OCR Geography A-Level

5.1: Quantitative Data Analysis - Statistical Tests

Extra Notes

Statistical Analysis and Tests

- **Standard Deviation**
- **Location Quotient**

- **Mann-Whitney U**
- **Chi-squared**
- **Student's T-test**
- **Spearman's Rank**

- **Measures of Central Tendency**
- **Measures of Dispersion**

Data Analysis Methods: Statistical Analysis

Standard Deviation

This shows by how much most piece of data vary from the mean.

- Find the **mean** of the data.
- Calculate, in a separate column, how **each piece of data differs** from the mean.
- **Square** this value.
- Use this equation:

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set,
N = number of values in the data set.

For the example below, the **mean** is **41.1** (3sf)

Data value	Variance from the mean (x - $\overline{x}$)	Variance from the mean squared (x - $\overline{x}$)^2
13	28.1	789.61
25	16.1	259.21
79	37.9	1436.41
82	40.9	1672.81
1	40.1	1608.01
45	3.9	15.21
49	7.9	62.41
45	3.9	15.21

67	25.9	670.81
45	3.9	15.21
1	40.1	1608.01
Sum: 8152.91 **Standard Deviation =** $\sqrt{\frac{8152.91}{11}}$ **= 27.2 (3sf)**		

Variance:

This shows how far each piece of data varies from the average. It is simply equal to the **square of the standard deviation**.

Location quotient:

The location quotient (LQ) is used to determine the **spatial distribution** (the extent of clustering/dispersal) of a phenomenon in a **subset** of data compared to the **total** data, for instance the concentration of an industry in a **region** compared to the **nation**. They are often used in demography, economics and any type of locational analysis.

1. Find the **proportion** of subset and the **total** with the phenomenon observed
2. **Divide** the **proportion** of the **subset** by the **proportion** of the **total**

E.g. Ethnic diversity- Proportion of people who are White British in England's regions

Region	**White British Population**	**Total population**	**Proportion**
South West	510800	536000	95.3%
England	42279236	53010000	79.8%

$\frac{95.3}{79.8}$ **= 1.19** **LQ = 1.19**

Interpretation of location quotient results:

If the LQ is **greater than 1**, this indicates a **high spatial concentration** for that subset compared to the total set.

If LQ = 1, the share of the total is **equal** for the subset and the total set

If the LQ is **less than 1**, this indicates a **low spatial concentration** for that subset compared to the total set.

Mann-Whitney U

Mann-Whitney U looks at the medians of two sets of data and decides whether there is a **significant difference** between the two.

It can be used on data that has the following characteristics:

- The 2 samples are independent
- The data is ordinal- it can be ranked
- There are at least 6 pairs of data
- It does not require a normal distribution
- It does not require there to be the same number of data sets

Method

1. Label one data set 'sample A' and the other 'sample B'

Sample A: 22, 18, 25, 33, 31, 28, 19, 24, 29 Na (number of data points in A) = 9
Sample B: 26, 18, 30, 16, 35, 21, 31, 17, 18, 27 Nb (number of data points in B) = 10

2. Rank all of the data points in sample A and Sample B as one set (order the data in each sample for ease)

A	**18**	**19**	**22**	**24**	**25**	**28**	**29**	**31**	**33**		
Rank(R)	4	6	8	9	10	13	14	16.5	18		ΣRa= 98.5
B	**16**	**17**	**18**	**18**	**21**	**26**	**27**	**30**	**31**	**35**	
Rank(R)	1	2	4	4	7	11	12	15	16.5	19	ΣRb = 91.5

Where ranks are tied, add up the corresponding ranks, divide by the number of tied ranks and give this rank to all the tied ranks.
E.g. **18, 18 and 18** are tied across ranks 3,4 and 5
(3+4+5) ÷ 3 = 4 so all the **18s** get a rank of 4. The next number in the ranking (19) gets a rank of 6 as 3, 4 and 5 have been used by the 18s.

E.g. **31 and 31** are tied across ranks 16 and 17.
(16 + 17) ÷ 2 = 16.5 so both **31s** get a rank of 16.5.

3. Sum up the ranks of sample A and sample B

4. Calculate the U values using the formula:

$$U_a = n_a n_b + \frac{n_a(n_a+1)}{2} - \sum R_a$$

$$Ua = (9 \times 10) + \frac{9(9+1)}{2} - 98.5$$

and

$$U_b = n_a n_b + \frac{n_b(n_b+1)}{2} - \sum R_b$$

$$Ub = (9 \times 10) + \frac{10(10+1)}{2} - 91.5$$

Ua = 36.5
Ub = 53.5

5. Select the smaller of the two U values

Smaller U value is Ua = 36.5

6. Look up the critical values in the table at the given level of significance.

Level of significance: 5% (P = 0.05)

Size of the largest sample (n_2)

Size of the smallest sample (n_1)	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23
6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30
7			8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38
8				13	15	17	19	22	24	26	29	31	34	36	38	41	43	45
9					17	20	23	26	28	31	34	37	39	42	45	48	50	53
10						23	26	29	33	36	39	42	45	48	52	55	58	61
11							30	33	37	40	44	47	51	55	58	62	65	69
12								37	41	45	49	53	57	61	65	69	73	77
13									45	50	54	59	63	67	72	76	80	85
14										55	59	64	67	74	78	83	88	93
15											64	70	75	80	85	90	96	101
16												75	81	86	92	98	103	109
17													87	93	99	105	111	117
18														99	106	112	119	125
19															113	119	126	133
20																127	134	141

Mann-Whitney U is an exception to the 'MRSA' rule for these statistical tests.

More than
Reject
Smaller than
Accept

If the smaller U is smaller than or equal to the critical value, reject the null hypothesis. There is a significant difference between the two data sets.
If the smaller U value is greater than the critical value, accept the null hypothesis. There is no significant difference between the two data sets.

E.g. As 36.5 is greater than 20, the null hypothesis is accepted. There is no significant difference between the two data sets.

Mann-Whitney U tests can be used to check the statistical significance in choropleth maps, line graphs and scatter graphs.

Chi-Squared Test

The Chi-Squared test looks at the relationship between a **set of data of interest** and a **theoretical/expected set of data** to decide whether the difference between the two is **significantly different**. It is used to see how closely the data collected or observed by the researcher fits with the widely accepted findings. This test only checks to see if there is an association between two sets of data, not what the nature of the relationship might be between those sets, nor the strength of any relationship.

It can be used on data which has the following characteristics:

- The data must be in the form of **frequencies** counted in a number of groups (% cannot be used).
- The total number of observations must be **> 20**.
- The observations must be **independent** (i.e. one observation must not influence another).
- The expected frequency in any one category must not normally be **> 5**.

Method

- State the **hypothesis** being tested – there is a significant difference between sample groups. It is convention to give a **null hypothesis**. For example, **there is no significant difference between the samples**.
- **Tabulate** the data as shown in the example below. The data being tested for significance is the '**observed**' frequency and the column headed '**O**'
- Calculate the '**expected**' number of frequencies that you would expect to find in the column headed 'E'.

- Calculate the statistic using the formula
- Calculate the degrees of freedom.

$$\Sigma = \frac{(O-E)^2}{E}$$

Degrees of freedom: number of rows - 1

Beach site	O Number of pebbles > 5 cm long	E Mean number of pebbles > 5 cm long	(O–E)	$(O–E)^2$	$(O–E)^2/E$
1	40	18	22	484	20.89
2	15	18	3	9	0.5
3	5	18	13	169	9.39
4	12	18	6	36	2
					Σ 38.78

- Compare the **calculated** figure with the **critical values** in the significance tables using the **appropriate degrees of freedom**. Read off the probability that the data frequencies you are testing could have occurred by chance.

If the calculated value **exceeds** the tabulated critical value for the correct number of degrees of freedom at the given confidence level (usually **95%**), then **reject** the null hypothesis. This means that the it can be stated with **99% confidence that there is a statistically significant difference** in the data sets, and this difference is not down to chance.

If the calculated Chi-Squared value is smaller than the critical value, **accept** the null hypothesis.

An easy way of remembering this: MRSA **M**ore than, **R**eject, **S**maller than, **A**ccept

Chi-squared tests are appropriate to use for bar charts and histograms.

T-test

The student's t-test looks at the **means of two sets of data** and decides whether there is a **significant difference between the two**. It looks at the **degree of overlap** between the two samples. It applies to data that is measured on an interval or ratio scale and for data that is normally distributed around the mean.

The **null hypothesis** is that the two data sets are the same (there is **no significant difference** between them). The alternate hypothesis is that there **is** a significant difference between the two data sets.

Method

1. Calculate the **mean** and **standard deviation** for the two sets
2. Plug the values into this formula:

X̄1 = Mean of sample 1
X̄2 = Mean of sample 2
S1= Standard deviation of sample 1
S2 = Standard deviation of sample 2
N1= Number of subjects in sample 1
N2 = Number of subjects in sample 2

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

To see if your t value is **significant** you will need to **calculate the degrees of freedom** and compare your **calculated t value** to the **appropriate critical value**.

These critical values give **95% confidence**. This means that if your calculated t value is the **same or higher than the critical value**, you can be 95% confident that you have a significant difference between your two sets of data.

Degrees of freedom = $n_1 + n_2 - 2$
If calculated t ≧ critical t you **reject** your null hypothesis and accept your alternative hypothesis.
If calculated t **<** critical t you **accept** your null hypothesis and reject your alternative hypothesis.

An easy way of remembering this: MRSA
More than
Reject
Smaller than
Accept

Spearman's Rank

Spearman's Rank tests the **relationship (correlation) between two sets of data**. For example, it could test the correlation between age of respondents and the score for their perception of the city centre, or the sediment size along a coast and the rate of erosion there. Completing Spearman's Rank is best in a **table,** as shown below with a series of steps to follow:

1st set of data	r_1 - ranks for 1st set of data	2nd set of data	r_2 - ranks for 2nd set of data	d - difference between ranks	d^2 - difference squared

1. List a set of data (e.g age of respondent) in the first column. Then **rank** each piece of data relative to each other in the second column - for example, the youngest person will rank 1, the second youngest is 2, etc.

2. List your second set of data and **rank** each piece (similar to your first set) in the following columns. If there is more than one respondent with the same answer (for example the same score or same age) then you may rank them consecutively in any order. Ensure that you do not skip any rank; as a check, ensure that your lowest/ worst rank number is the same as your sample size (e.g. 20th is the last rank, and there are 20 people in your sample)

3. Calculate the **difference between the two ranks** - along one row, take the second rank from the first rank ($r_2 - r_1$).

4. **Square** this difference and record the value.

5. Repeat steps 3 and 4 for each row. Add up all values in the final column.

6. Complete these two word equations with your own values, remembering to calculate the brackets first:

   (6 x the sum of the final column) ÷ (n x n x (n - 1))

   Spearman's Rank = 1 - (the value you calculated above)

   As a check, your value must be between 1 and -1. Alternatively (if you're more maths-y!) the actual equation is :

$$R = 1 - \frac{6 \Sigma d^2}{n(n-1)}$$

7. To finish, you must describe the **correlation** between your data.
   - If the value (ignoring the sign) you calculated is **greater than 0.5**, then your data has a **strong correlation**. Or if the value you calculated is **smaller than 0.5**, then your data has a **weak correlation**.
   - If your Spearman's Rank is **positive**, then your correlation is **positive**. A **negative** correlation will cause a **negative** Spearman's Rank.

Spearman's Rank tests work well when testing the significance of trends on scatter graphs.

Measures of Central Tendency

'Measures of central tendency' refers to a group of **statistical tests**. These statistical tests describe data distribution in relation to the **'middle'** value to indicate the **concentration** of the values in the **central part** of the **distribution of frequencies of the whole data**.

The numbers below will be used as an example for each measure of central tendency.
13 25 79 82 1 45 49 45 67 45 1

Mean:
The mean is calculated by **adding all the data** and dividing by the **number of data items**. For example, using the numbers above, the sum would be 452 and there 11 numbers, so the mean would equal 41.1 to 3sf.

Mode:
The **most appearing number**. In the example above, the mode is 45.

Median:
The median is the **midpoint** value. The data needs to be ranked first from lowest to highest value.
1 1 13 25 45 45 45 49 67 79 82

- When there is an **odd number** of data items, the **median is a whole number**. As in the example above, there are 11 data items, so the median is 45.
- When there is an **even number** of data items, the median lies across the **two items** at the midpoint. The median is therefore an **average** (mean) of the **two middle items**.

Measures of Dispersion

'Measures of dispersion' refers to a group of statistical tests which describe data distribution.

Range:
The range describes the **spread of the data**. Simply, **subtract the highest number from the lowest number**. In the example above, the range would be: 82-1 = 81

Interquartile range (IQR):
The interquartile range shows where the middle 50% of the data lie. Anomalies should be ignored in this calculation.

- Find the **median** using the method above. (45)
- Find the **lower quartile** by calculating the **median of the lower half of the data**. (13)
- Find the **upper quartile** by finding the **median of the upper half of the data**. (67)
- The **difference** between the lower and upper quartiles is the IQR. (67-13 = 54)