Bivariate Data

Response Variable: the variable, which is being influenced, the dependent variable (y axis)

Explanatory Variable: the variable, which is influencing, the independent variable (x axis)

Note: When investigating the correlation between two variables, the Explanatory Variable is the variable we expect to explain or predict the value of the Response Variable

Example:

Of the following pairs of variables, which are response, and which are explanatory?

	Explanatory	**Response**
Amount of alcohol consumed and reaction time	Amount of Alcohol	Reaction Time
Distance travelled, and time taken	Distance Travelled	Time Taken
Heart disease and amount of fat in diet	Amount of Fat	Heart Disease
Hours worked per week and salary	Hours worked	Salary

Two Way Frequency Table

- A statistical tool used to investigate associations between two categorical variables

  Example:

  According to the results summarized in the table, is there an association between support for banning mobile phones in cinemas and the sex of the respondent?

Ban mobile phones	*Sex*	
	Male	*Female*
Yes	87.9%	65.8%
No	12.1%	34.2%
Total	100.0%	100.0%

  Yes, the percentage of males in support of banning mobile phones in cinemas (87.9%) was much higher than for females (65.8%).

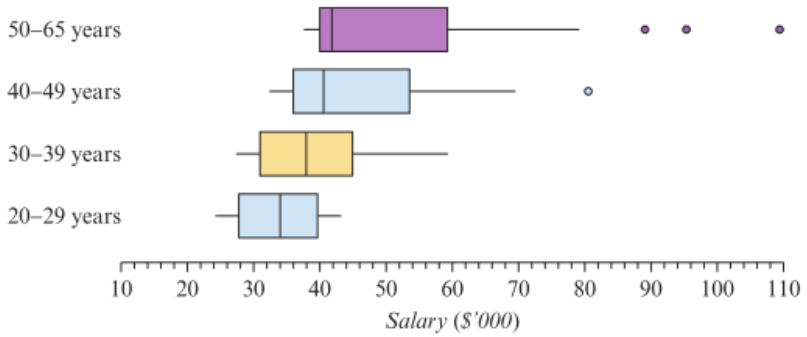
  Note. A difference of 5% is significant

Parallel Box Plots

- A statistical tool used for investigating associations between a numerical and categorical variable

  Example:

  The parallel box plots below compare the salary distribution for four different age groups: 20–29 years, 30–39 years, 40–49 years and 50–65 years.



Identify and Describing Associations

- Median

  Example:

  The parallel box plots show that median salaries and age group are associated because median salaries increase with age group. For example, the median salary increased from $34 000 for 20−29 year-olds to $42000 for 50−65 year-olds.

- IQR and/or ranges

  Example:

  From the parallel box plots we can see that the spread of salaries is associated with age group. For example, the IQR increased from around $12000 for 20−29-year-olds to around $20 000 for 50−65-year-olds.

- Shape

  Example:

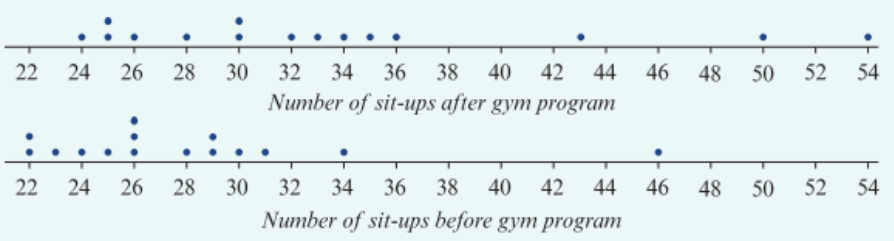
  From the parallel box plots we can see that the shape of the distribution of salaries is associated with age group because of the distribution, which is symmetric for 20−29-year-olds, and becomes progressively more positively skewed as age increases. Outliers also begin to appear.

Parallel Dot Plots

- Used to investigate associations between numerical and categorical variables for small data sets

Example:

Do the parallel dot plots support the contention that the number of sit-ups performed is associated with completing the gym program? Write a brief explanation that compares medians.



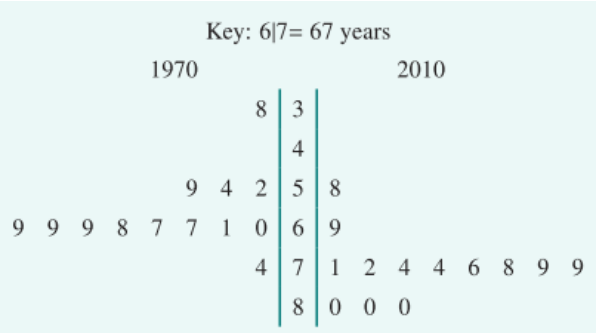
Yes; the median number of sit-ups performed after attending the gym program (M = 32) is considerably higher than the number of sit-ups performed before attending the gym program (M = 26). This indicates that the number of sit-ups performed is associated with completing the gym program.

Back to Back Stem Plots

- Used to investigate associations between numerical and categorical variables for small data sets

Example:

The back-to-back stem plot below displays the distribution of life expectancy (in years) for 13 countries in 2010 and 1970. Do the back-to-back stem plots support the contention that life expectancy is increasing over time? Write a brief explanation based on your comparisons of the two medians.



Yes: the median life expectancy in 2010 (M = 76 years) is considerably higher than the median life expectancy in 1970 (M = 67 years). This indicates that life expectancy is increasing over time.

Scatterplots

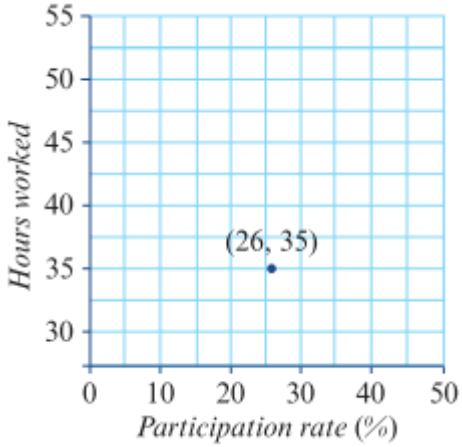
- Used to investigate associations between two numerical variables

Direction and Outliers >>> Positive, Negative, No association
Form >>> Linear or Non-linear
Strength >>> Strong, Moderate, Weak, None

Example:

Construct a scatterplot using the data shown below.

Participation rate (%)	26	20	36	1	25	9	30	3	55
Hours worked	35	43	38	50	40	50	40	53	35



Which graph – two variables?

Response variable	Explanatory variable	Graph
Categorical	Categorical	Segmented bar chart Parallel bar chart Two-way frequency
Numerical	Categorical	Parallel box plot Parallel dot plot
Numerical	Categorical (two categories only)	Back-to-back stem plot Parallel box plot Parallel dot plot
Numerical	Numerical	Scatterplot

Pearson's Correlation Coefficient

- $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)S_xS_y}$
- Assumes that:
    - Variables are numeric
    - Association is linear
    - No outliers in the data set
- When converting r^2 to r, check whether the gradient is positive or negative

Strength of a Linear Relationship

Strong positive association: r between 0.75 and 0.99
Moderate positive association: r between 0.5 and 0.74
Weak positive association: r between 0.25 and 0.49
No association: r between −0.24 and +0.24
Weak negative association: r between −0.25 and −0.49
Moderate negative association: r between −0.5 and −0.74
Strong negative association: r between −0.75 and −0.99

Correlation of Determination

- Represented as r^2, may be expressed as a decimal or percentage
- The coefficient of determination (as a percentage) tells us the variation in the response variable that is explained by the variation in the explanatory variable

Correlation and Causality:

- Correlation tells you about the strength of association instead of the source or cause
- Finding out if one variable causes the other variable to occur
- Causation cannot exist without correlation; correlation can exist without causation

Non-Casual Explanation for Association

Common Response: association with a common third variable

Confounding Variables: two possible explanations for association but no way to detangle their affects

Coincidence: association occurs by chance

Least Squares Regression Line

Fitting a straight line to bivariate data, minimising the sum of the squares of the residual

Residual: vertical distance between the actual data point and the regression line

- (Residual = Actual Data Value – Predicted Data Value)
- Takes into account every point on the scatterplot and is affected by outliers

The equation of the least squares regression line

The equation of the least squares regression line is given by $y = a + bx$,* where:

the **slope** (b) is given by: $b = \frac{rs_y}{s_x}$

and

the **intercept** (a) is then given by: $a = \bar{y} - b\bar{x}$

Here:

- r is the correlation coefficient
- s_x and s_y are the standard deviations of x and y
- $\bar{x}$ and $\bar{y}$ are the mean values of x and y

- When fitting a least squares regression line, it is assumed that:
    - Variables are numeric
    - Association is linear
    - No outliers in the data set