The big picture

Have you ever thought to yourself, 'Every time _____ happens, _____ seems to happen too'? You can fill in the blanks with almost anything, because it is natural to recognise patterns in life. When it is rainy, you feel more tired. The tallest players on the basketball court tend to be the best. When you get a good night's sleep, you do better on assessments. These patterns show a potential **relationship** between two variables, but it might just be coincidence. How can we determine whether the relationship really exists?

Important

If we can quantify the variables, then we can determine whether the relationship actually exists and how strong it is. We call this correlation .

Be aware

Correlation does not imply a cause-and-effect relationship between variables.

By recognising this and discussing other contributing variables within the situation, you can demonstrate critical reflection and understanding of mathematics, two of the criteria your IA is assessed for.

While a relationship may exist, that does *not* mean it is a cause-and-effect relationship. We may say that one variable is independent and the other dependent, but this is only in a mathematical sense and it does not indicate cause and effect. There are much more sophisticated tests that attempt to establish causal relationships, but we will not study those in this course.

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Before exploring the concept of correlation in detail, we need to define a few terms to set the context of our study. We will look at correlation of bivariate data, that is, data that are collected in pairs, which we usually write as points, (x, y). We call the x-coordinate the **independent variable**. It is plotted along the horizontal axis and is the data you would input in order to determine y. The y-coordinate is called the **dependent variable** and is plotted along the vertical axis.

Since we are looking for some kind of mathematical relationship, there should be a logical connection between the two variables. They should relate to the same scenario or group of people, as in the relationships mentioned at the start of the section. Would it be meaningful to try to find a relationship between the number of children in 20 families living in Queensland, Australia, and the number of bicycles sold in 20 local businesses in Stockholm, Sweden? You might find sets of data that show a mathematical relationship, but you must use discretion to appropriately interpret the relationship in context.

Concept

This subtopic is all about determining **relationships** between variables. How can you determine whether there is a significant relationship between two variables? If there is, how can you **model** that relationship in order to use it to make predictions?

4. Probability and statistics / 4.4 Linear correlation of bivariate data

An introduction to linear correlation

Recognising linear relationships

Scatter plots

A group of students are asked to jump from standing, first horizontally and then vertically. The vertical measurement is the distance between their reach and their jump. The following measurements were obtained:

	Horizontal jump (cm)	Vertical jump (cm)
Student 1	215	46

Student view

	Horizontal jump (cm)	Vertical jump (cm)
Student 2	200	43
Student 3	168	46
Student 4	191	57
Student 5	240	49
Student 6	128	23
Student 7	112	25
Student 8	121	28
Student 9	150	34
Student 10	140	30
Student 11	261	62
Student 12	170	37
Student 13	232	58
Student 14	212	52
Student 15	162	23

Is there any relationship between the horizontal and vertical jump distances? To investigate this, you can make a **scatter plot** of the data.

Activity

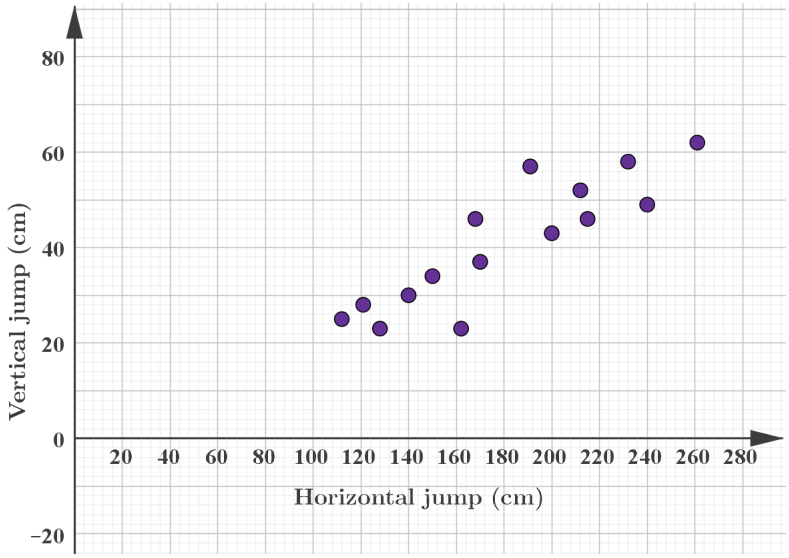
Plot the jump data on a sheet of graph paper, taking horizontal jump distance as the x-coordinate and vertical jump distance as the y-coordinate. When constructing the graph, make sure you clearly label both axes with the variable each represents and clearly mark the scale.

For this situation we have made the horizontal jump distance the independent variable, because we will eventually try to predict someone's vertical jump from their horizontal jump.

Important

Choose your independent and dependent variables carefully. The primary purpose of finding relationships is to create models for **prediction**. The variable you are trying to predict is the dependent variable (y), and the variable you use to make the prediction is the independent variable (x).

Overview (/study/app/math-aa-hl/sid-134-cid-761926/o



The scatter plot of horizontal jump distance against vertical jump distance.

More information

The graph is a scatter plot comparing horizontal jump distance to vertical jump distance. The X-axis represents horizontal jump distance in centimeters, with markers at intervals of 20 from 0 to 140 cm. The Y-axis represents vertical jump distance, also in centimeters, with markers at intervals of 20 from 0 to 80 cm. Several purple dots are scattered across the graph, indicating individual data points. The general pattern shows an upward trend, suggesting that as horizontal jump distance increases, vertical jump distance also tends to increase.

[Generated by AI]

Activity

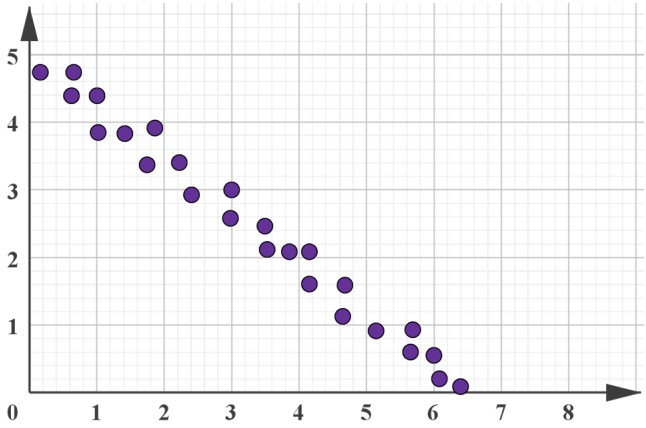
Draw a straight line that is as close as possible to all the points on your scatter plot. Is it easy to draw the line? What do you notice about the line? What does this tell you about the data?

You can draw a straight line that is quite close to many of the data points, meaning there is likely a linear correlation. More specifically, that line has a general upward trend, so you would describe it as a positive linear correlation, or just **positive correlation**. With this pattern, you can hypothesise that the further a student can jump horizontally, the higher they can jump vertically. Later we will see how likely this hypothesis is to be true, and in section 4.4.2 (/study/app/math-aa-hl/sid-134-cid-761926/book/pearsons-r-correlation-coefficient-id-25528/) we will construct a model to use for prediction.

While the scatter plot shown above shows a positive linear correlation, there are other trends we can identify as well.

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

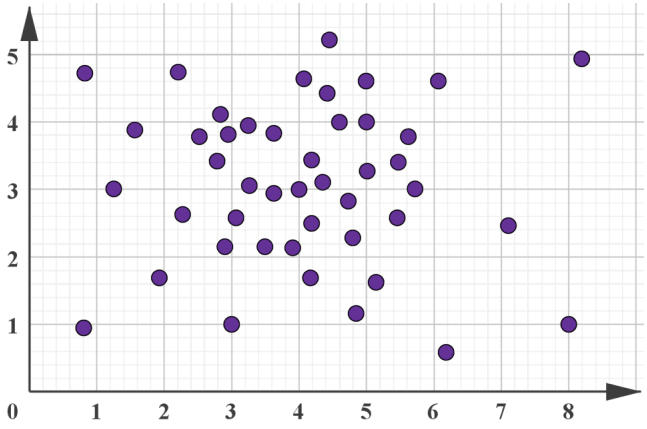


An example of negative linear correlation. The data has a downward trend.

More information

This image is a scatter plot graph illustrating a negative linear correlation. The X-axis ranges from 0 to 8, while the Y-axis ranges from 0 to 5. Purple data points are plotted across the grid, forming a downward trend from the upper-left to the bottom-right of the plot. This downward movement indicates a negative correlation between the variables indicated on the axes. Specific data points decrease as you move from left to right, showing a consistent downward trend in the data. This visual representation of data points highlights a pattern where one variable decreases as the other increases.

[Generated by AI]



An example of no correlation. There is no recognisable pattern.

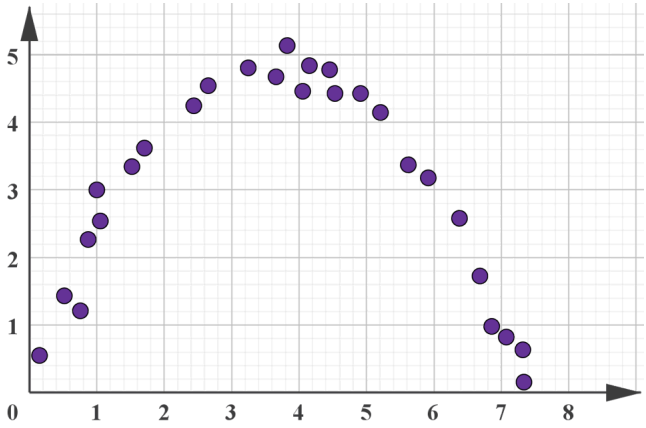
More information

The image is a scatter plot on a grid background. The X-axis ranges from 0 to 8, and the Y-axis ranges from 0 to 5. Numerous purple dots are scattered across the grid with no discernible pattern, signifying no correlation between the variables. Each dot represents a data point without showing any noticeable trend or relationship. The grid remains consistent throughout the plot, providing uniform spacing for the data points.

[Generated by AI]

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

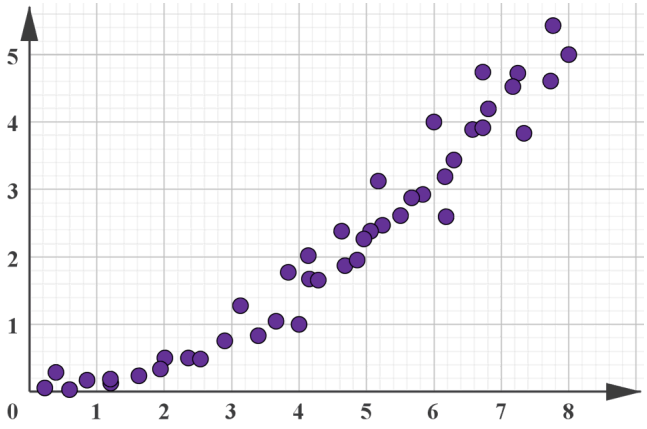


An example of a quadratic trend. The points form the shape of a parabola.

More information

The image is a graph displaying a set of data points arranged in a quadratic trend, resembling a parabola. The X-axis has values ranging from 0 to 8, while the Y-axis ranges from 0 to 5. The points begin near the bottom, increasing along the Y-axis up to a peak around the X-axis value of 3 to 4, and then begin descending symmetrically, forming a smooth curve. Purple dots represent the data points, following a clear symmetrical rise and fall pattern along the parabolic path.

[Generated by AI]



An example of an exponential trend. The points follow the shape of an exponential function.

More information

The graph displays an exponential trend with data points following an upward curve. The X-axis ranges from 0 to 8, representing an undefined variable, and the Y-axis ranges from 0 to 5, also representing an undefined variable. The data points are marked with purple dots distributed in a manner that creates an exponential trajectory, starting from the lower left to the upper right of the graph. The points increase gradually and then rapidly as they move along the curve, indicating an exponential growth pattern.

[Generated by AI]

Student view

Important

There are many types of functions, and data that you plot could show a trend similar to any of them. It is important to make sure the data you use when calculating linear correlation actually has a linear trend, or your model will not be reliable.

Making connections

While we are studying **linear** relationships, we have just seen that it is possible to identify other relationships. Thus, it is essential to examine the scatter plot of the data first to determine the most appropriate model to use. For example, exponential growth and decay models are used quite frequently in the natural sciences as scientists examine things that are continually growing proportionally. The methods we will explore to create and interpret linear models are similar to those employed to create other models.

Estimating the line of best fit by eye

Let's revisit the jumping data from the earlier table. On your scatter plot you drew a line to show that there is likely a linear correlation between the variables. We call this line a **line of best fit**. Can you think of a way that we could improve the accuracy of this line of best fit?

One way might be to draw the line **through** as many points as possible. But it might be impossible to find a line that goes through more than two points, and those two points might not be typical of the rest of the data. Instead, we draw the line through a single point that is most typical of the data as a whole. We call this point the mean point.

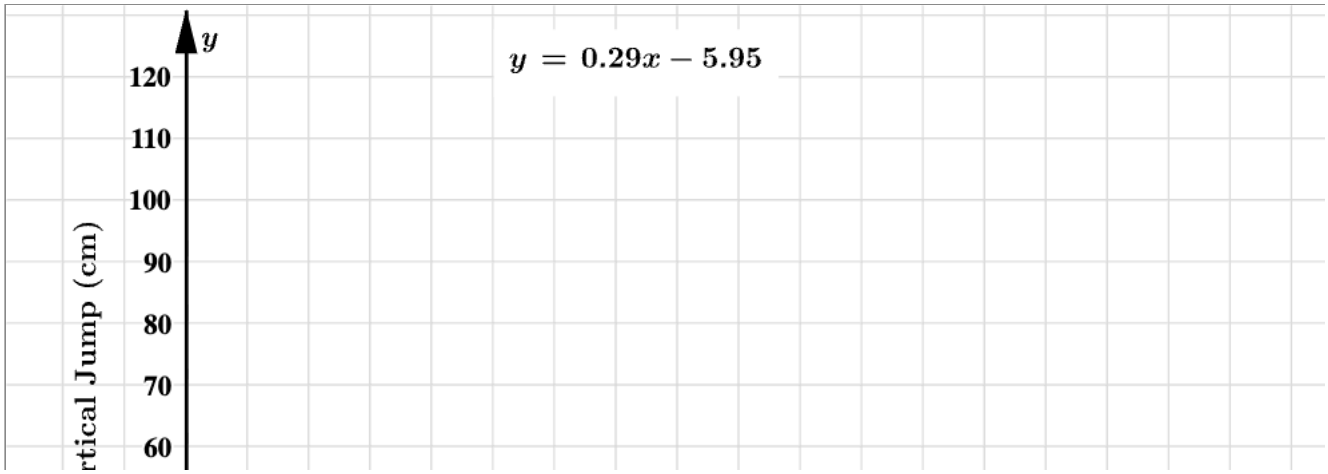
To find the mean point, begin by finding the mean of all the x-coordinates, $\bar{x}$, and the mean of all the y-coordinates, $\bar{y}$. Then $(\bar{x}, \bar{y})$ is the mean point . In our example, $\bar{x} = 180$ and $\bar{y} = 40.9$, so the mean point is $(180, 40.9)$. Plot this point and label it M, as shown in the scatter diagram below.

Our task now is to draw a single line that goes through the mean point and is as close as possible to each of the data points. When this is done properly, you should have approximately the same number of points above the line as there are below it.

Activity

Experiment with the applet below to see where you would put the line. Does it look the same as the line you drew on your scatter plot? The applet shows the equation of the line. Is the equation of the line similar to the equation given below?

Overview (/study/ap aa-hl/sid-134-cid-761926/o



Interactive 1. Equation of Line Plot.

More information for interactive 1

This interactive allows users to explore scatter plots and understand the concept of the **line of best fit**. The plot represents a dataset where:

- The **x-axis** (horizontal) represents the independent variable—**Horizontal Jump (cm)** (the forward jump distance).
- The **y-axis** (vertical) represents the dependent variable—**Vertical Jump (cm)** (the upward jump height).

A scatter plot of individual data points is displayed, showing how vertical jump height varies with horizontal jump distance. The interactive also highlights the **mean point** $(\bar{x}, \bar{y})$, , which is the average of all x and y values.

The **line of best fit** is displayed, calculated using the least-squares regression method. The equation of the line is: $y = 0.29x - 5.95$

This equation shows that for every additional centimeter in horizontal jump distance, the vertical jump increases by approximately **0.29 cm** on average.

The intercept(-5.95) suggests a theoretical vertical jump when the horizontal jump is zero, though it may not have a practical interpretation in this context.

A specific point, **M**$(180,\ 40.9)$ is highlighted in **green** on the graph. This represents an individual data point, showing that a person with a **horizontal jump of 180 cm** achieved a **vertical jump of 40.9 cm**. The relationship between this point and the best-fit line helps users understand **prediction accuracy and deviations (residuals)** from the model.

By analyzing the relationship between horizontal and vertical jumps, this tool helps users understand scatter plots, identify trends, and learn the principles of linear regression and the line of best fit in a practical way.

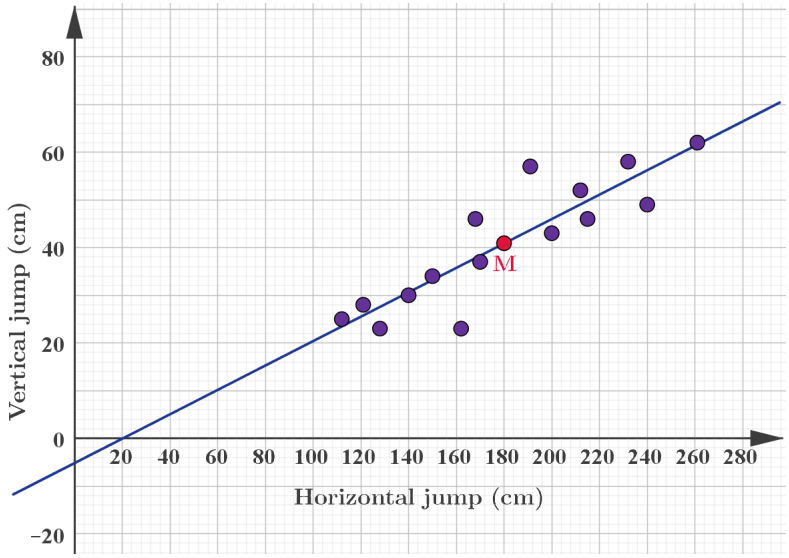
The exact equation of the best-fitting line is determined by the least-squares regression. This is a method to find the equation of a line in which the square of the vertical distance between the line and each data point is minimised for all data points. In section 4.4.3 (/study/app/math-aa-hl/sid-134-cid-761926/book/predicton-id-25529/), we will see how to calculate the slope and y -intercept to give the equation of the line.

In our example, the least-squares regression line has equation $y = 0.256x - 5.24$. It is shown in the figure below.

The line of best fit, when drawn by hand, must go through the mean data point and attempt to minimise the vertical distance between the line and all points.

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o



Scatter diagram with line of best fit.

More information

The image is a scatter diagram with axes and a line of best fit. The X-axis represents the 'Horizontal jump (cm)' and is labeled with values ranging from 0 to 300 in increments of 50. The Y-axis represents 'Vertical jump (cm)' with a scale labeled from -20 to 80 in increments of 20. Purple data points are scattered across the graph, and there is a blue line of best fit that trends upwards, indicating a positive correlation between the horizontal and vertical jumps. In the middle of the graph, one particular data point is marked with a red circle and labeled 'M', representing the mean data point. The points are distributed around the line with varying distances, reflecting the spread of the data.

[Generated by AI]

3 section questions

Question 1

Difficulty: ★☆☆

René wants to determine if there is a relationship between the number of snacks someone eats each day (x) and the number of times they work out in a week (y). He asks four of his friends and records his data in the table below.

x	y
2	5
4	3
7	6
9	2

René decides to first approximate a line of best fit for the data. He knows that this line passes through the mean point, $(\bar{x}, \bar{y})$, for his data. Find this mean point.

Write your answer as an ordered pair in parentheses with no spaces.

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

(5.5,4) ✓

Accepted answers

(5.5,4), 5.5,4

Explanation

The mean point, $(\bar{x}, \bar{y})$, has these coordinates:

$$\bar{x} = \text{mean of the x-coordinates} = \frac{2+4+7+9}{4} = 5.5$$

$$\bar{y} = \text{mean of the y-coordinates} = \frac{5+3+6+2}{4} = 4$$

Therefore, the mean point is $(5.5, 4)$.

Question 2

Difficulty: ★☆☆

To draw a line of best fit on a scatter graph using a pencil and ruler, the line must

1 Be a straight line that minimises the total vertical distance to all points and go through the mean point. ✓

2 Join up all the points in a zigzag line and go through the mean point.

3 Join all the points of the data set in a zigzag line.

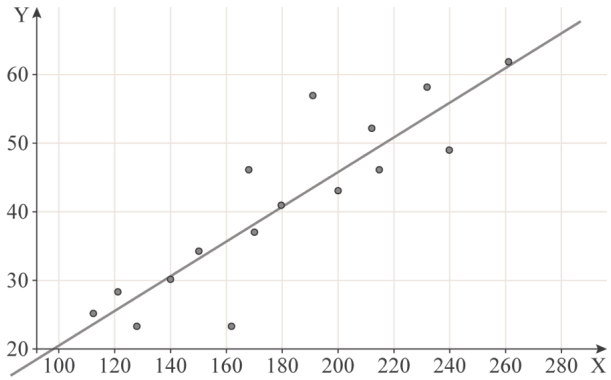
4 Be a straight line that goes through as many points as possible, including the mean.

Explanation

The line of best fit is drawn to be as close as possible to all points.

This might mean that it only goes through the mean value. It may go through some other points, but the aim is proximity rather than going through points.

Here is the scatter diagram of the jump data with the line of best fit.



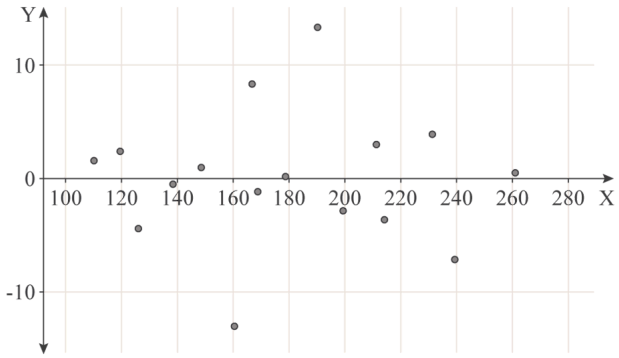
More information

Student view

Notice that the line touches only three points.

Overview (/study/ap aa-hl/sid-134-cid-761926/o

The same data set is shown in the graph below. However, this graph is called a 'residual graph'. Each data point is plotted with its distance from the line of best fit. The best line is the line such that the sum of all these distances is minimised.



More information

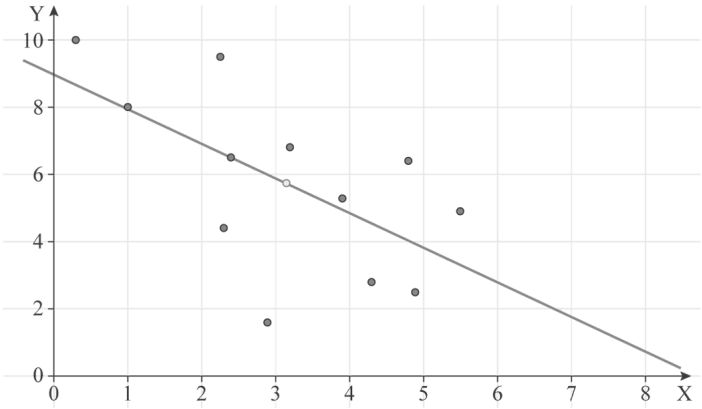
Question 3

Difficulty: ★★☆

A line of best fit is drawn on a data set. The point with values $(\bar{x}, \bar{y})$ is shown in white.

Which of the following shows the line of best fit?

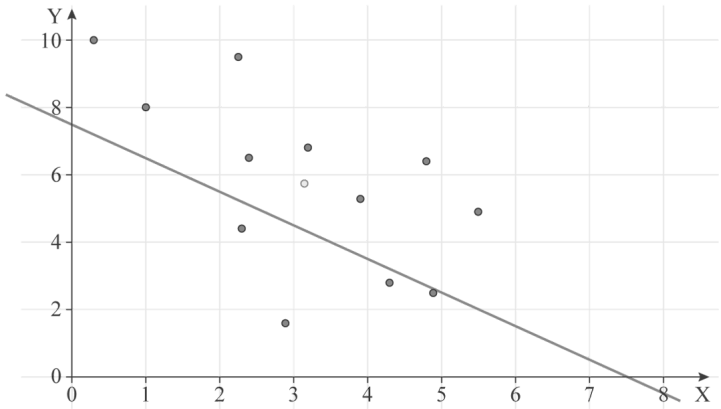
1



✓

More information

2

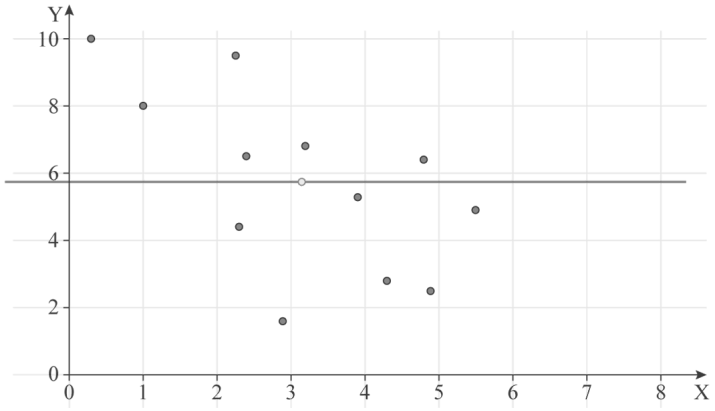


More information

Student view

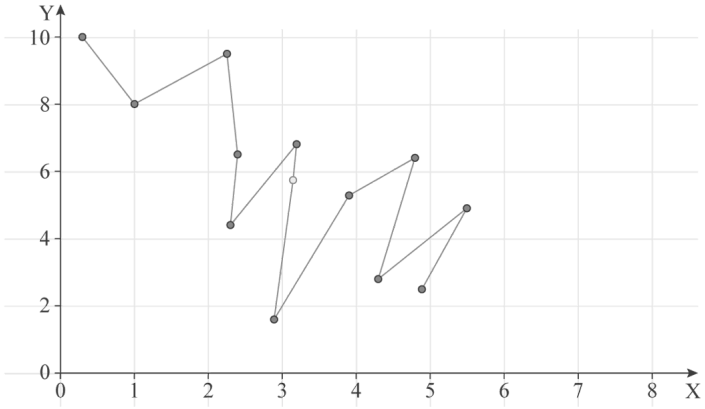
Overview (/study/ap aa-hl/sid-134-cid-761926/o

3



More information

4



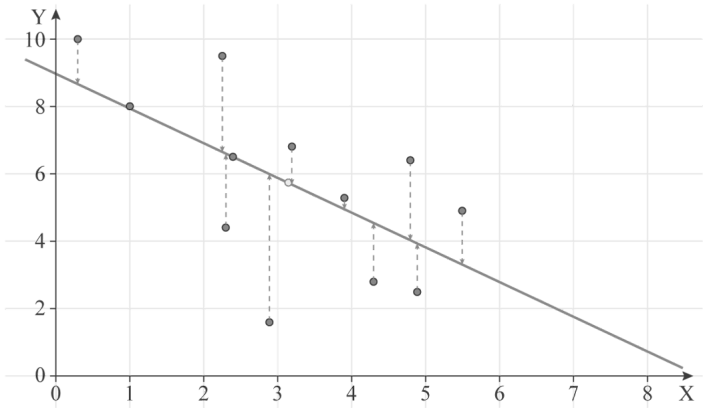
More information

Explanation

The line goes through the point $(\bar{x}, \bar{y})$ (the white point on this diagram).

Answer 2 does not go through the white point.

The job of the line is to go as close as possible to as many points as possible — to minimise the vertical distance:



More information

This is the best of the four options.

In the next subtopic, we will calculate this line using a formula.

Student view

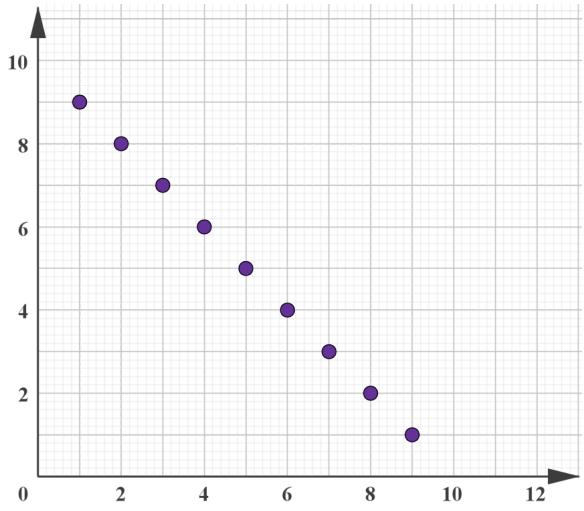
Pearson's r correlation coefficient

As you considered the concept of correlation as a linear trend in the previous section, you may have wondered how **reliable** the line of best fit really is. The line of best fit seeks to minimise the vertical distance of each point from the line, but one set of data might be clustered much closer to its line of best fit than another set of data. To analyse this difference, we seek to measure the correlation strength using the Pearson product-moment correlation coefficient, denoted by the variable r. This is often referred to simply as Pearson's r.

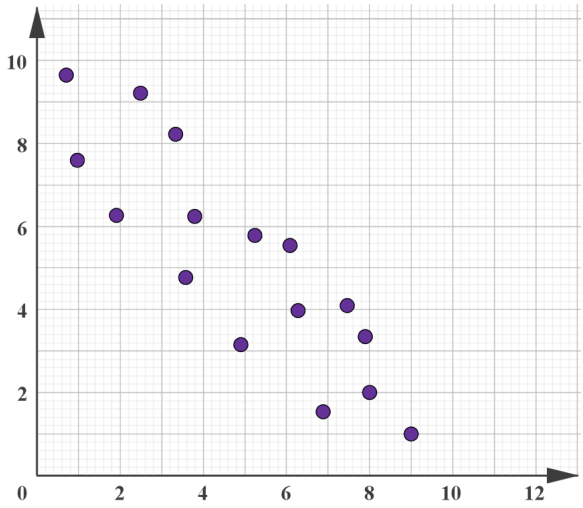
The value of the product-moment correlation coefficient lies in the range: $-1 \leqslant r \leqslant 1$, where

- $r = 1$is a perfect positive linear correlation: all the data points lie on the line
- $r = -1$is a perfect negative linear correlation: all the data points lie on the line
- $r = 0$is no linear correlation: no best-fitting line can be drawn.

The closer to $r = \pm 1$, the stronger the linear correlation there is, e.g. $r = 0.8$ $r = -0.8$is said to be a strong positive (negative) linear correlation, while $r = -0.5$indicates a moderate negative correlation and $r = 0.3$a weak positive linear correlation. Some typical r values with associated scatter plots are shown in the figures below.

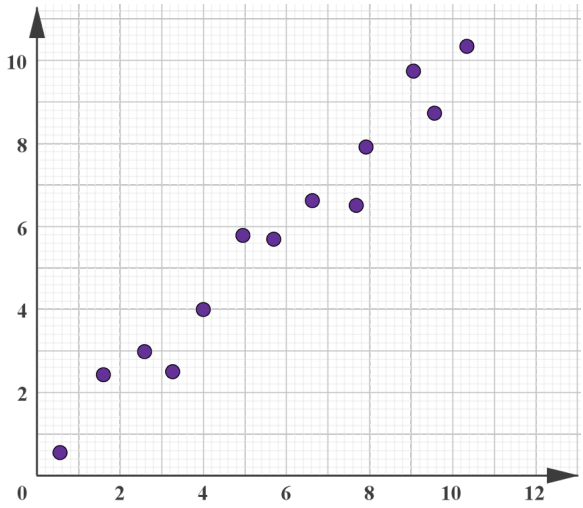
Step	Explanation
A perfect negative linear correlation, $r = -1$.	 More information The image is a scatter plot graph showing a perfect negative linear correlation. The X-axis represents numerical values from 0 to 12, increasing at intervals of 2, while the Y-axis represents numerical values from 0 to 10, increasing at intervals of 2. The points form a straight line from the top-left to the bottom-right, indicating a perfect negative linear correlation with a correlation coefficient of (r=-1). The data points start at approximately (1, 9) and descend evenly to about (11, 1). [Generated by AI]

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Step	Explanation
A moderate negative linear correlation, $r = -0.5$.	 More information The image is a scatter plot graph showing a negative linear correlation with a correlation coefficient of r = -0.5. • **X-axis**: The horizontal axis ranges from 0 to 12, with increments of 2. It likely represents the independent variable or measurement. • **Y-axis**: The vertical axis ranges from 0 to 10, with increments of 2. It represents the dependent variable or measurement. Data points are scattered across the plot. They are distributed in a downhill trend from left to right, illustrating a negative correlation. As the values on the X-axis increase, the values on the Y-axis tend to decrease, indicating that as one variable increases, the other decreases. [Generated by AI]

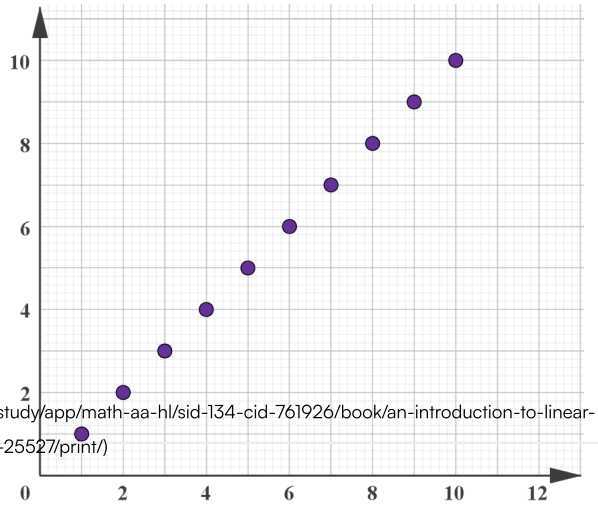
Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
A strong positive linear correlation, $r = 0.8$.	 More information This graph shows a scatterplot with a high positive linear correlation. • **X-Axis:** The X-axis represents values from 0 to 12. • **Y-Axis:** The Y-axis displays values ranging from 0 to 10. • **Data Points:** There are several points arranged in a linear upward trend. The data points start lower on the left side of the graph and increase towards the right, showing a positive correlation with an approximate correlation coefficient, r, of 0.8. This indicates that as the values on the X-axis increase, the values on the Y-axis also increase, following a linear pattern. [Generated by AI]

Student view

Overview (/study/aa-hl/sid-134-cid-761926/o

Step	Explanation
A perfect positive linear correlation, $r = 1$.	

Section Student... (0/0) Feedback Print (/study/app/math-aa-hl/sid-134-cid-761926/book/an-introduction-to-linear-correlation-id-25527/print/) Assign

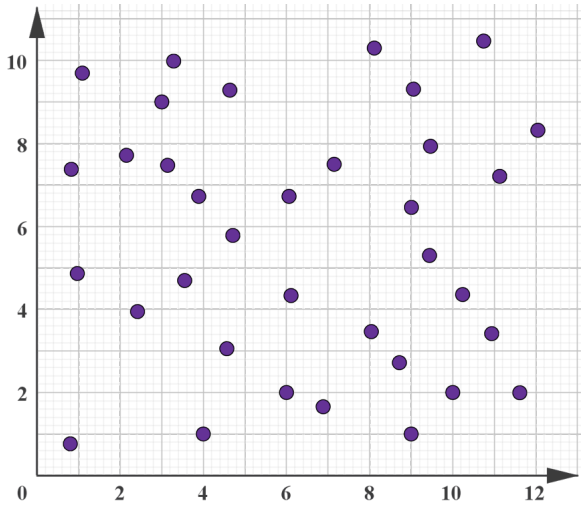
More information

The graph shows a perfect positive linear correlation with a correlation coefficient of (r=1). The X-axis represents values from 0 to 12 in intervals of 2. The Y-axis represents values from 0 to 10 in intervals of 1. The data points form a straight line, with each Y value being a perfect linear increment of 1 as the X value increases by 1. The points are evenly spaced, indicating a consistent and perfect linear relationship between the variables.

[Generated by AI]

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
No (linear) correlation, $r = 0$.	 More information The image is a scatter plot graph showing no linear correlation with a correlation coefficient (r) of 0. The graph has a vertical Y-axis and a horizontal X-axis, both numbered from 0 to 12. Data points are scattered randomly across the graph, with no visible pattern or trend indicating any relationship between the variables represented on the X-axis and Y-axis. The grid lines are evenly spaced, and the data points are represented as purple dots. The randomness of the plot points supports the description of no linear correlation, meaning changes in one variable do not predict changes in the other. [Generated by AI]

The strength of a correlation can fall into four general categories: strong, moderate, weak or no correlation. While it is nearly impossible to find data that has **exactly** $r = 0$, we often describe data with r very close to 0 as having no correlation, as in the table below.

Ranges of Pearson product-moment correlation.

Value of r	Description
$0.7 < r \leqslant 1$	Strong positive correlation
$0.3 < r \leqslant 0.7$	Weak to moderate positive correlation
$-0.3 < r \leqslant 0.3$	No correlation
$-0.7 < r \leqslant -0.3$	Weak to moderate negative correlation
$-1 \leqslant r \leqslant -0.7$	Strong negative correlation

Student view

Overview (/study/app/math-aa-hl/sid-134-cid-761926/o

Exam tip

Note that there is no universal agreement on the cutoff values given above. On an exam, if you are asked to classify a relationship based on the correlation coefficient, the critical values will be given.

Important

The strength of a correlation is only one part of this process. Another factor is **significance**, which takes into account the number of points in the data set. Which set of data do you think would be more useful for making predictions: a set of 5 data points with $r = 0.9$ or a set of 50 data points with $r = 0.75$?

Calculating Pearson's r

The Pearson product-moment correlation coefficient can be found by hand, but it can be a very tedious process and, like standard deviation, is a prime example of the benefits of using technology.

The value of r is calculated using the following formula:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

An alternative version is the formula:

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)}\sqrt{(\sum y^2 - n\bar{y}^2)}}$$

Exam tip

In exams, you will always use a calculator to find r. In an Internal Assessment project, detailing the formula can demonstrate mathematical skill and can be a nice complement to the task.

Consider the two bivariate data sets in **Tables 1 and 2** below.

Student view

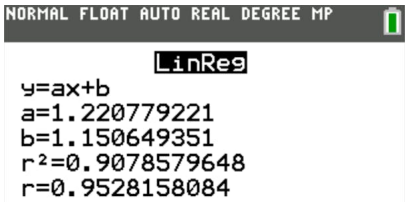
Overview (/study/aa-hl/sid-134-cid-761926/o

Table 1

x	y
5	4
7	12
12	18
16	21
20	24

Here is guidance on how to find the regression line and the Pearson correlation coefficient on some calculators.

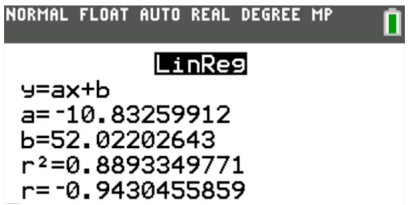
Pearson's r correlation coefficients for the data in **Tables 1 and 2** above are shown in the figures below. Note that the sign of the correlation coefficient is the same as the sign of gradient of the best fit line.



More information

The image shows the display of a calculator screen with the title 'LinReg' indicating a linear regression result. It includes the linear equation format 'y=ax+b' along with calculated values: 'a=1.220779221', 'b=1.150649351', 'r²=0.9078579648', and 'r=0.9528158084'. Above these results, there is a menu bar with options: NORMAL, FLOAT, AUTO, REAL, DEGREE, MP, and a battery indicator showing full charge. These values are the outputs of the linear regression computation performed by the calculator.

[Generated by AI]



More information

Student view

Overview (/study/aa-hl/sid-134-cid-761926/o

The image shows a calculator display with the results of a linear regression. The header reads 'NORMAL FLOAT AUTO REAL DEGREE MP' and there is a battery icon. Below is a box with 'LinReg' inside. The linear regression equation is displayed as 'y = ax + b', with specific values for coefficients and statistical indicators as follows: 'a = -10.83259912', 'b = 52.02202643', 'r^2 = 0.8893349771', and 'r = 0.9430455859'.

[Generated by AI]

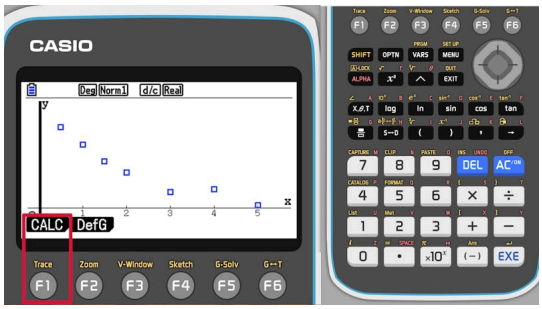
The output of the linear regression function on the TI-84 with the data of **Table 1**, showing a correlation coefficient of 0.953, indicating a **high positive linear correlation**.

The output of the linear regression function on the TI-84 with the data of **Table 2**, showing a correlation coefficient of -0.943, indicating a **high negative linear correlation**.

Step	Explanation
These instructions show you how to find the Pearson correlation coefficient, the equation of the regression line and how to display the scatter plot and the regression line. Press 2 to open the statistics mode.	
Enter the data (note that on this screenshot only the last few lines are visible). These instructions use the second data set from the example above. Once done, press F1 to start the process of drawing the scatter plot.	

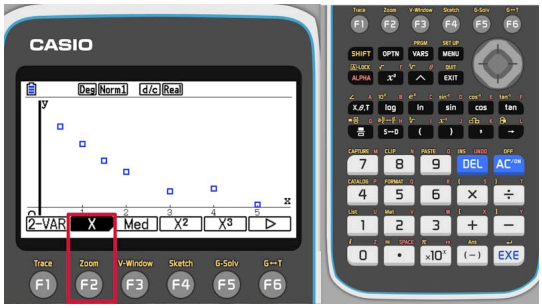
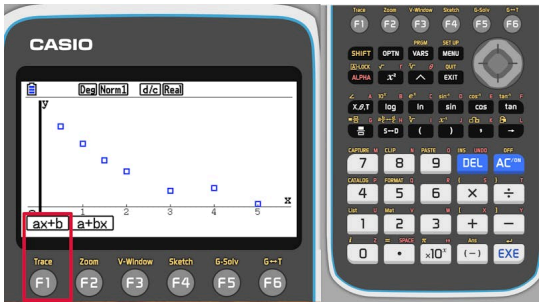
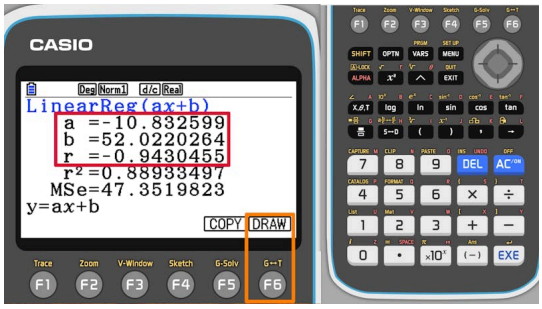
Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
You need to tell the calculator where the data is stored, so press F6.	
Make sure that every line is filled correctly. • You want to see a scatter plot • The x and y lists are stored in List1 and List2 • There are no frequencies involved, so make sure the frequency is set to 1 instead of a list. Once done, press EXE to confirm and F1 to see the scatter plot.	
To find the correlation coefficient and the regression line, press F1.	

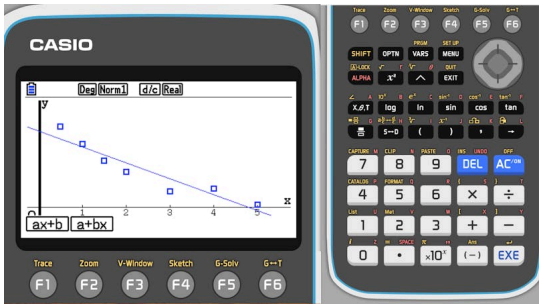
Student view

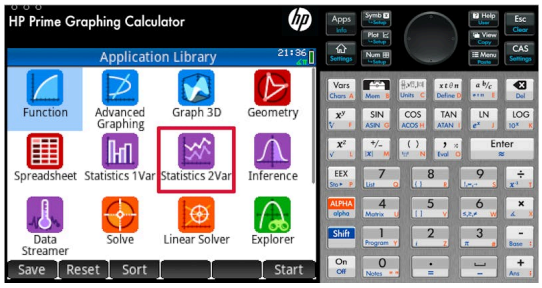
Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
You are looking for a linear regression, so press F2. If you are interested, at this point you can of course experiment with other models, too.	
Press F1 to choose the type.	
The calculator shows you the coefficients of the regression line (a and b) and the Pearson correlation coefficient (r). Press F6 to see the regression line and the scatter plot on the same diagram.	

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

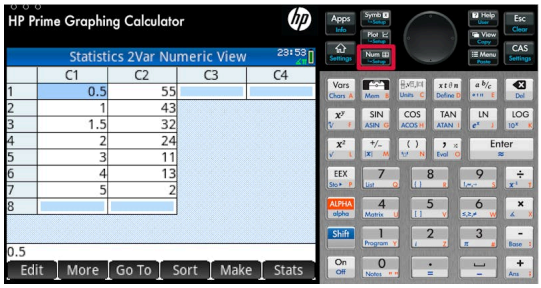
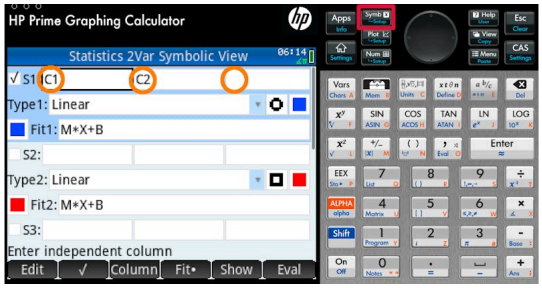
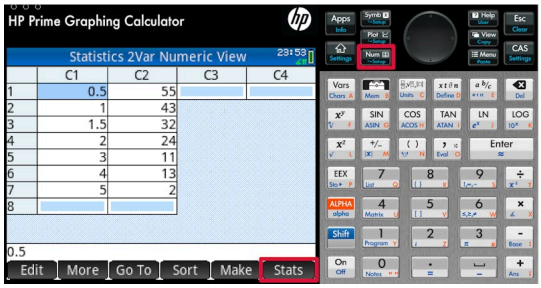
Step	Explanation
	

Step	Explanation
These instructions show you how to find the Pearson correlation coefficient, the equation of the regression line and how to display the scatter plot and the regression line. Choose the two variable statistics application.	

Section Student... (0/0) Feedback Print (/study/app/math-aa-hl/sid-134-cid-761926/book/pearsons-r-correlation-coefficient-id-25528/print/) Assign

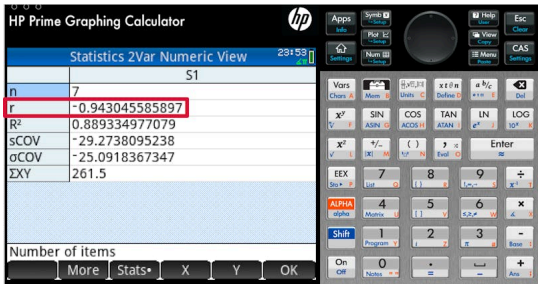
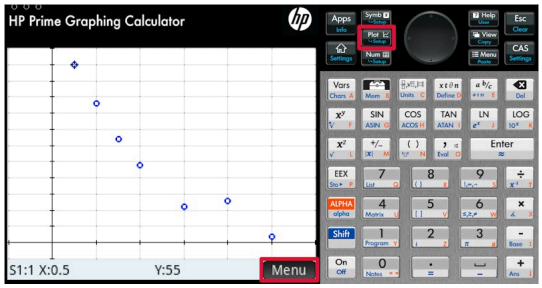
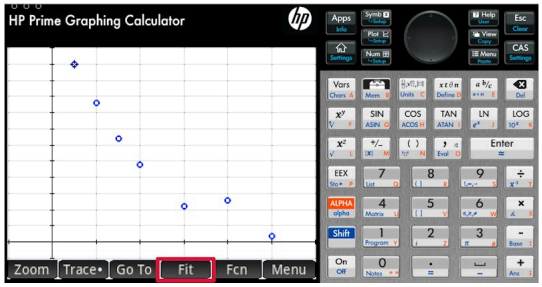
Student view

Overview (/study/aa-hl/sid-134-cid-761926/o

Step	Explanation
In numeric view, enter the data. These instructions use the second data set from the example above.	
Make sure that every field is filled correctly. • You are looking for a linear model (the type of equation is displayed as $y = mx + b$). • The x and y lists are stored in C1 and C2 • There are no frequencies involved, so make sure the frequency field is left empty.	
Go back to numeric view and press stats to view the result of the calculation.	

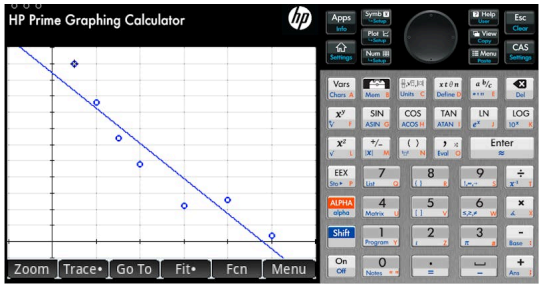
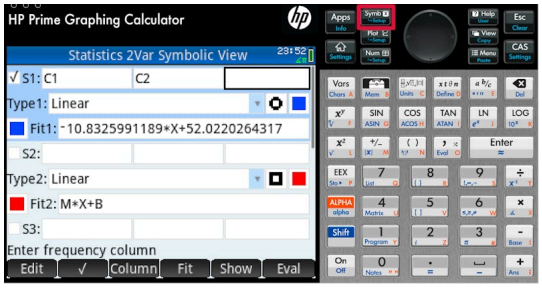
Student view

Overview (/study/aa-hl/sid-134-cid-761926/o

Step	Explanation
Among other information, the calculator shows you the Pearson correlation coefficient (r).	
In plot view you can see the scatter plot (you may need to adjust the viewing window in plot setup). Depending on your previous work with the calculator, you may also see the regression line. If not, press menu.	
To see the regression line, press fit.	

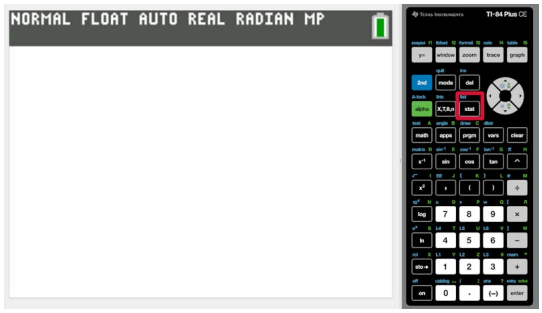
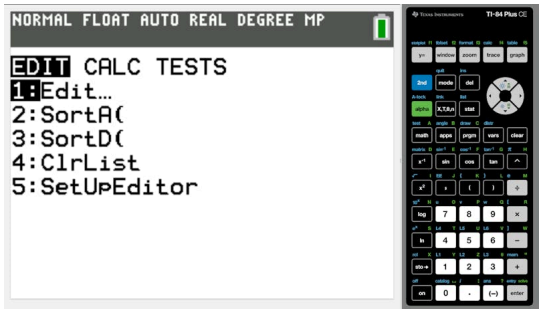
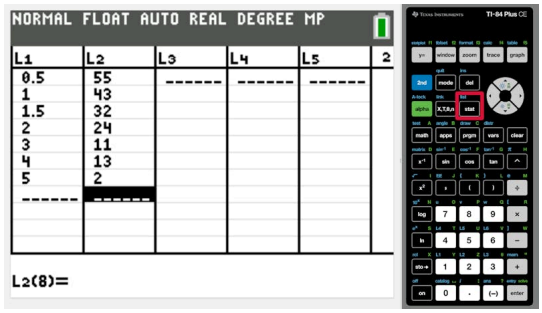
Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
	
In symbolic view, you can now also see the equation of the regression line.	

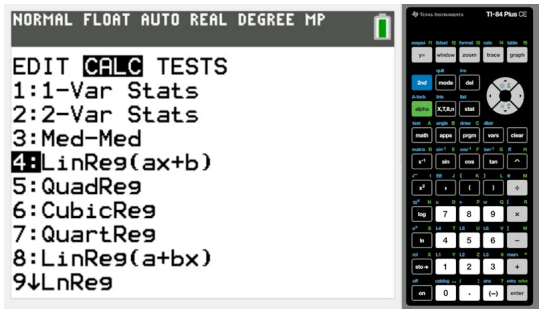
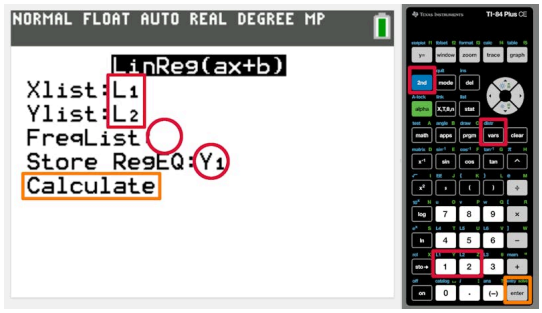
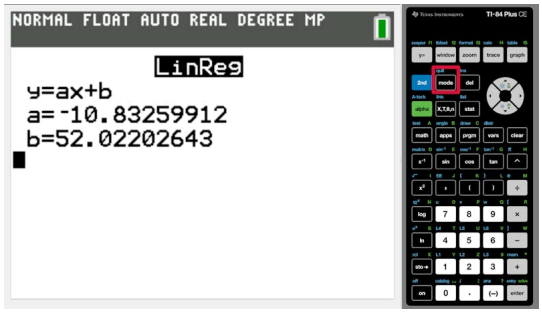
Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Step	Explanation
These instructions show you how to find the Pearson correlation coefficient, the equation of the regression line and how to display the scatter plot and the regression line. Press stat to work with data.	
First you need to tell the calculator the data, so choose the edit option.	
Enter the data. These instructions use the second data set from the example above. Once done, press stat again.	

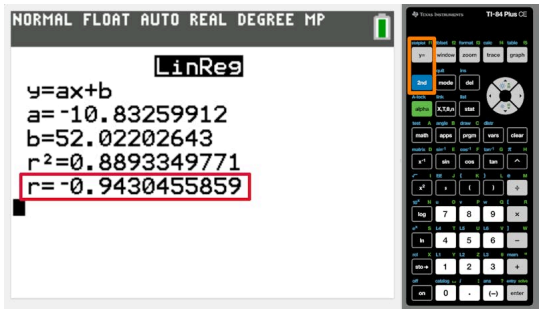
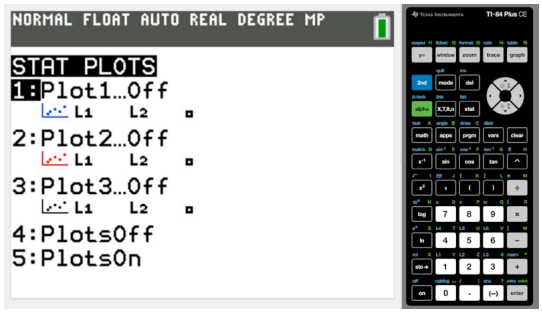
Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
Choose the option to calculate the linear regression.	
Make sure that every line is filled correctly. • The x and y lists are stored in L1 and L2. • There are no frequencies involved, so make sure the frequency list is left empty. • You can choose to store the equation of the regression line to any y-variable. This is useful when you want to display it. You can access the function variable names through the vars button. Once done, scroll down to calculate and press enter.	
The calculator displays the coefficients of the regression line. You may also see the correlation coefficient. If not (like on this screenshot), you need to change the mode.	

Student view

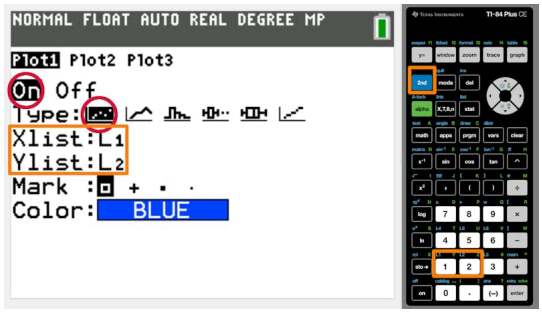
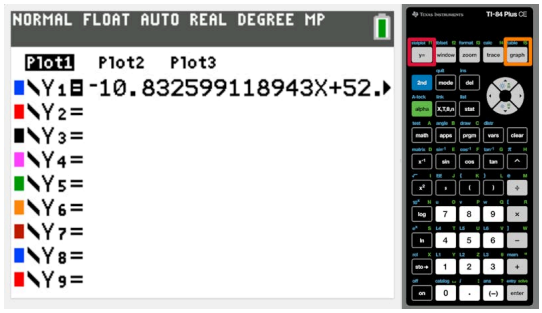
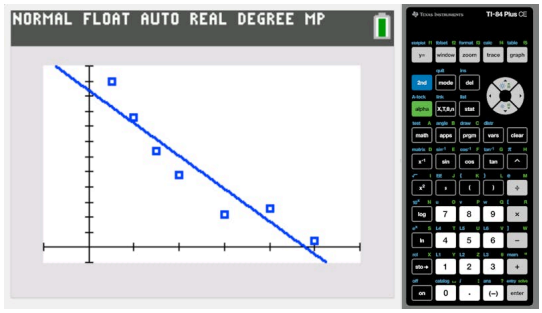
Overview (/study/ap aa-hl/sid-134-cid-761926/o

Step	Explanation
To see the correlation coefficient, statistics diagnostics needs to be turned on.	
Doing the process again, the calculator now also shows you the Pearson correlation coefficient (r). To see the scatter plot and the regression line, bring up the statistical plot options.	
Choose any of the plots.	

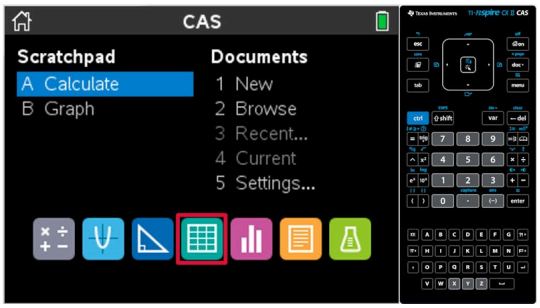
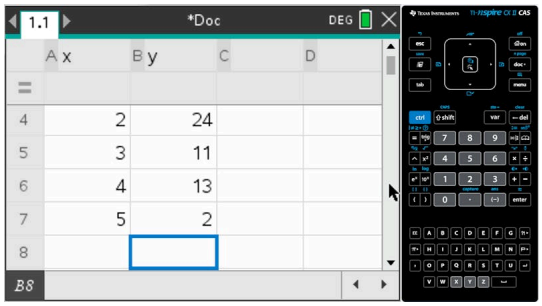
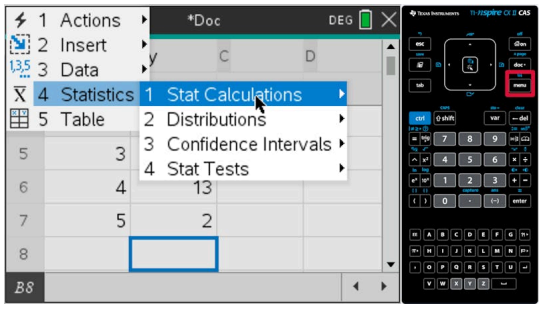
Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Student view

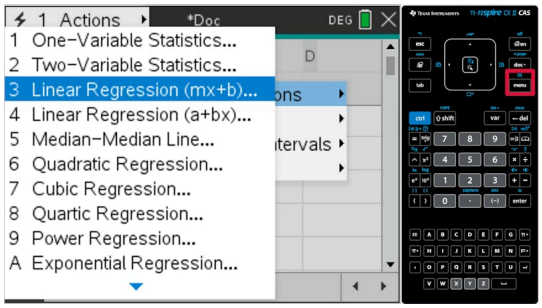
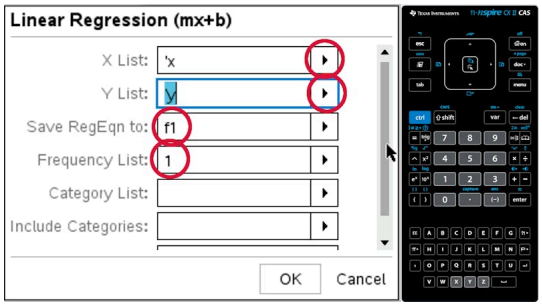
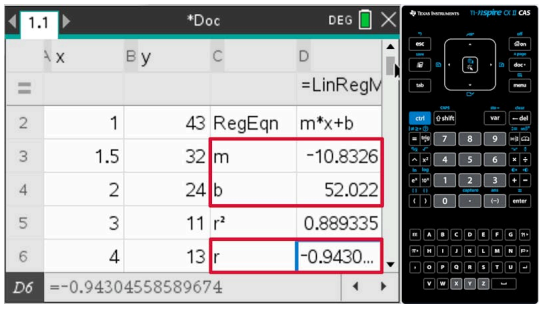
Step	Explanation
Turn the plot on, choose the scatter plot type and make sure that the correct list names are displayed for the x and y lists.	
In the function definition view notice that the equation of the regression line is stored already and that the first statistical plot is also highlighted to be drawn. Adjust the window and press draw to see the graphs.	
	

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
These instructions show you how to find the Pearson correlation coefficient, the equation of the regression line and how to display the scatter plot and the regression line. To start, open a spreadsheet page.	
Enter the data (note that on this screenshot only the last few lines are visible). These instructions use the second data set from the example above. You can also give a name for the columns. In this example the names x and y are used, but of course you can give more meaningful names depending on the context.	
To find the correlation coefficient, open the menu and choose statistical calculations ...	

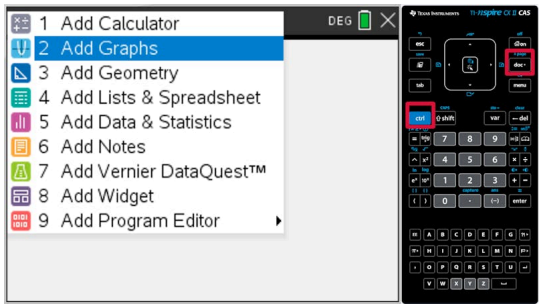
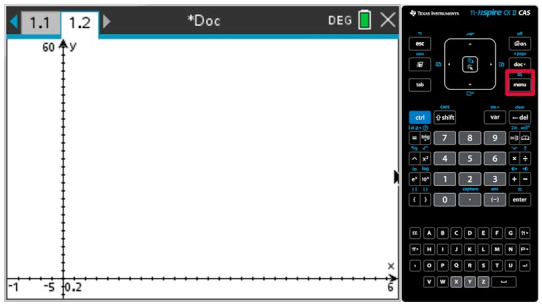
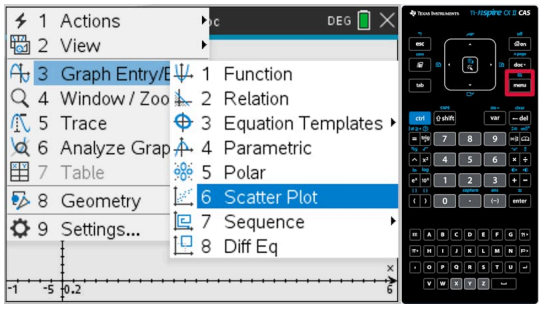
Student view

Overview (/study/aa-hl/sid-134-cid-761926/o

Step	Explanation
... and then the option to find the linear regression line.	
Make sure that every line is filled correctly. • Use the names you set for the x and y lists. • There are no frequencies involved, so make sure the frequency list displays the number 1. • You can choose to store the equation of the regression line. You can use any function name. This is useful when you want to display it. Once done, press OK.	
The calculator shows you the coefficients of the regression line (m and b) and the Pearson correlation coefficient (r).	

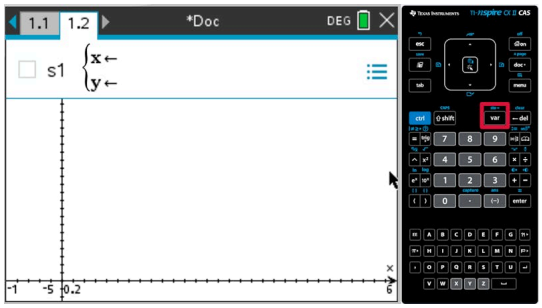
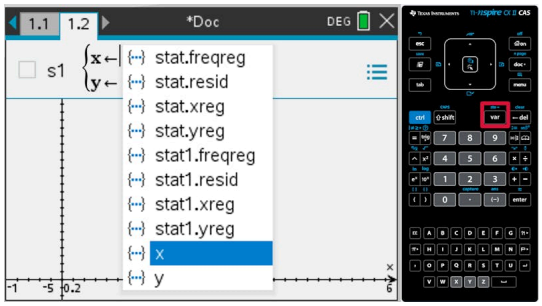
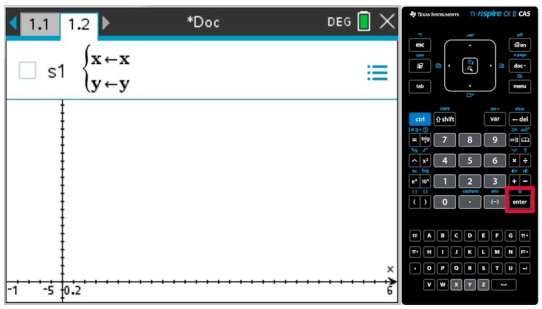
Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Step	Explanation
To see the scatter plot and the regression line, add a graph page to the document.	
Set the window and open the menu.	
First you can add the scatter plot.	

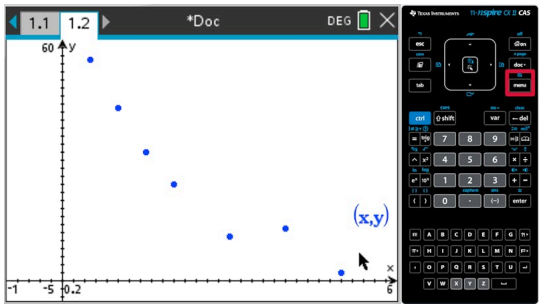
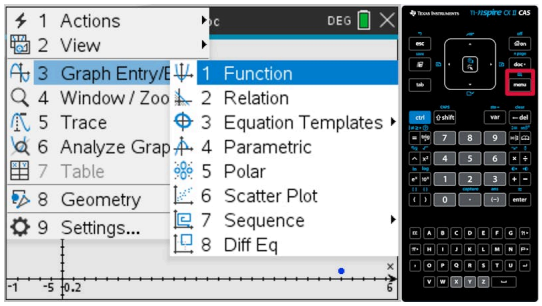
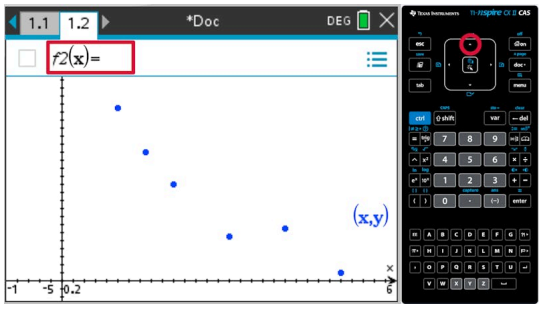
Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Step	Explanation
You need to tell the calculator, where the data is stored. Press the var button to get access to the name you set for your data.	
Choose the name you set for the independent variable ...	
... and for the dependent variable. Press enter to confirm your choice and to see the scatter plot.	

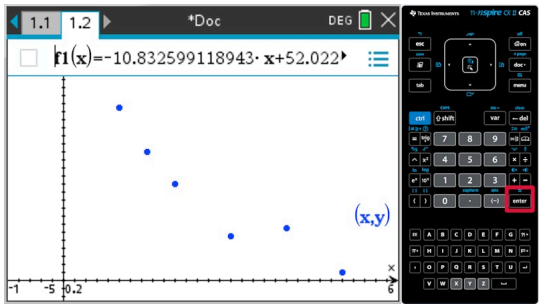
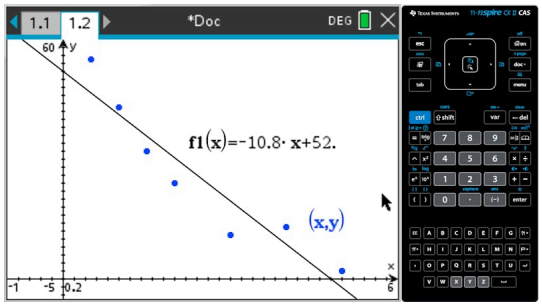
Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Step	Explanation
To add the regression line, open the menu again.	
This time, choose to add a function.	
The calculator asks for the definition of function $f2$ to display. Since you chose to store the equatin of the regression line in $f1$, you can bring it up by pressing the up arrow.	

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Step	Explanation
Once you see the definition, pressing enter will diplay the graph.	
	

Activity

Consider once again the data set from the horizontal and vertical jump exercise that we examined in section 4.4.1 (/study/app/math-aa-hl/sid-134-cid-761926/book/an-introduction-to-linear-correlation-id-25527/). Try to find the value of Pearson's r correlation coefficient for the data by hand using the formula given above. Use the calculator to check your answer. Can you describe the relationship better with r than you did in the previous section?

Example 1

Consider the following bivariate data.

X	72.9	74.2	76.8	65.2	81.2	82.0	70.7	84.2	72.6	75.4
Y	73.2	73.1	76.4	63.8	80.3	81.7	69.9	82.4	71.4	75.4

Find the Pearson correlation coefficient.

Student view

Overview (/study/aa-hl/sid-134-cid-761926/o

Calculators have a built-in app to find the correlation of paired data. Make sure you know where to find and how to use this app.

The following explanation demonstrates the long way of finding the correlation coefficient using the formula. Note that on exams you will not be expected to use this method.

Using the two-variable statistics capabilities of a calculator, you can find the values of the expressions needed in the formula:

- $n = 10$
- $\overline{x} = 75.52$
- $\sum_{i=1}^{n} x_i^2 = 57330.82$
- $\overline{y} = 74.76$
- $\sum_{i=1}^{n} y_i^2 = 56191.12$
- $\sum_{i=1}^{n} x_i y_i = 56756.15$

All you have to do now is to use these values in the formula:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}} \\ &= \frac{56756.15 - 10 \times 75.52 \times 74.76}{\sqrt{(57330.82 - 10 \times 75.52^2)(56191.12 - 10 \times 74.76^2)}} \\ &= \frac{297.398}{\sqrt{298.116 \times 300.554}} \\ &\approx 0.994. \end{aligned}$$

International Mindedness

Imagine you found a correlation between hours of sleep and academic performance. Does that mean getting more sleep will cause a student to perform better in school? That might be the case in some circumstances, but numerous factors can play a role in school performance. A student living in a poorer area may have to go to work after school and to choose between sleep and studying. A student living in the northern part of Canada may find it hard to sleep in the summer when it can be daylight for over 20 hours a day.

What other factors might affect sleep and school performance? What impact could those factors have on the strength of the correlation?

The broader the span of cultures, socioeconomic strata and environments your data are collected from, the more underlying causes may exist for the relationships that you find.

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Example 2

Matt and Evelyn are measuring their reaction times. They need to respond to the same input as fast as they can and a computer records their response times. The table below summarises the result of 10 such trials. The times are given in milliseconds.

Matt	224	220	242	219	204	239	243	201	188	209
Evelyn	229	219	244	242	222	250	239	201	212	211

a) Find r, the Pearson correlation coefficient.

b) Interpret the result.

a) You can use the appropriate app of a calculator to find the r value:

$$r \approx 0.83456317$$

b) This r value is close to 1, so the association is strong positive. In this sample, a longer reaction time for Matt tends to correspond to a longer reaction time for Evelyn.

Note, that although the association between the reaction times is strong positive, this does not mean causation. It does not mean that Evelyn reacts slower because Matt reacts slower (or the other way around). A more plausible explanation is that some responses have slower or faster reaction times for both of them.

3 section questions

Question 1

Difficulty: ★★☆

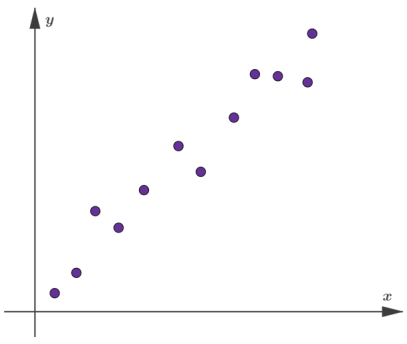
Match the scatter plots, **Figures 1—4**, with the following correlation coefficients:

A: $r = -0.81$,

B: $r = -0.50$,

C: $r = 0.55$

D: $r = 0.96$.



Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Figure 1.

More information for figure 1

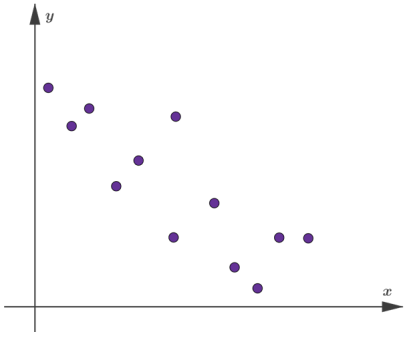


Figure 2.

More information for figure 2

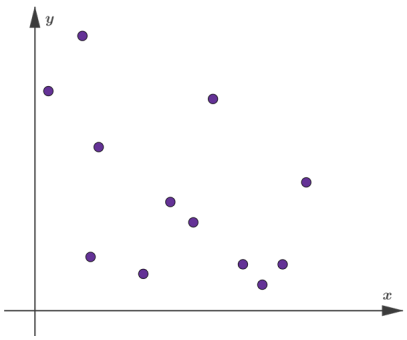


Figure 3.

More information for figure 3

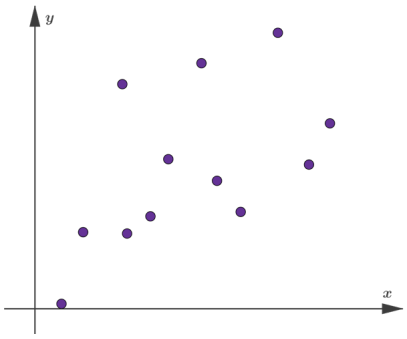


Figure 4.

More information for figure 4

1 A: **Figure 2**, B: **Figure 3**, C: **Figure 4**, D: **Figure 1** ✓

2 A: **Figure 3**, B: **Figure 2**, C: **Figure 4**, D: **Figure 1**

3 A: **Figure 2**, B: **Figure 3**, C: **Figure 1**, D: **Figure 4**

4 A: **Figure 2**, B: **Figure 4**, C: **Figure 3**, D: **Figure 1**

Student view

Explanation

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Figures 1 and **4** have a positive correlation, $r > 0$, with **Figure 1** having a much higher positive correlation as its data points are much closer to the best-fitting line.

Figures 2 and **3** have a negative correlation, $r < 0$, with **Figure 2** having a higher correlation as its data points are closer to the best-fitting line.

Question 2

Difficulty: ★★☆

Alejandro is exploring the connection between the heights (x) in centimetres of his berry bushes and the kilograms of berries (y) he harvested from each. The data he collected is shown in the table below. Find the correlation coefficient for the data.

x (cm)	y (kg)
90	1.1
100	2.8
110	4.3
120	5.9
130	8.1
140	9.9

1 0.998 ✓

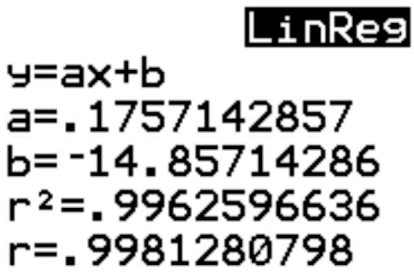
2 0.176

3 0.996

4 -14.9

Explanation

On the calculator, enter the data and use the Linear Regression function. The results are as follows:



The correlation coefficient is the value of r.

Student view

Question 3

★★☆

Overview (/study/ap aa-hl/sid-134-cid-761926/o

René wants to determine if there is a relationship between the number of snacks someone eats each day (x) and the number of times they work out in a week (y). He asks four of his friends and records his data in the table below.

x	y
2	5
4	3
7	6
9	2

Find the correlation coefficient for this data. Give your answer as a decimal, rounded to three significant figures.

-0.352 ✓

Accepted answers
-0.352, -0.352, -.352, r=-0.352, r=-0.352, r=-.352

Explanation
Using your calculator, enter the data and use the linear regression tool to find the correlation coefficient.

$$r = -0.352332\ldots \approx -0.352$$

4. Probability and statistics / 4.4 Linear correlation of bivariate data

Prediction

Section Student... (0/0) Feedback Print (/study/app/math-aa-hl/sid-134-cid-761926/book/predicton-id-25529/print/) Assign

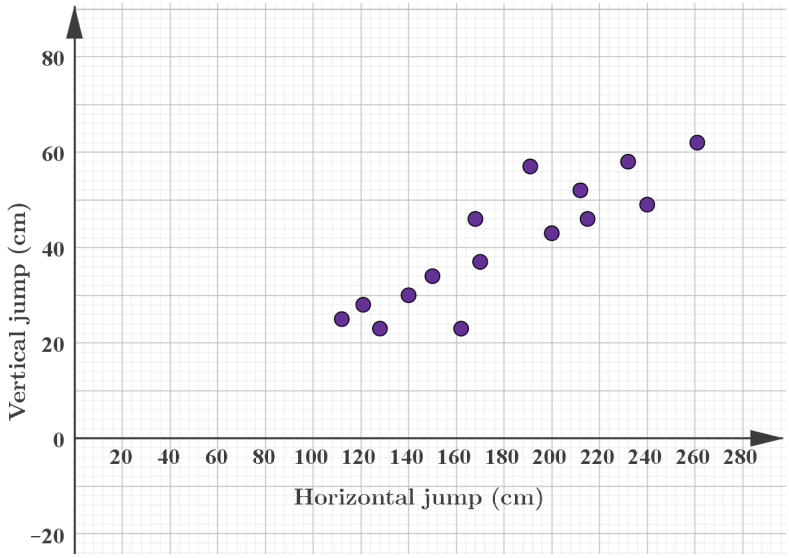
The least-squares regression line

Finding the regression line

Let us return to the scatter diagram for the jump data, a paired data set allowing a comparison between how far and how high a student can jump.

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

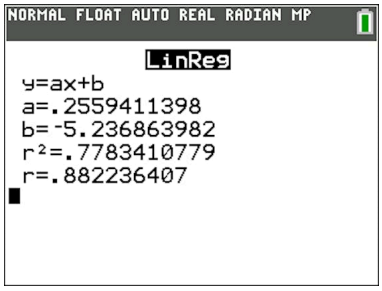


More information

The image is a scatter diagram showing a comparison between vertical and horizontal jump distances for students. The X-axis represents the horizontal jump in centimeters, with labels and values ranging from 0 to approximately 130 cm. The Y-axis represents the vertical jump in centimeters, with labels and values spanning from -20 to approximately 80 cm. Each point on the chart represents a specific student's jump data, showing a general upward trend indicating a possible correlation between larger horizontal and vertical jump distances. The points are scattered, but there is a noticeable positive trend suggesting that as the horizontal jump distance increases, the vertical jump distance also tends to increase.

[Generated by AI]

We know that the line of best fit is a straight line with equation $y = mx + b$ where m is the gradient and b is the y-intercept. Indeed, we already found the values of m and b in the calculator instructions in section 4.4.2 (/study/app/math-aa-hl/sid-134-cid-761926/book/pearsons-r-correlation-coefficient-id-25528/) as we were finding the Pearson correlation coefficient, r. The results are shown below.



The calculator also gives us the least-squares regression line.

More information

The image is a screenshot of a graphing calculator displaying results from a linear regression analysis under the heading 'LinReg'. The text reads:

- 'y=ax+b'
- 'a=0.2559411398'
- 'b=-5.236863982'
- 'r²=0.7783410779'

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

- 'r=0.882236407'.

These values represent the equation of the least-squares regression line, with 'a' as the slope and 'b' as the y-intercept. The coefficients of determination (r^2) and correlation (r) suggest a measure of how well the line fits the data. The calculator settings include 'NORMAL', 'FLOAT', 'AUTO', 'REAL', 'RADIAN', and 'MP' at the top, indicating current modes and settings.

[Generated by AI]

We can see from the figure that the calculator gives us a gradient (here, a rather than m) and a y-intercept b. Since we used **linear regression** to find this information by minimising the square of the differences, the best fit line is also known as the least-squares regression line. From the results shown above, we know the equation of this line is $y = 0.256x - 5.24$, rounded to 3 significant figures.

Be aware

You might see variations of the linear equation $y = mx + b$. Other variations include, but are not limited to, $mx + c$, $ax + b$ and $kx + b$. The key is to remember that the coefficient of x is the slope and the constant is the y -intercept.

Making connections

The term regression is derived from a research study published by Sir Francis Galton, an English geneticist, in the late 19th century. He found that the descendants of very tall people tended to regress back towards an average height. Hence, we have the phrase ‘regression to the mean’. Galton published his results in a paper called ‘Regression towards mediocrity in hereditary stature’.

Interpreting the regression line

In the example above, we can see that the slope is 0.256 and the y-intercept is -5.24, but what do these mean in the context of the problem?

Consider what x and y represent: x is the horizontal jump distance and y is the vertical height of the jump. With this in mind, we can interpret the slope this way:

$$\text{slope} = \frac{0.256 \text{ cm vertical jump}}{1 \text{ cm horizontal jump}}$$

This means that for every 1 cm further a student can jump horizontally, the student can jump approximately 0.256 cm higher. Think about the y-intercept and what it means in this situation: if a student cannot jump horizontally, then they jump -5.24 cm up. Does this make sense? Of course not. This indicates that there must be limitations on the model we just found. So how can you use the least-squares regression line?

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Correlation and/or causation?

When two variables are correlated and there is a strong relationship (often linear) between them, it is often tempting to say that one of the variables *causes* the other variable. For example, the claim has been widely made that smoking *causes* lung cancer.

We must again consider the role of the statistician versus the role of the scientist, the health worker, the psychologist, the economist, the engineer, etc. The role of a statistician is to test for a *statistical relationship* , that is, a correlation. The role of the other professionals is to look for *a reason* for the relationship, which might then lead to a claim that one variable causes the other variable. The reason could be considered a *hidden* variable in the original context. In other words, there may be something else that both variables are related to that causes one of the variables but creates a statistical link connecting all three variables.

For example, a student might prepare a data set that examines pairs of data points that compare coffee consumption with blood pressure. They might see a strong positive linear trend. Does that mean that coffee consumption causes high blood pressure? Is it possible that the data set shows a group of coffee and high-salt-diet people and another group of no-coffee and low-salt people, and possibly their blood pressure has more to do with their salt intake rather than their coffee intake? Or even more difficult, is it possible that people who have high blood pressure tend to enjoy drinking coffee?

Following a two-variable statistical analysis, the job of a statistician is to look for correlation only. They might suggest that further study may yield a causal relationship, but a scatter diagram and a linear trend alone are not enough to conclude causation. At most, the statistical study *can suggest reasons* for a correlation, but they *cannot conclude proof* of causation.

Interpolation

Interpolation is when the model (the equation of the linear regression) is used to predict outcomes whose values are between the minimum and maximum values of the data set studied.

For example, predicting the salary of someone who has worked in a field for 8 years using a data set of workers with 1–15 years of experience is interpolation, and therefore acceptable. Predicting the salary of a someone with 30 years of experience from the same data set is called extrapolation and would not necessarily yield an accurate result.

Important

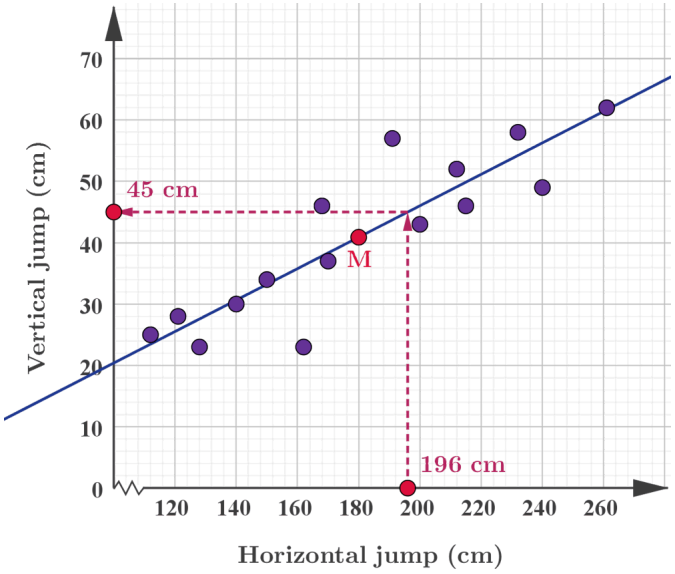
The primary purpose of finding correlations between variables is so we can create models to predict the dependent variable from the independent variable. For this reason, learning how to interpolate values from the model is essential.

Suppose a student had to leave class before they were able to attempt the vertical jump. However, the student's horizontal jump measured 196 cm. What would you predict for the student's vertical jump?

Roughly speaking, we can look at and read from the best fit line on the scatter diagram.

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o



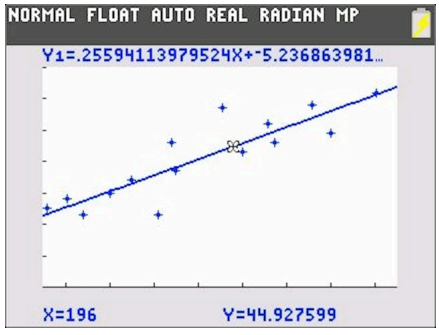
More information

The image shows a scatter plot graph depicting the relationship between vertical and horizontal jumps measured in centimeters. The X-axis represents the horizontal jump ranging from 100 to 260 cm, while the Y-axis represents the vertical jump ranging from 0 to 70 cm. Several data points are plotted, denoting various jump measurements.

A best fit line, shown in blue, runs diagonally across the graph, indicating a positive correlation between the horizontal and vertical jump distances. Two notable data points have annotations. One point at approximately (180, 40) cm is marked with a vertical red dashed line extending to the X-axis indicating a horizontal jump of 196 cm. Another red dashed line extends horizontally from around 50 cm on the Y-axis to align with a data point. This horizontal line is marked as 45 cm. Together, these lines illustrate specific horizontal and vertical measurements on the graph.

[Generated by AI]

More precisely, we can calculate it using the linear regression equation.



More information

The image is a graph depicting a linear regression analysis. The X-axis represents an unspecified variable with a marked point at X=196. The Y-axis represents another variable with a value at Y=44.927599. The graph includes scattered data points around a central linear trend line labeled with the equation Y=1.2559411397524X+5.236863981. The points appear to be evenly distributed around this line, indicating a good fit of the trend line with the data. The linear regression analysis suggests a positive relationship between the X and Y variables.

[Generated by AI]

Student view

In the figure above , we use the graph already drawn (in the previous section) and type 2ND CALC VALUE X = 196 , which will return us the $y$value $44.9\ \text{cm}$) .

Alternatively, we can use the equation and substitute the value $x = 196$:

$$y = 0.256x - 5.24 = 0.256 \times 196 - 5.24 = 44.9\,\text{cm}.$$

That is, if a student can jump $196\ \text{cm}$ horizontally, we expect that they will be able to jump $44.9\ \text{cm}$ vertically.

Now the closer the value of $r$is to $+1$or to -1, the more closely the points will gather around the line. That is, the stronger the correlation, the more reliable the predictions we make using the equation of the line.

In summary, predictions made with the line of best fit are good when

- there is a strong correlation.
- the value used for the prediction is between $x_{\min}$and $x_{\max}$of the data set.

Extrapolation

Extrapolation is when the model (the equation of the linear regression) is used to predict outcomes for a data value that is not between the minimum and maximum of the data set. As stated above, it would be unreliable to predict the salary of someone with 30years of experience from a data set of workers with $1–15$years of experience , but it would be entirely reasonable to predict the salary of someone with 17years of experience from this data set. We must simply be careful to mention that the data set does not extend to 17years of experience. This is also why the y-intercept of the model predicting vertical jump from the horizontal jump represented an impossible value. The minimum $x$in the set is $x = 110$, so clearly $x = 0$is extrapolation far outside of the range.

In summary :

- It can sometimes be reasonable to make predictions using the line of best fit for values that are smaller than $x_{\min}$or larger than $x_{\max}$.
- Data values calculated by extrapolating from the data set should be accompanied by a disclaimer such as: ‘This value was extrapolated from the data set . Use at your own risk.’

Activity

Find the winning times for the women’s 100 m (https://www.olympic.org/athletics/100m-women) at the past several Summer Olympic Games. Using x for year and y for time, find the least-squares regression line. What do the intercepts represent? Were people really that slow 2000 years ago? Did you just discover when teleportation will be invented?

Predicting x from y

It is tempting to try and use the regression line to predict a value of x from a value of y. While it is possible mathematically to substitute a value of y into the equation and solve for x, this will not always be a reliable approximation. We choose to make x the independent variable because we want to use it to predict y. To predict x from y

Overview (/study/app/aa-hl/sid-134-cid-761926/o

, we would need to pick y as the independent variable.

Theory of Knowledge

It has been said that humans are 'pattern-seeking animals' because we identify and attribute causal relationships where they don't actually exist. A key skill in mathematics is the ability to recognise patterns and express them mathematically. However, as a TOK student, you should ask whether or not the expression of the pattern makes it valid. In short, if you can describe a relationship between two variables using mathematics, does that mean the relationship exists? Or does it mean you simply described something?

For example, one can describe a place where money grows on trees, the beaches are pristine, and there is no crime or inequity whatsoever. The description however does not make the existence of this idyllic place true. The same must be considered in mathematics. Statistics can be said to be simply descriptions rather than objective truths. An interesting example is the failure of statisticians to predict that Donald Trump would win the 2016 United States presidential election. How could the mathematics be so wrong? Read this article (https://fivethirtyeight.com/features/the-real-story-of-2016/) written by one of the statisticians whose predictive models failed and discuss the **limitations** of mathematical **application**.

4 section questions

Question 1

Difficulty: ★★☆

The best-fitting line to a data set is given by $y = 5.25x - 2.84$. If the mean for y is $\bar{y} = 8.75$, what is the mean for x?

1 $\bar{x} = 2.21$ ✓

2 $\bar{x} = -2.21$

3 $\bar{x} = 1.13$

4 $\bar{x} = 43.1$

Explanation

If $y = 5.25x - 2.84$, then

$$x = \frac{y + 2.84}{5.25}.$$

Thus, if $\bar{y} = 8.75$, then

$$\bar{x} = \frac{\bar{y} + 2.84}{5.25} = \frac{8.75 + 2.84}{5.25} \approx 2.21.$$

Question 2

Difficulty: ★★☆

Student view

The scores of ten students are collected from their English and their history exams, and are shown below.

Overview (/study/aa-hl/sid-134-cid-761926/o

English	54	59	50	75	82	67	71	73	44	60
History	65	71	63	78	87	66	70	76	58	65

A student sits the English exam and scores 62%. Unfortunately, the student was absent on the day of the history exam.

Use the line of best fit to predict what that student might have scored on the history exam, to the nearest per cent.

1 69% ✓

2 62%

3 65%

4 59%

Explanation

We first enter the data into list 1 and list 2.

We then calculate the line of best fit, which is $y = 0.636x + 29.5$.

We now substitute $x = 62$ to find that $y = 0.636 \times 62 + 29.5 = 68.932 = 69$ (nearest per cent).

Question 3

Difficulty: ★☆☆

The equation of a line of best fit can be used to predict the value of a second variable when:

1 ✓
- The correlation coefficient is close to positive one or close to negative one.
- The value being used is between the minimum and maximum data value studied.

2
- The precise value already occurs in the data set.

3
- The scatter diagram looks more like a cloud.

4
- The second variable is caused by the first variable.
- The second variable is less than 100.

Explanation

When the value being used to predict a second variable is between the minimum and maximum of the data values studied, we say that the data is interpolated. That is, we are interpreting within the boundaries of the data studied. This is more reliable than extrapolating.

Secondly, if the correlation coefficient is close to either 1 or -1, then the points are close to the line, so the values calculated using the line resemble the data set already collected.

Student view

Question 4

Difficulty: ★★☆

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Given the bivariate data in the table below, what is the value of y for the mean of x, i.e. $\bar{x}$, and does this require interpolation or extrapolation?

x	y
1	9
2	12
3	7
4	6
5	3
6	4
7	3
8	4
9	2
10	1
11	2
12	3

1 $y = 4.67$, interpolation ✓

2 $y = 5.04$, interpolation

3 $y = 4.1$, interpolation

4 $y = -4.67$, extrapolation

Explanation

Using our GDC, we obtain the best-fitting line as $y = -0.755245x + 9.57576$. See the figure below. For the data set, $\bar{x} = \frac{1 + 2 + 3 + \ldots + 11 + 12}{12} = 6.5$. Hence, substituting $x = 6.5$ in the equation of the best-fitting line, we obtain, $y = -0.755245 \times 6.5 + 9.57576 \approx 4.67 = \bar{y}$.

There is no $x = 6.5$, but it lies within the data range, so this is interpolation.

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

	=LinRegMx(a[],l
RegEqn	m*x+b
m	-0.755245
b	9.57576
r^2	0.699141
r	-0.836146

More information

4. Probability and statistics / 4.4 Linear correlation of bivariate data

Checklist

Section Student... (0/0) Feedback Print (/study/app/math-aa-hl/sid-134-cid-761926/book/checklist-id-25530/print/) Assign

What you should know

By the end of this subtopic you should be able to:

- recognise when a scatter plot shows data that has a linear trend
- identify whether a scatter plot indicates a correlation that is positive or negative and whether it is strong, moderate or weak
- find the mean point and sketch an estimate of the line of best fit by hand
- find the Pearson product-moment correlation coefficient, r, and the equation of the least-squares regression line with your GDC
- interpret the slope of the least-squares regression line in terms of the data
- use the least-squares regression line to interpolate the value of the dependent variable from a known value of the independent variable
- understand for which values extrapolation might be reliable and the danger of extrapolation in general.

4. Probability and statistics / 4.4 Linear correlation of bivariate data

Investigation

Section Student... (0/0) Feedback Print (/study/app/math-aa-hl/sid-134-cid-761926/book/investigation-id-25531/print/) Assign

In subtopic 4.3 (/study/app/math-aa-hl/sid-134-cid-761926/book/the-big-picture-id-25517/), you saw that when you transform a set of data with multiplication or addition, you can calculate what the new mean, variance and standard deviation are without doing all the work over again. Do you think correlation works in a similar way?

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

For this investigation, start with any set of bivariate data and use your GDC to calculate r. Then transform the data (both coordinates) and recalculate r for the new data. Did r change?

Transform the data several times in a variety of ways and record the results to see if you can find some sort of pattern. What if you only change the dependent variables or just the independent variables? Maybe you can even discover a new theorem.

Activity

Explore how to modify data using a spreadsheet or lists on your calculator to save yourself time as you change the sets of data. Learning to use technology efficiently is essential to becoming a great statistician.

Student view