Correlation

-Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Correlation

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

Properties of Correlation coefficient

- The correlation coefficient lies between -1 & +1 symbolically $(- 1 \leq r \geq 1)$
- The correlation coefficient is independent of the change of origin & scale.
- The coefficient of correlation is the geometric mean of two regression coefficient.

$$r = \sqrt{bxy * byx}$$

- The one regression coefficient is (+ve) other regression coefficient is also (+ve) correlation coefficient is (+ve)

(i.e. Same sign)

Methods of Studying Correlation

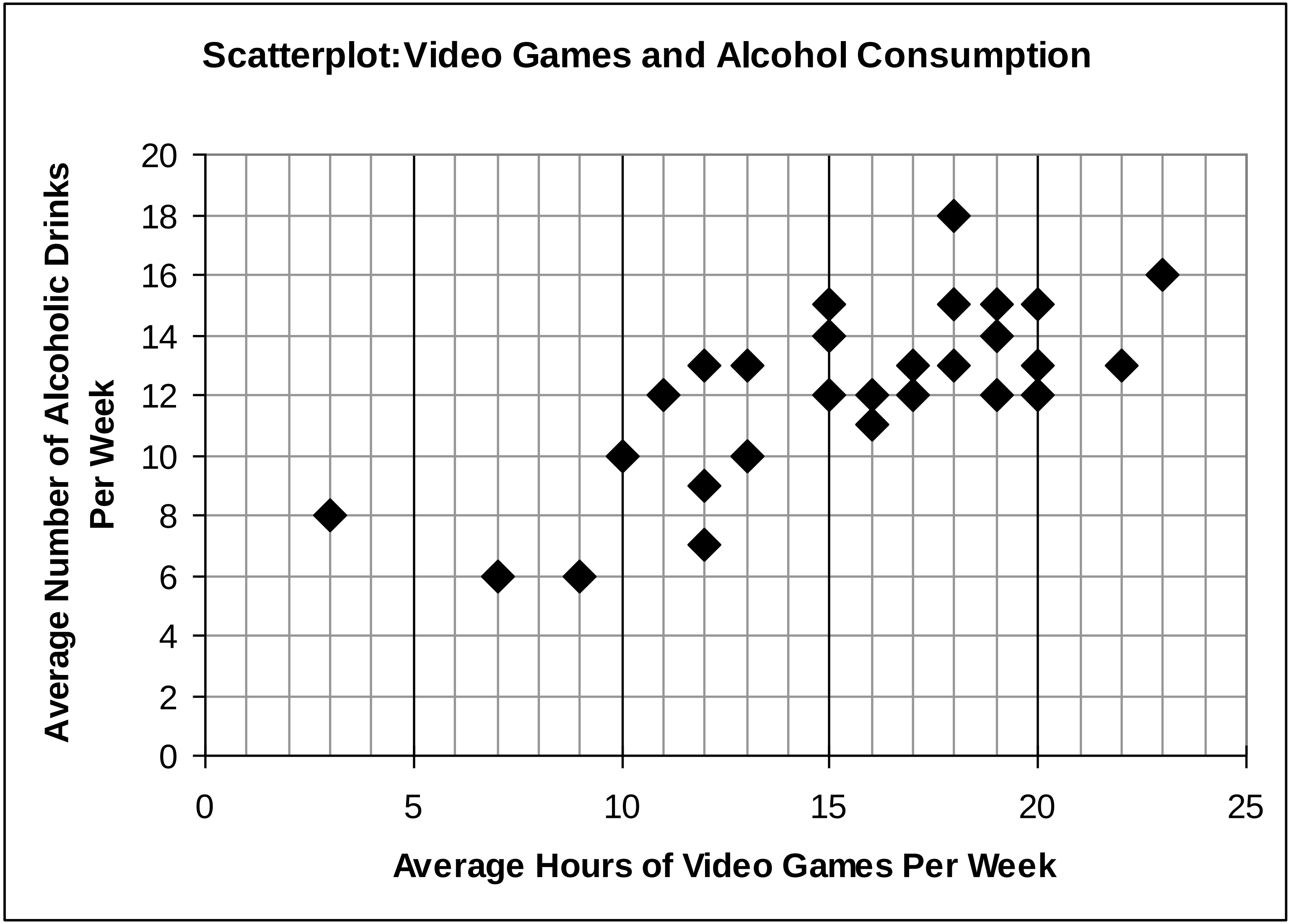
- Scatter Diagram Method
- Karl Pearson’s Coefficient of Correlation
- Spearman’s Rank Correlation
- Kendall rank correlation coefficient

1. Scatter diagram

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

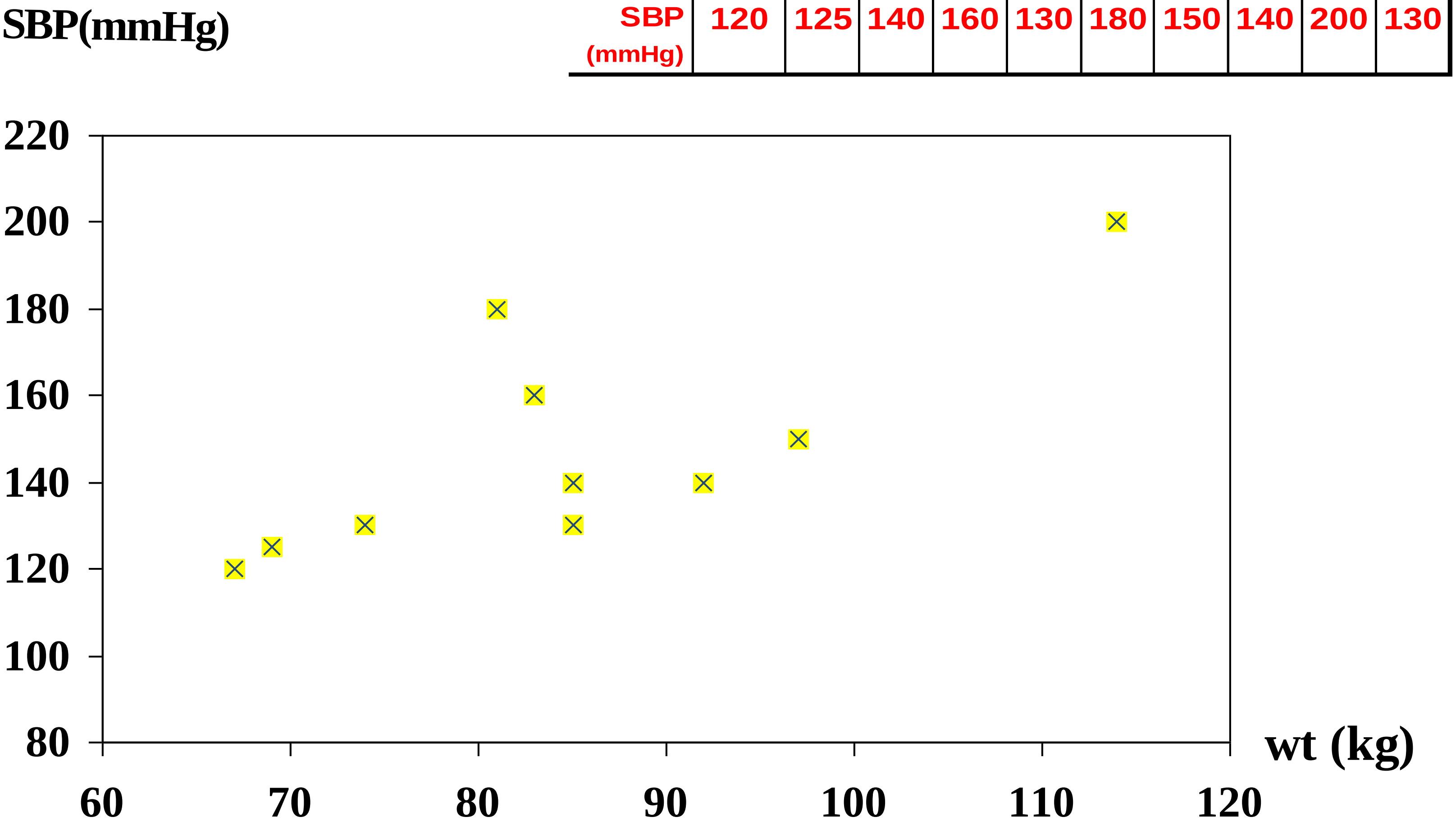
Example of Scatter Plot



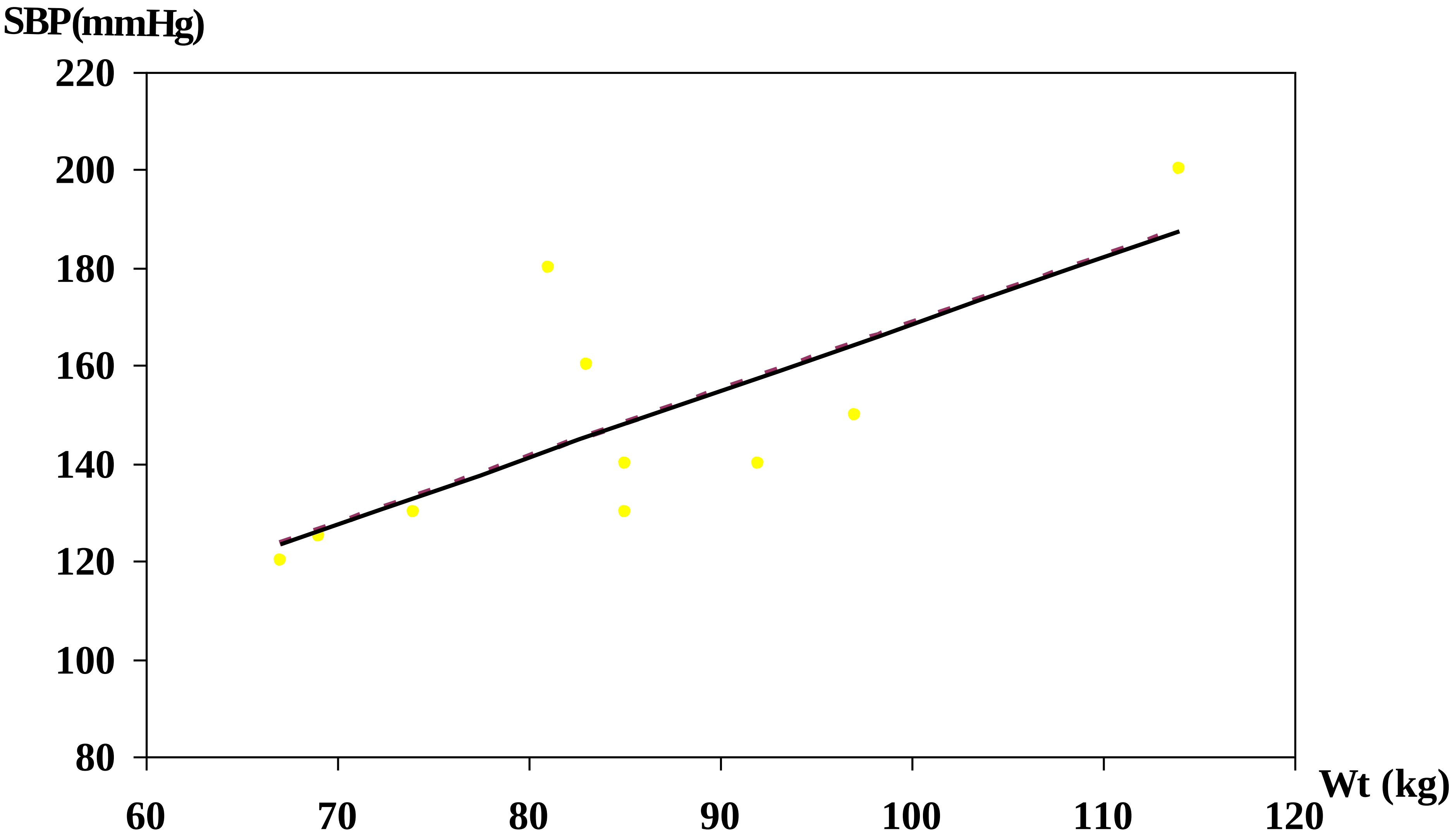
Example

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

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



Scatter diagram of weight and systolic blood pressure



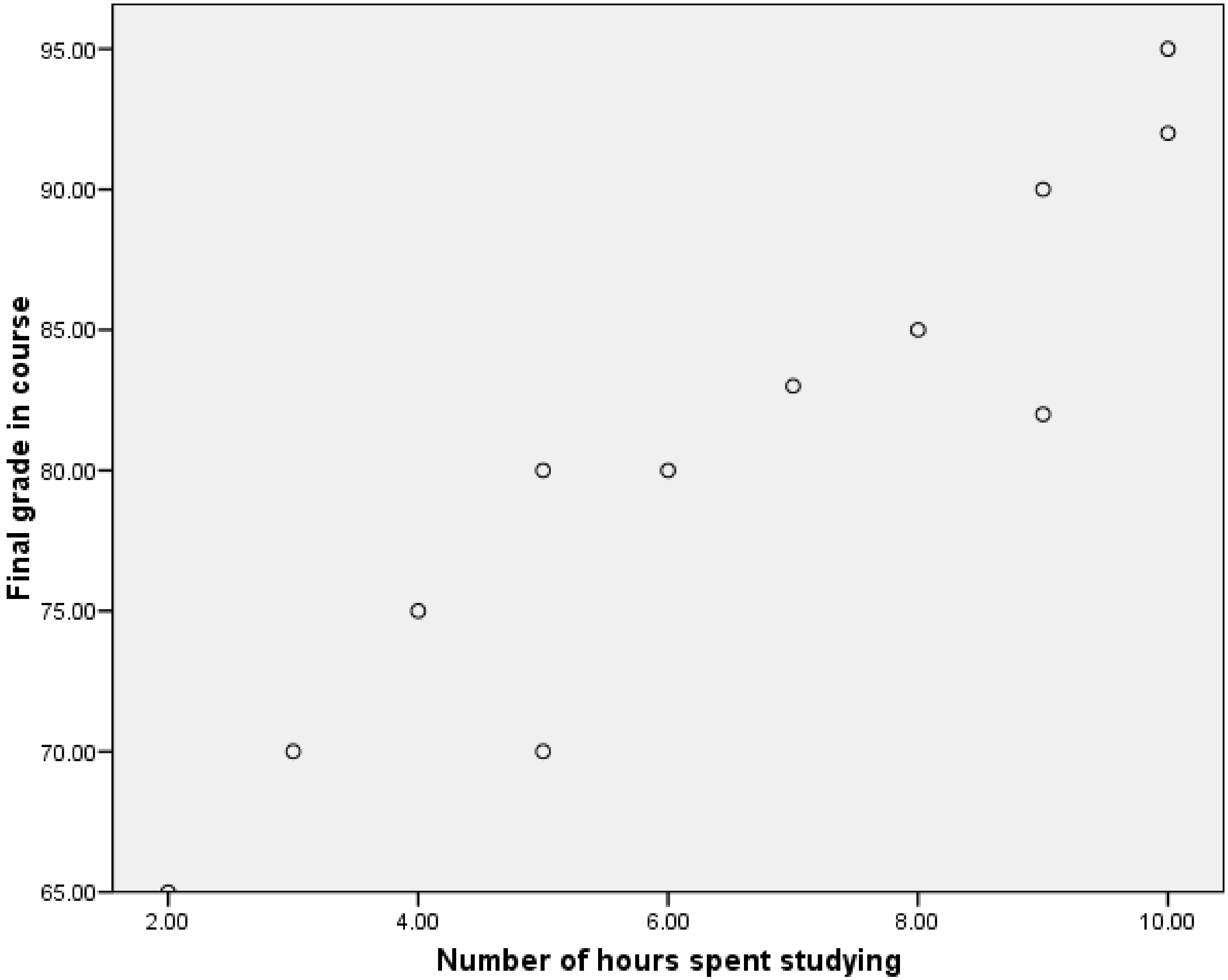
Scatter diagram of weight and systolic blood pressure

Scatter plots

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

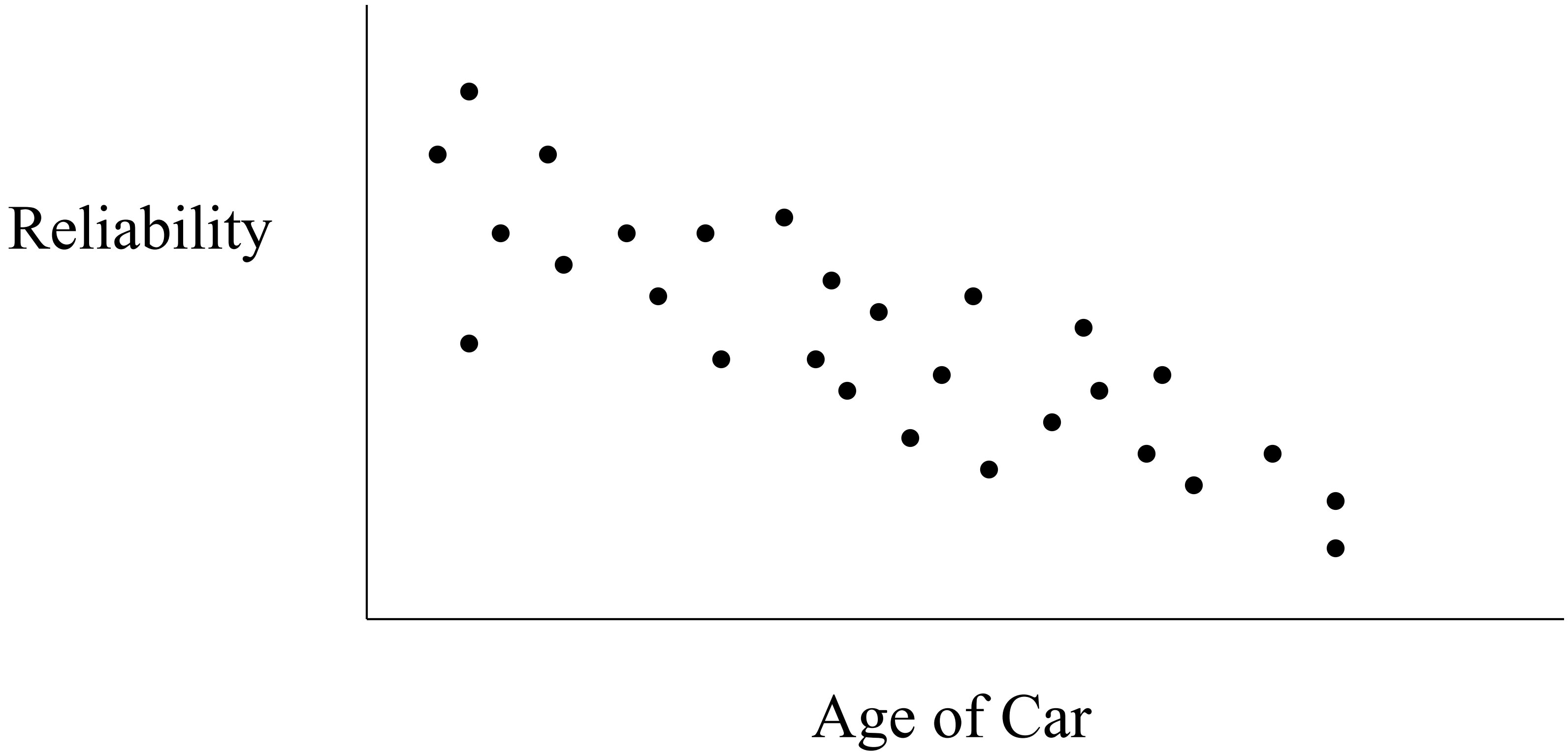
- positive relationship
- negative relationship
- no relationship

Positive relationship



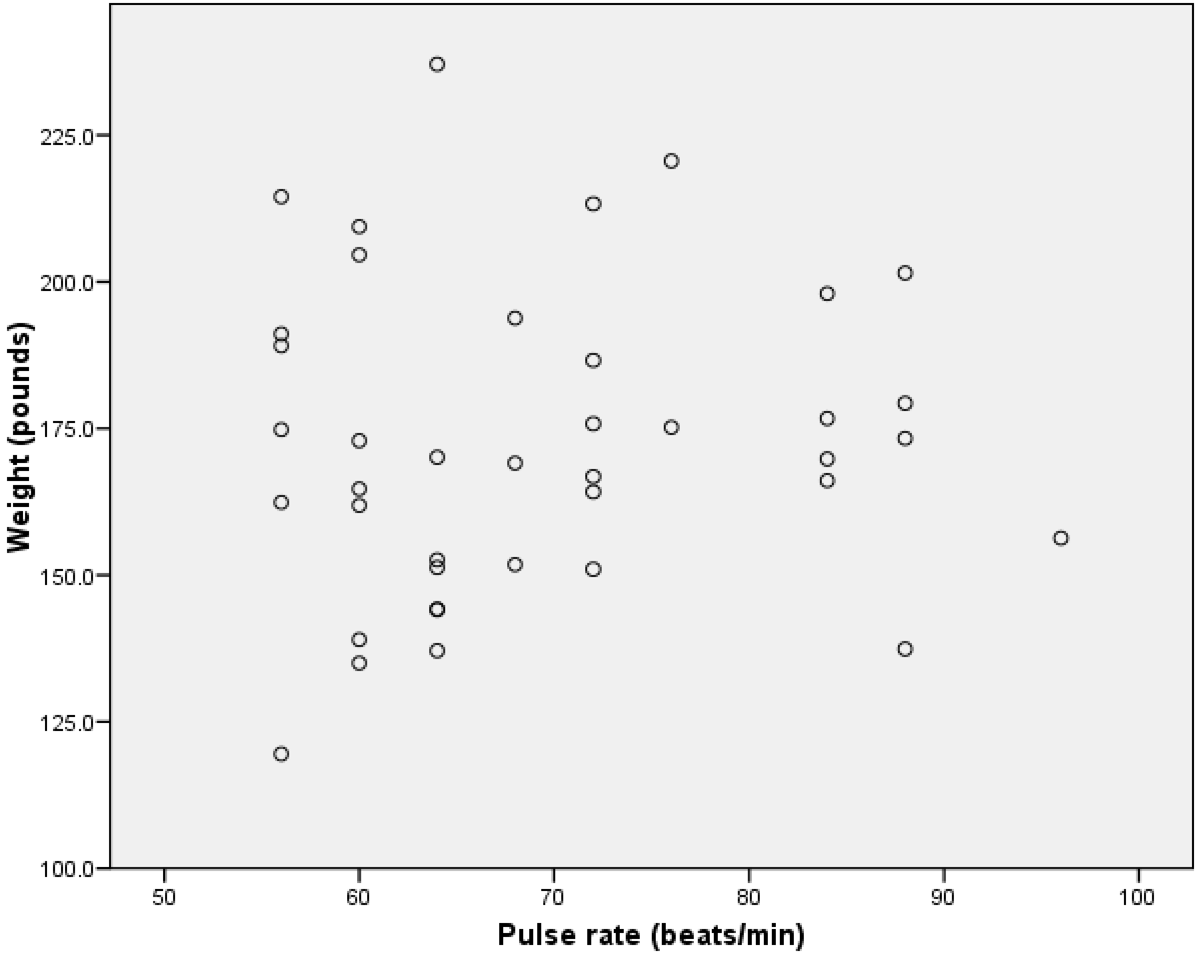
Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Negative relationship



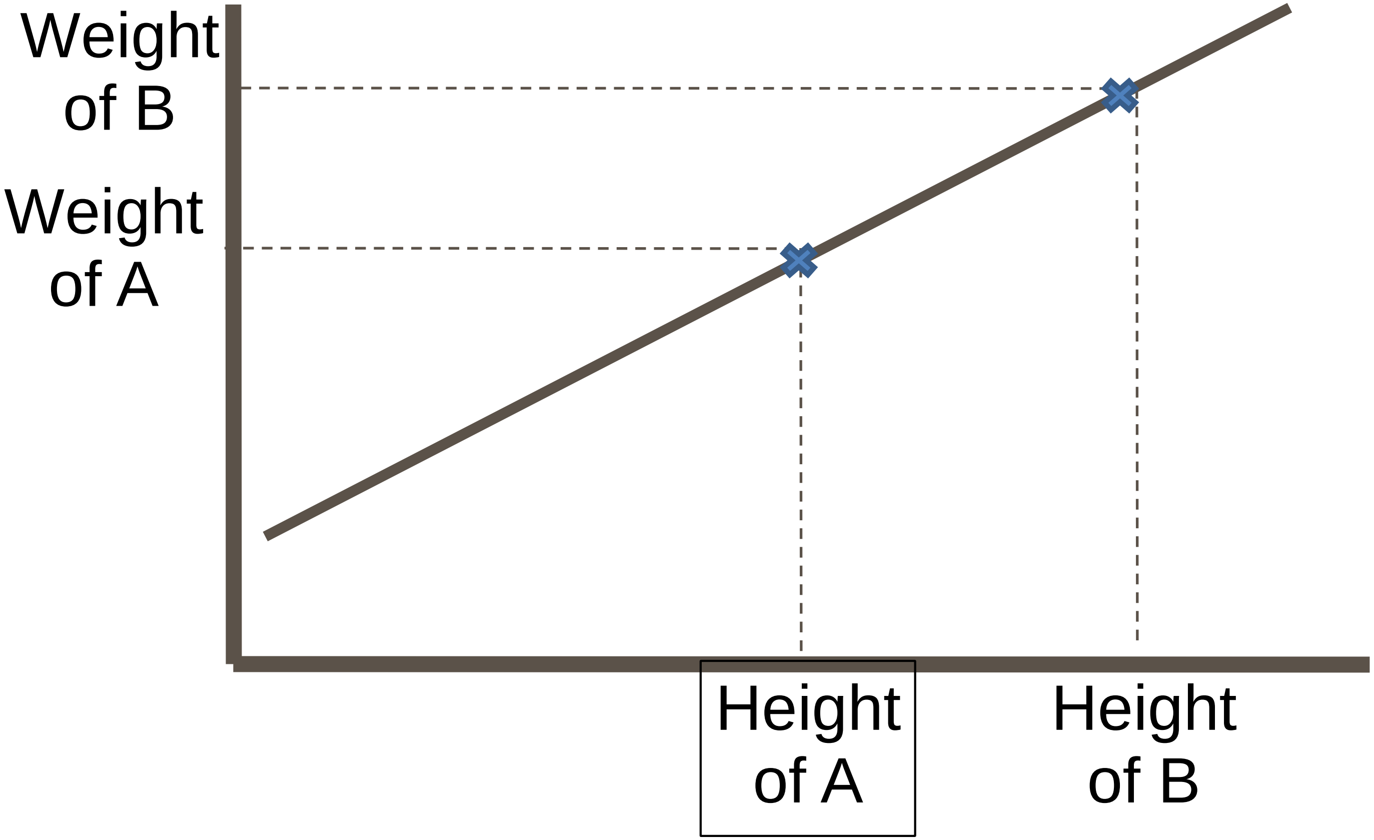
Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

No relation



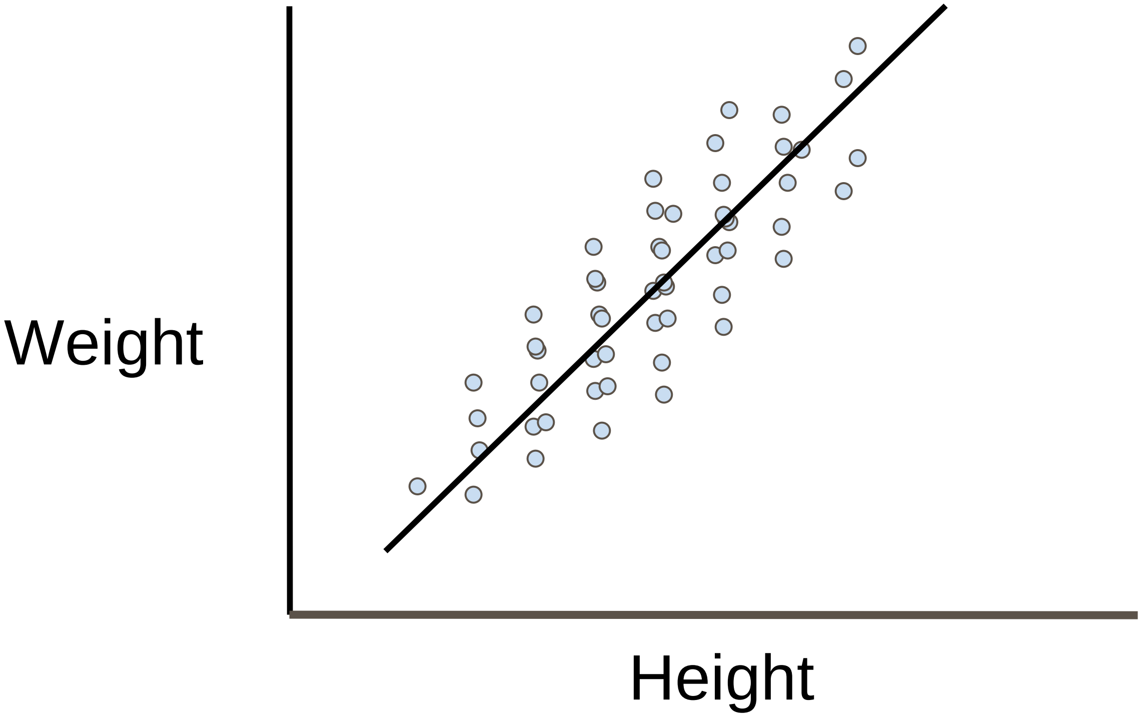
Weight (pounds)
225.0
200.0
175.0
150.0
125.0
100.0
50
60
70
80
90
100
Pulse rate (beats/min)

A perfect positive correlation



High Degree of positive correlation

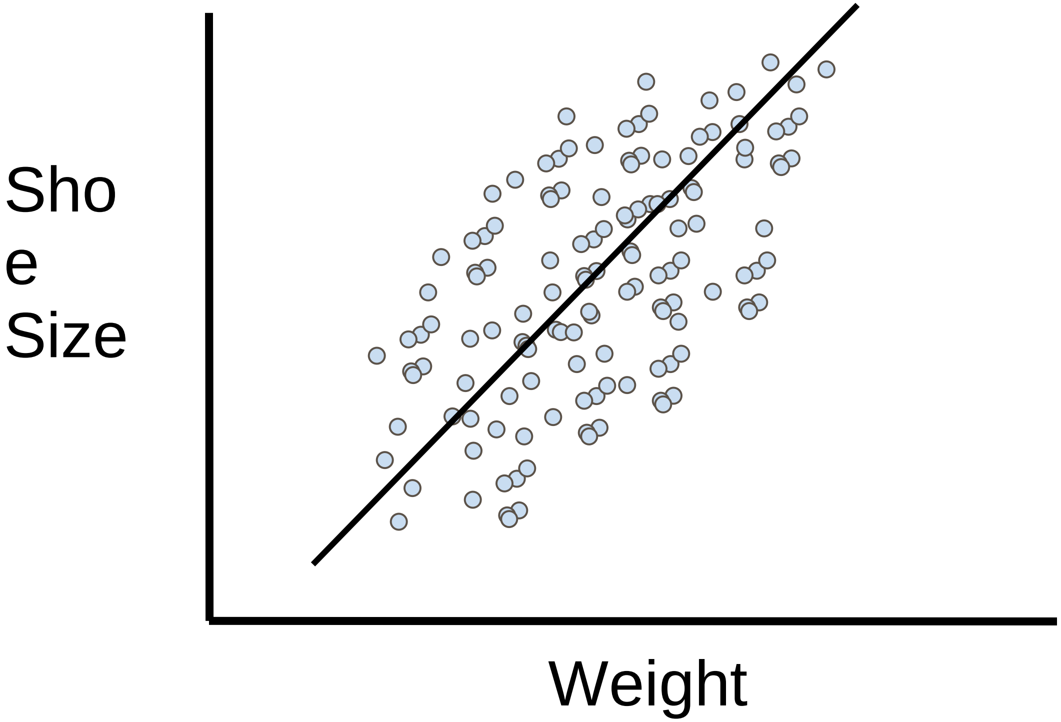
- Positive relationship



r = +.80

Degree of correlation

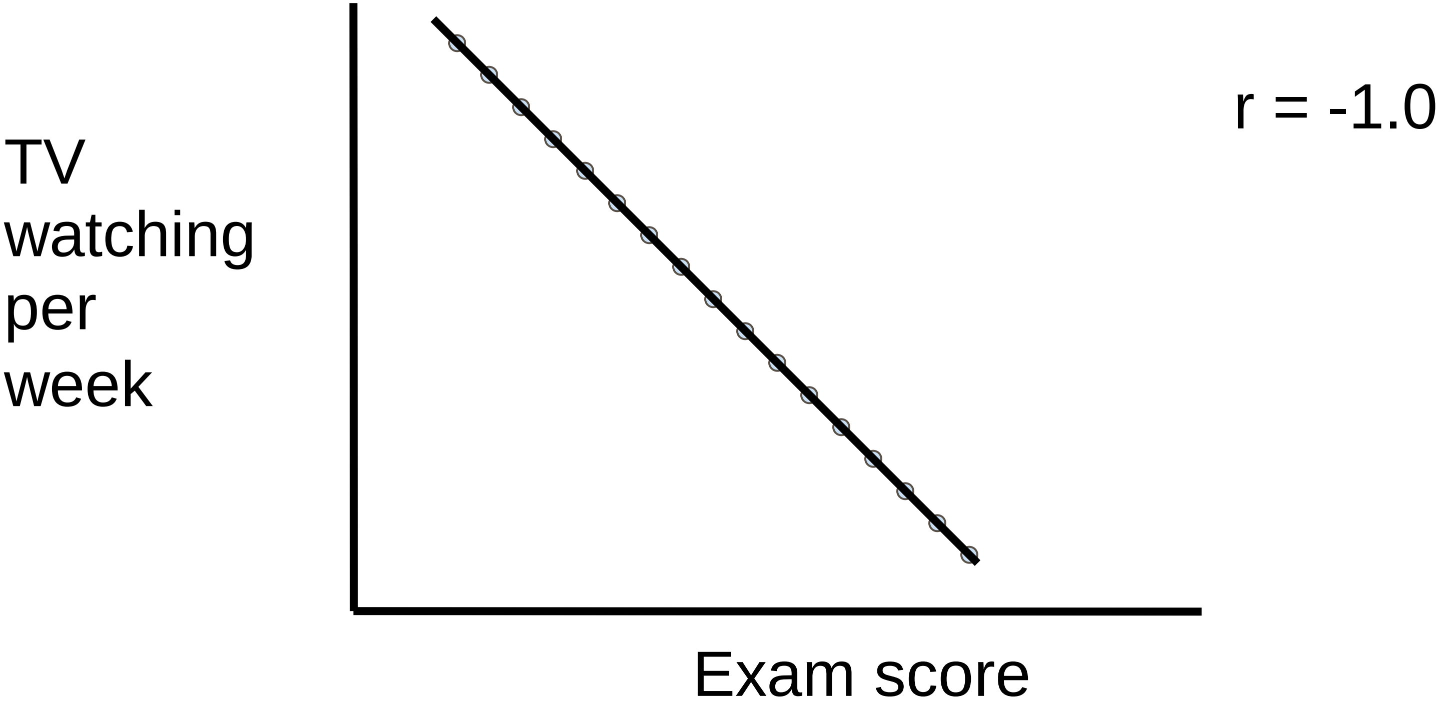
- **Moderate Positive Correlation**



r = + 0.4

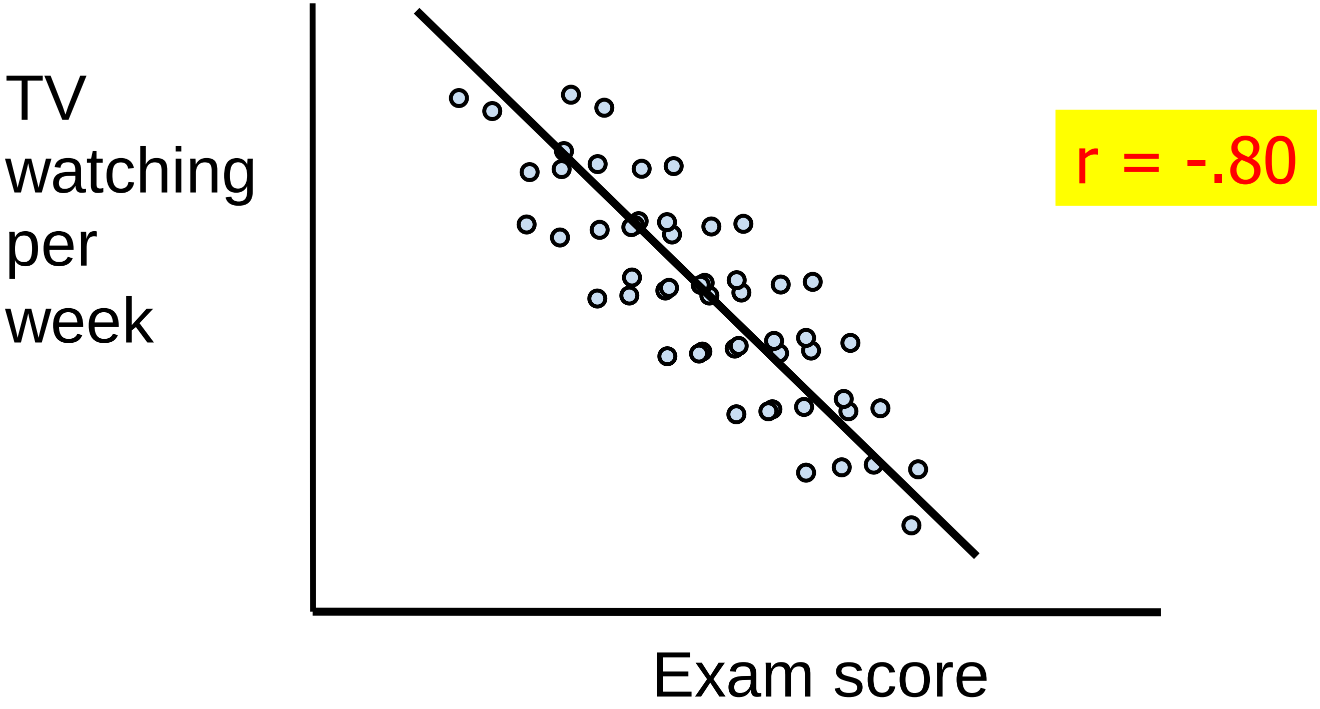
Degree of correlation

- **Perfect Negative Correlation**



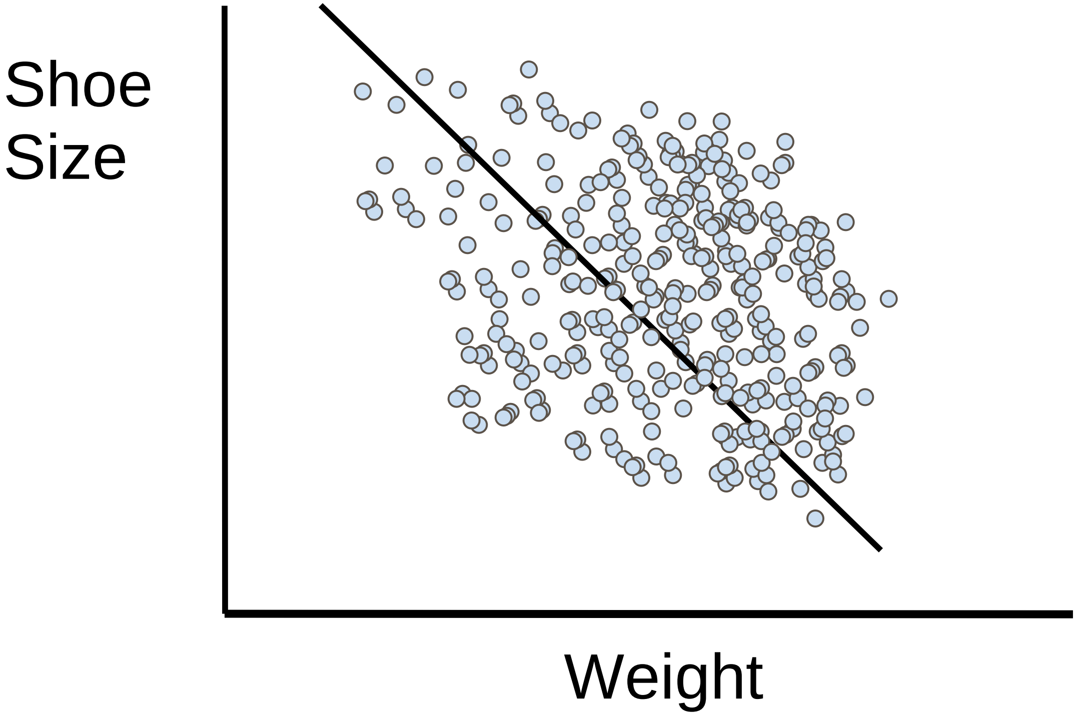
Degree of correlation

- **Moderate Negative Correlation**



Degree of correlation

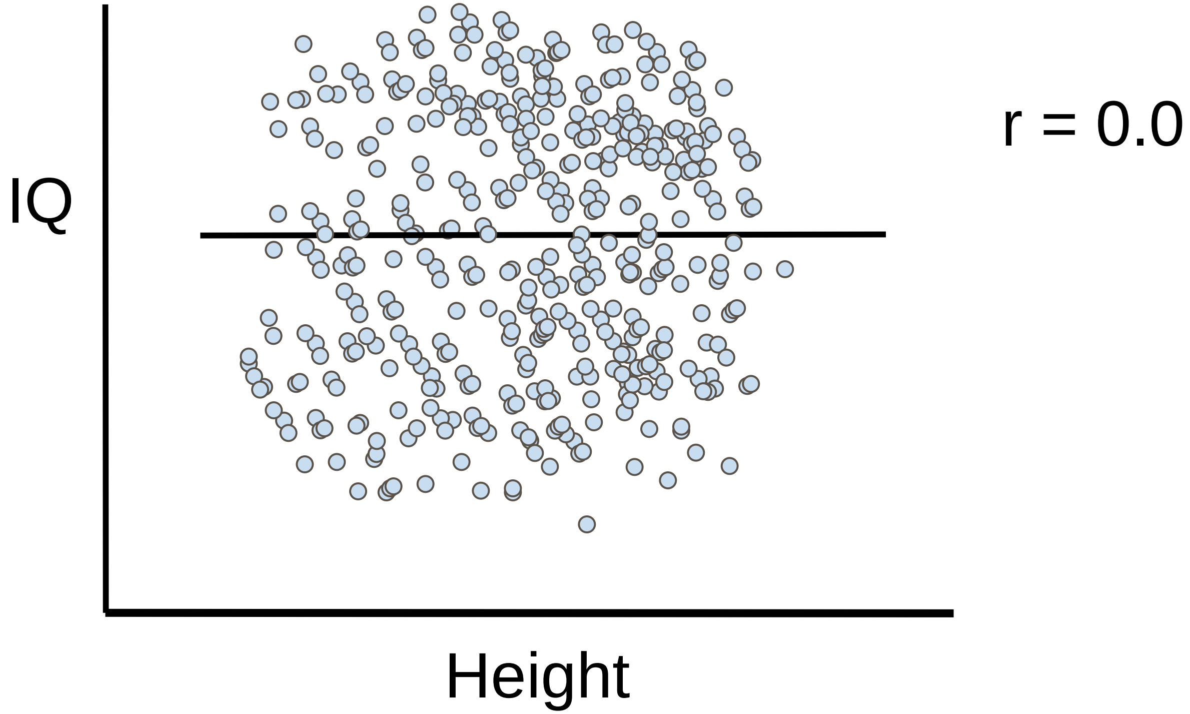
- **Weak negative Correlation**



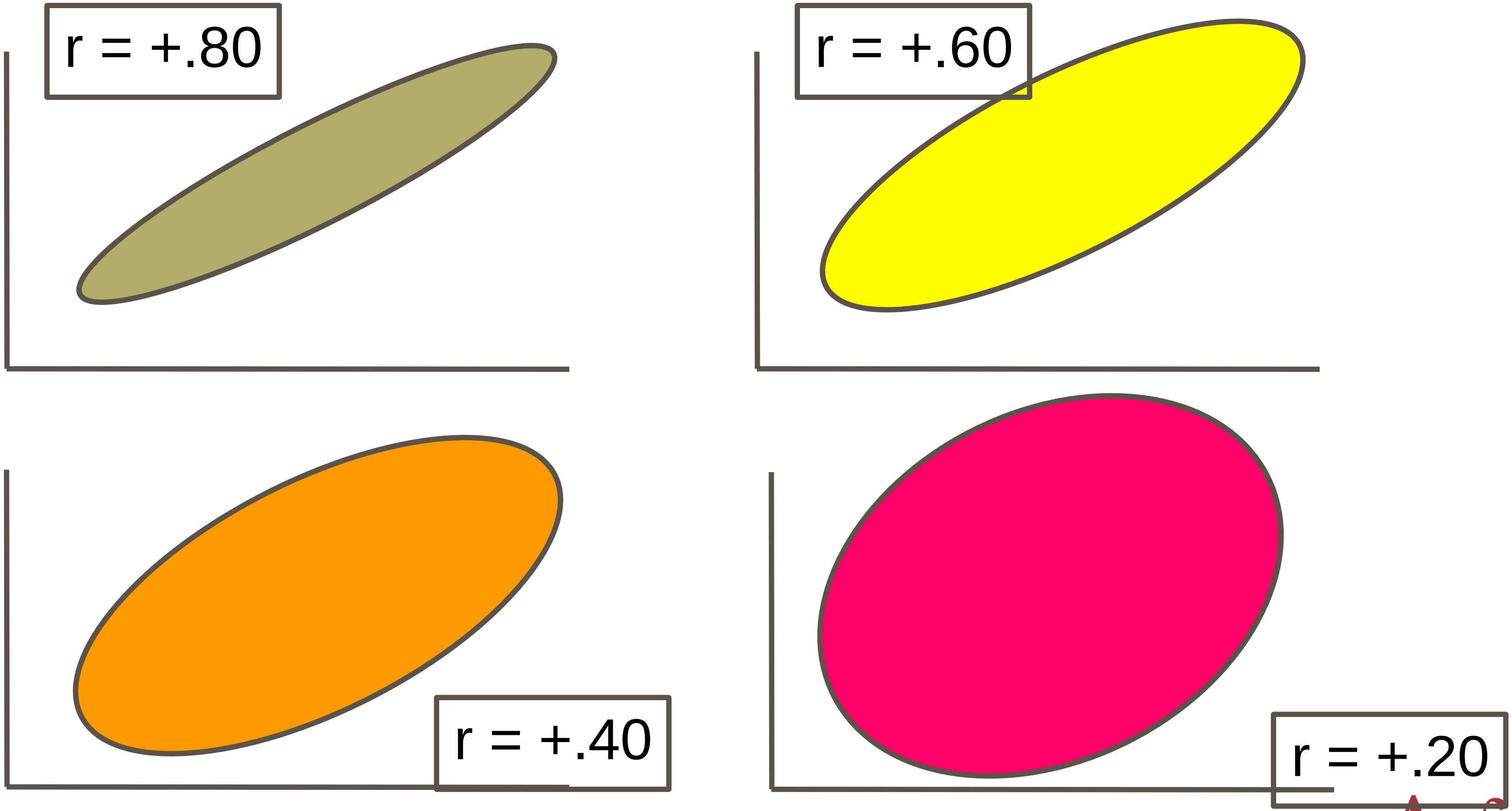
r = - 0.2

Degree of correlation

- **No Correlation (horizontal line)**



Degree of correlation (r)



Correlation

High positive correlation

Zero correlation

High negative correlation

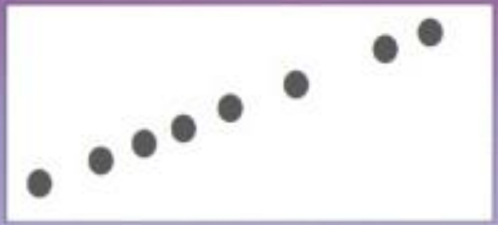
stronger

weaker

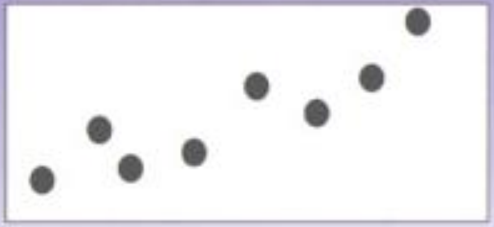
weaker

stronger

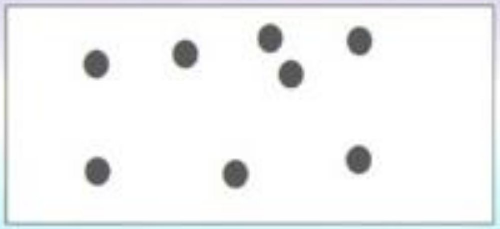
+1.00 **perfect positive**
as one event increases, the second exactly increases



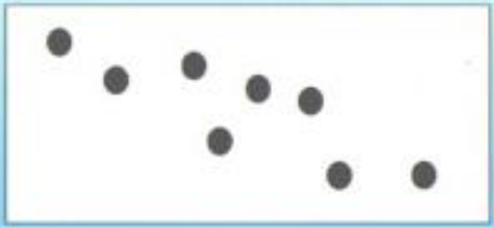
+.50 **positive**
as one event increases, the second sometimes increases



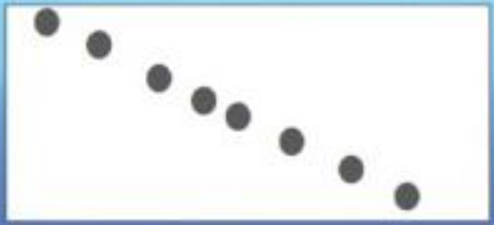
0 **zero correlation**
no relationship between the events



-.50 **negative**
as one event increases, the second sometimes decreases



-1.00 **perfect negative**
as one event increases, the second exactly decreases



Spurious/Non-sense Correlation:

- The correlation in absence of causation is called Spurious or Non-sense Correlation.

- Ex. Correlation between ***Marks of Student*** and ***Gold Prices***.

Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

Disadvantage of scatter diagram

Can not adopt the an exact degree of correlation

1st way of classification:

Types of Correlation

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Pub. Exp. & sales, Height & weight.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & qty. demanded.

More examples

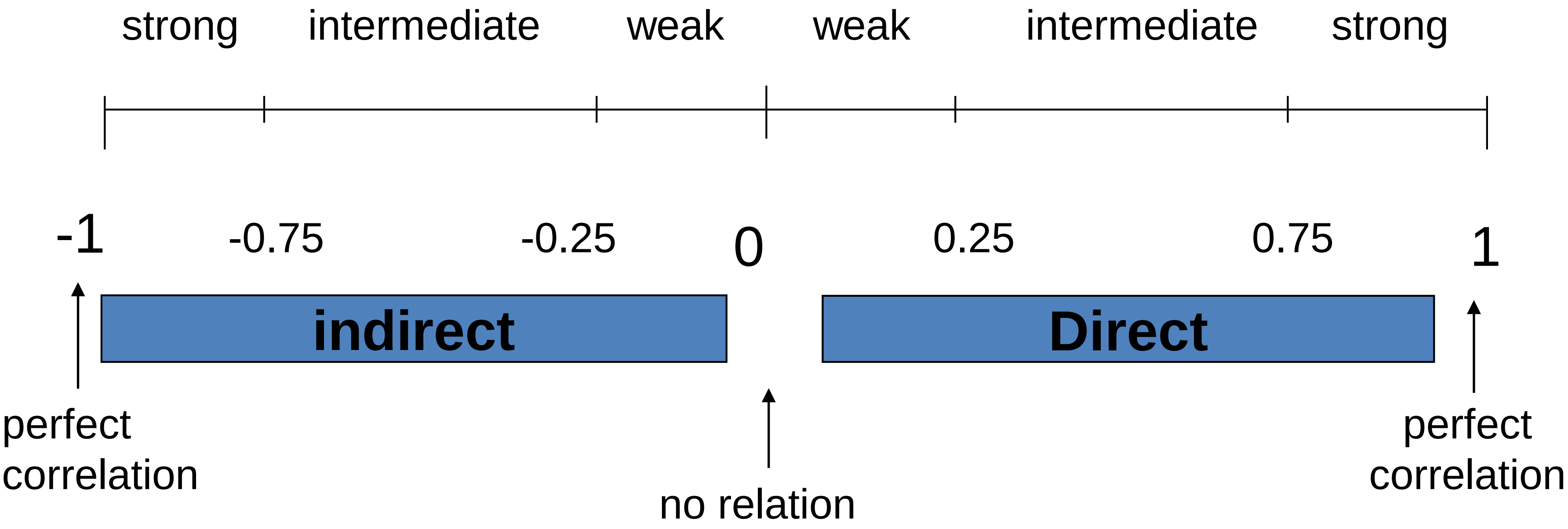
- **Positive relationships**
- water consumption and temperature.
- study time and grades.

- **Negative relationships**:
- alcohol consumption and driving ability.
- Price & quantity demanded

2nd way of classification:

Types of Correlation

- **Simple correlation:** Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied.
- **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.



strong intermediate weak weak intermediate strong

-1 -0.75 -0.25 0 0.25 0.75 1

indirect

Direct

perfect correlation

no relation

perfect correlation

2. Karl Pearson's Coefficient of Correlation

Karl Pearson's Coefficient of Correlation

- **Formula**

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x) * var(y)}}$$

where, $cov(x,y) = \frac{\sum(x-\overline{x})\,(y-\overline{y})}{(n-1)}$

OR

$$r_{xy} = \frac{\sum(x-\overline{x})\,(y-\overline{y})}{\sqrt{[\sum(x-\overline{x})^2][\sum(y-\overline{y})^2]}}$$

Simplified formula for **Ungrouped data**

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Simplified formula for **Grouped data**

$$r_{xy} = \frac{N\sum\sum f_{xy}xy - (\sum xf_x)(\sum yf_y)}{\sqrt{[N\sum x^2 f_x - (\sum xf_x)^2][n\sum y^2 f_y - (\sum yf_y)^2]}}$$

Advantages of Pearson’s Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson’s Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming method

3. Spearman's Rank Coefficient of Correlation

Spearman's Rank Coefficient of Correlation

- When statistical series arranged in serial order, in such situation Spearman Rank correlation can be used.

$$\rho_{xy} = 1 - \frac{6\sum d^2}{n^3 - n}$$

where d_i=R_1-R_2

- R = Rank correlation coefficient
- D = Difference of rank between paired item in two series.
- N = Total number of observation.

Rank Correlation Coefficient (R)

a) Steps after finding ranks:

1) Calculate the difference ‘D’ of two Ranks i.e. (R1 – R2).
2) Square the difference & calculate the sum of the difference i.e. $\sum D^2$
3) Substitute the values obtained in the formula.

Rank Correlation Coefficient (R)

- **Equal Ranks or tie in Ranks:**

In such cases average ranks should be assigned to each individual.

$$\rho_{xy} = 1 - \frac{6\sum(d^2 + CF)}{n^3 - n}$$

and

$$CF = \frac{1}{12\,(m_1{}^3 - m_1)} + \frac{1}{12\,(m_2{}^3 - m_2)} + \cdots$$

m = The number of time an item is repeated

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Merits Spearman’s Rank Correlation

- This method is simpler to understand and easier to apply compared to karl pearson’s correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitative term)
- This method is to use where the initial data in the form of ranks.

Limitation Spearman’s Correlation

- Cannot be used for finding out correlation in a grouped frequency distribution.

- This method should be applied where N exceeds 30.

Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study.
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect co relational data

Disadvantages of correlation studies

- Can’t assume that a cause-effect relationship exists
- Little or no control (experimental manipulation) of the variables is possible
- Relationships may be accidental or due to a third, unmeasured factor common to the 2 variables that are measured

3. Kendall rank correlation coefficient (Kendall’s Tau)

Home work!!!

Examples