# Tutorial 4 - Statistics Part 2

**Section A: Graphs** Good Data visualization – clear labels on axis etc. (see lecture notes)

1. Using the following data:

| Category | Frequency | Relative Frequency |
|---|---|---|
| 3.51 – 4.00 | 1 | |
| 4.01 – 4.50 | 0 | |
| 4.51 – 5.00 | 2 | |
| 5.01 – 5.50 | 13 | |
| 5.51 – 6.00 | 7 | |
| 6.01 – 6.50 | 22 | |
| 6.51 – 7.00 | 24 | |
| 7.01 – 7.50 | 30 | |
| 7.51 – 8.00 | 10 | |
| 8.01 – 8.50 | 6 | |
| 8.51 – 9.00 | 3 | |

Construct:

i. A histogram of the data

Notes: Can use Frequency or Relative Frequency when drawing the histogram. However, as we are drawing by hand – it may be easier to manage units on the *y*-axis by using Frequency.

ii. Comment on the skewness of the distribution of the histogram

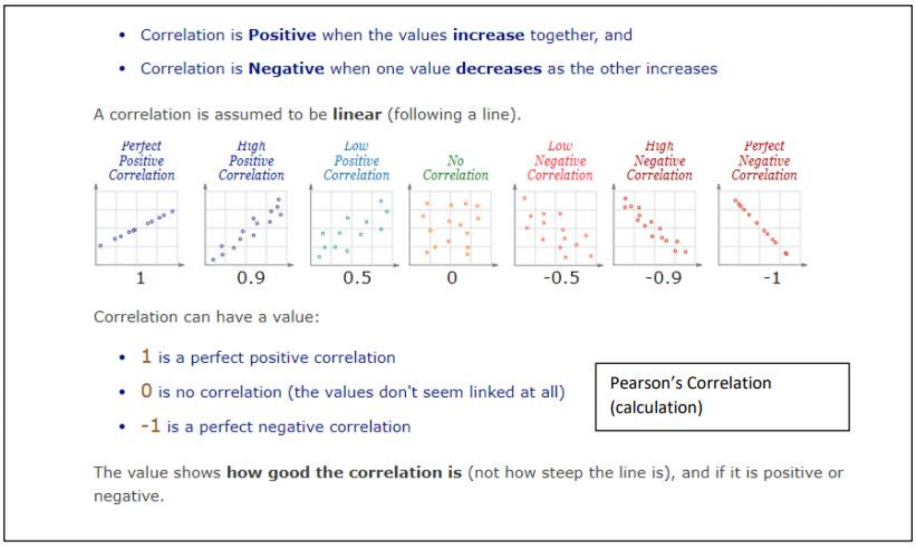
**Skewed Distribution**: A histogram that does not have a peak in the centre but is peaked on one side.
***skewed to the left*** – longer tail on the left, & the peak is on the right
***skewed to the right*** – longer tail on the right, & the peak is on the left
**Symmetric**: When there is no obvious skew, we say that the distribution is roughly symmetric

2. The data below contains details of the number of students taking computing at third level between 1970 and 2000.

| **Year** | **Number of Students** |
|---|---|
| 1970 | 10,000 |
| 1975 | 12,000 |
| 1980 | 13,500 |
| 1985 | 11,500 |
| 1990 | 12,250 |
| 1995 | 14,000 |
| 2000 | 17,000 |

Construct a *time series plot* to display the data

Sequential Data: **Time Series** Data **-** graph usually has time (months or years) on the horizontal axis and the values of interest on the vertical axis. Plot each point and join the dots to reveal the trend line…



- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

A correlation is assumed to be **linear** (following a line).

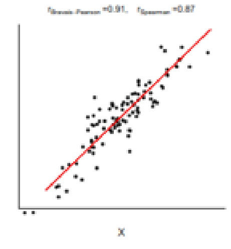
Perfect Positive Correlation
1
High Positive Correlation
0.9
Low Positive Correlation
0.5
No Correlation
0
Low Negative Correlation
-0.5
High Negative Correlation
-0.9
Perfect Negative Correlation
-1

Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

Pearson's Correlation (calculation)

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.



(a) Strong positive linear relationship

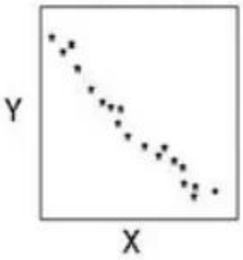
*Figure 1* *Figure 2*

**Scatter plots:**

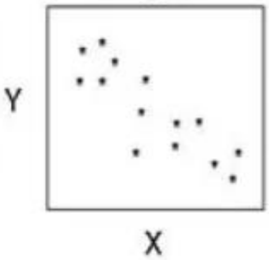
- Way to graphically summarize the association between two continuous variables
- Plot of paired observations of two variables in an *n*-dimensional coordinate system
- Reveals possible relationships (**correlation**) and trends between two variables
- Attributes values determine the position

3. For each of the scatter plots below, state whether there is a correlation, and if so, the type and strength (i.e. *perfect positive correlation*, *high/low positive correlation*, *no correlation*, *high/low negative correlation*, *perfect negative correlation*) of the correlation.

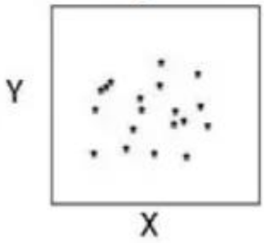
A



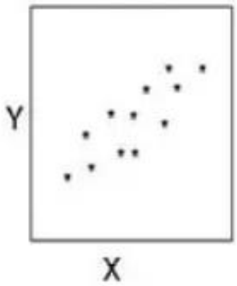
B



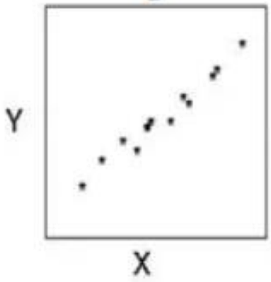
C



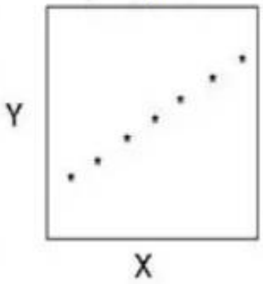
D



E



F



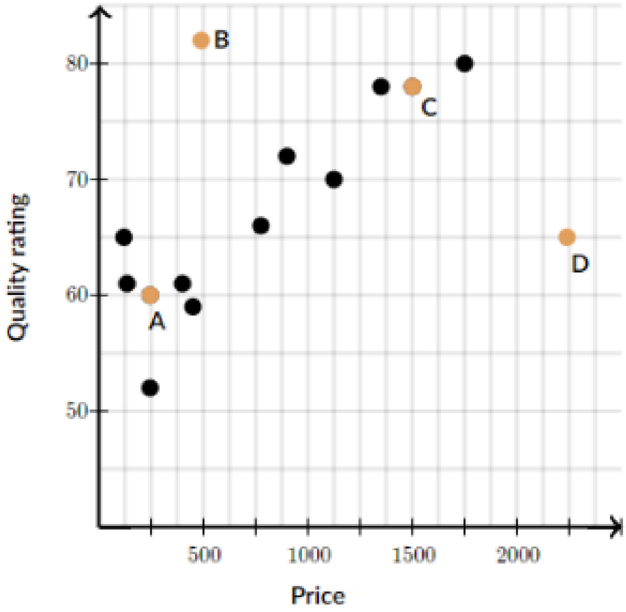
4. Use a scatter plot to display the following:

| Temperature*(Degrees Celsius)* | 14.2 | 16.4 | 11.9 | 15.2 | 18.5 | 22.1 | 19.4 | 25.1 | 23.4 | 18.1 | 22.6 | 17.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ice-cream Sales (Euro) | 215 | 325 | 185 | 332 | 406 | 522 | 412 | 614 | 544 | 421 | 445 | 408 |

State whether there is correlation and if so the type and strength (i.e. *perfect positive correlation, high/low positive correlation, no correlation, high/low negative correlation, perfect negative correlation*) of the correlation.

Which variable goes on which axis? – When constructing a scatter plot, consider which variable probably influences the other variable and put it on the *x*-axis (i.e. the independent variable). Here, *Temperature* (independent variable) can be used on the *x*-axis and *Ice-cream sales* (dependent variable) on the *y*-axis. Plot as *x-y* pairs e.g. *point 1* is *(14.2, 215)*. Unlike time series plots, we do not join the dots on a scatter plot. Instead, we want to see if there is a relationship (correlation) between the variable represented on the *x*-axis and the variable represented on the *y*-axis. It's a linear relationship between the variables if a change in one variable corresponds to a consistent change in the other (see Figure 1). Mathematically, this is expressed as the equation of a straight-line $y = mx + b$ where $x$ is the independent variable and $y$ is the dependent variable. This line is sometimes referred to as the best-fit line through the data (see Figure 2).

5. Mary is researching different computers to buy for college. She looked up prices (in Euro) and quality ratings for a sample of computers. Her data is shown in the scatter plot below, where each data point is a computer.



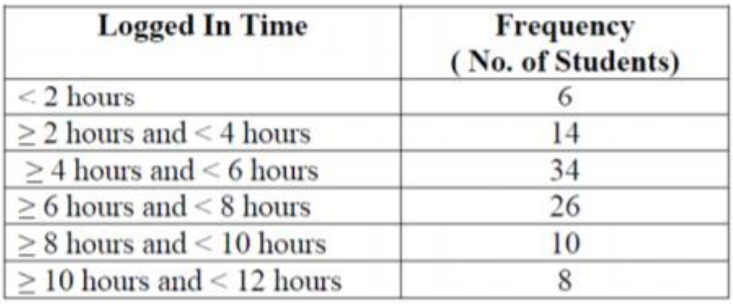
i. Using the plot above – how would you identify exceptions in the data? In your answer discuss any exceptions in the data represented above. Note: an exception (outlier) is any point(s) away from the main trend/cluster of points.

ii. Mary wants to buy a computer whose quality rating is far higher than the pattern would predict based on its price. Which of the labelled points (A, B, C, or D) represents a computer that Mary wants to buy and why? What can we say about points B and D in the plot?

## Section B: Numerical Calculations – GROUPED DATA

The formula for *mean*, *standard deviation*, *median*, and *quartiles* are different for **grouped data** (see the formula at the end of the tutorial sheet). Question 6 is demonstrated for you. Follow these steps to complete questions 7 and 8.

6. The system administrator in a college is interested in determining the amount of time that computing students spend logged in for each day. To do this the administrator monitors 98 randomly selected computing students for an entire day. The study provides the following information:



| Logged In Time | Frequency<br>( No. of Students) |
|---|---|
| < 2 hours | 6 |
| ≥ 2 hours and < 4 hours | 14 |
| ≥ 4 hours and < 6 hours | 34 |
| ≥ 6 hours and < 8 hours | 26 |
| ≥ 8 hours and < 10 hours | 10 |
| ≥ 10 hours and < 12 hours | 8 |

(i) Calculate the mean $\bar{x}$ of the daily logged in time for these students

**Use: (Grouped) Mean:** $\bar{x} = \dfrac{\sum_i f_i m_i}{\sum_i f_i}$

where $f$ = frequency and $m$ = midpoint (of the class intervals) $\sum$ *means SUM of*

| Logged in Time | Frequency $f$ | $m_i$ | $f_i m_i$ |
|---|---|---|---|
| 0-2 | 6 | 1 | (6x1) = 6 |
| 2-4 | 14 | 3 | (14x3) = 42 |
| 4-6 | 34 | 5 | (34x5) = 170 |
| 6-8 | 26 | 7 | (26x7) = 182 |
| 8-10 | 10 | 9 | (10x9) = 90 |
| 10-12 | 8 | 11 | (8x11) = 88 |
|  | $(\sum f_i)$ = **98** |  | $\sum f_i m_i$ = **578** |

$\bar{x} = \frac{578}{98}$ = 5.897 = 5.9 (to one dec place)

(ii) Calculate the standard deviation *s* of the daily logged in time

Use: $s^2 = \dfrac{\sum_{i=1}^{M} f_i m_i^2 - M(\bar{x})^2}{M-1}$ where $M = \sum f_i$

| Logged in Time | Frequency $f$ | $m_i$ | $m_i^2$ | $f_i m_i^2$ |
|---|---|---|---|---|
| 0-2 | 6 | 1 | $(1^2) = 1$ | (6x1) = 6 |
| 2-4 | 14 | 3 | $(3^2) = 9$ | (14x9) = 126 |
| 4-6 | 34 | 5 | $(5^2) = 25$ | (34x25) = 850 |
| 6-8 | 26 | 7 | $(7^2) = 49$ | (26x49) = 1274 |
| 8-10 | 10 | 9 | $(9^2) = 81$ | (10x81) = 810 |
| 10-12 | 8 | 11 | $(11^2) = 121$ | (8x121) = 968 |
| | $(\sum f_i)$ = **98** | | | $\sum f_i m_i^2$ = **4034** |

$$s^2 = \frac{\sum_{i=1}^{M} f_i m_i^2 - M(\bar{x})^2}{M-1} = \frac{4034 - 98(5.9)^2}{98-1} = \frac{4034 - 3411.38}{97} = 6.42$$

$$(standard\ deviation)\ s = \sqrt{6.42} = 2.53$$

7. Evaluate the following for the grouped data given below,
   i. mean
   ii. standard deviation

| Class | Frequency |
|---|---|
| 0.0 – 4.0 | 2 |
| 4.0 – 8.0 | 6 |
| 8.0 – 12.0 | 12 |
| 12.0 – 16.0 | 6 |
| 16.0 – 20.0 | 2 |

8. The following is the grouped frequency distribution of the number of compile errors returned for a tutorial group of 20 students on writing their first Java assignment.

| Number of compile errors | Number of students |
|---|---|
| 0 and < 2 | 8 |
| 2 and < 4 | 5 |
| 4 and < 6 | 4 |
| 6 and < 8 | 2 |
| 8 and < 10 | 1 |

i. Draw a suitable *diagram* of the data given in the table above
ii. Calculate the *mean* of the grouped data
iii. Calculate the *standard deviation* of the grouped data
iv. Comment of the *distribution* of the data (symmetric/skewed)

Try: Compute the *median* and *quartiles* for the grouped data in question 8. See lecture notes.

## Formula - Grouped Data

**Mean:** $\bar{x} = \dfrac{\sum_i f_i m_i}{\sum_i f_i}$

**Standard Deviation:** $s^2 = \dfrac{\sum_i f_i (m_i - \bar{x})^2}{\sum_i f_i - 1}$ $\quad s^2 = \dfrac{\sum_{i=1}^{M} f_i m_i^{\,2} - M\left(\bar{x}\right)^2}{M - 1}$

$$median = \left(\frac{n}{2} - cf\right)\frac{w}{f} + L$$

$$Q_1 = \left(\frac{n}{4} - cf\right)\frac{w}{f} + L$$

$$Q_3 = \left(\frac{3n}{4} - cf\right)\frac{w}{f} + L$$