Investigating the association between a numerical and a categorical variable

Tuesday, 26 February 2019 5:55 PM

★ By the end of the lesson I would hope that you have an understanding and be able to apply to questions the following concepts:

- Know what it means by the term **associated**
- Know how to use parallel box plots to identify and describe associations
- Use parallel dot plots and back to back stem plots to identify and describe associations
- Know how to write a short, concise report about associations
- Remember how to construct and interpret a box plot.

RECAP:

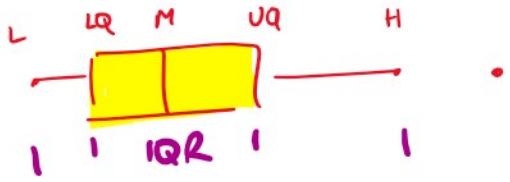
In the previous lessons we have learnt about response and explanatory variables and investigating associations between two categorical variables. In this lesson we are going to look at how we can investigate relationships between numerical and categorical variables.

Numerical variable: A variable which takes number values.
Categorical variable: A variable which groups into categories.

Recap about box plots

A box plot is a way to grouping data into quarters.
It is also a way to express the following information:

- Lowest data item
- Lower Quartile
- Median
- Upper Quartile
- Highest data item



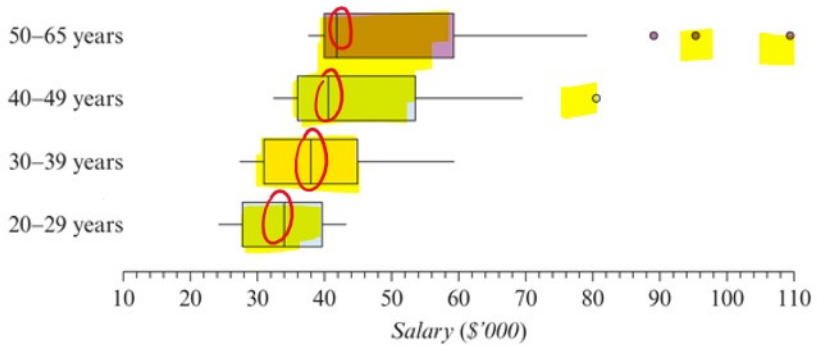
The above is also known as the 5-number summary.

Box plots come into their own when two or more are drawn together.
It lets us compare the data.

Example
Extracted from the Cambridge Further Mathematics Textbook with permission

Questions to ask:

- Which is the categorical variable?
- Which is the numerical variable?
- Is there anything about the scales which might trick us?
- What do the coloured in dots mean?
- Which is the explanatory variable and which is the response?
- Which data points do we use to compare and decide if there is an association?



Comparing medians

We might compare the medians to decide if there is an association.

The parallel box plots show that median salaries and age group are associated because median salaries increase with age group. For example, the median salary increased from $34000 for 20–29 year-olds to $42000 for 50–65 year-olds.

increase with age group. For example, the median salary increased from $34000 for 20–29 year-olds to $42000 for 50–65 year-olds.

Comparing IQR's and/or ranges

We might compare the Interquartile ranges

From the parallel box plots we can see that the spread of salaries is associated with age group. For example, the IQR increased from around $12000 for 20–29 year-olds to around $20000 for 50–65 year-olds.

Comparing shapes

We might compare the shapes.

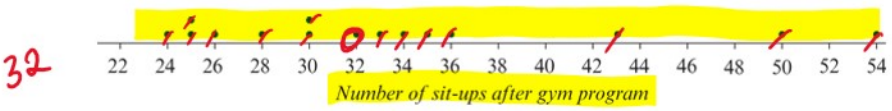
From the parallel box plots we can see that the shape of the distribution of salaries is associated with age group because of the distribution, which is symmetric for 20–29 year-olds, and becomes progressively more positively skewed as age increases. Outliers also begin to appear.

Using dot plots to compare categorical and numerical data.

We can use the raw data provided in a dot plot to compare and decide if numerical and categorical data is related!

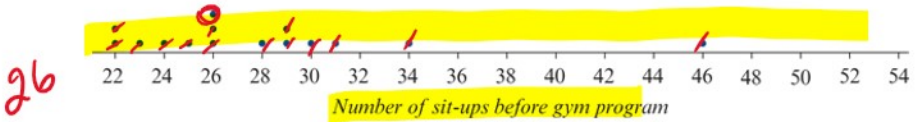
Example
Extracted from the Cambridge Further Mathematics Textbook with permission

The parallel dot plots below display the distribution of the number of sit-ups performed by 15 people before and after they had completed a gym program.



Number of sit-ups after gym program

Do the parallel dot plots support the contention that the number of sit-ups performed is associated with completing the gym program? Write a brief explanation that compares medians.



Number of sit-ups before gym program

As the question has given you the item of data it wishes you to compare (median!), then we need to find the medians, and then write a brief report about our findings.

Report from the data:
Yes; the median number of sit-ups performed after attending the gym program (M=32) is considerably higher than the number of sit-ups performed before attending the gym program (M=26). This indicates that the number of sit-ups performed is associated with completing the gym program.

Using back to back stem plots to compare categorical and numerical data.

We can use the raw data provided in a back to back stem plot to compare and decide if numerical and categorical data is related!

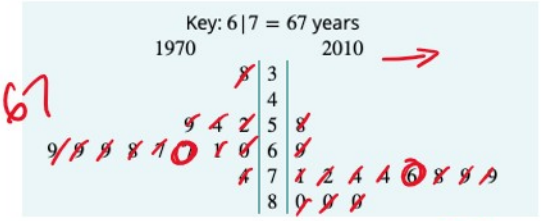
Example
Extracted from the Cambridge Further Mathematics Textbook with permission

The back-to-back stem plot below displays the distribution of life expectancy (in years) for 13 countries in 2010 and 1970.

Do the back-to-back stem plots support the contention that life expectancy is increasing over time? Write a brief explanation based on your comparisons of the two medians.

2010 and 1970.

Do the back-to-back stem plots support the contention that life expectancy is increasing over time? Write a brief explanation based on your comparisons of the two medians.



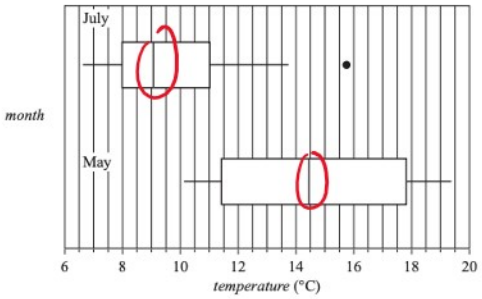
Again, the question has asked you to compare medians!

So, find the two medians and then write a concise report.

Report from the data:
Yes: the median life expectancy in 2010 (M=76 years) is considerably higher than the median life expectancy in 1970 (M=67 years. This indicates that life expectancy is increasing over time.

VCAA Exam Question on this concept
2016 Paper 2

b. The boxplots below display the distribution of maximum daily *temperature* for the months of May and July.



iii. Using the information from the boxplots, explain why the maximum daily *temperature* is associated with the *month* of the year. Quote the values of appropriate statistics in your response. 1 mark

_____

_____

_____

_____



VCAA Exam Question on this concept
2016 Paper 1

Question 8

Parallel boxplots would be an appropriate graphical tool to investigate the association between the monthly median rainfall, in millimetres, and the

~~A.~~ monthly median wind speed, in kilometres per hour.
~~B.~~ monthly median temperature, in degrees Celsius.
C. month of the year (January, February, March, etc.).
~~D.~~ monthly sunshine time, in hours.
~~E.~~ annual rainfall, in millimetres.