VISIONARY TUTORING

Year 12 Applications ATAR:
Intro to Bivariate Data

Categorical vs Numerical Data
Two-Way Tables, Scatter Graphs

Name: _____

Bivariate Data – Finding associations between 2 variables

- Categorical Data
- Numerical Data

CATEGORICAL DATA

The two variables are represented by words or labels and is divided into groups

e.g. Football teams, Brands, Colors

The 2 variables in categorical bivariate date can be split into explanatory and response variables

- Explanatory = cause
- Response = effect

Apply – Calc Free

1 A survey investigating whether preferred film type can be explained by the gender of the person asked to give their preference.

9 Whether or not a person has a particular mental disorder and whether they claim to 'hear voices'.

14 Lack of sleep and level of tiredness.

15 Stress levels and heart disease.

Two-way Tables

	Horror	Romance	Comedy	**Total**
Male	30	8	12	
Female	13	67	20	
Total				

- The explanatory variable is converted to percentages

Row/column percentages

	Horror	Romance	Comedy	**Total**
Male				100%
Female				100%
Total	100%	100%	100%	

Identifying an association between the variables;

- Do the percentages vary across the explanatory variables?
- If so, *there appears to be an association between variable 1 and variable 2*
- E.g. There appears to be an association between gender and favourite movie genre, since the percentages vary across the two genders.
  For example, 60% of males enjoy horror, while only 13% of females enjoy horror.

Apply – Calc Free

Q1. Participants at a conference were categorised by district they worked in and main area of interest. The table below shows the number of participants in these categories.

		Main area of interest			
		Technology	Science	Engineering	**Total**
District	Metropolitan	13		4	**25**
	Regional	7	14		
	Total		**22**		**65**

1. Complete the missing values in the table
2. Use the below table to complete the row/column percentages, rounding entries to the nearest whole number.

		Main area of interest			
		Technology	Science	Engineering	
District	Metropolitan				**100%**
	Regional				**100%**
		100%	**100%**	**100%**	

3. Explain whether the percentage table above suggest the presence of an association between district worked in and main area of interest for the participants.

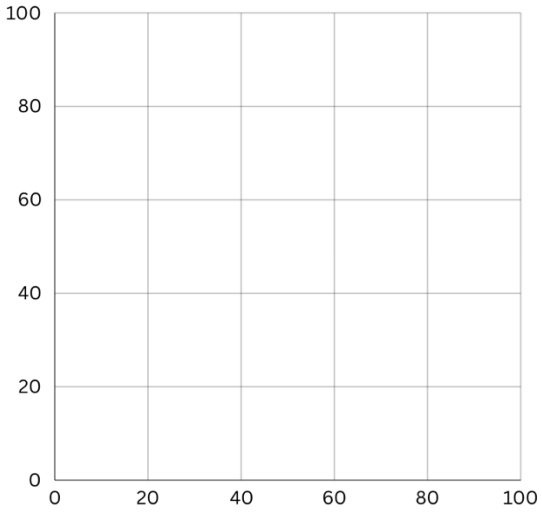
NUMERICAL DATA

The variables are represented by numerical values and numbers
E.g. Height, population, age

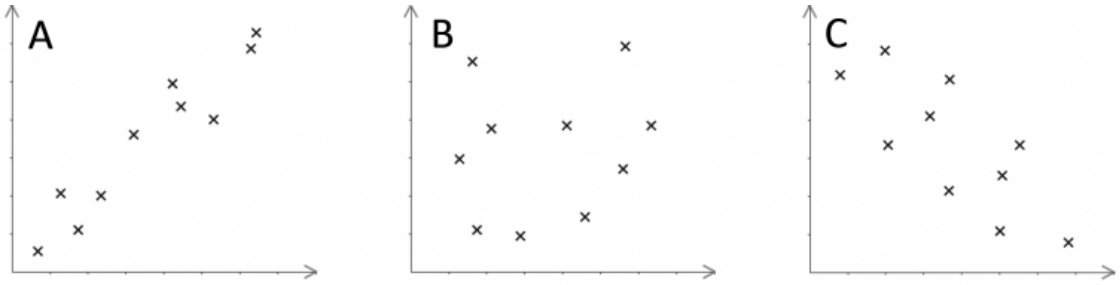
Scatter Graphs

Explanatory variable is plotted on the x-axis, and response variable is plotted on the y-axis.

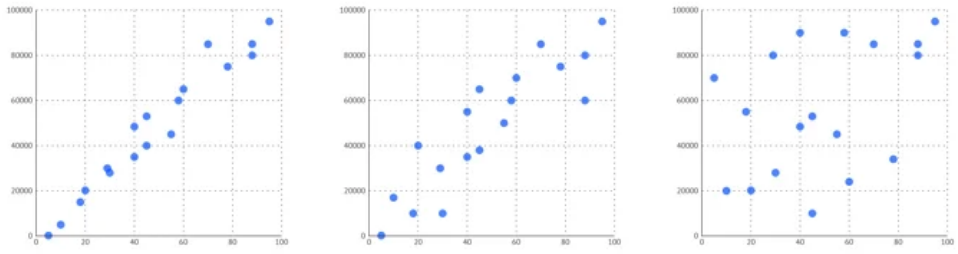
Response variable



Explanatory variable



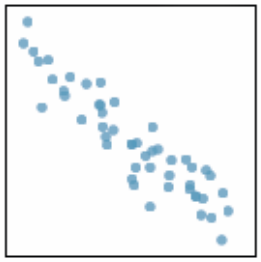
Positive Correlation | No Correlation | Negative Correlation

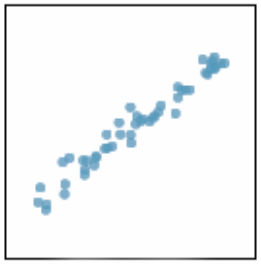


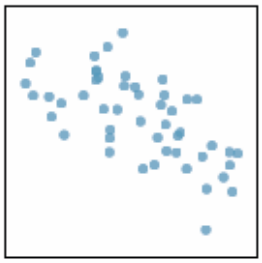
Strong Correlation | Moderate Correlation | Weak/No Correlation

Apply – Calc Assumed

Q2. Label the following scatter graphs as *strong/moderate/weak*, and *positive/negative*.



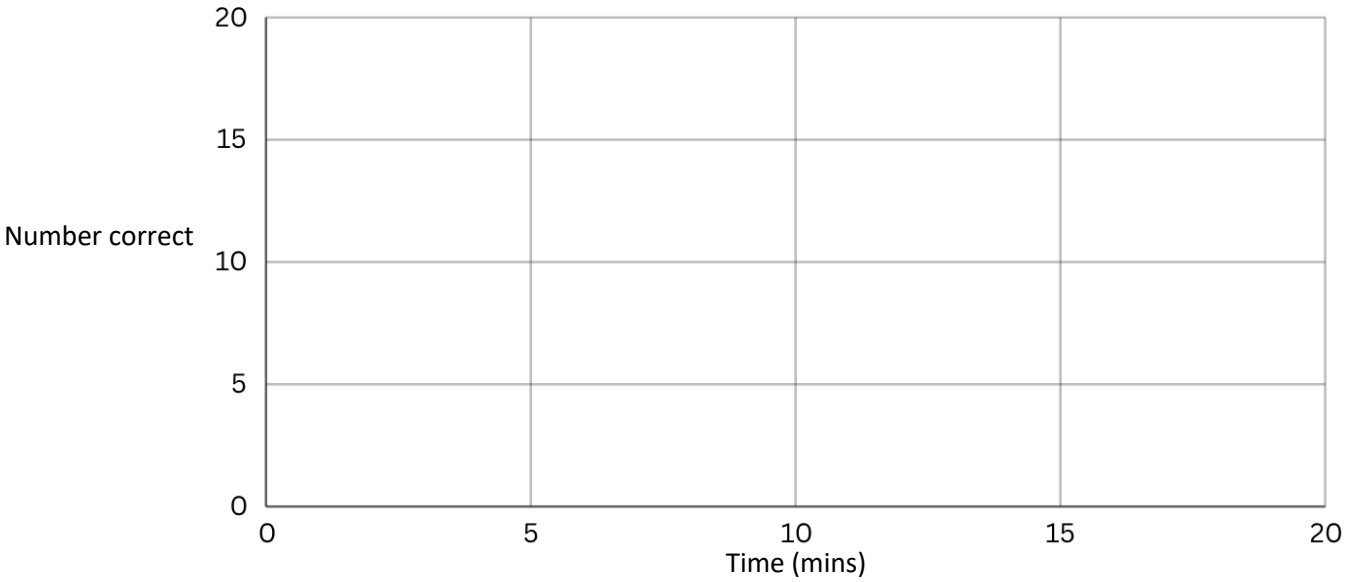




Q3. a) A student recorded the time taken and the number of correct answers made when completing nine multiple choice tests, each with 20 different questions, in the table below.

Time, t minutes	10	9	15	7	4	13	12
Number correct, c	17	14	17	12	10	18	16

Construct a scatter plot of the data on the axes below.



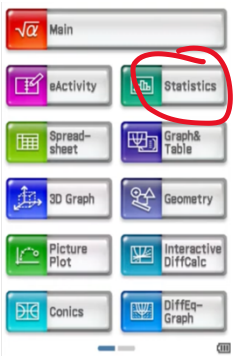
Using the scatter graph, describe the association between the given variables

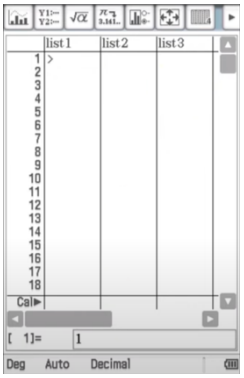
Least Squares Regression Line
aka. Line of best fit, regression line

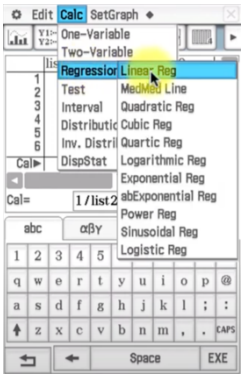
A straight line drawn on a scatter graph that is used to identify the strength + direction of the association between the variables, as well as to make predictions between the two variables

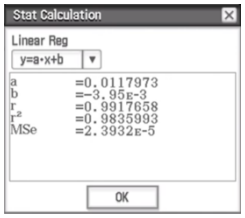
How to find the least squares regression line

- Calc Free – equation will be given in the form $\hat{y} = mx + c$, where $\hat{y}$ represents the predicted value
- Calc Assumed –– ClassPad must be used…









Apply – Calc Assumed

Q3. b) Find and state the equation for the least squares regression line of the following data and plot the line on the previous scatter graph.

Time, t minutes	10	9	15	7	4	13	12
Number correct, c	17	14	17	12	10	18	16

Interpreting the least squares regression line

Correlation coefficient – aka. "r value", and takes a value between -1 and 1

Direction

- If the r value is >1 (positive value), the direction is positive
- If the r value is <1 (negative value), the direction is negative
- May also use gradient of regression line – positive value means positive association, vice versa

Strength

- Closer the r value is to -1 or 1, the stronger the association is

Absolute value	**Strength**
$0 < r < 0.2$	No linear association
$0.3 < r < 0.4$	Weak linear association
$0.5 < r < 0.7$	Moderate linear association
$0.8 < r < 1$	Strong linear association

Coefficient of determination – aka. "r^2", and takes a value between 0 and 1

- Found by squaring the correlation coefficient (r value)
- Will always be positive, since squaring a negative results in a positive number
- r^2 % of variation in the [response variable] can be explained by the variation in the [explanatory variable]
- e.g. Let $r^2 = 0.8$. 80% of variation in y can be explained by the variation in x

Apply – Calc Assumed

Q3. c) Find and interpret the correlation coefficient of the following data

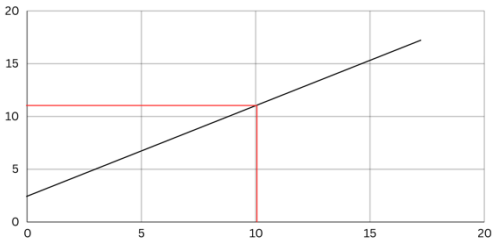
Time, t minutes	10	9	15	7	4	13	12
Number correct, c	17	14	17	12	10	18	16

What percentage of the variation in the number of correct answers can be explained by the variation in the time taken?

Applying the least squares regression line

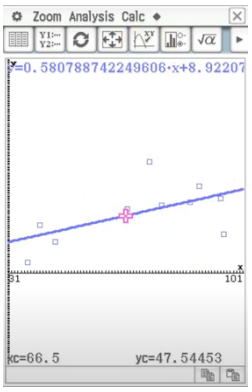
The least squares regression line can be used to make predictions between the two variables.

e.g. estimate the [y value] for an [x value] of 10.



Making estimations

- Substitute value into least squares equation, solving to find missing value. (using ClassPad solve, or algebra)
- Draw construction lines on graph to make an estimation
- Use trace function on ClassPad to make estimation



Apply – Calc Assumed

Q3. d) Using the previous least squares regression line, estimate the number of correct answers for an additional student who took 11 minutes to answer.

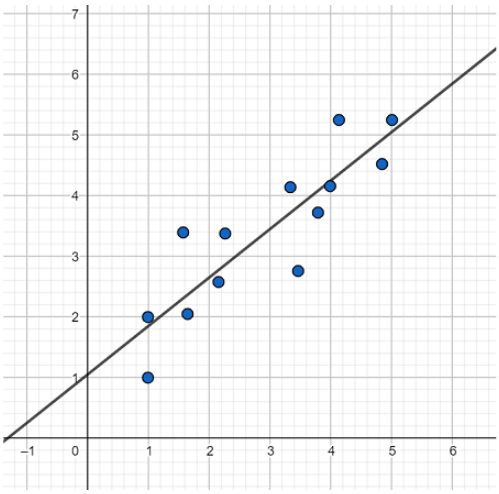
Time, t minutes	10	9	15	7	4	13	12
Number correct, c	17	14	17	12	10	18	16

Evaluating the reliability of an estimation

- Is the correlation coefficient close to -1 or 1? Is there a strong correlation?
- Is the estimate interpolation? Is it within known data points?
- How many data points are provided?

Residuals

The differences between the predicted values (ŷ) and the actual values (y) of the response variable.



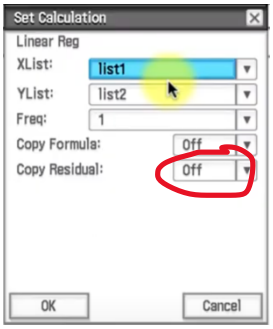
To find the residual, the following formula is used:

Residual = y - ŷ

e.g. (5,6) has a residual of 1

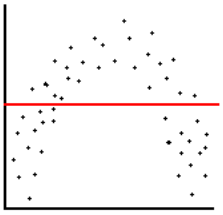
OR

Use ClassPad

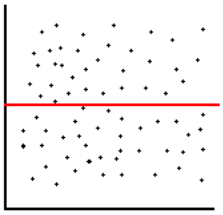


By plotting residuals on a scatter graph, you can identify if a linear model is suitable for the relationship between the 2 variables.

- If a pattern is present in the residual plot, a linear model is not suitable.
- If the residuals appear randomly scattered, a linear model is suitable.



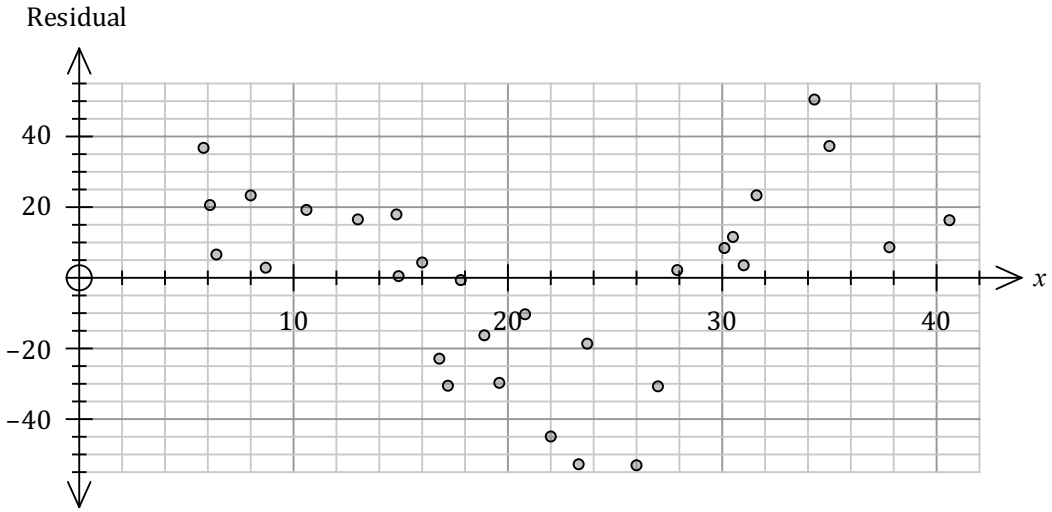
Not suitable



Suitable

Apply – Calc Free

The linear model fitted to a data set has equation $\hat{y} = 15.65x - 52.5$. The residual plot for the linear model is shown below.



The residual for the data point $(39, 596)$ is not shown. Determine the residual for this point and add it to the residual plot.

Use the residual plot to assess the appropriateness of fitting a linear model to the data.

Other important notes

Common question: Does the information gathered indicate that the explanatory variable causes the change in the response variable?

How to answer
No. Just because there appears to be a mathematical correlation, you cannot assume there is a cause-and-effect relationship between the variables. There may be a lurking variable present.

Explanation
Correlation ≠ Causation. There may be another variable causing the visible association.
E.g. Intelligence vs Height