Name of skill: Mean, mode, median

Category of skill: Statistical

Example of skill:

For the following set of numbers calculate the mean, median and mode

4	9	37	11	11
15	20	3	9	5
10	11	12	6	16

Mean = ...

Median = ...

Mode = ...

Improvements/alternatives:

The mean, median and mode give a useful summary value for a set of data but give no information about the spread of values around the "average" figure. As such, this summary value can be misleading and give an untrue picture of reality. The spread, or deviation, from a central value can be measured giving a fairer picture about the set of data.

Measure	Method	Evaluation
Mean	Values are added together and then the total is divided by the number of values in the data set.	(+) All data is considered and the results can be used for further analysis. (-) Can be misleading if using a small data set or there are very high/low values which can distort the mean.
Median	The central value when the values are ranked in order (if an uneven number of values, the midpoint between placed values is used).	(+) Not affected by extreme values. (-) Cannot be used for further mathematical processing, but best used in reference to interquartile range.
Mode	Most frequently occurring value in a data set.	(+) Quick to calculate and not affected by extreme values. (-) Can only be used when individual values are known and cannot be used for further mathematical processing.

Justify (why use this technique?):

- To summarise a large amount of data into a single value.
- Indicates that there is some variability around this single value within the original data.
- Average is a commonly used term that is easily understood by most.

How does this improve my geographical understanding?

- It allows us to superficially accept/reject our hypothesis.
- Explaining this and relating to theory will solidify this understanding.
- Further investigation could be undertaken to explain.
- Explaining anomalies will lead to deeper geographical understanding.

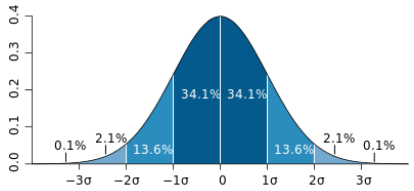
Name of skill: Range, interquartile range & standard deviation

Category of skill: Statistical

What is normal distribution?

50% of values less than the mean and 50% greater than the mean.

We can then complete a standard deviation which shows how spread out the numbers are.



We can work out how far away the values are from the mean.

Justify (why use this technique?):

The mean, median and mode give a useful summary value for a set of data but give no information about the spread of values around the "average" figure. As such, this summary value can be misleading and give an untrue picture of reality. The spread, or deviation, from a central value can be measured giving a fairer picture about the set of data.

Because it measures the spread of data around the mean – this is useful because it shows whether there is consistency rather than lots of anomalies.

A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data is spread out over a large range of values.

Measure	Method	Evaluation
Range	- Difference between the highest and lowest value. - Often used for things like describing climate figures.	(+) Easy and quick to calculate. (-) Only considers extreme values and does not make any reference to other values.
Interquartile range	- Difference between the 25th and 75th percentiles. - A higher interquartile range means the spread of the values around the median is greater.	(+) Fairly simple to calculate. (+) Represents the spread of the middle 50% of values, so more representative of entire data set. (+)Extreme values are not considered. (-) Not all data is considered.
Standard deviation	- Indicates the degree of clustering of each data values about the mean. - Calculated by measuring the deviation of each value from the mean. - When standard deviation is low data is clustered around the mean. When it is high the data set is widely spaced, with some much higher or lower figures.	(+) Measures spread of data around a central value, as includes all data in the set. (+) Allows comparisons of the distribution of the values. (+) Results can be used for further analysis. (-) Fairly complicated to calculate as lots of steps to follow.

How does this improve my geographical understanding?

- Allows us to see whether our data is clustered around the mean - this will then help us establish whether our data is reliable and ultimately whether our conclusions are representative.
- Standard deviation is commonly used to measure confidence in statistical conclusions - we can state with confidence whether we can accept or reject our hypotheses.

Name of skill: Spearman's rank correlation co-efficient

Category of skill: Statistical

Advantages:

- Shows if relationship between two data sets are statistically significant (e.g. 95% confidence level).
- It can confirm or reject a hypothesis allowing for further areas of investigation.
- It summarises large amounts of raw data into one piece of more manageable data (i.e. Rs).
- It reduces the impact of anomalous results which could have led to invalid conclusions.

Disadvantages:

- Large sample sizes can detect small significances and magnify them.
- Dealing with large amounts of data could result in human error if completed by hand. This could impact on the interpretation of these results.
- Ethics in statistics- is it sensible to accept/reject your research hypothesis based on one number (Rs)?
- Use of ranked data means raw data is lost.

How does this improve my geographical understanding?

If you accepted the null hypothesis (statistically no relationship between data sets), why not? Trying to explain will improve geographical understanding. If you accepted the alternative hypotheses (statistically there is a relationship between data sets), why? Explaining/justifying improves geographical understanding.

Describe how to construct…

1. You must first construct a scatter graph to see if there is a visual relationship between the data. If your graph indicates no correlation, the spearman's rank test is almost certain to confirm this, similarly, if the graph indicates a positive or negative relationship.
2. State the null hypothesis (H_0) and the alternative hypothesis (H_1).
3. In the first column record each site/location/person where the data comes from.
4. Complete the second column by recording the first set of data that you would like to test, e.g. distance downstream or number of years lived in a place.
5. Complete the **R1** column by ranking the first set of data from the highest (1) to the lowest.
6. Complete the forth column by recording the second set of data that you found.
7. Work out **R2** by ranking the second data set from the highest (1) to the lowest.
8. **Note**: If you have 2 numbers the same, e.g. site 3 and site 4 both had a width of 6m you would need to rank them equally. If they are the 7th and 8th widest sites you would give both sites a rank of 7.5 as this is half way between rank 7 and rank 8. Your next smallest stream should then be ranked 9. Make sure the last rank number you give is the same as the number of sites you have. If it isn't you've gone wrong!
9. Calculate **d** by subtracting the second rank from the first rank to give the difference.
10. Now square the differences (i.e. multiply each figure by itself) to give $\mathbf{d^2}$ – this is done to remove any negative values.
11. Add up all the figures in the final column to get $\mathbf{\Sigma d^2}$
12. Now substitute this value into the formula. You should end up with a number between -1 and +1. If it is a negative number it shows there is a negative correlation and if there is a positive number it indicates a positive correlation – but we still don't know if the relationship between the two variables is significant.

$$1 - \left(\frac{6\Sigma d^2}{n\,(n^2-1)}\right)$$

Remember: closer to +1 = strong positive correlation, closer to -1 = strong negative. Result needs to be higher than critical value.

Justify (why use this technique?):

The most commonly used test for measuring correlation, it can only be used to examine a linear relationship and it is a nonparametric test (no assumptions are made about the data being normally distributed). If the answer is yes to the following questions, a spearman's test can be performed to analyse your data.

- Are you investigating a relationship between two variables?
- Do you have 10 or more pairs of data (it is not reliable with fewer)?
- Do you have fewer than 30 pairs of data (it becomes arithmetically difficult with more)?
- Are you assuming the data is not normally distributed?

How does this improve my geographical understanding?

If you accepted the null hypothesis (statistically no relationship between data sets) why not? Trying to explain will improve geographical understanding. If you accepted the alternative hypotheses (statistically there is a relationship between data sets) why? Explaining/justifying improves geographical understanding.

Name of skill: Chi-squared
Category of skill: Statistical

Advantages:

- Useful to compare the median or means of TWO sets of data.
- Comparative test supports own ideas on geographical significance of theory.

Disadvantages:

- Data must be ordinal, in order, for the test to be applied.
- Limited number of data sets can be analysed.

How does this improve my geographical understanding?

If you accepted the null hypothesis (statistically no difference between mean/median of the data sets) why not? Trying to explain will improve geographical understanding.
If you reject the null hypotheses (statistically there is a difference between the mean/median) why? Explaining/justifying improves geographical understanding.

Describe how to construct…

1. The first task is to generate a **null hypothesis (H_o):**
2. Create a table that contains all of your variables in the columns. The expected data (e) is simply taken as being the mean. It is calculated by adding up all of the observed data (o) and then dividing by the number of categories. This gives an expected frequency for each category.

	Variable a	Variable b
Observed (o)		
Expected (e)		
Deviation (d) = o-e		
Deviation ^{2} = d^2		
d^2/e		
$x^2 = \sum d^2/e$		

3. Calculate **degrees of freedom**:

- **df = n-1** (where **n** is the no. of categories)

Use a *critical values* table to work out the *significance* of your result.

Significance tells us *how confidently* we can disprove the null hypothesis

4. **Compare this to the critical value**. For example, for 4 degrees of freedom the critical value for 0.05 level of significance is 9.49. In order to reject the null hypothesis your chi squared score must be greater than the critical value at the 0.05 level of significance.

Justify (why use this technique?):

Chi-squared is used to examine differences between what you actually find in your study and what you expected to find. Look at the list of questions below. If the answer is yes to each question, a chi-squared test is appropriate:

- Are you trying to see if there is a difference between what you have found and what would be found in a random pattern?
- Is the data gathered organised into a set of categories?
- In each category, is the data displayed as frequencies (not percentages)?
- Does the total amount of data collected (observed data) add up to more than 20?
- Does the expected data for each category exceed four?

Summarising statistical tests.

Type of test	Why would you use this test?	What type of data is needed?	Draw a copy of the table you would use…	Do you rank separately, together or not at all?	Do you rank from high to low or not at all?	Does your value need to be above or below significance level to be accepted?
Spearman's rank						
Chi-squared test						