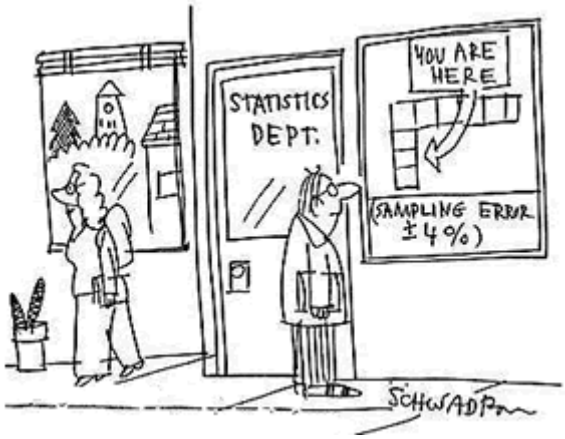
Statistics

Mr Booth's guide for AS Geographers



Contents:

Statistics – The basics!

Levels of Data

- **Nominal:** simplest level – categories and frequencies
  Number of male / female participants
- **Ordinal:** ordered scale of data – no indication as to interval between categories
  Quality of water – very poor, poor, clean, very clean;
  Weather – freezing, cold, mild, warm, hot
- **Interval:** scale with defined intervals
  Temperature in °C Date in day / month / year
- **Ratio level:** continuous data, with defined zero point and no negative values
  Weight in gram Height in metres

N O I R

Mean, Median, Mode and Inter-Quartile Range

• **Mean** - This is sometimes known as the arithmetic mean. You need to add all the data in the set (x) and divide by the number of items of data (n). Make sure you use a calculator and check your answer.

The formula for a mean is: $\bar{\chi} = \frac{\chi}{n}$

• **Median** - This is the central or **middle value**. If there are two middle values (if there is an even number of samples), the median lies midway between those two values.

The formula for a median is $\frac{n+1}{2}$ when the values are in rank order.

• **Mode** - This is the figure in a set of data that occurs **most often**.

• **Quartiles** - A quartile is one of the four equal divisions in a dispersion diagram of a set of data. The upper quartile is the median of the values above the median and the lower quartile is the median of the values below the median.

• **Inter-Quartile Range** If the number of values of ranked data is divided into four equal parts then the lines marking each division are quartiles. The inter-quartile range is the difference between the values of the upper quartile and lower quartile. The closer the clustering of values around the median, the smaller the inter-quartile range. This can be important when comparing two or more sets of data.

$$\text{Inter - quartile range} = Q_3 - Q_1$$

Practice Exercises

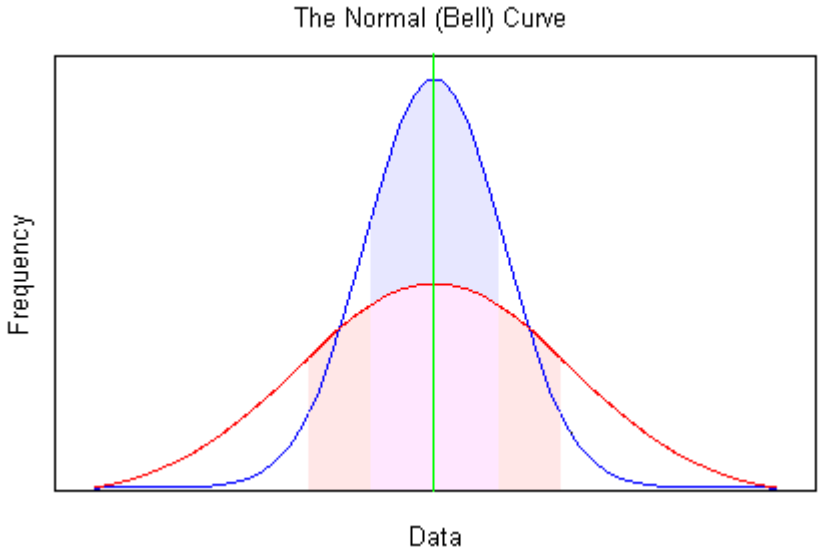
Look at the data below showing the population densities of a group of parishes.
1. Determine the **mean** and **median** and **modal** values.
2. Determine the **upper quartile**, the **lower quartile** and the **inter-quartile range**

34.2	4.9	1.1	28.1	3.3	1.7	19.1
3.0	13.4	5.2	6.1	19.1	15.3	11.1
		8.8	6.5	19.1		

The Basics – Part II

Standard Deviation

- The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range.
- For example, the mean of the following two is the same:
  - 15, 15, 15, 14, 16
  - 2, 7, 14, 22, 30.
- However, the second is clearly more spread out. If a set has a low standard deviation, the values are not spread out that much.



Standard deviation can only be used when the distribution of data either side of the mean is normal.

The formula is simple and is just a case of plugging the numbers in:

$$\sigma = \sqrt{\frac{\Sigma[x - \bar{x}]^2}{n-1}}$$

Top tip: The standard deviation can usually be calculated much more easily with a calculator. With some calculators, you go into the standard deviation mode (often mode '.'). Then type in the first value, press 'data', type in the second value, press 'data'. Do this until you have typed in all the values and then press the standard deviation button (it will probably have a lower case sigma on it). Check your calculator's manual to see how to calculate it on yours.

Practice exercises

i. Find the standard deviation of 4, 9, 11, 12, 17, 5, 8, 12, 14
ii. Find the standard deviation of 8, 9, 24, 29, 5, 9, 14, 17, 17
iii. Compare the 2 results, what can you conclude about the data?

Measuring Dispersion

- We are concerned in geography with distributions in space
- Previously distribution has used terms such as dense, sparse, nucleated or dispersed
- There are essentially 3 types of distribution
  - Uniform
  - Random
  - Clustered

Dispersion Diagrams

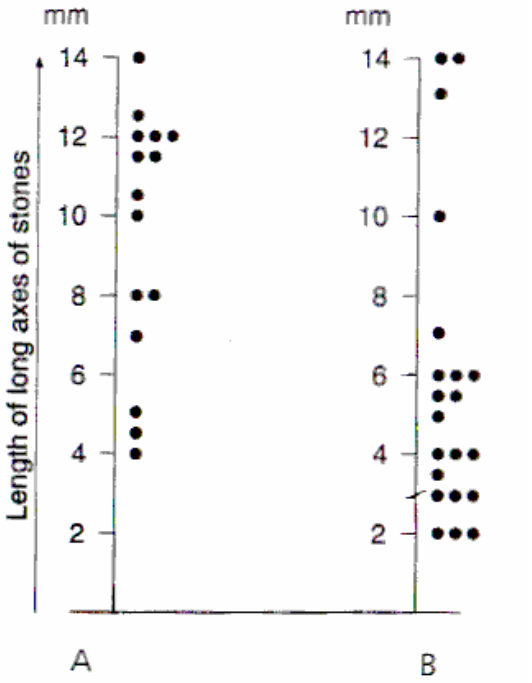
- A plot along a single vertical axis to show the spread of values in a distribution.
- Each value is shown as a dot. Often used to help identify **median** and **modal** values, the **quartiles** and **inter-quartile range.**
- Dispersion diagrams can also be used to identify positive or negative **skew** in a set of data.

Practice Exercise

A dispersion diagram showing a fieldwork sample of data collected by a group of students from glacial deposits in two different areas, A and B is shown to the right

i. Work out the **mean, median** and **mode** for each sample?

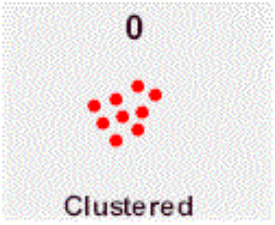
ii. How are the two samples **skewed**?

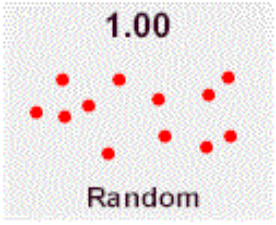


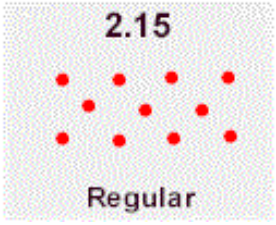
Nearest Neighbour Analysis

- The formula produced by the nearest neighbour analysis produces a figure expressed as R (the nearest neighbour index) which measures the extent to which the pattern is clustered, random or regular.

- **Clustered:** R = 0 All the dots are close to the same point.
- **Random:** R = 1.0 There is no pattern.
- **Regular:** R = 2.15 There is a perfectly uniform pattern where each dot is equidistant from its neighbours.

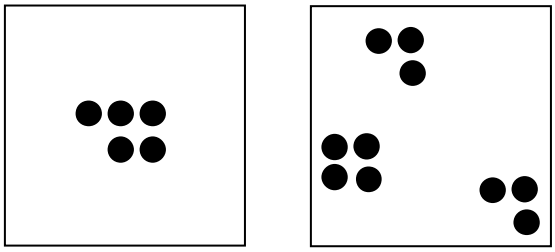






Nearest Neighbour pitfalls…

- It cannot always distinguish between a single and multi-clustered distribution. These two areas have an almost identical R value but are visually different

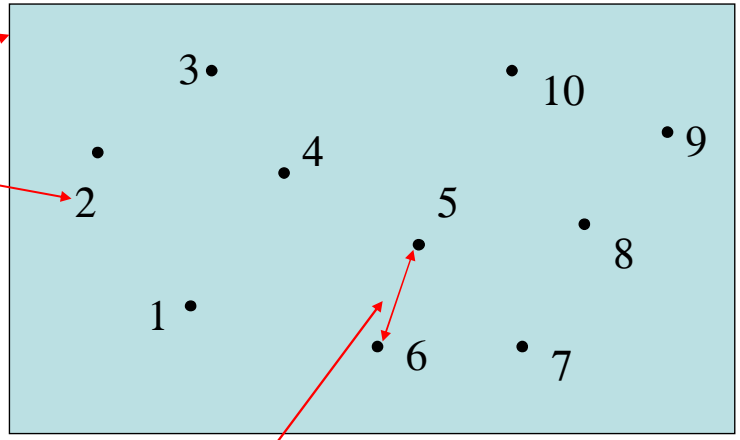


- An index of 1.0 does not always mean a totally random distribution. 2 patterns on the map when combined could give a false impression of randomness
- Sometimes a value of 1.0 is not caused by chance but an underlying factor e.g. the random location of springs causing the location of villages
- The NNI may depend on the area chosen. When comparing 2 populations you should choose the same scale and same size sample areas

Nearest Neighbour Analysis

1 – The first task is to define the boundary of the study area – take care over this as it will have a dramatic impact on the outcome and calculate its area - A

2 – Number on the points within the study area – it does not matter how you do this



Point number on map	Number of nearest neighbour	Distance between the 2 points (cm)
1	4	2.6
2	3	2.1
3	4	2
4	3	2
5	6	1.8
6	5	1.8
7	8	2.1
8	9	2
9	8	2
10	9	2.6
		Total ($\sum D$) = 21

3 – Draw up a table as shown. Measure accurately the distance from each point to its nearest neighbour, i.e. the nearest other point to it. Record this in the 3rd column. It is possible that one point may be the nearest neighbour of several other points – this does not matter

4 – Calculate the mean of the distances by adding up all the distances and dividing by the total number of points

$$\overline{D} = \frac{\sum D}{n} = 2.1$$

5 – Calculate the Nearest Neighbour Index (NNI) using the formula. Take care with units.

$$R = 2\overline{D}\sqrt{\frac{n}{A}}$$

$$R = 2 \times 2.1\sqrt{\frac{10}{84}} = 1.45$$

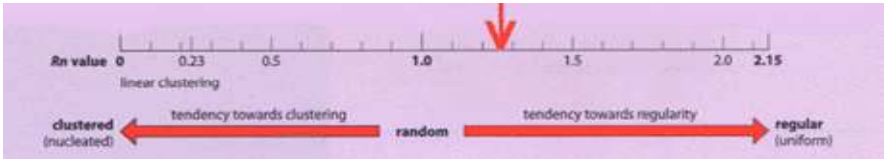
6 –
To test for clustering: reject H0 if the calculated value of **R** is less than the critical value at the chosen significance level

To test for dispersion: reject H0 if the calculated value of **R** is greater than the critical value at the chosen significance level.

Having obtained our result for **R** we need to determine its significance. There are 2 methods available – to use the tables of critical values or to compare our result to the diagrams shown

Table of Critical Values for the Nearest Neighbour Index for between 2 and 25 points and at the 0.05 and 0.01 probability levels

	Clustered pattern		Dispersed pattern	
n	0.05	0.01	0.05	0.01
2	0.392	0.14	1.608	1.86
3	0.504	0.298	1.497	1.702
4	0.57	0.392	1.43	1.608
5	0.616	0.456	1.385	1.544
6	0.649	0.504	1.351	1.497
7	0.675	0.54	1.325	1.46
8	0.696	0.57	1.304	1.43
9	0.713	0.595	1.287	1.406
10	0.728	0.615	1.272	1.385
11	0.741	0.633	1.259	1.367
12	0.752	0.649	1.248	1.351
13	0.762	0.663	1.239	1.337
14	0.77	0.675	1.23	1.325
15	0.778	0.686	1.222	1.314
16	0.785	0.696	1.215	1.304
17	0.792	0.705	1.209	1.295
18	0.797	0.713	1.203	1.287
19	0.803	0.721	1.197	1.279
20	0.808	0.728	1.192	1.272
21	0.812	0.735	1.188	1.266
22	0.817	0.741	1.183	1.259
23	0.821	0.746	1.179	1.254
24	0.825	0.752	1.176	1.248
25	0.828	0.757	1.172	1.243



The calculated value of **R** is more than the critical value for 10 points (0.728) for clustering so therefore we can not reject the null hypothesis and there is no evidence of clustering. However the calculated value is more than the critical value (1.272) for dispersion so we can reject the null hypothesis and conclude there is evidence of dispersion in the pattern of settlements. The diagram also supports the proposition that there is a tendency towards regularity in the spatial distribution of the settlements.

Exercise

1. The diagram below shows a part of a student's field sketchbook of the distribution of newsagents and banks/building societies in a small town in Southern England.

i. Draw in arrows showing the correct nearest neighbours for newsagents.
ii. construct the null and alternative hypothesis for both urban functions
iii. Carry out the tests
iv. Determine the significance levels
v. Draw a conclusion and reject or accept the null hypothesis suggesting reasons for any differences you have found

Hypothesis Testing

- The statistical comparison of places, areas, and areal distributions
- The identification of differences, similarities, and associations
- Those which operate consistently and from which predictions can be made e.g. difference between land use in West and East Anglia is partly a function of rainfall
- Those which are irregular i.e. characterised by random factors e.g. eccentricities of individual farmers

There are a range of hypothesis tests which you can conduct. The tricky part is choosing the correct one for the data you have got. See the next page for a structured explanation of which test to use

The premise…. The Null Hypothesis

- Starting point of most statistical tests
- Negative proposition formulated for the test with the anticipation of being rejected
- To calculate the probability that chance alone might yield the given result
- E.g. there is no significant difference in land use between West and East Anglia

The general procedure

This holds for all hypothesis testing

- Formulate H0, Formulate an alternative hypothesis, H1
- Decide upon a rejection level α (significance level). Often set at 0.05 or 5% (can be 1%) i.e. there is a 5% chance that the data could have occurred randomly under H0
- Carry out test
- Reject or accept null hypothesis
- Indicates the likelihood of there being some difference or correlation in the population. It says nothing about the magnitude of correlation/difference.

The Chi-squared Test

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

- Used to compare counted data, individual observations assigned to categories
- See method sheet on next page

Requirements for test:

- The data must be in the form of frequencies
- The frequency data must have a precise numerical value and must be organised into categories or groups.
- The expected frequency in any one cell of the table must be greater than 5.
- The total number of observations must be greater than 20.

Practice exercises

1. Aim: To identify any preferred orientation of corries in Snowdonia

Method: Orientations of corries were recorded in four categories according to compass quadrants, see below

Orientation from true north	Frequency of corries	Expected	O-E	$\frac{(O-E)^2}{E}$	X^2
0° - 89°	30				
90° - 179°	5				
180° - 269°	6				
270° - 359°	11				
Totals					

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

2. Aim: To detect differences in snail-shell patterns in 2 distinct habitats

	Light shells	Dark shells
Limestone pavement	153	112
Limestone woodland	96	120

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

Stats for Geography

Chi-squared Test

1 - **Observed data** collected from a survey of villages inside and outside of the National Park. Each house was categorised as being either inside or outside the National Park (the location of the Village) and as being one of 5 pre-determined age categories.

Expected values are calculated using:

$$\text{Expected Value} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

For example the expected number of houses which were modern and located inside the National Park would be:

$$\text{Expected Value} = \frac{141 \times 545}{1038} = 74.0$$

Q - Is there an association between age of housing and location (inside or outside of the National Park)?

	Location				Total
	Inside National Park		Outside National Park		
	Observed	Expected	Observed	Expected	
Pre-Victorian	47	**37.8**	25	**34.2**	72
Victorian	254	**199.5**	126	**180.5**	380
Inter-War	157	**157.0**	142	**142.0**	299
Post-War	48	**76.7**	98	**69.3**	146
Modern	39	**74.0**	102	**67.0**	141
Total	545		493		1038

2 - the Goodness of Fit of the Observed to Expected values is given by:

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

and this value is compared against the critical value for: (columns - 1) x (rows - 1) degrees of freedom.

So for our example the final value of Chi² (= 93.53) needs to be compared against the critical value for: (2 - 1) x (5 - 1) degrees of freedom. The Null Hypothesis (that there is no association between location and age of housing) is rejected if the value of Chi² is greater than or equal to the critical value

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(47-37.8)^2}{37.8} + \frac{(254-199.5)^2}{199.5} + \frac{(157-157)^2}{157} + \frac{(48-76.7)^2}{76.7} + \frac{(39-74)^2}{74} + \frac{(25-34.2)^2}{34.2} + \frac{(126-180.5)^2}{180.5} + \frac{(142-142)^2}{142} + \frac{(98-69.3)^2}{69.3} + \frac{(102-67)^2}{67}$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(9.2)^2}{37.8} + \frac{(54.5)^2}{199.5} + \frac{(0)^2}{157} + \frac{(-28.7)^2}{76.7} + \frac{(-35)^2}{74} + \frac{(-9.2)^2}{34.2} + \frac{(-54.5)^2}{180.5} + \frac{(0)^2}{142} + \frac{(28.7)^2}{69.3} + \frac{(35)^2}{67}$$

$$= 2.24 + 14.89 + 0 + 10.74 + 16.55 + 2.47 + 16.46 + 0 + 11.89 + 18.29 = 93.53$$

As the value of Chi-squared (93.53) is greater than the critical value (13.28 for 4 degrees of freedom) at the 99% certainty level (p=0.01) we can reject the null hypothesis and say that the goodness of fit between our observed and expected values is very poor. The major causes of this seem to be relatively more Vicorian and Pre-Victorian houses were present in (as compared to outside) the National Park whilst the situation for Post-war and Modern Houses was the opposite - there were more of these than we expected outide the Park.

Alternative Use of Chi-squared - the Chi² test can also be used to test the Goodness of Fit of Observed to Expected data where the expected values are those from a previous survey. For example a GOAD map was used to assess land use categories for buildings in a town in 1984. The town was resurveyed in 2005 and the question was raised as to how well the 2005 data (observed) matched those from 1984 (expected).

The degrees of freedom is given by: no. of categories - 1. As the buildings were categorised as belonging to one of 12 land use types the degrees of freedom would be 11 and the 5% critical value is 19.68.

The Chi² value obtained was 38.69 which means that we can reject the Null Hypothesis that the 2005 data do not differ from the 1984 data. The biggest increases were in Entertainment, Grocers and Department stores, whilst Clothing, Commercial and Electrical decreased.

Q - How well do land use categories for buildings surveyed in 2005 match those taken from GOAD maps for the same town dated 1984.

	Finance & Admin. Services	Clothing & Footwear	Supermarkets	Electrical, Computer & Phones	DIY, Hardware & Furniture	Commercial Services	Entertainment & Catering	Department & Specialist Stores	Grocers & General Convenience	Second Hand & Charity	Vacant premises	Dwellings & Other Use	Total
Total 1984 (E)	**16**	**34**	**5**	**8**	**12**	**27**	**25**	**31**	**5**	**2**	**13**	**23**	201
Total 2005 (O)	17	19	4	4	9	18	36	49	11	4	13	17	201
$\frac{(O-E)^2}{E}$	0.06	6.62	0.20	2.00	0.75	3.00	4.84	10.45	7.20	2.00	0.00	1.57	38.69

Table of Critical Values for the Chi² Test for between 1 and 20 degrees of freedom and at the 0.05 and 0.01 probability levels

Degrees of Freedom	Critical Value	
	p = 0.05	p = 0.01
1	3.84	6.64
2	5.99	9.21
3	7.81	11.35
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.73
12	21.02	26.22
13	22.36	27.69
14	23.69	29.14
15	24.99	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57

Student's t Test

- Allows a comparison of 2 means
- Do not use when there are more than 30 individuals in a sample
- Uses the standard error of the differences in the mean using the 't' distribution not the normal distribution

Requirements for the test:
The data needs to approximate a normal distribution
The two data sets should have a similar standard deviation
Not appropriate for data collected as percentages, proportions or simple counts

Practice exercises

1. Aim: Is there a significant difference between the number of plant species supported by acid moorland and limestone upland?

Quadrat No.	Acid X_1	Limestone X_2
1	6	14
2	8	12
3	9	6
4	4	11
5	7	15
6	11	14
7	7	17
8	6	8
9	8	
10	7	

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

2. Aim: To determine whether large towns and small towns experienced significantly different rates of population change during the period 1991-2001

Large Towns		**Small Towns**	
Nottingham	-0.40	Preston	-1.51
Leicester	-0.16	Darlington	0.16
Stoke	-0.44	Barnsley	0.08
Derby	0.31	Hemel Hempstead	2.30
Portsmouth	-0.87	Cheadle and Hulme	2.89
Swansea	0.31	Cannock	1.85
Warley	-0.41	St. Albans	0.35
Bolton	-0.43	Margate	0.92

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

Stats for Geography

1 - The data collected needs to be checked to see if they approximate to a normal distribution. Do this by setting up a tally chart.

- First identify the smallest and largest values.
- The difference between these gives the range.
- Divide that by 10 to give a rough size class interval.
- Round that up or down to match the data.
- Start you first size class just below your smallest value and then add on the size class interval.
- Construct the other size classes similarly.

2 - As the data seem to approximate to a normal distribution (more or less symmetrically distributed with a bell shaped peak in the middle) we can use the means to summarise the data. Mean ($\bar{x}$) = $\Sigma x/n$.

3 - The standard deviation gives a measure of spread (or dispersion) around the mean value. We calculate it using the formula:

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

To do this we need to calculate the squares of each of the individual values of x. This has been done in the 3rd and 5th columns of the main table. We then sum each of columns 2, 3, 4 & 5 to give the values used in the equation.

So for the 1st set of data:

$$s_1 = \sqrt{\frac{1448.62 - \frac{143.00^2}{15}}{15-1}} = \sqrt{\frac{1448.62 - \frac{20449}{15}}{15-1}} = 2.47$$

Higher levels of significance - As a convention a result is accepted as significant if it could occur by chance 1 time in 20. This corresponds to a probability of 0.05 or a 95% certainty. In critical situations you might not want to be 95% certain but 99% (p=0.01) or 99.9% (p=0.001) certain. Critical values exist for these probability levels as well.

Smallest = 5.1
Largest = 17.6
Range = 12.5
Size class = 1.25 (1.3 to 1 d.p.)

Size class	x_1	x_2
5.0 - 6.3	|	|
6.4 - 7.7	||	|
7.8 - 9.1	||||	|
9.2 - 10.5	|||	|||
10.6 - 11.9	|||	||
12.0 - 13.3	|	|||
13.4 - 14.7		||
14.8 - 16.1	|	
16.2 - 17.5		|
17.6 - 18.9		|

n	x_1	3rd x_1^2	x_2	5th x_2^2
1	7.1	50.41	10.8	116.64
2	9.7	94.09	10.1	102.01
3	9.7	94.09	5.9	34.81
4	6.9	47.61	7.1	50.41
5	5.1	26.01	9.4	88.36
6	7.9	62.41	12.4	153.76
7	8.7	75.69	11.3	127.69
8	9.3	86.49	13.9	193.21
9	11.4	129.96	12.5	156.25
10	12.1	146.41	13.7	187.69
11	10.8	116.64	16.7	278.89
12	8.4	70.56	12.8	163.84
13	9.0	81.00	8.6	73.96
14	15.1	228.01	17.6	309.76
15	11.8	139.24	10.4	108.16
Column Totals	Σx_1	Σx_1^2	Σx_2	Σx_2^2
	143.00	1448.62	173.20	2145.44

Mean ($\bar{x}$)	9.5	11.5
Standard Deviation (s)	2.47	3.22

The tally chart suggest that the data are approximating to a normal distribution so it is safe to use the mean and standard devaition to summarise each data set. The means of the two sets of data are clearly different but the tally charts indicate that the data overlap considerably so we can use the t -Test to measure how much overlap there is and give a mathematical statement of the probability that the two means are different.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|9.5 - 11.5|}{\sqrt{\frac{2.47^2}{15} + \frac{3.22^2}{15}}}$$

$$t = \frac{2.0}{\sqrt{\frac{6.100}{15} + \frac{10.368}{15}}} = \frac{2.0}{\sqrt{0.407 + 0.691}} = \frac{2.0}{\sqrt{1.098}} = \frac{2.0}{1.048} = 1.91$$

The Criitcal Value at the 5% level for 28 Degrees of Freedom ($n_1 + n_2 - 2 = 28$) is 2.05. As the value of t (1.91) is less than this critical value we have to accept that there is no significant difference between the two means. A difference between the two sets of data this good could occur by chance more than 1 time in 20, we have to accept the Null Hypothesis.

4 - Calculate t using this equation:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

It looks complicated but think of it as being a way of measuring the overlap between two sets of data.

This overlap depends on how far apart the means are (the top row of the equation) and on how spread out the data are around those means (the bottom part of the equation uses the standard deviations (s_1 and s_2)).

The smaller the value of t - the more overlap we have; and the bigger the value of t - the better separated the two sets are. To draw a conclusion we compare our value of t (1.91) against the critical value for the appropriate sample size. The sample size is calculate as Degrees of Freedom (D.F.) = $n_1 + n_2 - 2 = 15 + 15 - 2 = 28$

To be significant the value of t needs to be bigger than the critical value at at least the 5% level (p=0.05) for the correct Degrees of Freedom

Student's t Test

Table of Critical Values for the Student's t-Test for degrees of freedom between 20 and 40 and at the 0.05, 0.01 and 0.001 probability levels

p = D.F.	0.001	0.01	0.05
20	3.85	2.85	2.09
21	3.82	2.83	2.08
22	3.79	2.82	2.07
23	3.77	2.81	2.07
24	3.75	2.80	2.06
25	3.73	2.79	2.06
26	3.71	2.78	2.06
27	3.69	2.77	2.05
28	3.67	2.76	2.05
29	3.66	2.76	2.05
30	3.65	2.75	2.04
31	3.63	2.74	2.04
32	3.62	2.74	2.04
33	3.61	2.73	2.03
34	3.60	2.73	2.03
35	3.59	2.72	2.03
36	3.58	2.72	2.03
37	3.57	2.72	2.03
38	3.57	2.71	2.02
39	3.56	2.71	2.02
40	3.55	2.70	2.02

Mann-Whitney U Test

- Similar to the t-test in determining whether the data obtained belongs to two separate populations.
- However, the important difference is that this test determines whether there is a significant difference between the two medians
- Advantage of the test is that there is no assumption that the data is normally distributed but they need a similar distribution
- Used when sample sizes are below 20 and above 5

Practice exercises

1. Aim: In a study of the River Noe with its junction with Grinds Brook 2 random samples of stones were taken from near the mouth of Grinds Brook and from a river gravel bank slightly downstream. The object was to determine whether the bank at this point was fed by stones brought down by the river itself or from Grinds Brook

Stone roundness coefficients	
Grinds Brook Mouth	Gravel bank
51	58
32	10
23	52
6	47
13	23
13	27
19	38
47	30
27	57
20	16
7	47
12	45
13	
19	

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

2. Aim: You are conducting an investigation to determine whether Colchester fits one of the urban land use models so have determined the age of buildings in two separate areas of the town. Your task is to determine whether these form two separate areas

Lexden Road	New Town
90	42
21	81
86	17
70	22
5	76
1	8
45	35
73	3
111	77
24	12
78	31
63	90
101	7
69	47
24	16

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

Stats for Geography

Mann-Whitney U Test

1 - Arrange the values from your data sets into two columns - x_1 and x_2

2 - Calculate the **Medians**
This gives the first evidence to suggest that the Null Hypothesis is wrong. Putting the values of x1 in order:

0 3 12 12 **13** **17** 19 19 25 29

The median is the middle value, with 10 measurements then it will be halfway between the 5th and 6th values in this case 15

3 - **Line Plot** - Arranging your data along a line graph can help to visualise the overlap between the two data sets and make it easier to rank the values

We have 5 0's with ranks 1 - 5 available, the average rank awarded is 3
(1 + 2 + 3 + 4 + 5 = 15, 15 / 5 = 3)

10 + 11 = 21 so the 3's get 10.5

Rank given (R_1)			3						10.5
Ranks available			*4*						*11*
x_1			**0**						**3**
x_2	**0**	**0**	**0**	**0**	**1**	**2**	**2**	**2**	**3**
Ranks available	*1*	*2*	*3*	*5*	*6*	*7*	*8*	*9*	*10*
Rank given (R_2)	3	3	3	3	6	8	8	8	10.5

There are 3 2's with ranks 7, 8 & 9 available, the average rank awarded is 8 (7 + 8 + 9 = 24, 24 / 3 = 8)

Table of Critical Values for the Mann-Whitney U Test for values of n_1 and n_2 from 5 to 15 (at the 5% probability level ($p = 0.05$)) .

p = 0.05	n_1 (must be larger sample)										
	5	6	7	8	9	10	11	12	13	14	15
5	2	3	5	6	7	8	9	11	12	13	14
6		5	6	8	10	11	13	14	16	17	19
7			8	10	12	14	16	18	20	22	24
8				13	15	17	19	22	24	26	29
9					17	20	23	26	28	31	34
10						23	26	29	33	36	39
11							30	33	37	40	44
12								37	41	45	49
13									45	50	54
14										55	59
15											64

n	x_1	x_2	Rank(R_1)	Rank(R_2)
1	3	0	10.5	3
2	13	3	15	10.5
3	17	2	16	8
4	19	0	17.5	3
5	19	0	17.5	3
6	29	1	20	6
7	12	2	13.5	8
8	12	4	13.5	12
9	25	2	19	8
10	0	0	3	3
Median	15	1.5	$\sum R_1$ = 145.5	$\sum R_2$ = 64.5

4 - We now need to **rank those measurements**. As we have 20 values in total our Ranks will range from 1 to 20.

The smallest value gets the lowest rank - 1, and the largest value the highest rank (20 in this case). If values are tied they need the average rank of those available.

5 - Calculate the sum of the ranks for each data set.

$\sum R_1$ = 10.5 + 15 + 16 + 17.5 + ------- + 19 + 3 = 145.5

$\sum R_2$ = 3 + 10.5 + 8 + 3 + ---------- + 8 + 3 = 64.5

As the medians were different it suggested that the Null Hypothesis (that there is **no difference** between the two medians) may be wrong. A line plot was then drawn to examine the overlap between the two sets of data and to make it easier to rank the measurements.

Rank given (R_1)			3						10.5	13.5	13.5	15	16	17.5	17.5	19	20
Ranks available			*4*						*11*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
x_1			**0**						**3**	**12**	**12**	**13**	**17**	**19**	**19**	**25**	**29**
x_2	**0**	**0**	**0**	**0**	**1**	**2**	**2**	**2**	**3**	**4**							
Ranks available	*1*	*2*	*3*	*5*	*6*	*7*	*8*	*9*	*10*	*12*							
Rank given (R_2)	3	3	3	3	6	8	8	8	10.5	12							

As the two sets of data overlap we couldn't be certain that there was a significant difference between the two medians. The Mann-Whitney U test was applied to the data to get a mathematical probability of just how certain we could be that the medians were different.

$U_1 = n_1 \times n_2 + 0.5n_2(n_2 + 1) - \sum R_2$

$U_1 = 10 \times 10 + 0.5 \times 10(10 + 1) - 64.5$

$U_1 = 100 + 55 - 64.5 = 90.5$

$U_2 = n_1 \times n_2 + 0.5n_1(n_1 + 1) - \sum R_1$

$U_2 = 10 \times 10 + 0.5 \times 10(10 + 1) - 145.5$

$U_2 = 100 + 55 - 145.5 = 9.5$

As the smallest U value ($U_2 = 9.5$) is less than or equal to the critical value (23 for values of $n_1 = 10$ and $n_2 = 10$ at the 5% probability level) we can reject the Null Hypothesis. The median value for the first data set is significantly higher than that for the second data set. A difference this good could occur by chance less than one time in twenty ($p < 0.05$) so we can be 95% certain that there is a real difference between the two median values.

6 - Calculate the values of U_1 and U_2

$$U_1 = n_1 \times n_2 + \frac{1}{2}n_2(n_2 + 1) - \sum R_2$$

$$U_2 = n_1 \times n_2 + \frac{1}{2}n_1(n_1 + 1) - \sum R_1$$

Where n_1 is the number of values of x_1 and n_2 is the number of values of x_2 that you have collected. In this example both n_1 and n_2 are 10.

7 - Compare the smaller of the two U values against the critical value for the correct values of n_1 and n_2.

If the smallest U value is less than or equal to the critical value then you can reject the Null Hypothesis and accept that there is a significant difference between the two medians. There would be less than a 5% chance that two medians were different simply due to random variation within the data sets.

This is usually expressed as: $p < 0.05$

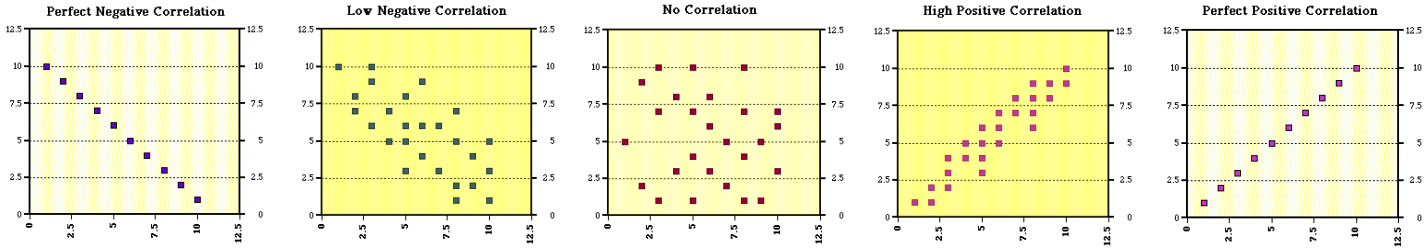
Correlation Statistics

- 2 tests are available: Spearman's Rank and Pearson Product Moment. Pearson is more advanced and allows the calculation of a line of best fit
- These tests can be used to see if there is a ***correlation*** between two sets of data. A ***correlation*** means that as one set of data changes, the other set seems to change with it. There are two types of correlation, either positive or negative.
- If one set of data ***rises*** as the other one rises, this is a ***positive correlation***. If one set of data ***falls*** as the other one rises, this is a ***negative correlation***.

WARNING
Detection of a significant correlation does not imply a causal relationship
WARNING

Spearman's Rank Correlation Coefficient

- Avoids rigid requirements of Pearson Product Moment correlation and quicker and easier to apply!
- Called a rank correlation because rank order is used to determine the association not the actual values



Practice exercises

Settlement Population	Number of services
220	4
350	3
1016	11
2362	19
4981	35
5632	41
6781	73
6793	43
7982	81
8763	72
10714	87
15739	114

1. Aim: To investigate whether there is a relationship between settlement size and number of services.

i. construct the null and alternative hypothesis
ii. Carry out test
iii. Calculate degrees of freedom and significance level
iv. Draw a conclusion and reject or accept the null hypothesis

Stats for Geography

Spearman Rank Correlation

1 - Enter the data into the 2nd and 4th columns. By convention we usually put the **independent variable** as the x-variable and the **dependent variable** as the y-variable. Here we are saying that cross-sectional area of a stream is likely to depend on the gradient of that stream rather than the other way round.

For statistical purposes we are going to test the **Null Hypothesis** that there is no relationship between the two variables. It's a bit like the idea that underlies our legal system - innocent until proven guilty, it's a safe way of looking at things.

2 - The first bit of evidence we will present to argue that the Null Hypothesis is wrong is a Scattergraph showing the strength of the correlation between the two variables.

	2nd	3rd	4th	5th	6th	7th
	Gradient		Cross Sectional Area		Diff.(D)	D^2
n	x	Rank (R_x)	y	Rank (R_y)	(R_x-R_y)	(R_x-R_y)^2
1	0.018	6	0.70	3	3	9
2	0.037	9	0.54	2	7	49
3	0.059	11	0.43	1	10	100
4	0.081	12	0.97	5	7	49
5	0.045	10	1.08	8	2	4
6	0.025	8	1.01	7	1	1
7	0.016	3.5	0.98	6	-2.5	6.25
8	0.016	3.5	0.86	4	-0.5	0.25
9	0.007	2	2.05	11	-9	81
10	0.017	5	1.78	9	-4	16
11	0.006	1	2.55	12	-11	121
12	0.021	7	2.01	10	-3	9

3 - Ranking the data - for a Spearman Rank Correlation it doesn't matter if you give the smallest value the lowest or the highest rank. For other tests, such as the Mann-Whitney U test, you have to do smallest gets lowest rank so it's, perhaps, best to adopt this for the Spearman Test as well. For our Gradient data (2nd Column):

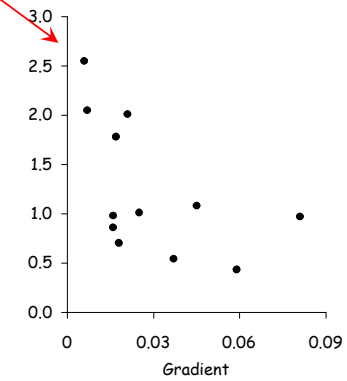
- the lowest value of 0.006 gets Rank 1 (column 3).
- the next lowest is 0.007 - Rank 2
- then we have two at 0.016 - they get the average Rank of 3.5 using up ranks 3 and 4
- so the next value 0.017 gets the next available rank of 5 and so on for the rest of the data.

4 - If there was a perfect +ve correlation then the ranks for each variable would agree perfectly. Similarly if there was a perfect -ve correlation then the site with the lowest ranked gradient would have the highest ranked cross-sectional area. By calculating the differences (column 6) between the ranks for the two variables we can start to look at this relationship.

If you try and add all the values in column 6 you will find the answer will be 0, all positive differences are cancelled exactly by the negative differences. To get rid of the sign we square the differences (column 7) and we can now calculate the sum of the squared differences ($\sum D^2$) and use this in the formula to calculate the Spearman Rank Correlation Coefficient:

n is the number of pairs of measurements we have

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$



Looking at the scattergraph which shows the relationship between cross-sectional area and gradient it does seem as if there is some suggestion of a negative correlation between the two variables. It isn't, however, a perfect correlation and it would not be safe to reject the Null Hypothesis on the basis of this evidence alone. For this reason we will use the Spearman rank Correlation test to measure the strength of the correlation and to give a mathematical statement on the strength of the relationship.

$\sum D^2 = 445.5$

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6 \times 445.5}{12(12^2-1)} = 1 - \frac{2673}{1716} = 1 - 0.558 = -0.558$$

5 - The final step is to compare the calculated value of r_s against the critical value corresponding to the number of pairs of measurements you looked at. A perfect correlation would have a value of +1 (+ve correlation) or -1 (-ve correlation) so to be significant your vale of r_s needs to be greater than or equal to the appropriate critical value (ignoring any sign) at at least the 5% probability level (p=0.05).

The calculated value of r_s (0.558 - ignoring the sign) is less than the critical value for 12 pairs of measurements (0.587) at a probability of 0.05. This means that a correlation this good could occur by chance more than 1 time in 20 just simply due to random variation in the variables. As a result we will have to accept the Null Hypothesis - there is no correlation between cross-sectional area and gradient.

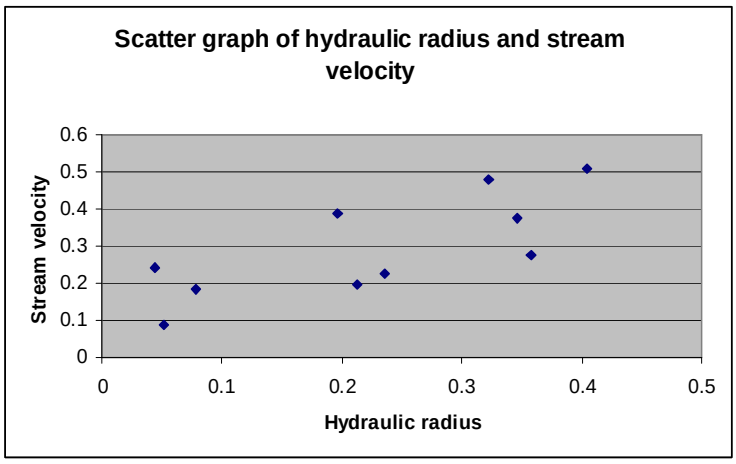
Table of Critical Values for the Spearman Rank Correlation test for between 5 and 24 pairs of measurements and at the 0.05 and 0.01 probability levels

No. of pairs of measurements	Critiical Value	
	p =0.01	*p =0.05*
5	n/a	1.000
6	1.000	0.886
7	0.929	0.786
8	0.881	0.738
9	0.833	0.700
10	0.794	0.648
11	0.755	0.618
12	0.727	0.587
13	0.703	0.560
14	0.679	0.538
15	0.654	0.521
16	0.635	0.503
17	0.618	0.488
18	0.600	0.472
19	0.584	0.460
20	0.570	0.447
21	0.556	0.436
22	0.544	0.425
23	0.532	0.416
24	0.521	0.407

Higher levels of significance - As a convention a result is accepted as significant if it could occur by chance 1 time in 20, corresponding to a probability of 0.05 or a 95% certainty. In critical situations you might not want to be 95% certain but 99% (p=0.01). Critical values exist for this probability level as well.

2. Aim: Having conducted a detailed field investigation of the River Lemon you wish to test whether there is a correlation between hydraulic radius and velocity of the stream

Hydraulic Radius	Stream velocity
0.044	0.241
0.052	0.086
0.078	0.184
0.197	0.389
0.235	0.225
0.213	0.197
0.322	0.479
0.358	0.273
0.346	0.373
0.404	0.510



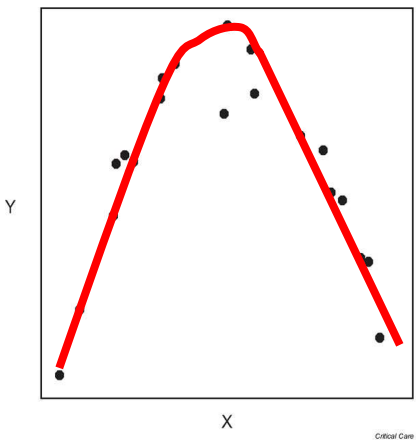
vi. construct the null and alternative hypothesis
vii. Carry out test
viii. Calculate degrees of freedom and significance level
ix. Draw a conclusion and reject or accept the null hypothesis

Pearson Product-Moment Correlation Coefficient

- More sophisticated test than the Spearman Rank test.
- More accurate result as it uses actual measured values of the data rather than their relative rankings
- It allows the calculation of a regression line, line of best fit, which can be useful for predictive purposes
- However, the data MUST come from a normally distributed population
- If unsure use the Spearman Rank Correlation
- The method sheet is included for reference

Cautionary correlation tales…

- Nonsense correlations are possible
  - It has been shown that the number of storks sighted correlates positively and significantly with births in Sweden. Only perform analysis where there is a possible and sensible relationship
- Correlation is for linear relationships ONLY
  - You may need to draw a scatter graph to ensure this is the case or you may end up with a non-linear relationship which is significant…



Stats for Geography

Pearson Product Moment Correlation

1 - Enter the data into the 2nd and 4th columns. By convention we usually put the **independent variable** as the x-variable and the **dependent variable** as the y-variable. Here we are saying that the temperature depends on altitude. The altitudes were selected at roughly 50m intervals from so that there was uniform distribution of values along the x-axis

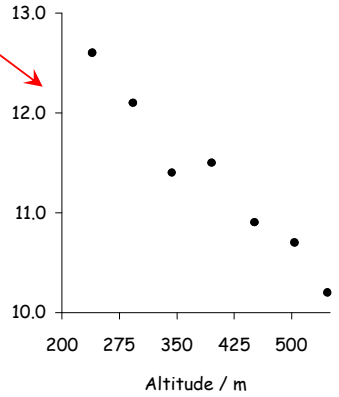
We are going to test, firstly, the **Null Hypothesis** that there is no relationship between the two variables. If we reject that we can calculate a line of best fit and use it to predict temperatures for other altitudes.

	2nd	3rd	4th	5th	6th
	Altitude / m		Temperature /°C		Product
n	x	x^2	y	y^2	xy
1	240	57600	12.6	158.76	3024.0
2	293	85849	12.1	146.41	3545.3
3	344	118336	11.4	129.96	3921.6
4	396	156816	11.5	132.25	4554.0
5	452	204304	10.9	118.81	4926.8
6	504	254016	10.7	114.49	5392.8
7	547	299209	10.2	104.04	5579.4
	Σx	Σx^2	Σy	Σy^2	Σxy
	2776	1176130	79.4	904.72	30943.9

3 - The formula used to calculate the Pearson Product Moment Correlation Coefficient is almost as long as the name of the test. To calculate it though is fairly straightforward. What we need are values of x^2, y^2 and x.y these are calculated in **columns 3**, **5** and **6** respectively. The next step is to sum each of the columns and those are the values which go into the formula.

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

2 - The first bit of evidence we will present to argue that the Null Hypothesis is wrong is a Scattergraph showing the strength of the correlation between the two variables.



Looking at the scattergraph which shows the relationship between temeprature and altitude it does seem as if there is a strong negative correlation between the two variables with an almost perfect lenear relationship. We wanted to be able to predict the temperature at other altitudes. For this reason we will use the Pearson Product Moment test to measure the strength of the correlation then Regression Analysis to produce the equation of the line of best fit and from this predict temperatures at 600m.

4 - The equation for any straight line has the form:

$y = a + bx$ - a is the intercept and b the gradient

The intercept is calculated as: $$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

And the gradient: $$a = \bar{y} - b\bar{x}$$

Table of Critical Values for the Pearson Product Moment correlation for between 5 and 24 pairs of measurements and at the 0.05 and 0.01 probability levels

Degrees of freedom (n-2)	Critiical Value	
	p =0.01	*p =0.05*
5	0.874	0.754
6	0.834	0.707
7	0.798	0.666
8	0.765	0.632
9	0.735	0.602
10	0.708	0.576
11	0.684	0.553
12	0.661	0.532
13	0.641	0.514
14	0.623	0.497
15	0.606	0.482
16	,590	0.468
17	0.575	0.456
18	0.561	0.444
19	0.549	0.433
20	0.537	0.423
21	0.526	0.413
22	0.515	0.404
23	0.505	0.396
24	0.496	0.388

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}} = \frac{30943.9 - \frac{2776 \times 79.4}{7}}{\sqrt{\left(1176130 - \frac{2776^2}{7}\right)\left(904.72 - \frac{79.4^2}{7}\right)}}$$

The calculated value of r ends up as being -0.9795. As this is greater than the critical value (ignoring the sign) for n - 2 = 5 degrees of freedom we can reject the Null Hypothesis and accept that temperature is negatively correlated with altitude. As you increase in altitude the temperature goes down. To find the temperature at 600m we need to use: **Regression Analysis** - calculating the intercept and gradient gives an equation for the line of best fit:

$$y = 14.2 - 0.0072x$$

Substituting an x value of 600 gives: $y = 14.2 - 0.0072 \times 600 = 14.2 - 4.3 = 9.9$

So we predict that the temperature at 600m would be 9.9°C

Example - Calculating b (the gradient)

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{30943.9 - \frac{2776 \times 79.4}{7}}{1176130 - \frac{2776^2}{7}}$$

$$b = \frac{30943.9 - 31487.8}{1176130 - 1180882.3} = \frac{-543.8}{75247.7} = -0.0072$$

and a (the intercept)

$$a = \bar{y} - b\bar{x} = \frac{79.4}{7} - \left(-0.0072 \times \frac{2776}{7}\right)$$

$$a = 11.34 - (-2.86) = 14.2$$

Statistics for Geographical Investigations

Suitable Statistical Tests

If you are planning to collect a series of **replicated** measurements of one variable

- then you need to make sure you collect at least six replicates and if you end up with
  - less than 15 → you can summarise the data using the median and range.
    - Then → to compare this set of data against a second set
      - use → **the Mann Whitney U Test.**
        - tests for the difference between the medians of two sets of data
        - needs at least 6 replicates per data set and ideally 8 or 10
        - if you have more than twenty measurements you will have to use the t-Test
  - more than 15 → you will need to check that the data are **normally distributed** using a **tally chart**.
    - If they're not normal → you can summarise the data using the median and range.
    - If they are → you can summarise the data using the **mean** and a measure of **dispersion.** (eg **standard deviation**)
      - Then → to compare this set of data against a second set
        - use → **the Student's t-Test.**
          - tests for the difference between the means of two sets of normally distributed data
          - needs at least 12-15 replicates per data set before you can start to look for a normal distribution
          - is unlikely to be appropriate for data in the form of percentages, proportions or counts of animals

If not → then are you going to collect two sets of data to compare. (eg soil pH readings from two sites)

- If so → you need to make sure you collect at least six replicates and if you end up with (see above)

If not → then will you be looking for a **correlation** between two variables. (eg between the hydraulic radius and velocity of sites along a stream)

- If so → then you will need twelve to fifteen pairs of measurements and most of the time you can
  - use → **the Spearman Rank Correlation Test.**
    - tests the strength of the correlation between two variables
    - ideally needs at least 12-15 pairs of data with values more or less evenly spread along the *x*-axis
    - can't be used for scattergrams with a "U" or inverted "U" shaped pattern
  - But → if you want to draw a **line of best fit** or predict values of the **dependent variable** from the **independent variable** then
    - use → **the Pearson Product Moment Correlation** and **Regression Analysis.**
      - needs at least 12-15 pairs measurements which are normally distributed along the *x* and *y* axes
      - data needs to show a clearly linear relationship
      - enables a line of best fit to be calculated and for values of the dependent variable (*y*-axis) to be predicted from values of the dependent variable (*x*-axis)

If not → then you might be want to match one set of data against a set of **expected values**

- If so then use → **the Chi Squared Test.**
  - is probably the most misused test
  - requires the calculation of expected values which can come from:
    - a theoretical model,
    - a previously collected set of data
    - probabilities calculated from row and column totals in a contingency table

If not → **then you you need to think very carefully about whether your investigation will generate data suitable for statistical analysis.**

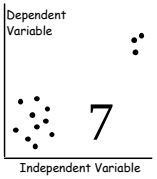
Glossary

Expected Values - are generated as a result of a geographical theory. You might start with a Null Hypothesis which predicts, for example, an even spread of results across a series of categorical measurements. Sometimes you can use a set of data from a previous study as expected values and test the goodness of fit of a second set to that model. In other situations you might have to calculate expected values by using the row and column totals in a contingency table.

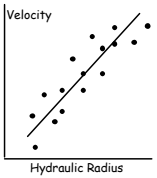
More details of all these can be found on the Chi Squared page.

Independent Variables - a variable which you think may be important in "causing" variation in a second - **Dependent Variable**.

Ideally you need to have a more or less even spread of values along your *x*-axis.



Line of Best Fit - a line drawn through a scattergram to emphasise the relationship between the two variables. Lines of best fit can either be drawn by eye or (preferably) calculated mathematically using regression analysis.



Glossary

Replicates - there is no guarantee that a single measurement of anything will be typical of the whole population or site. So every variable you want to measure will need to be **replicated**. To avoid biasing your results then it is worth considering whether some form of random sampling is needed. The other main consideration is the number of replicates needed. Each statistical test has a minimum number of replicates it requires before it will let you make a decision. Your plan needs to ensure that in the time you have available you can collect at least this minimum number of samples.

Mean - one of the averages available to summarise a set of data. To calculate the mean add all the measurements together and divide by the number of samples collected. The mean can only be used if the data approximate to a **normal distribution**.

Example:

12 23 21 17 24 31 29 25 26 24

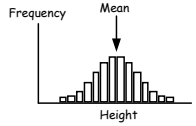
Sum = 232 No. of samples = 10

Therefore, mean = 23.2

The mean is usually denoted as:

$\bar{x}$ (pronounced x bar)

Normal Distribution - the symmetrical bell shaped distribution you get if you take a large series of measurements of lets say heights of people of the same sex and age and then plot a frequency histogram. The mean of the measurements will in the middle of the distribution with an equal number of smaller and larger values either side of it.



Tally Chart - a simple way of deciding if a set of measurements are normally distributed:

- calculate the range of your set of measurements
- divide this range into 10 size classes
- tally up against each size class.

Size Class	Tallies	Tallies
12 - 15mm	I	IIIIII
16 - 19mm	II	IIIIIIIIII
20 - 23mm	IIII	IIIIIIIIIIII
24 - 27mm	IIIIIII	IIIIIII
28 - 31mm	IIIIIIII	II
32 - 35mm	IIIIIII	
36 - 39mm	IIIIII	I
40 - 43mm	IIII	I
44 - 47mm	II	I
48 - 51mm	I	I

12 23 21 17 24 31 29 25 26 24

Arrange in order:

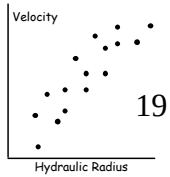
12 17 21 23 **24 24** 25 26 29 31

Dispersion - the spread of data around the average value. If the data are normally distributed use the semi inter-quartile range or standard deviation. The usual way of expressing the measure of dispersion is as:

mean ± semi inter-quartile range or
mean ± standard deviation

For data which are not normally distributed the median and range are the best summary.

Correlation - a relationship between two variables where as one variable changes then there is a corresponding change in the second variable. Correlations can either be +ve or -ve and are plotted as scattergrams with the independent variable plotted on the *x*-axis and the dependent on the *y*-axis.



Which Statistics Test?

Choose which statistical test (if any) you would use for the following data

1. Velocity and hydraulic radius readings at 12 sites along a stream.
2. Eight velocity readings at sites along 2 different streams.
3. Total numbers of Freshwater Invertebrates in 8 Riffles and 8 Pools.
4. Numbers of Carnivores, Herbivores and Detrivores in Riffles and Pools.
5. Soil depth readings at 15 sites along a transect down a slope.
6. Soil depth readings at 8 sites in a woodland area and 6 sites in a grassland area.
7. A questionnaire on traffic problems with 5 possible choices of answer. Both locals and tourists were asked.
8. The lengths of 20 randomly selected pebbles from each of 2 sites.
9. Traffic and pedestrian survey results from 10 locations within a town.
10. Tourist count from 5 locations along a footpath.
11. Spatial distribution of different Land Use data radiating out from a CBD.
12. Comparing the length of 18 limestone and 18 gritstone pebbles from a single site.
13. Comparing Land Use in 1990 with Land Use in 2005 using GOAD maps.
14. Wind velocity and temperature at 6 sites in a woodland.
15. Environmental Quality from 3 villages inside and 3 villages outside the National Park.
16. House Style data from 3 villages inside and 3 villages outside the National Park.
17. Soil depth and soil pH readings at 4 sites on managed grassland and 4 sites on rough grassland.
18. Soil depth and soil pH readings at 10 sites on managed grassland and 15 sites on rough grassland.