Statistical Applications

3: Correlation Coefficient and Regression Line

Chapter 5: p224 - 232

Measuring Correlation

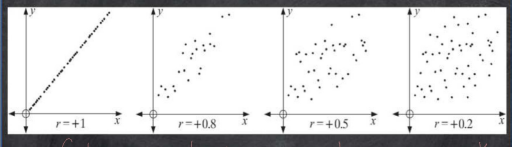
- with linear association, we can use the correlation coefficient to measure the strength and direction of association.
- the correlation coefficient (r) is a numerical measure of correlation - instead of "guessing" strength.
- the r value can have a value between -1 and 1.
- an r value of 0 suggests there is no linear association present (zero correlations)

Positive Correlation

- Positive correlation occurs when an increase in one variable results in the other increasing in an approximately linear manner
- the strength of the linear association is best measured with the Pearson's product-moment correlation coefficient (r).
- an r value of 1 suggests there is perfect linear association present (perfect positive correlation)

Positive Correlation

- r values between 0 & 1 represent varying degrees of linearity. (correlation)



perfect strong weak very weak

Negative Correlation

- Negative correlation occurs when an increase in one variable results in the other decreasing in an approximately linear manner
- an r value of -1 suggests there is perfect linear association present (perfect negative correlation)

r-value	Correlation
$0 < \|r\| \le 0.25$	Very weak
$0.25 < \|r\| \le 0.5$	Weak
$0.5 < \|r\| \le 0.75$	Moderate
$0.75 < \|r\| \le 1$	Strong

Learn this

→ find a line of best fit

Using the Calculator

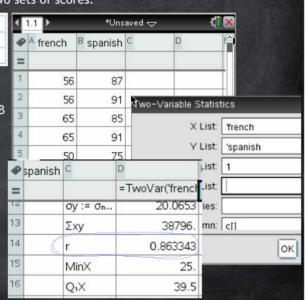
Finding correlation coefficient r

1. Nine students sat a French and a Spanish test. The table gives their results. Find the r-value and describe the correlation between the two sets of scores.

Subject	A	B	C	D	E	F	G	H	I
French	56	56	65	65	50	25	87	44	35
Spanish	87	91	85	91	75	28	92	66	58

Open a Lists & Spreadsheets page

- Type *french* in cell A and *spanish* in cell B
- Enter x and y values in rows 1 - 9
  NOTE it is important that you don't mix up the pairs
- Press MENU → 4:Statistics → 1: Stat Calculations → 2:Two-Variable Statistics
- Choose *french* for x and *spanish* for y. Frequency list stays at 1 & categories stay blank. Press OK
- Scroll to bottom to find r value



r = 0·863 (3 sf)

Correlation is positive and strong

Example to try:

1. Mo had always been told to stop playing computer games and get on with some work, so he decided to conduct a survey of 10 friends to see the effect on GPA. The results are in the table below.

GPA	2.7	3.8	1.5	3.6	2.2	3.8	2.0	1.9	2.5	3.0
Game time (h/week)	10	24	25	17	5	26	14	30	22	7

a) Find the r-value $r = 0.0262$

b) Describe the correlation (very) weak positive

c) Based on the survey, would Mo's grade increase if his game time decrease?

No, (almost) no correlation

Linear Regression

- A regression line is mathematically calculated so that the distances from the data points to the line is minimised.
- We always find the equation of the regression line with the GDC.

  (But in the IA, it will need to be calculated using a formula.)
- The equation of the line is usually written in the form of $y = mx + c$.

Using the Calculator
Finding the regression line equation

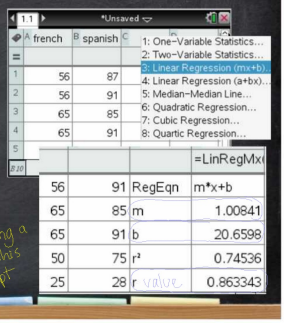
1. Nine students sat a French and a Spanish test. The table gives their results.
   b) Find the equation of the regression line.

Subject	A	B	C	D	E	F	G	H	I
French	56	56	65	65	50	25	87	44	35
Spanish	87	91	85	91	75	28	92	66	58

Open a Lists & Spreadsheets page

- Type *french* in cell A and *spanish* in cell B
- Enter x and y values in rows 1 - 9 NOTE it is important that you don't mix up the pairs
- Press MENU → 4:Statistics → 1: Stat Calculations → 3:**Linear Regression (mx+b)**
- Choose *french* for x and *spanish* for y. Frequency list stays at 1 & categories stay blank. Press OK
- The **m** and **b** values are given, and the **r value** is ALSO here

$y = 1.01x + 20.7$ if plotting a line use this y intercept



Interpolation & Extrapolation

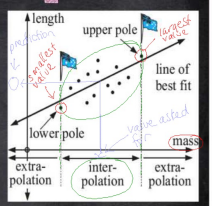
- A line of best fit allows predictions to be made.
- If we use the equation of the regression line to predict values between the smallest data value & the largest data value, we are interpolating ("between the poles")
- If we predict values outside the smallest & the largest data value, we are extrapolating ("outside the poles")

Interpolation & Extrapolation

This is a common IB Q

- The accuracy of interpolation depends on the strength of correlation (r value)

  *only reliable if r = strong/moderate



- Extrapolating data is more problematic as you assume the trends will continue (but have no data to support this).
- Generally it is safest not to make predictions by extrapolating. *Always say → "unreliable" (regardless of "r")

Example to try:

2. A patient is given medicine by a drip feed and its concentration in his blood is measured at hourly intervals. The doctors believe that a linear relationship will exist between the variables.

Time x (hours)	0	1	2	3	4	5	6
Concentration y	2.4	4.3	5.0	6.9	9.1	11.4	13.5

a) Find the r value $r = 0.992$

b) Describe the correlation strong positive

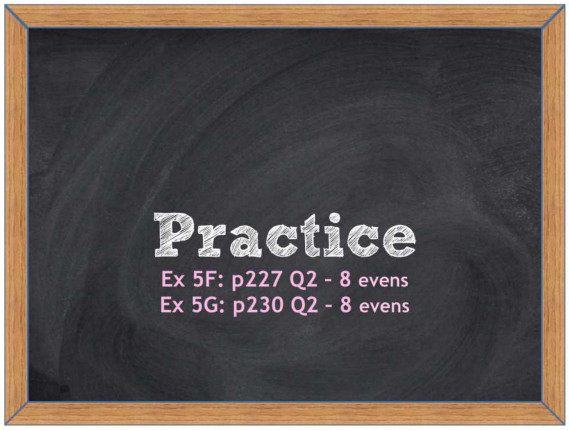
c) Write down the equatoin of the line $y = 1.98x + 1.99$

d) Find the concentration of the medicine in the blood after 3.5 hours. Is this a reliable prediction? Conc = 8.43 Yes, because r is strong

e) Find the concentration of the medicine after 8 hours. Is this a reliable prediction?

Conc = 16.71 = 16.7

No, because it is extrapolation


Practice
Ex 5F: p227 Q2 - 8 evens
Ex 5G: p230 Q2 - 8 evens