01 Zero & One

Unit 10: Data Analysis

Programming & AI, 1st secondary (Introduction to ICT)



Engineer Rahim

10-1 Types of Data and Analysis

- **Quantitative data**
  - **Interval:** the gap between two numbers means something, but there is **no absolute zero**, so ratios mean nothing
  - **Ratio:** it has an **absolute zero**, so both gaps and ratios mean something
- **Qualitative data**
  - **Nominal:** just names that tell categories apart, **with no order**
  - **Ordinal: the order means something**, but the gaps between levels do not
- **A quick way to decide**
  - Ask: **can I add or subtract these numbers?** If not (like phone numbers), the data is **qualitative**.
  - Qualitative: **is there an order?** Yes = ordinal, no = nominal.
  - Quantitative: **does zero mean "none"?** Yes = ratio (0 kg = no weight), no = interval (0 °C does not mean no heat).
- **Ways to analyse data**
  - **Descriptive statistics:** summarises and organises existing data so we understand what the sample looks like.
  - **Measures of central tendency:** the **mean, median and mode**
  - **Measures of spread:** the **range, variance and standard deviation**
  - **Inferential statistics:** we draw conclusions, predict and generalise about the **whole population** from a **sample**

Quantitative data	Data written as numbers that can be measured, where the size of the number has meaning.
Qualitative data	Data that describes a property or category, even if it is written with digits.
Nominal	Qualitative data that only names categories, with no order (e.g. blood type).
Ordinal	Qualitative data whose order matters but whose gaps do not (e.g. test ranking).
Interval	Quantitative data with meaningful gaps but no absolute zero (e.g. temperature in °C).
Ratio	Quantitative data with an absolute zero, so ratios have meaning (e.g. weight).
Descriptive statistics	Summarising and organising existing data to describe a sample.
Inferential statistics	Drawing conclusions about a whole population from a sample, e.g. with hypothesis tests.
Measures of central tendency	Single values that represent the data: mean, median and mode.
Measures of spread	Values that show how scattered the data is: range, variance and standard deviation.

10-2 Data Analysis [1]

- **Frequency distribution table**
  - **Class:** the interval itself, such as "16 or higher and lower than 19".
  - **Class width:** the length of the interval = upper limit – lower limit (19 – 16 = 3).
  - **Class mark:** the middle of the class = (lower + upper) ÷ 2, so (16 + 19) ÷ 2 = 17.5.
  - **Frequency:** how many values fall in the class
- **Histogram**: The frequency distribution table drawn as **bars**: classes on the horizontal axis, frequency on the vertical one, and the bars touch because the classes follow each other.
- **Measures of central tendency**
  - **Mean:** sum of the values ÷ how many there are
  - **Median:** sort the data in **ascending** order and take the middle one
  - **Mode:** the **most frequent** value
- **Worked example**
  - **Mean:** sum 84 ÷ 12 = **7**.
  - **Median:** sorted 3, 3, 4, 6, 6, **6, 7**, 8, 9, 10, 10, 12
  - **Mode:** 6 appears 3 times = **6**.

Frequency distribution table	A table that counts how many data values fall in each class.
Class	One interval of a frequency distribution table.
Class width	The size of a class: upper limit minus lower limit.
Class mark	The middle value of a class.
Frequency	The number of data values inside a class.
Histogram	A frequency distribution table drawn as touching bars.
Mean	The sum of the values divided by how many there are (AVERAGE).
Median	The middle value of the sorted data, or the mean of the two middle values (MEDIAN).
Mode	The value that appears most often (MODE).

10-3 Data Analysis [2]

- **Quartiles**: Numbers that split data sorted in ascending order into **4 equal parts**: the **first quartile (Q1)**, the **second quartile (Q2) = the median**, and the **third quartile (Q3)**
- **Steps to find the quartiles**
  - **1):** Sort the data in **ascending** order and find the **median** (this is Q2).
  - **2):** Split the data at the median into two halves: the **lower half** and the **upper half**.
  - **3): Median of the lower half = Q1**, and **median of the upper half = Q3**.
- **Box-and-whisker plot**
  - **1):** Find the five numbers and draw a short line at each one.
  - **2):** Draw a **box** from **Q1 to Q3**
  - **3):** Draw the **whiskers**: a line from one end of the box to the **minimum value** and from the other end to the **maximum value**.
  - It can be drawn horizontally or vertically.
- **Reading the plot**
  - The **box** holds roughly the **middle half of the data**

- With **31 values**: the median is value number **16**, Q1 is number **8**, and Q3 is number **24**.
- If the **median is below 60**, then **at least half** of the students scored 60 or less.
- **Range = maximum value – minimum value:** (the length of the whole plot).

Quartile	One of the values that split sorted data into four equal parts.
First quartile	Q1: the median of the lower half of the data.
Second quartile	Q2: the median of the whole data set.
Third quartile	Q3: the median of the upper half of the data.
Box-and-whisker plot	A figure showing the minimum, Q1, median, Q3 and maximum with a box and two whiskers.
Minimum value	The smallest value in the data; the end of one whisker.
Maximum value	The largest value in the data; the end of the other whisker.

10-4 Data Analysis [3]

- **Why variance and standard deviation?**: Two classes can have the same **mean**, yet one has marks close together and the other has marks spread out
- **Definitions**
  - **Deviation:** value – mean
  - **Variance:** the **mean of the squared deviations**
  - **Standard deviation:** the **positive square root of the variance**, s = √variance
- **Steps (make a table)**
  - **1):** Find the **mean**.
  - **2):** For each value find the **deviation** = value – mean.
  - **3): Square** each deviation (so minus and plus values do not cancel out).
  - **4): Variance** = the mean of the squares (their sum ÷ the number of values).
  - **5): Standard deviation** = √variance.
- **Worked example**
  - Mean = 40 ÷ 5 = **8**.
  - Deviations: −4, −2, 0, 2, 4 → squares 16, 4, 0, 4, 16, with sum 40.
  - Variance = 40 ÷ 5 = **8**, and standard deviation = $\sqrt{8} = 2\sqrt{2} \approx 2 \times 1.414 \approx$ **2.83**.

Deviation	A data value minus the mean.
Variance	The mean of the squared deviations; larger means more spread (VAR.P).
Standard deviation	The positive square root of the variance, in the data's own unit (STDEV.P).
Spread	How far the data values are scattered around the mean.
VAR.P	The Excel function for variance (VARP in Google Sheets).
STDEV.P	The spreadsheet function for standard deviation.

10-5 Data Analysis [4]

- **Scatter plot**
  - **Positive correlation:** x goes up → y goes up
  - **Negative correlation:** x goes up → y goes down
  - **No correlation:** the points are scattered with no direction.

- **Correlation coefficient r**
  - Its value is always **$-1 \leq r \leq 1$**.
  - **Close to 1:** = **strong** positive correlation, and **close to −1** = **strong** negative correlation.
  - **Close to 0:** = almost no correlation
  - In spreadsheets: the `CORREL` function.
- **Causation and pseudo-correlation**
  - **Causal relationship:** one thing is the **cause** and the other is the **effect**
  - **Pseudo-correlation:** two things **seem** to be cause and effect, but they are not
  - So **correlation alone is not proof of causation**.
- **Cross tabulation**: A table that compares data across **two or more categories** at once (such as gender × opinion)

Scatter plot	A graph that shows each pair of values (x, y) as a point.
Positive correlation	As x increases, y also tends to increase.
Negative correlation	As x increases, y tends to decrease.
No correlation	There is no clear upward or downward trend between x and y.
Correlation coefficient	r, between -1 and 1, measuring how strong a correlation is (CORREL).
Causal relationship	A relationship where one thing is the cause and the other is the effect.
Pseudo-correlation	Two things that seem to be cause and effect but are not.
Cross tabulation	A table that compares data across two or more categories.

10-6 Data Analysis [5]

- **Regression analysis**
  - **Simple regression analysis:** we predict the **result variable y** from **one causal variable x**.
- **Regression line**
  - **a:** = the slope: if **positive** the line goes up (positive correlation), if **negative** it goes down (negative correlation).
  - **b:** = the value of y when x = 0.
- **Least squares method**: **Residual:** the vertical distance between a real point and the line
- **Predicting with the equation**
  - $y = 0.0287 \times 5{,}000 + 297.86 = 143.5 + 297.86 = 441.36 \approx$ **441 minutes**.

Regression analysis	A method that clarifies how a result value depends on a causal value, used for prediction.
Simple regression analysis	Predicting the result variable y from one causal variable x.
Regression line	The straight line on a scatter plot used to find predicted values.
Regression line equation	The first-degree equation of the regression line, $y = ax + b$.
Least squares method	Choosing the line that makes the squared residuals as small as possible.
Residual	The vertical distance between a data point and the regression line.

Study the whole course free

Every lesson, questions with answers, a ministry-style bank and a timed mock exam

zerooneacademy.com

Engineer Rahim