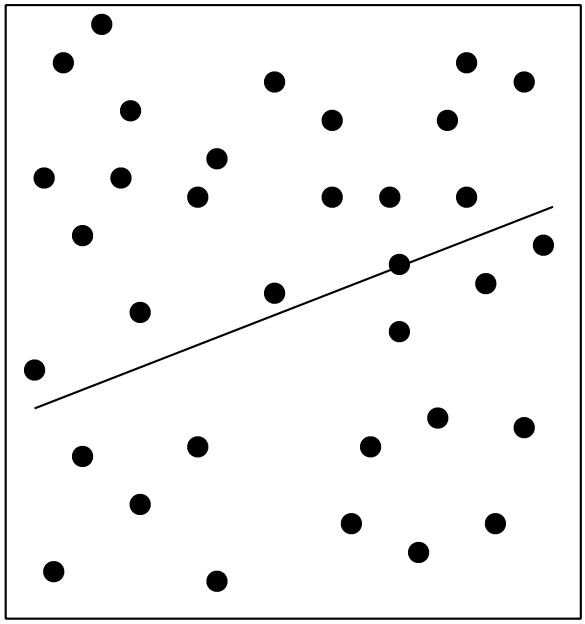
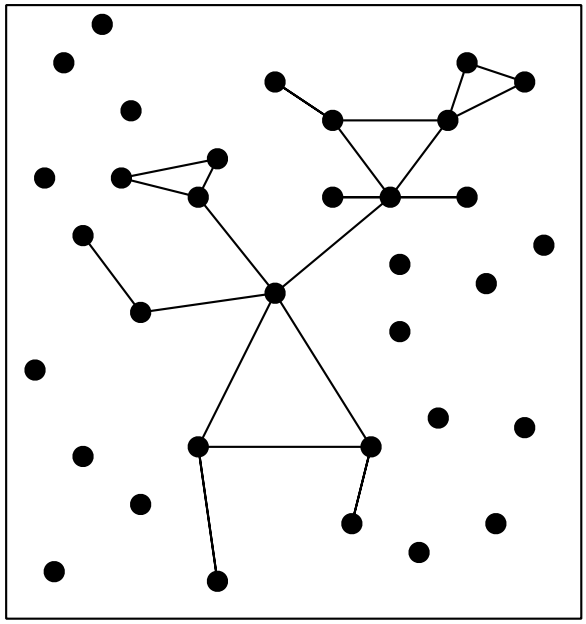
# Year 1

# ~ Statistics ~

# C O R R E L A T I O N I



**Proposed line of best fit**



**Proposed new constellation "Rex-Thor, the dog-bearer"**

**"When it's far easier to find new constellations in the scatter graph than decide on a line of best fit, there is negligible correlation"**

**Adapted from an xkcd cartoon by Randall Munroe**

STATISTICS

YEAR 1

# C O R R E L A T I O N  I

**Lesson 1**

**A-level Applied Mathematics**

**Statistics : Correlation I : Year 1**

## 1.1 Correlation Does Not Imply Causation

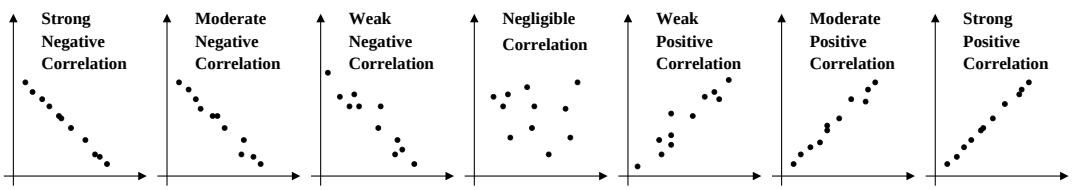
Two variables have a ***causal*** relationship if changing either ***causes*** the other to change. In trying to establish if a causal relationship exists between two variables we would like to be able to vary one and measure what the other does. For example, the temperature in a cage might be varied and the activity of a group of mice measured. Sometimes, we're not able to vary one variable at will, such as how much sunshine there is during the day, but we can still gather data on the number of hours of sunshine each day for a month along with how active the mice are. In both cases the resulting ***bivariate data*** can be plotted as points on a scatter graph. The graph may then being studied for ***correlation***. Finding correlation only suggests that there may be a causal relationship, but on it's own is not enough. The famous phrase used is "***correlation does not imply causation***". An underlying reason for ***why*** the variables are related would also need to be found.


## 1.2 What Correlation Looks Like On A Scatter Graph

Bivariate data can usefully be plotted on a scatter graph.

The ***independent*** or ***explanatory*** variable is usually plotted on the horizontal axis.

The ***dependent*** or ***response*** variable is usually plotted on the vertical axis.

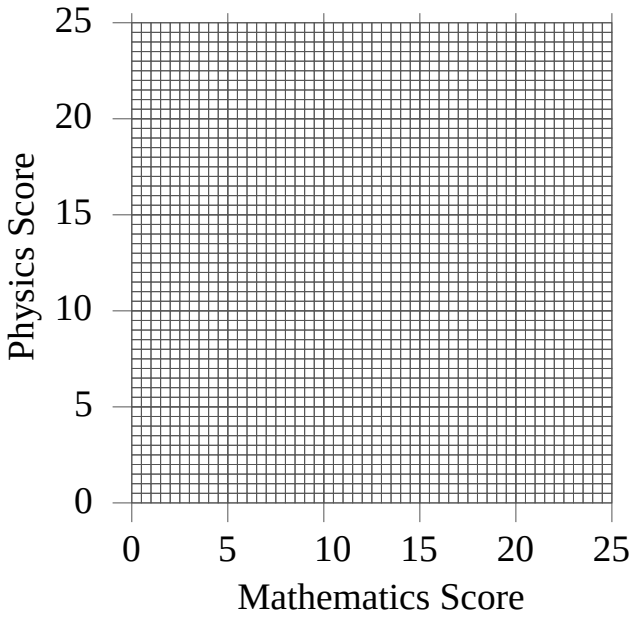


### 1.3 Scatter Graph Example

The scores of ten students in a Mathematics test and a Physics test, both marked out of 20, are presented below:

| Student | Dave | Pete | Pam | Liz | Joe | Bill | John | Sara | Zac | Zoe |
|---|---|---|---|---|---|---|---|---|---|---|
| Maths | 13 | 9 | 6 | 8 | 17 | 8 | 13 | 9 | 20 | 10 |
| Physics | 10 | 7 | 4 | 9 | 17 | 10 | 11 | 6 | 20 | 12 |

**( i )** Plot a scatter graph to represent these data




Let $\bar{m}$ be the mean mathematics score, and $\bar{p}$ be the mean Physics score

**( ii )** Calculate $\bar{m}$ and $\bar{p}$

**( iii )** Plot the point ( $\bar{m}$, $\bar{p}$ ) on the scatter graph

**( iv )** Draw a horizontal broken line and a vertical broken line through ( $\bar{m}$, $\bar{p}$ )

**( v )** Giving a reason, which of the following would best describe the correlation ?

**A** Positive **B** Negligible **C** Negative

**( vi )** The line of best fit has equation $p = -0.77 + m$

Add the line of best fit to the scatter graph by letting $m = 1$ in the regression line equation to get a point ( 1, ? ), then letting $m = 24$ to get another point ( 24, ? ) and finish by joining the two points.

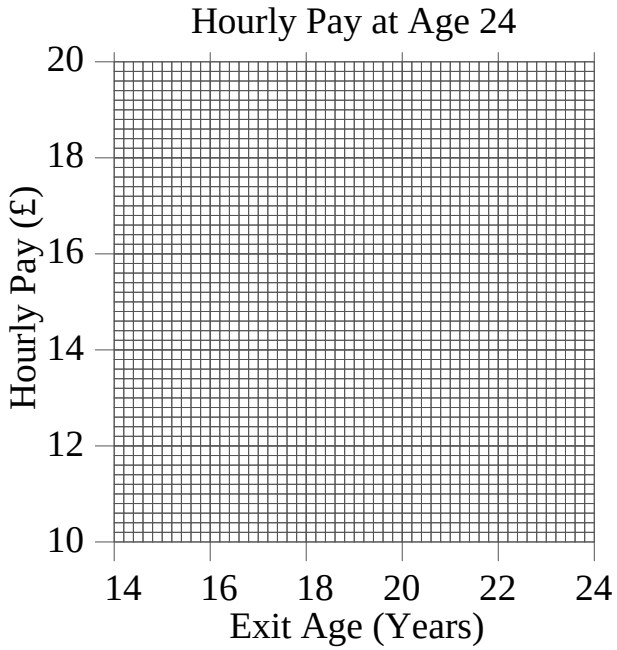
## 1.4 Exercise

**Question 1**

Natasha is investigating the correlation, if any, between what people earn and the age at which they left education or training, their "Exit Age".
Ten of Natasha's friends filled in an anonymous questionnaire. Here are the results;

| Friend | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Exit Age | 18 | 23 | 16 | 21 | 22 | 18 | 16 | 23 | 16 | 20 |
| Hourly Pay | 13.80 | 10.60 | 14.10 | 12.20 | 12.80 | 16.30 | 12.40 | 10.10 | 16.40 | 13.80 |

**( i )** Plot a scatter graph to represent these data




Let $\bar{e}$ be the mean Exit Age, and $\bar{p}$ be the mean Hourly Pay

**( ii )** Calculate $\bar{e}$ and $\bar{p}$

**( iii )** Plot the point ( $\bar{e}$, $\bar{p}$ ) on the scatter graph

**( iv )** Draw a horizontal broken line and a vertical broken line through ( $\bar{e}$, $\bar{p}$ )

**( v )** Giving a reason, which of the following would best describe the correlation ?

**A** Positive **B** Negligible **C** Negative

**( vi )** The line of best fit has equation $p = 23.4 - 0.528\,e$
Add the line of best fit to the scatter graph by letting $e = 14$ in the regression line equation to get a point ( 14, ? ), then letting $e = 24$ to get another point ( 24, ? ) and finish by joining the two points.

**Question 2**

A whisky distillery in Scotland matures its whisky in oak casks.
During the time it is maturing, evaporation takes place.
( The whisky so lost is called The Angels' Share; the subject of a film in 2012 )

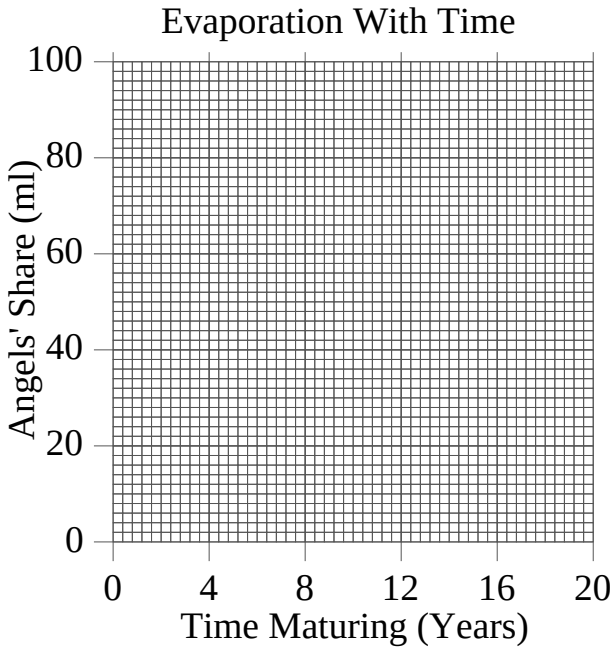
A random sample of 11 casks was taken and the time in storage, *x* years, and the evaporation loss, *y* ml, are shown in the following table;

| Cask ID | 412 | 531 | 487 | 583 | 417 | 418 | 434 | 438 | 512 | 576 | 433 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Age, *x* years | 3 | 15 | 12 | 18 | 4 | 5 | 8 | 10 | 13 | 16 | 6 |
| Angels' Share, *y* ml | 36 | 90 | 79 | 96 | 98 | 50 | 61 | 69 | 82 | 88 | 53 |

**( i )** What is the sampling frame being used ?
Be specific as to what details the sampling frame must contain.

**( ii )** Give an example of a unit from the sampling frame.

**( iii )** Plot a scatter graph to represent the data




After the sample has been taken, it is noticed that one of the casks has been leaking.

**( iv )** What is the cask ID of the cask that has leaked ?

**( v )** As a result of the leak, it is necessary to 'clean the data'.
Explain what this means will be done.

Let $\bar{x}$ be the mean Time Maturing, and $\bar{y}$ the mean Angels' Share for the cleaned data.

**( vi )** Calculate $\bar{x}$ and $\bar{y}$

**( vii )** Plot the point ( $\bar{x}$, $\bar{y}$ ) on the scatter graph

**( viii )** Draw a horizontal broken line and a vertical broken line through ( $\bar{x}$, $\bar{y}$ )

**( ix )** Giving a reason, which of the following would best describe the correlation ?

**A** Positive **B** Negligible **C** Negative

**( x )** Explain why it is appropriate to draw a line of best fit.
HINT : When would it not make sense to draw such a line ?

The equation of the regression line of $y$ on $x$ is $y = 29.02 + 3.9\,x$

**( xi )** Add this line to the scatter graph.

The distillery uses this model to predict the amount of evaporation that would take place after 20 years and after 30 years.

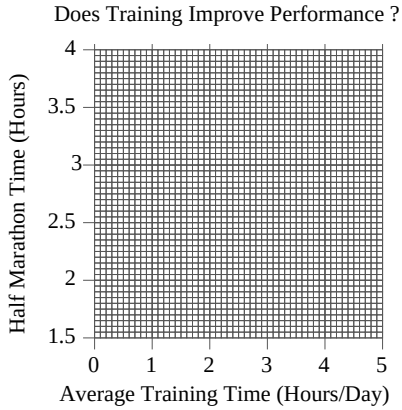
**( xii )** Comment, with a reason, on the reliability of each of these predictions.

**Question 3**

Tracy claims that 'cleaning the data' is the same as removing all outliers.
Explain why Tracy is wrong.

**Question 4**

Ian is gathering data to see if there is a correlation between how well a runner does in the Shrewsbury half marathon, and the average amount of training per day they did in the month ahead of the race.
He's about to plot a scatter graph.




Giving a reason, explain if you are expecting positive, negative or negligible correlation.

**Question 5**

Statisticians have found that in London violent crime is correlated with ice cream sales.

**( i )** Does this mean that there is something in ice cream that makes people more violent ?

**( ii )** The weather is a confounding variable.
Explain what a confounding variable is and how it might explains the correlation between violent crime and ice cream sales.

**Question 6**

A researcher is investigating if there is a correlation between how good a person's eyesight is and how long they spend playing computer games.

The researcher tests the eyesight of eleven volunteers using the LogMAR measure of visual acuity (clarity of vision) where

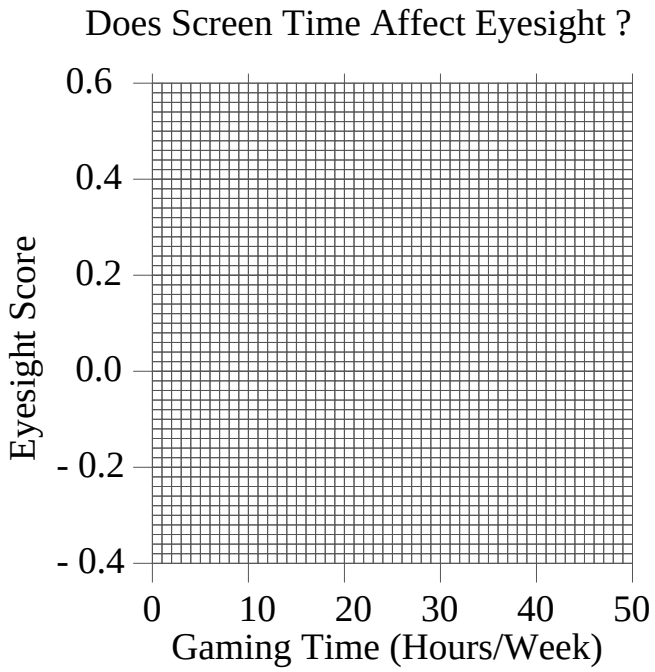
- 1.00 is poor eyesight 20/200 vision
- 0.00 is good eyesight 20/20 vision
- – 0.40 is excellent eyesight 20/8 vision

The researcher also records the number of hours each of the eleven volunteers spends playing computer games over a week. The results are presented below;

| Volunteer | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gaming time, $x$ hours | 4 | 23 | 35 | 12 | 44 | 0 | – 3 | 46 | 27 | 8 | 32 |
| Eyesight score, $y$ | 0.4 | 0.6 | – 0.3 | 0.3 | 0.3 | – 0.2 | 0.6 | – 0.1 | 0.1 | – 0.1 | 0.3 |

**( i )** Clean the data, justifying what you do.

**( ii )** Plot a scatter graph to represent the data.




**( iii )** A computer gives the line of best fit as $y = 0.143 - 0.000584\,x$

Explain why this line should not be added to the scatter graph.