Week 2: Sampling and Organising Data

- Method of organising data differs depending on the type of variable → need to first determine the type of variable (categorical or numerical), and then data can be organised and visualised with the appropriate method

Organising and Visualising Data
For categorical variables: - Summary table, contingency table - Bar chart, pie chart, Pareto chart, side-by-side bar chart
For numerical variables: - Ordered array, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution - Stem-and-leaf display, histogram, polygon, cumulative percentage polygon - Boxplot - Normal probability plot - Mean, median, mode, quartiles, geometric mean, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, kurtosis
For *two* numerical variables: - Scatter plot, time-series plot - Sparklines
For categorical and numerical variables considered together: - Multi-dimensional contingency tables, Pivot Tables, gauges, bullet graphs, and tree maps - Cluster analysis - Multi-dimensional scaling

ORGANISING CATEGORICAL VARIABLES

- Summary table
    - Tallies the values as frequencies or percentages for each categories
    - Helps you see the differences among the categories by displaying the frequency, amount or percentage of items in a set of categories
- Contingency table
    - Cross-tabulates or jointly tallies the values of two or more categorical variables → allows for the study of patterns existing between variables
        - Can be shown as a frequency, a percentage of overall total, row total or column total

Fund Risk Level	Number of Funds	Percentage of Funds
Low	212	67.09%
Average	91	28.80%
High	13	4.11%
Total	316	100.00%

Summary table

	RISK LEVEL			
FUND TYPE	Low	Average	High	Total
Growth	143	74	10	227
Value	69	17	3	89
Total	212	91	13	316

Contingency table

ORGANISING NUMERICAL VARIABLES

- Ordered array
  - Arranges the values of a numerical variable in rank order (from smallest to largest)
  - Helps to get a better understanding about the range of values → particularly useful for fewer values
- Frequency distribution
  - Tallies the values of a numerical variable into a set of numerically ordered classes
    - Each class groups a mutually exclusive range of values, called a *class interval* (each value can be assigned to only one class)
  - To create a useful frequency distribution, a suitable *width* needs to be determined for each class interval → class intervals defined by *class midpoints*
    - $\text{Interval width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}}$
  - Establish clearly defined *class boundaries* for each class

- Relative frequency distribution
  - Presents the relative frequency, or proportion, of the total for each group that each class represents
    - Relative frequency/proportion – equals the number of values in each class divided by the total number of values
- Percentage distribution
  - Presents the percentage of the total for each groups that each class represents

- Cumulative percentage distribution
  - Provides a way of presenting information about percentage of values less than a specific amount → use a percentage distribution as the basis to construct a cumulative percentage distribution
  - Rows of a cumulative distribution do not correspond to class intervals
    - Class intervals are mutually exclusive

VISUALISING CATEGORICAL VARIABLES

- Bar chart
  - Visualises a categorical variable as a series of bar (each bar represents the tallies for a single category)
  - Length represents either the frequency or percentage of values for a category
- Side-by-side bar chart
  - Sets of bars to show joint responses
- Pie chart
  - Uses parts of a circle to represent tallies of each category
  - Size of each part varies according to the percentage in each category
- Pareto chart
  - *Pareto principle:* observation that in many data sets, a few categories of a categorical variable represent the majority of the data, while other categories represent a relatively trivial amount of data
    → helps identify "vital few" categories
  - Tools to identify areas needing improvement (e.g. defective items)

Week 11/12: Simple Linear Regression

CORRELATION VS. REGRESSION

- A scatter plot can be used to show the *relationship* between two variables
- Correlation analysis used to measure the strength of the linear relationship between two variables
    - Correlation only concerned with strength and sign of the relationship
    - No causal effect or direction is implied by correlation

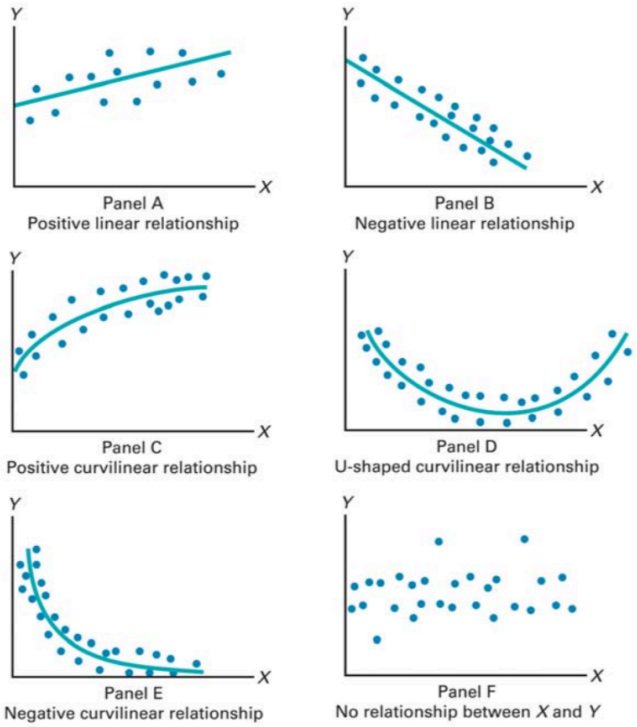
REGRESSION ANALYSIS

- Used to:
    - Predict the value of a dependent variable based on the value of at least one independent variable
    - Explain the impact of changes between variables
- Dependent variable: variable we wish to predict or explain (on the y-axis); hard to measure (i.e. output)
- Independent variable: variable used to predict or explain the dependent variable (on the x-axis)

SIMPLE LINEAR REGRESSION (SLR) MODEL

- Only *one* independent variable, X
- Relationship between X and Y is described by a linear function
    - Changes in Y are assumed to be related to changes in X
    - Direction X → Y is implicit in SLR
- Regression does not prove causation

Types of Relationships



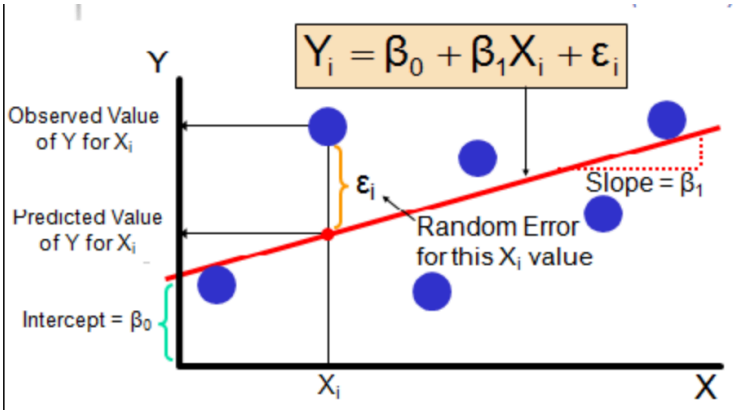
SIMPLE LINEAR REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

where

β_0 = Y intercept for the population
β_1 = slope for the population
ε_i = random error in Y for observation i
Y_i = dependent variable (sometimes referred to as the **response variable**) for observation i
X_i = independent variable (sometimes referred to as the predictor, or **explanatory variable**) for observation i

Random error component can be adjusted



DETERMINING THE SIMPLE LINEAR REGRESSION EQUATION

- The simple linear regression equation provides an <u>estimate</u> of the population regression line (i.e. the prediction line)

$$\hat{Y}_i = b_0 + b_1 X_i$$

where

$\hat{Y}_i$ = predicted value of Y for observation i
X_i = value of X for observation i
b_0 = sample Y intercept
b_1 = sample slope

The Least Squares Method

- b_0 and b_1 are obtained by finding the values of that minimise the sum of the squared differences between Y and Y-hat

$$\min \sum (y_i - \hat{y}_i)^2 = \min \sum (y_i - b_0 - b_1 x_i)^2$$

INTERPRETATION OF THE SLOPE AND INTERCEPT

- b_0 is the estimated mean value of Y when the value of X is zero (X=0)
- b_1 is the estimate change in the mean value of Y as a result of a one-unit increase in X
    - Small change in X → greater change in Y