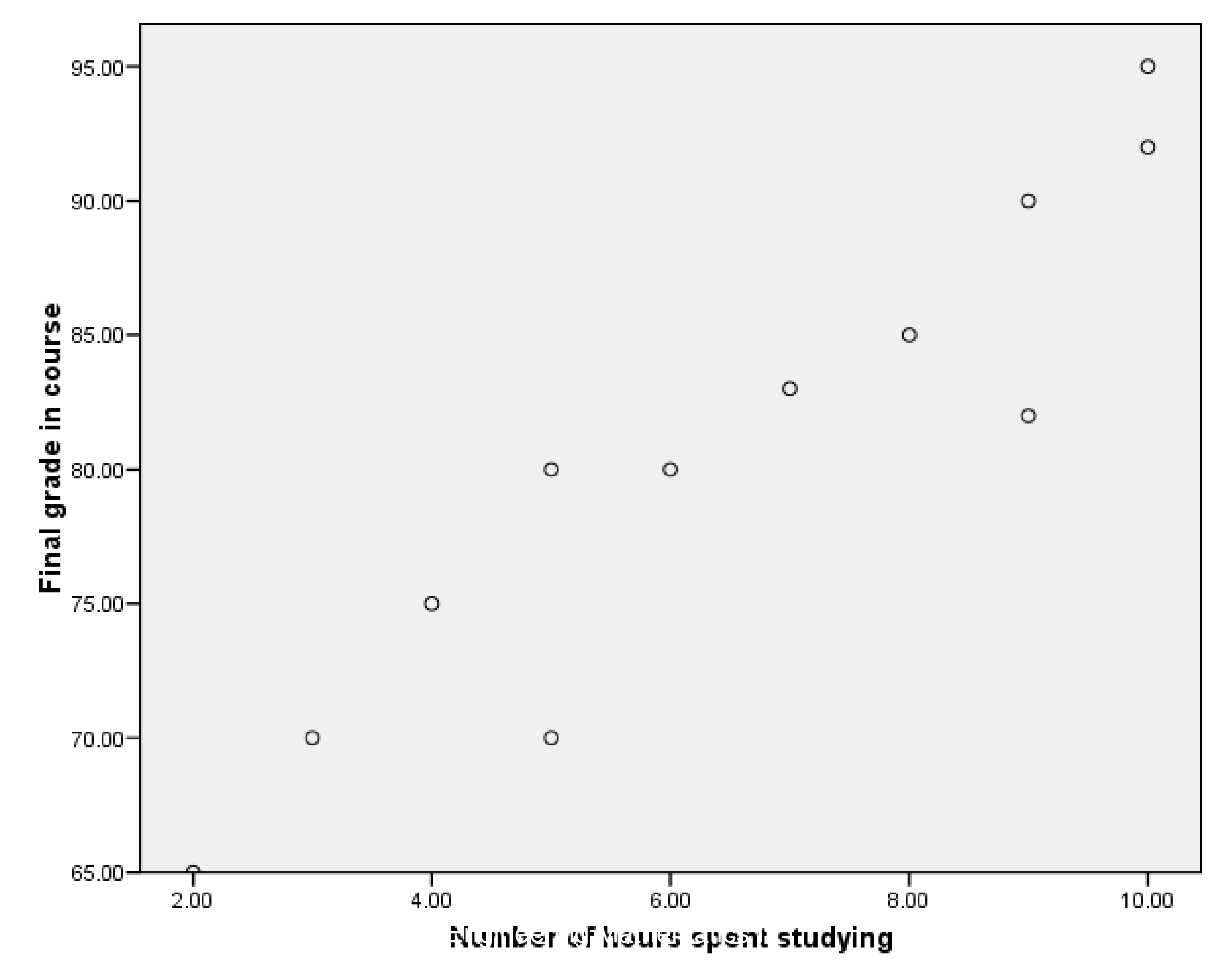
Scatter diagram of weight and systolic blood pressure

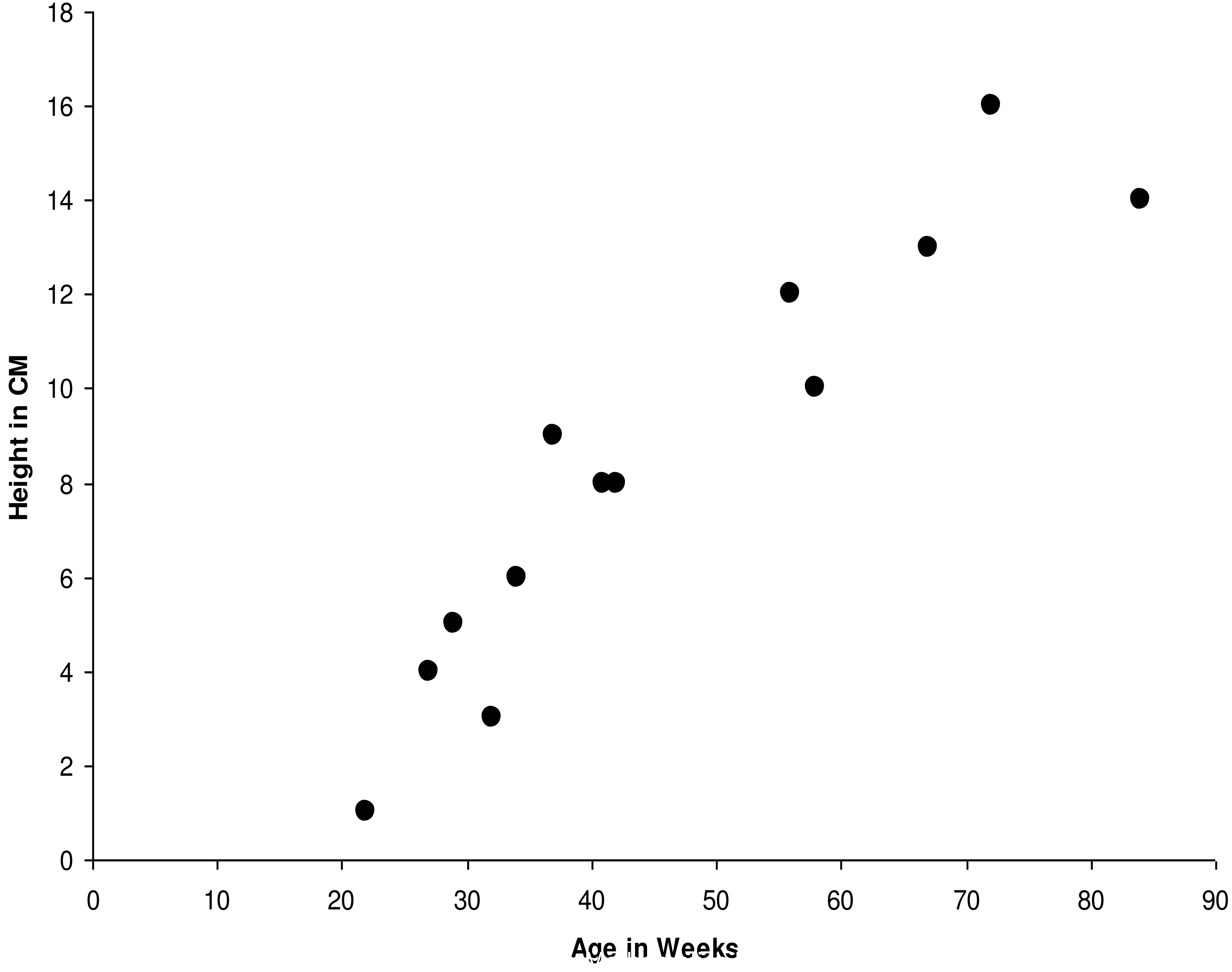
Scatter plots

The pattern of data is indicative of the type of relationship between your two variables:

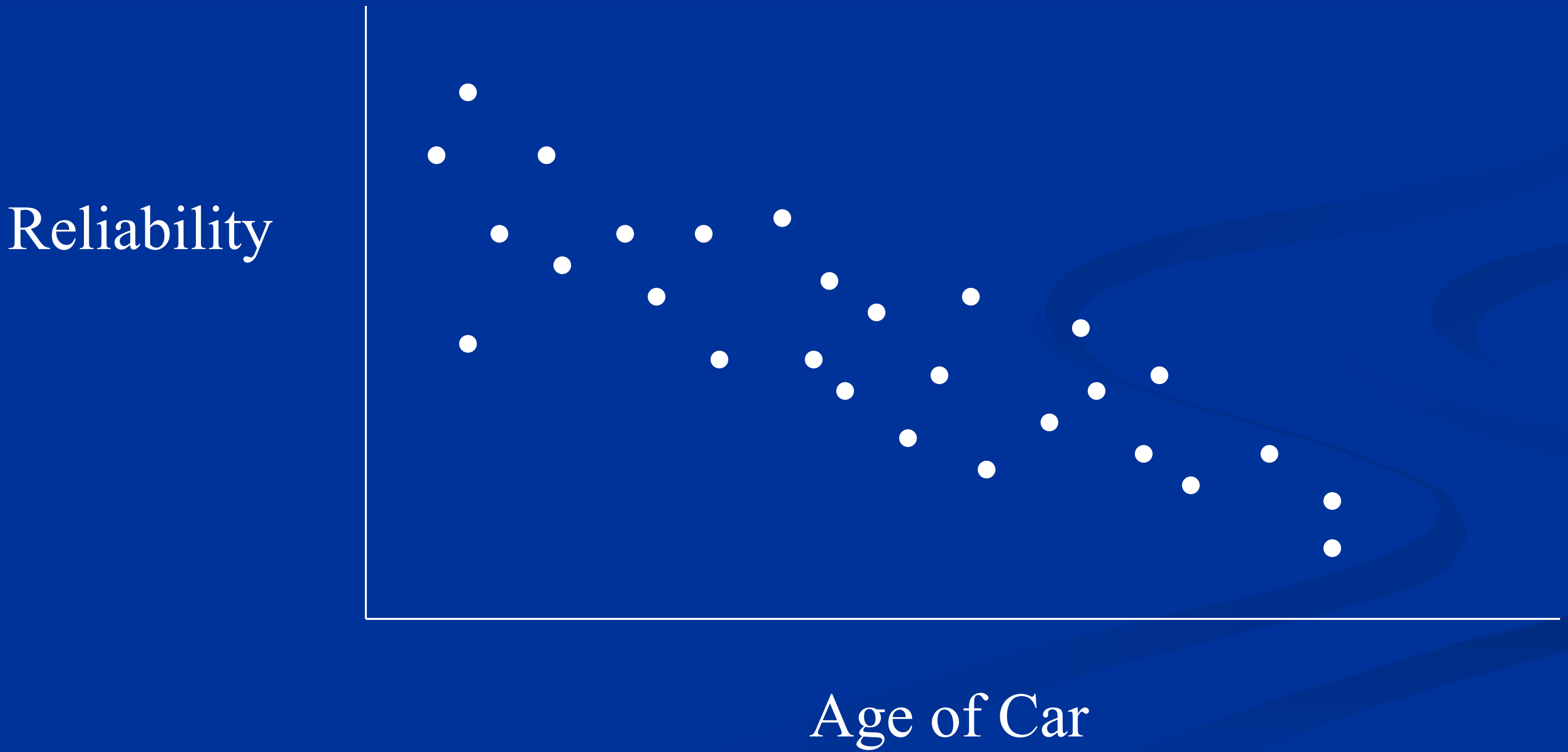
- positive relationship
- negative relationship
- no relationship

Positive relationship

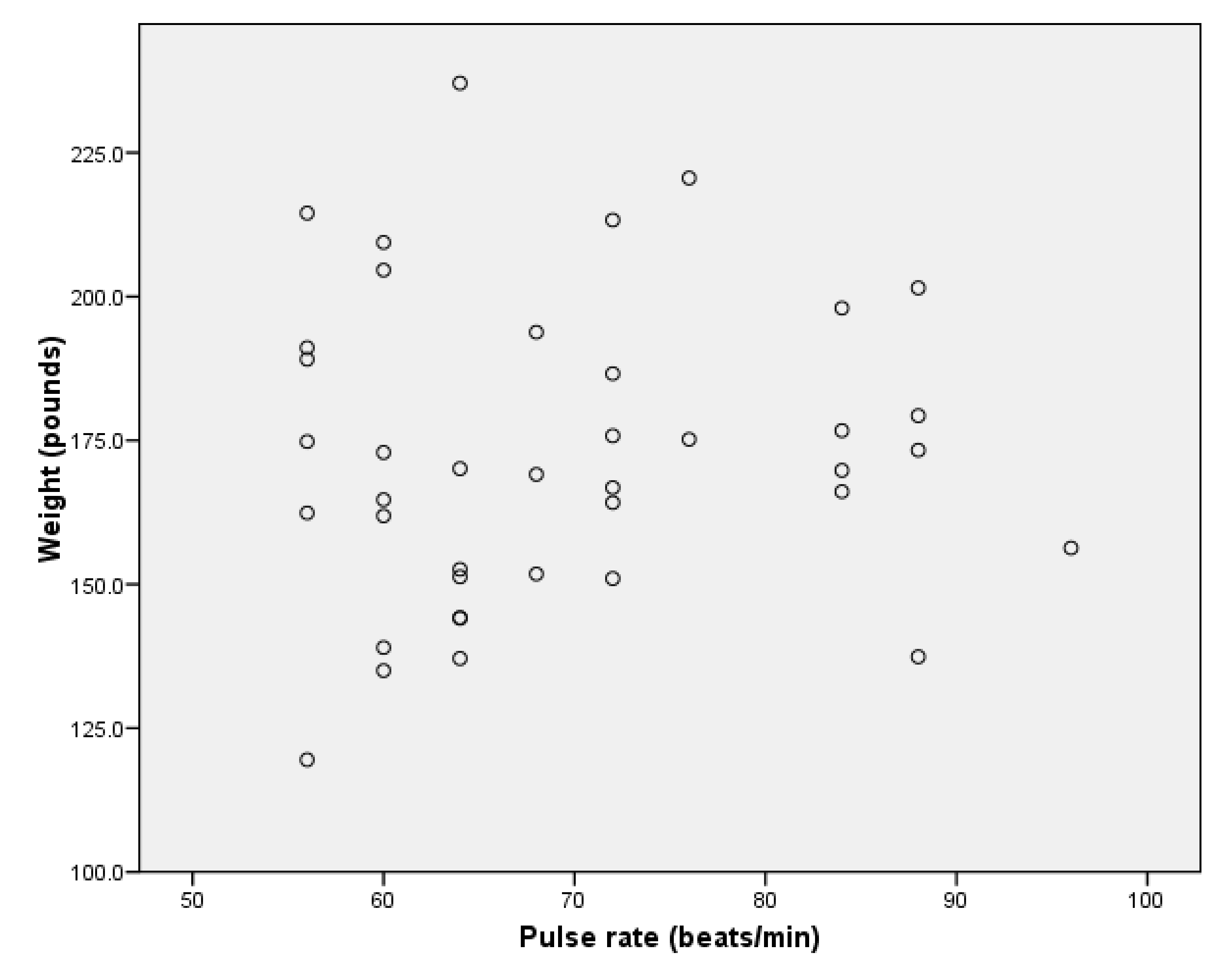



Height in CM
18
16
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60
70
80
90
Age in Weeks

Negative relationship



No relation



225.0
200.0
175.0
150.0
125.0
100.0
Weight (pounds)
50
60
70
80
90
100
Pulse rate (beats/min)

Correlation coefficient (r)

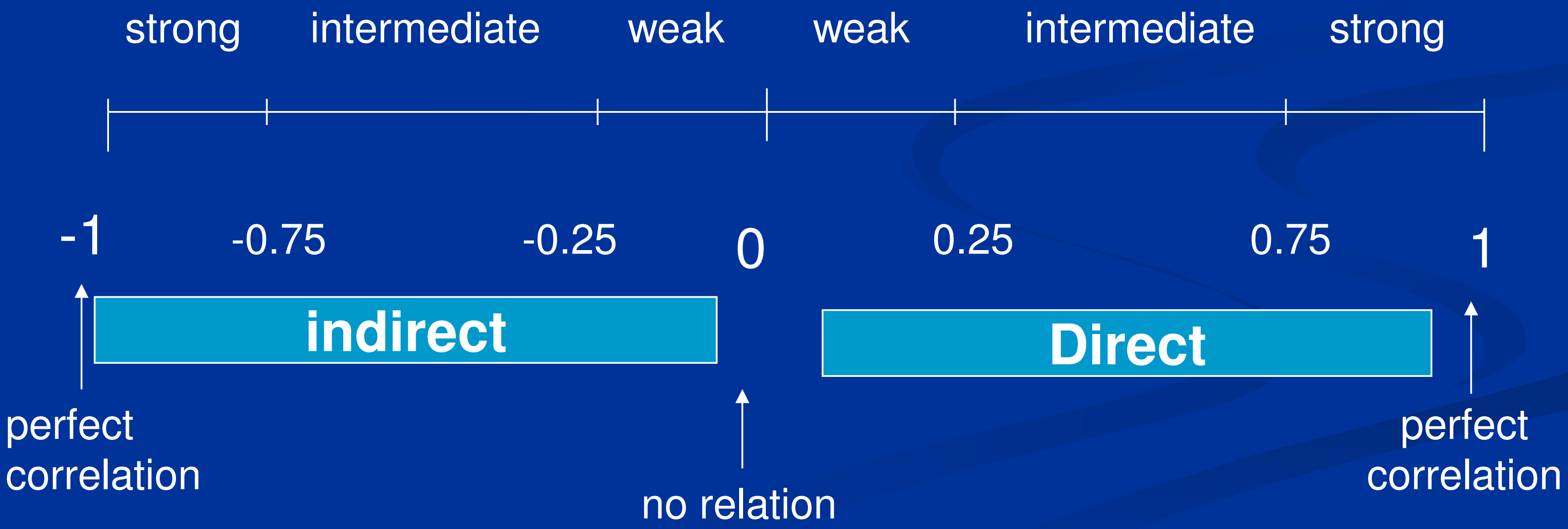
- It is also called Pearson's correlation or product moment correlation coefficient.
- It measures the nature and strength between two variables of the quantitative type.

- The sign of r denotes the nature of association
- while the value of r denotes the strength of association.

- If the sign is +ve this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).

- While if the sign is -ve this means an inverse or indirect relationship (which means an increase in one variable is associated with a decrease in the other).

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated
  by the following diagram.



- If r = Zero this means no association or correlation between the two variables.

- If $0 < r < 0.25$ = weak correlation.

- If $0.25 \leq r < 0.75$ = intermediate correlation.

- If $0.75 \leq r < 1$ = strong correlation.

- If $r = l$ = perfect correlation.

How to compute the simple correlation coefficient (r)

$$r = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \dfrac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \dfrac{(\sum y)^2}{n}\right)}}$$

Example:

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

serial No	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

These 2 variables are of the quantitative type, one variable (Age) is called the independent and denoted as (X) variable and the other (weight) is called the dependent and denoted as (Y) variables to find the relation between age and weight compute the simple correlation coefficient using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Serial n.	Age (years) (x)	Weight (Kg) (y)	xy	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	$\sum x=$ 41	$\sum y=$ 66	$\sum xy=$ 461	$\sum x2=$ 291	$\sum y2=$ 742

$$r = \frac{461 - \dfrac{41 \times 66}{6}}{\sqrt{\left[291 - \dfrac{(41)^2}{6}\right] . \left[742 - \dfrac{(66)^2}{6}\right]}}$$

r = 0.759

strong direct correlation

EXAMPLE: Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY=129$

Calculating Correlation Coefficient

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{\left(6(230) - 32^2\right)\left(6(204) - 32^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -.94$$

r = - 0.94

Indirect strong correlation

Spearman Rank Correlation Coefficient (r_s)

- It is a non-parametric measure of correlation.
- This procedure makes use of the two sets of ranks that may be assigned to the sample values of x and Y.
- Spearman Rank correlation coefficient could be computed in the following cases:
- Both variables are quantitative.
- Both variables are qualitative ordinal.
- One variable is quantitative and the other is qualitative ordinal.

Procedure:

1. Rank the values of X from 1 to n where n is the numbers of pairs of values of X and Y in the sample.
2. Rank the values of Y from 1 to n.
3. Compute the value of di for each pair of observation by subtracting the rank of Yi from the rank of Xi
4. Square each di and compute ∑di2 which is the sum of the squared values.

5. Apply the following formula

$$r_s = 1 - \frac{6\sum(di)^2}{n(n^2 - 1)}$$

- The value of r_s denotes the magnitude and nature of association giving the same interpretation as simple r.

Example

In a study of the relationship between level education and income the following data was obtained. Find the relationship between them and comment.

sample numbers	level education (X)	Income (Y)
A	Preparatory.	25
B	Primary.	10
C	University.	8
D	secondary	10
E	secondary	15
F	illiterate	50
G	University.	60

Answer:

	(X)	(Y)	Rank X	Rank Y	di	di^2
A	Preparatory	25	5	3	2	4
B	Primary.	10	6	5.5	0.5	0.25
C	University.	8	1.5	7	-5.5	30.25
D	secondary	10	3.5	5.5	-2	4
E	secondary	15	3.5	4	-0.5	0.25
F	illiterate	50	7	2	5	25
G	university.	60	1.5	1	0.5	0.25

$\sum di^2=64$

$$r_s = 1 - \frac{6 \times 64}{7(48)} = -0.1$$

Comment:

There is an indirect weak correlation between level of education and income.