Correlation

Imagine you determine there is a positive correlation between the number of times individuals visit doctors (over the course of their life span) and longevity (how many years the individuals live). Would it be appropriate to conclude that frequent visits to the doctor lead to a longer life span? Why or why not?

Two Variables: Which test?

		Explanatory variable	
		Categorical	**Numerical**
Response Variable	**Categorical**	•Contingency analysis	• Logistic regression • Survival analysis
	Numerical	• t-test • Analysis of variance	•Regression • Correlation

Correlation vs Regression

- Mathematical relationship between the two methods of analysis are close: share many computational steps which are similar or the same

- Appropriate method depends on the purpose of the investigator and the nature of the variables

- **Regression**:
  - **Intention:** describe dependence of a variable Y, on an independent variable X.
  - Regression equations are employed to support hypotheses regarding possible **causation** of changes in Y by changes in X; to predict Y in terms of X
  - A line (in linear regression) is fit to the data
  - *Study the **effects** of **X** on **Y***

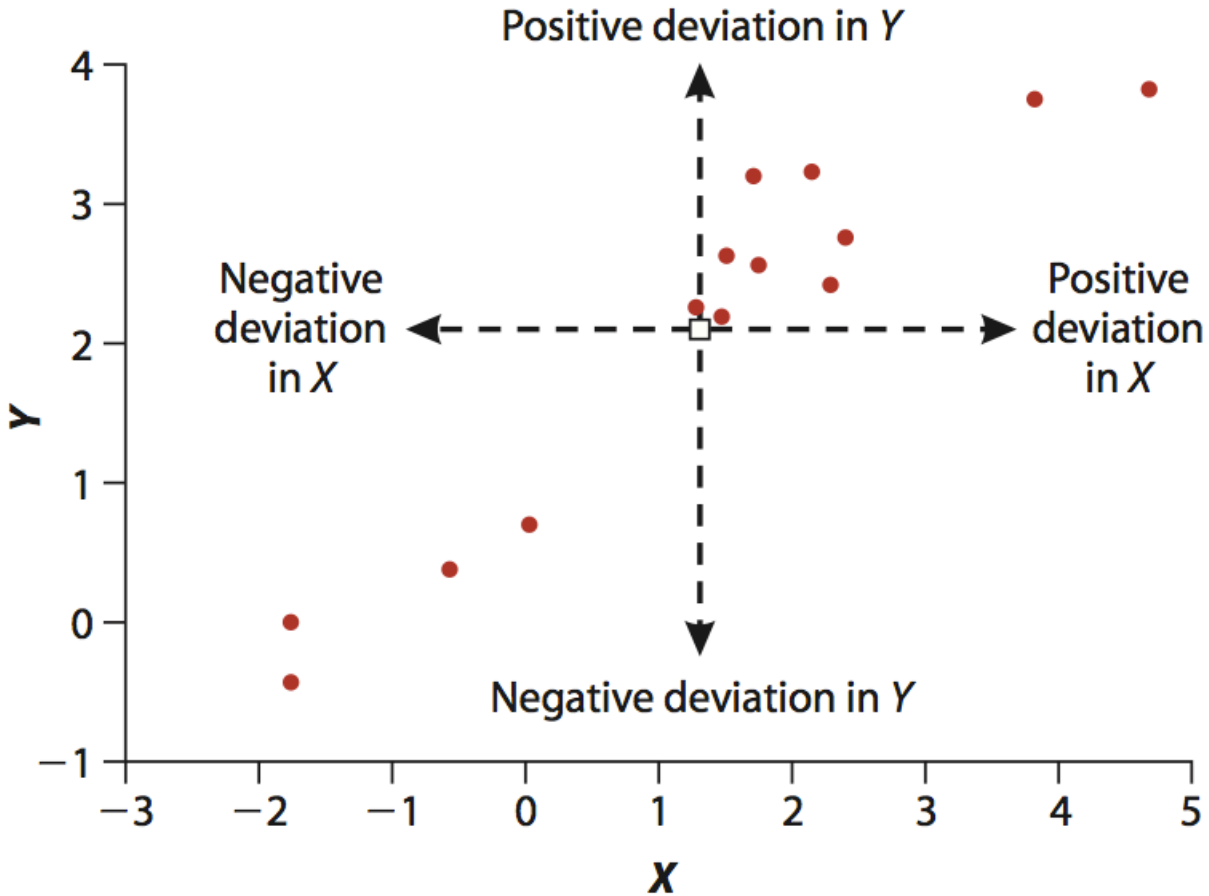
- **Correlation**:
  - Measures the amount/degree of linear association between two **numerical** variables
  - Estimate the degree to which variables **covary**
  - We do not express one variable as a function of the other variable
    - No distinction between dependent & independent variables
    - In fact, we usually assume that they both stem from a common cause -- with a pair of variables whose correlation is studied, one may very well be the cause of the other but we neither know nor assume!

(Pearson) Product-moment **Correlation Coefficient**

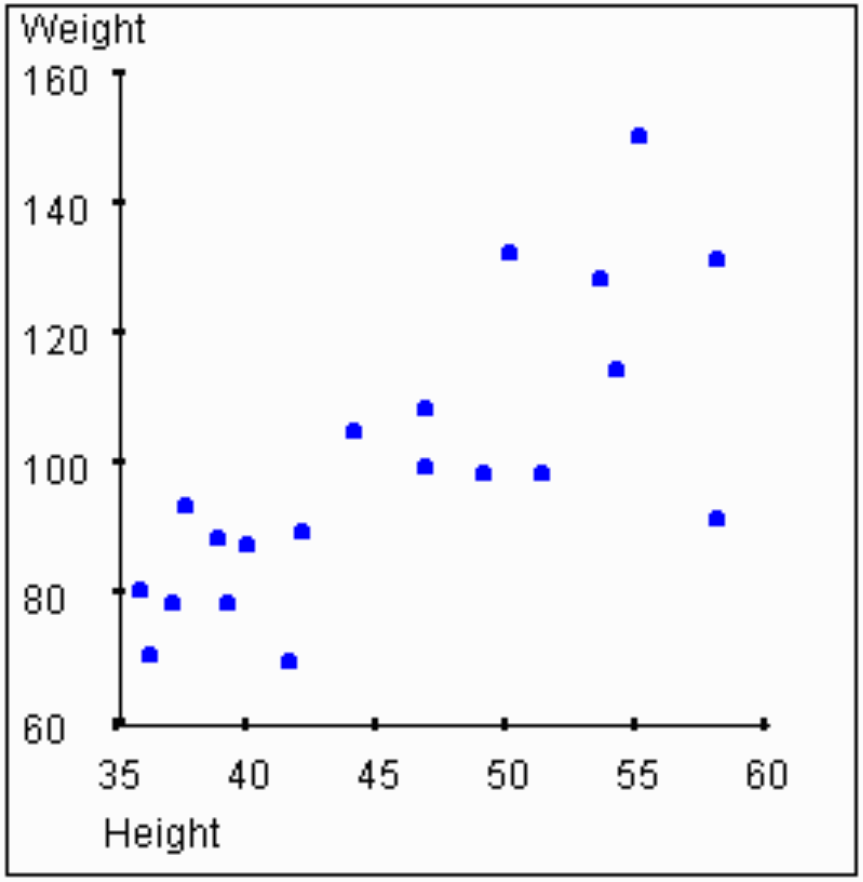
$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Co\,\mathrm{var}\,iance(X,Y)}{s_x s_y}$$

(Pearson) Correlation Coefficient

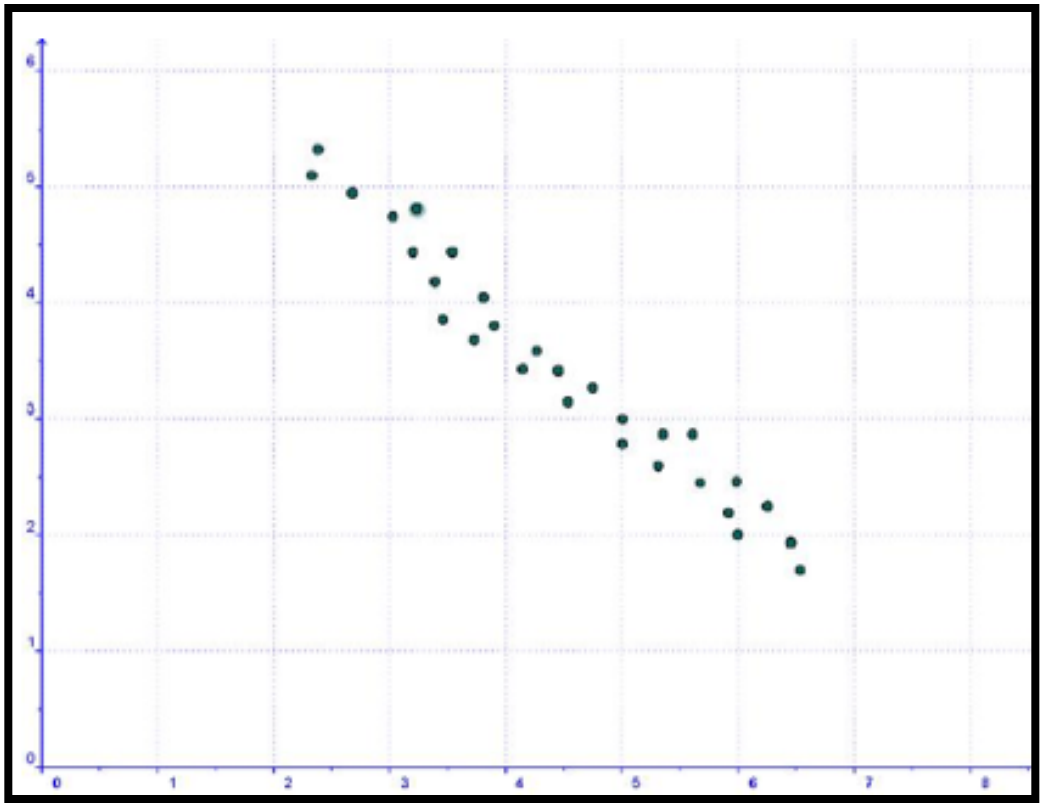
$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Co\,\mathrm{var}\,iance(X,Y)}{s_x s_y}$$



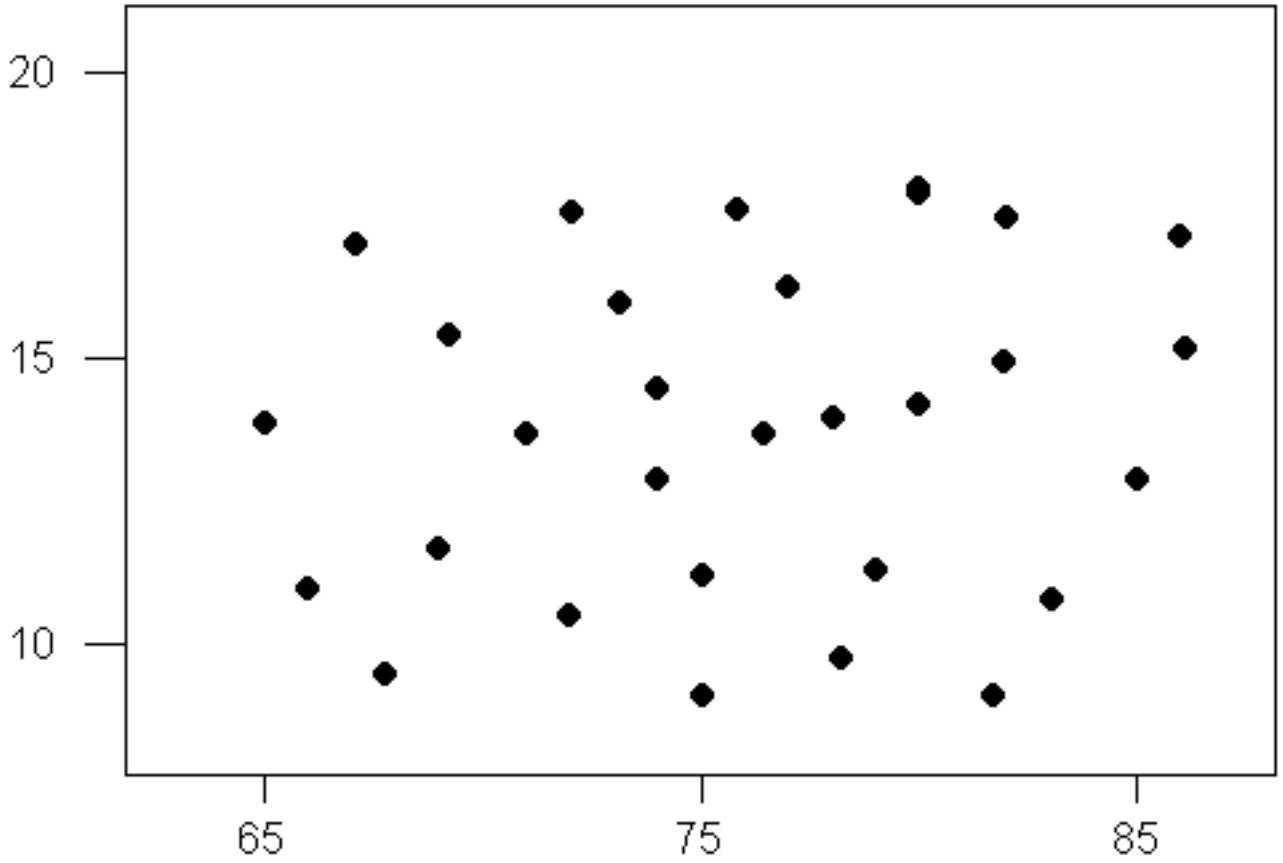
- Positive Correlation:



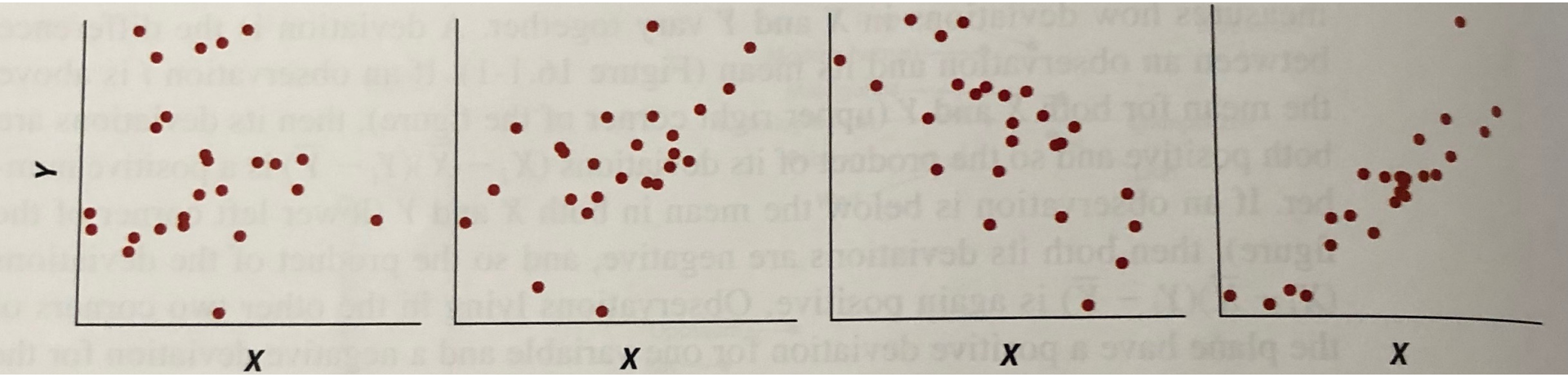
- Negative Correlation:



- No Correlation:



"Match the Graphs below with the most appropriate r value.



A) Graph 1: 0.5, Graph 2: 0.9, Graph 3: -0.7, Graph 4: 0.0

B) Graph 1: 0.0, Graph 2: 0.5, Graph 3: -0.7, Graph 4: 0.9

C) Graph 1: -0.7, Graph 2: 0.0, Graph 3: 0.5, Graph 4: 0.9

D) Graph 1: 0.0, Graph 2: 0.5, Graph 3: 0.9, Graph 4: -0.7”

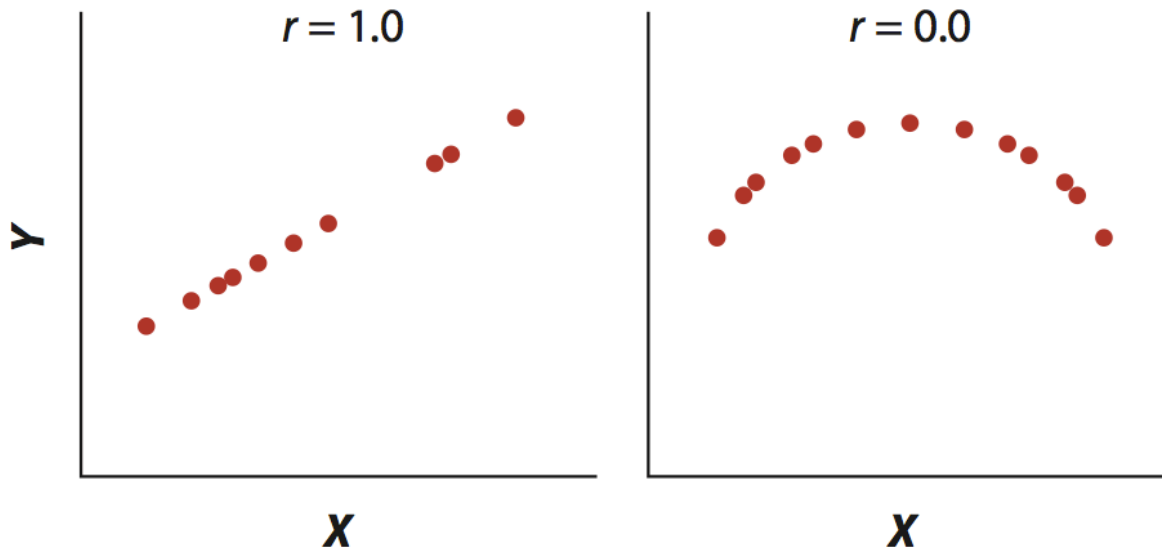
(Pearson) Correlation Coefficient

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Co\,\mathrm{var}\,iance(X,Y)}{s_x s_y}$$

$-1 < r, \rho < 1$

Warnings

Two variables might be strongly associated but have no correlation *if the relationship between them is nonlinear*



Correlation present at one scale may not exist at a different scale - correlation depends on **RANGE**

Standard Error:

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