B) Quantitative Interpretation of Correlation

The correlation will be represented by a number, called the correlation coefficient.

This coefficient will range from -1 to $+1$.

Its symbol is r.

Estimating the Correlation Coefficient

Example: Push-ups and Sit-ups

(27, 30), (26, 28), (38, 45), (52, 55), (35, 36),
(40, 54), (40, 50), (52, 46), (42, 55), (61, 62),
(35, 38), (45, 53), (38, 42), (63, 55), (55, 54),
(46, 46), (34, 36), (45, 45), (30, 34), (68, 62)

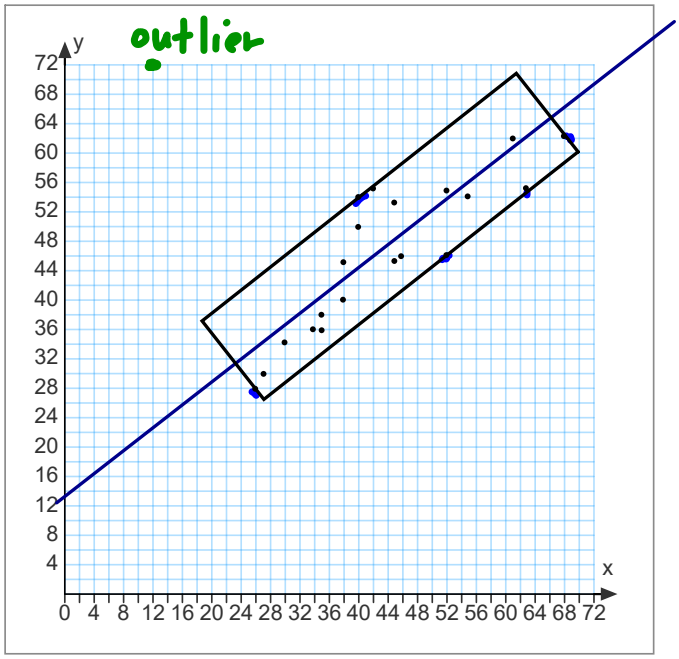
1) Draw a scatterplot (Done)

(27, 30), (26, 28), (38, 45), (52, 55), (35, 36),
(40, 54), (40, 50), (52, 46), (42, 55), (61, 62),
(35, 38), (45, 53), (38, 42), (63, 55), (55, 54),
(46, 46), (34, 36), (45, 45), (30, 34), (68, 62)

2) Draw the line of best fit, aka the regression line, that passes through the points.

3) Around the points, draw the smallest rectangle possible. Two of the sides should be parallel to the line.

4) Measure the dimensions of the rectangle.



short side: width = 2.4

long side: length = 7.9

5) Calculate the correlation coefficient, r.

choose the right sign

$$r = \pm\left(1 - \frac{width}{length}\right)$$

short / long

(+) if it's increasing,

(–) if it's decreasing

$$r = +\left(1 - \frac{2.4}{7.9}\right)$$

$$r = +(1 - 0.3)$$

$r = 0.7$ moderate

r	Meaning
Near 0	Zero correlation
Near ± 0.5	Weak correlation
Near ± 0.75	Moderate correlation
Near ± 0.87	Strong correlation
Near ± 1	Perfect correlation

The sign of the correlation coefficient has nothing to do with strength; it only indicates direction.

Note: The reason we estimate the correlation coefficient is because this is the actual calculation:

$$r = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2} \cdot \sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

Significant Correlation:

In order for a correlation to be considered significant (that is, a good indicator that there is a statistical relationship between the two variables), it must reach a certain value. This value is called a critical value.

If $|r| > c$, then the correlation is considered significant, (with a margin of error of 5%). This depends on the number, n, of values in the distribution.

In our example, $n = 20$; therefore r should be greater than 0.444.

The estimate was 0.7, so r is significant .

n	c
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
12	0.576
15	0.514
20	0.444
25	0.396
30	0.361
40	0.312
50	0.279
60	0.254
80	0.220
90	0.207
100	0.196

The Regression Line

Also called a line of best fit, a regression line is one that best represents (or passes through) the points of a scatterplot.

There are several different methods of finding the equation of a regression line.

Mayer Line Method

Example: (# of push-ups, # of sit-ups)

(27, 30), (26, 28), (38, 45), (52, 55), (35, 36),
(40, 54), (40, 50), (52, 46), (42, 55), (61, 62),
(35, 38), (45, 53), (38, 42), (63, 55), (55, 54),
(46, 46), (34, 36), (45, 45), (30, 34), (68, 62)

a. Organise the ordered pairs in ascending order based on the x-coordinates.

b. Divide the ordered pairs into two equal groups, if possible.

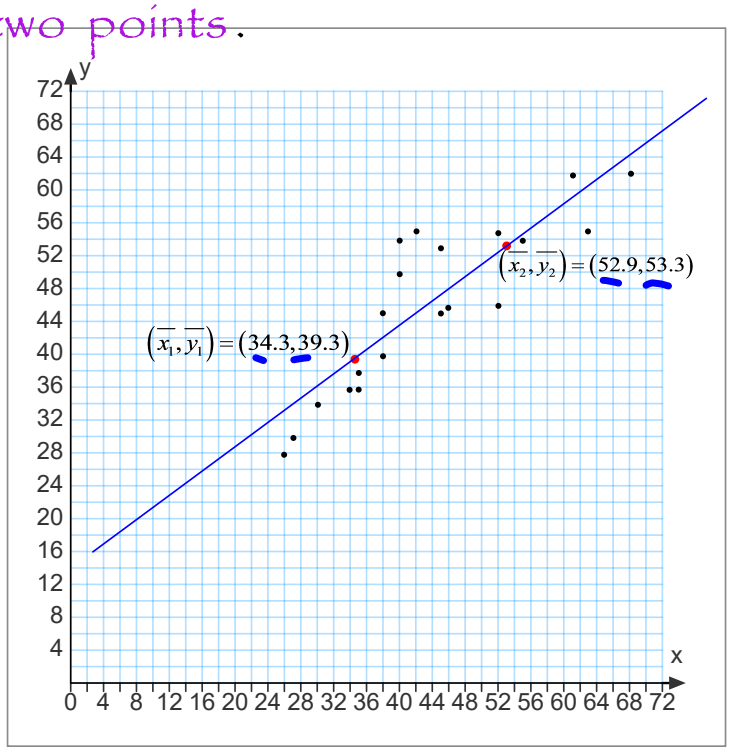
c. Calculate the means of the x-coordinates and the y-cordinates for each group.

x	y	
26	28	1
27	30	
30	34	
34	36	
35	36	
35	38	
38	42	
38	45	
40	50	
40	54	
42	55	2
45	45	
45	53	
46	46	
52	46	
52	55	
55	54	
61	62	
63	55	
68	62	

P_1: $(\overline{x_1}, \overline{y_1}) = (34.3, 39.3)$

P_2: $(\overline{x_2}, \overline{y_2}) = (52.9, 53.3)$

d. We can plot the points on the graph. The regression line is the line that passes through these two points.



Using the Regression Line

Besides playing a role in estimating the correlation coefficient, the regression line may be used to predict values that do not appear in the distribution.
If we have the value of one variable we can predict the value of the other.

Sometimes we can read the values from the graph.
Other times we need the equation of the regression line.

The reliability of the prediction depends on the strength of the correlation.

Recall the equation of a line...

$$y = mx + b$$

use the two average points

$$m = \frac{53.3 - 39.3}{52.9 - 34.3} = \frac{14}{18.6} \doteq .753$$

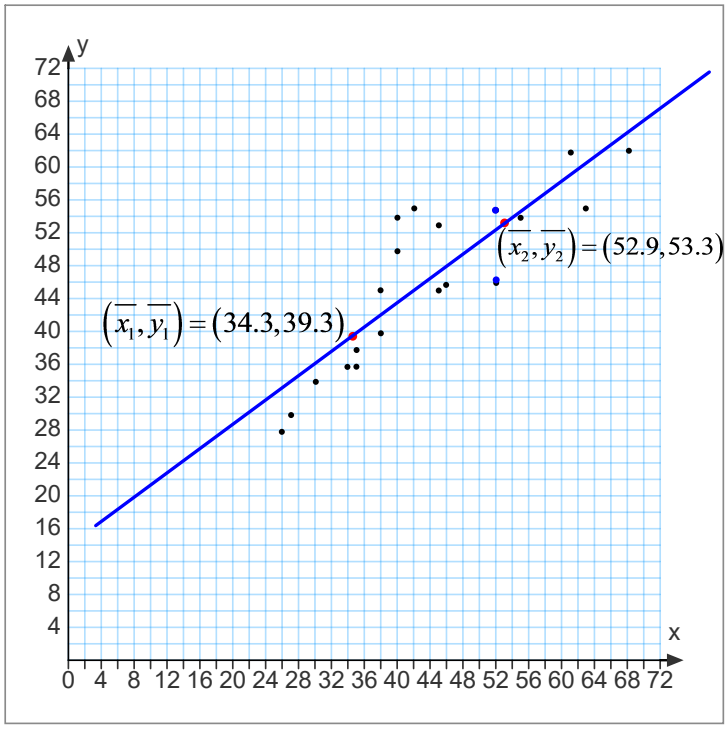
$$y = 0.753x + b$$

$$53.3 = 0.753(52.9) + b$$

$$53.3 = 39.817 + b$$

$$b = 13.48$$

$$\therefore y = 0.753x + 13.48$$



Example: $y = 0.753x + 13.48$ $r = 0.7$ estimated

$r = 0.87$ actual

x= # of push-ups

y= # of sit-ups

Interpolation Extrapolation

a) Predict the number of sit-ups a student can do if she can do 49 push-ups.

$y = 0.753(49) + 13.48$ (x = 49)

$y =$ =

b) Predict the number of push-ups a student can do if she can do 75 sit-ups.

$75 = 0.753x + 13.48$ (y = 75)

Interpreting a Correlation

A strong correlation indicates that there is a statistical relationship between two variables. But we have to be careful not to be too quick to interpret what the nature of that relationship is.

"Correlation Is Not Causation". This means is that a correlation does not prove one thing causes the other:

- One thing might cause the other
- The other might cause the first to happen
- They may be linked by a different thing
- Or it could be random chance!

There can be many reasons the data has a good correlation.

Interpretation	Example
• The link between two variables can be one of cause and effect: that is when with one of the variables has a direct effect on the other. In such cases, the correlation is perfect and the relation between the two variables is defined by a rule.	The correlation between altitude and temperature is perfect since the temperature varies in direct relation to altitude.
• The correlation between two variables can be significant without the two variables being directly linked to each other. They can both depend on a third variable which, as it varies, generates variations for the first two variables.	In the summer, it may seem that there is a strong correlation between the number of ice cream cones sold and the number of air conditioning units sold in a given city while in fact these two variables depend on another variable, is, the temperature.
• Considering a correlation as being linear while another model would be more appropriate.	The population growth of a major city can be studied according to a linear correlation. However, using an exponential model would be more appropriate.
• It sometimes may happen that there is a correlation between two variables only over a given interval.	Over the interval [5, 10] years, the correlation between a person's age and his or her height is linear. However, before and after this interval, the linear model is not the best fit.
• A two-variable distribution may include outlier data, notably due to manipulation or measurement errors.	The degree of precision of the instrument used during data collection is poor.