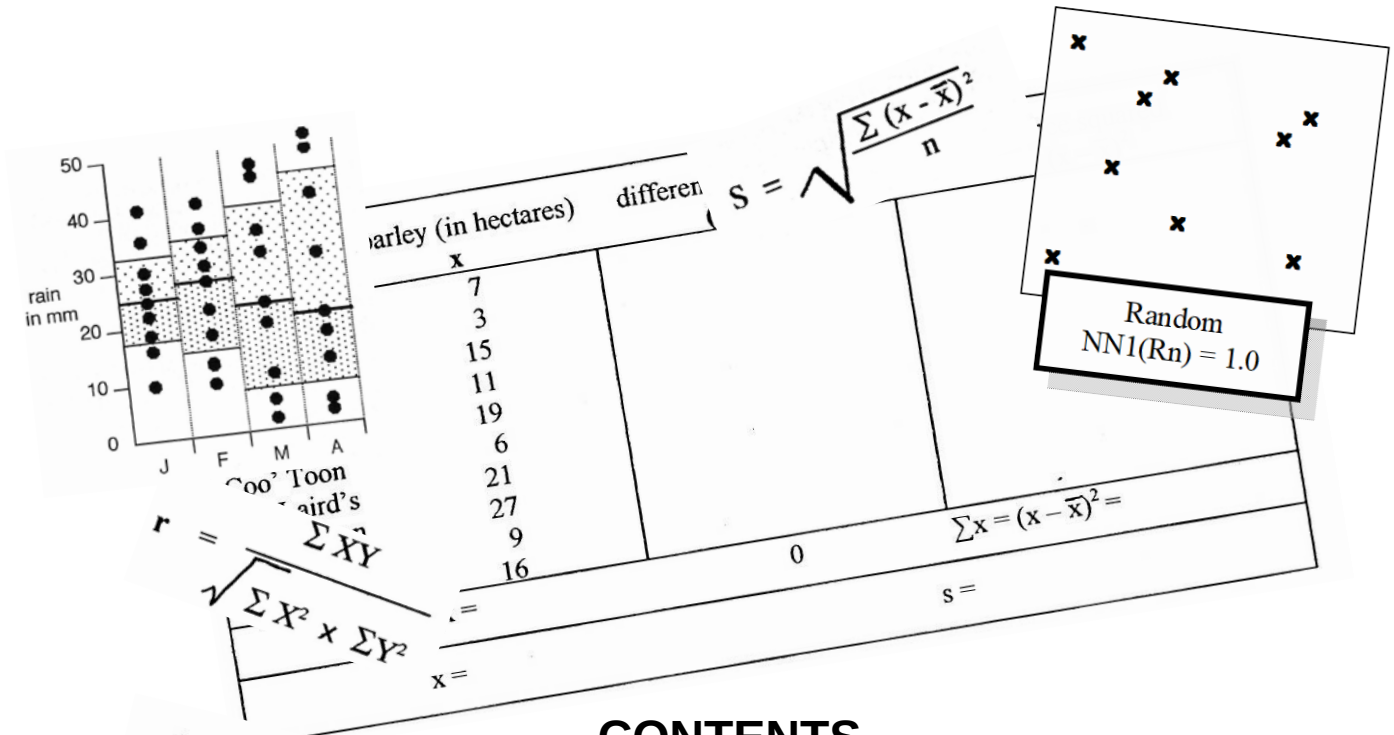


Adv. Higher Geography

Statistics

for

Geographical Methods & Techniques



CONTENTS

Introduction

3

Types of data

4

Descriptive statistics

4

Measures of centrality

4

Measures of dispersion

6

Sampling

9

Testing for sampling errors

11

Inferential statistics

12

Chi² test

12

Correlation and correlation coefficients

17

$$Rs = 1 - \frac{6 \sum d^2}{n^3 - n}$$

Pa

Spearman's Rank correlation coefficient (Rs)	18
Pearson's Product Moment correlation coefficient (r)	21
Linear regression	24
Non-linear relationships	27
Excel-ing statistics	31
Nearest Neighbour Analysis	32
<u>Significance Tables for Chi², 'r' and 'Rs'</u>	33
<u>Answers</u>	34

PREFACE

Statistics form one of the five areas of assessment in the Geographical Methods & Techniques unit of the SQA Advanced Higher Geography course. They are assessed as part of the unit and examined in the final exam along with graphical techniques. It is essential, therefore, that the student gets to grips with statistics if wishing to do well in the course.

Within the notes reference is made to the textbook '**Skills & Techniques for Geography A-level**' by Garrett Nagle and Michael Witherick (EPICS series : Stanley & Thornes) and students are encouraged to refer to it also. 'Answers' are provided at the end, but should not be referred to until the activity or question attempted is completed. If on checking the answers the student finds a error, then do try to work out what has gone wrong. If stuck, ask the teacher or tutor for help. Tables of significance are also provided.

If tackled positively, statistics open avenues of research which can enrich the learning experience and encourage the development of a scientific approach to the study of geography.

Statistics for Geog. Methods & Techniques

Introduction

Any geographical study involves the gathering of information 'in the field', from textual sources (books/ journals/newspapers) or from maps and atlases. In the course of collecting and recording information, the researcher frequently ends up with masses of numerical data in tabulated form. The problem then arises as to what to do with it.

Tabulated numerical facts - e.g. census data, climate figures, traffic flows - are **statistics**. But '**Statistics**' also refers to the science of data handling, presentation and interpretation. Primarily it is this latter meaning that involves us here, for it is the geographer's task to interpret any statistics accurately and to present them in the clearest and simplest way possible. For the geographer, then, a sound working knowledge of both statistical and graphical techniques are essential.

There are two branches of statistics – **descriptive** and **inferential**. **Descriptive statistics** seek to summarise data into manageable and easily understood formats, so that valid comparisons can be made between sets of data. **Inferential statistics** seek to identify relationships between sets of data in order to test hypotheses (proposed explanations).

But why are statistics so necessary? Consider for a moment two possible extracts from an imaginary report drawn up by the Planning Dept. of Highland Council for submission to the Scottish Office, seeking Scottish Executive funding for a major road improvement.

Extract 1

The roundabout at the south end of the Kessock Bridge on the A9 requires upgrading. The volumes of commuter traffic entering the city have increased steadily over the years leading to tailbacks and gridlock at peak times. The economy of Inverness is threatened.

Extract 2

The roundabout at the south end of the Kessock Bridge on the A9 requires upgrading. Originally built to deal with a maximum 30 vehicles per minute, recent traffic surveys have revealed average flows of 37 vehicles per minute and peak time flows of 61 to 73 vehicles per minute.

ACTIVITY 1

1. Which of the two extracts is more likely to attract Scottish Executive funding? Justify your answer.

Evidence is essential wherever and whenever proof is required. It substantiates what you wish to justify. The problem for the geographer, however, is that in the real world the evidence is frequently not clear-cut. So often we have to deal with **probabilities** rather than **absolutes**. This is why we, as geographers, need to turn to statistical techniques in order to describe and/or test geographical phenomena. They are then a valuable tool in geographical investigation.

The purpose of this unit is to :

- develop your skills in the manipulation and presentation of your own statistics gathered during fieldwork
- improve your ability to understand and interpret statistical information from other sources

Types of Data

While the choice of analytical technique used may involve a degree of personal preference, initially the type of data present limits the choice of statistical or graphical technique that can be applied. There are four main types of data :

nominal data	-	data with ' names ' e.g. numbers of hamlets, villages, towns or the dates of good summers (1955, 1975, 1976, 1985, 1998) ; the data is usually gathered in ' frequencies '.
ordinal data	-	data which has been rank ordered e.g. 1 st , 2 nd , 3 rd , 4 th ... ; medians/ quartiles and Spearman's Rank Correlation Coefficient depend on ordinal data.
interval data	-	real number data e.g. measurements of daily rainfall in mm or the summit heights of mountains in metres. Interval data and ratio data (see below) are both open to the use of the most advanced statistical techniques.
ratio data	-	ratios are also real numbers but differ from interval data in having a true zero ; they include proportions (<i>midges are 5 times bigger than 1st years</i>) and percentages (<i>1st years are 20% smaller than midges</i>).

ACTIVITIES 2

1. You are testing the idea that meander size is a consequence of stream discharge. Which branch of statistics will you use?

2. You want to compare the mean temperatures at the base and the top of Ben Nevis. Which branch of statistics will you use?
3. What types of data are the following :
 - (a) 21, 33, 36, 49, 54, 55, 67
 - (b) the numbers of igneous, sedimentary and metamorphic pebbles found in a stream bed
 - (c) 45mm, 0mm, 3mm, 7mm, 0mm, 14mm rainfall
 - (d) 1066, 1314, 1513, 1746, 1815
 - (e) 10th, 9th, 8th, 7th, . . .
 - (f) 4.5, -2.27, 0.12, 6.03, 8.47
 - (g) 12%, 17%, 22%, 49%
 - (h) the numbers of primary, secondary and tertiary industries

EXTENSION

Read page 8/9 of Skills & Techniques for Geography A-level (EPICS). Do the Review (p.9).

Descriptive Statistics

Descriptive statistics deal with the description and comparison of data sets in order to seek out the similarities and differences between them. For example, you may have data from a number of climate stations which you wish to compare. So you might look for the **highest** and **lowest** rainfall totals (i.e. the **max.** and **min.** values). You might want to find the **average** or **mean** air temperatures. You may even wish to go further and examine the **variability** of rainfall or temperature at each climate station. In all three cases you are seeking to summarise and present the data in a form that permits direct comparison between the climate stations. So what statistical techniques are available?

Measures of central tendency :

Let us begin by looking at '**measures of central tendency**'. There are three possibilities – the **mode**, the **median** and the **mean**.

MODE : the **mode** is the group or class that occurs most often. Where more than one modal group occurs, data set may be described as **bi-modal** (2 modes) or **tri-modal** (3 modes) or similar.

MEDIAN : the **median** is the middle value of a set of rank ordered data.

MEAN : the **mean** is the arithmetic average of a set of data. It is calculated by summing the values (x) and dividing the total Σx by the number of values (n)

i.e.

$$x_1 + x_2 + x_3 + x_4 + x_5 / \Sigma x_{1-5}$$

e.g.

$$4 + 5 + 8 + 6 + 7 = 30$$

$$\Sigma x_{1-5} = 30 \qquad n = 5$$

$$\text{mean}(x) \text{ or } \bar{x} = 30/5$$

$$\bar{x} = \underline{6}$$

So, when is it best to use one in preference to the other two? All three are **measures of central tendency** around which the other individual values are dispersed. The **mean** would seem to be the most accurate, but there are times when its use is not suitable. For example, an average of 12.34 people per day climb Ben Wyvis ; when did you last see 0.34 of a person walking about? Then again, a set of data may be clearly bi-modal, i.e. have two classes that occur most often ; neither the mean nor the median could identify that.

ACTIVITIES 3

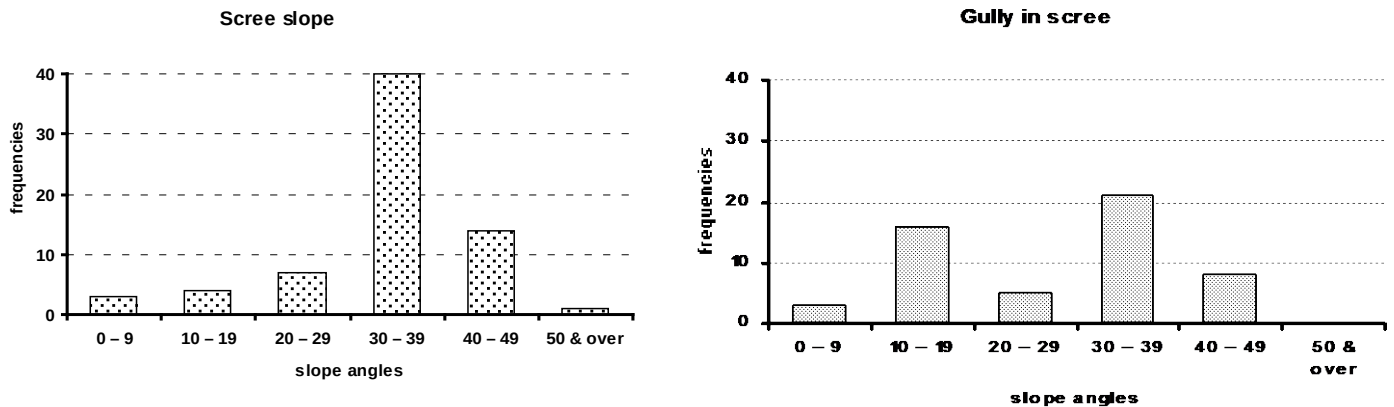
Examine Table 1 and Fig.1 :

1. What is the **modal** slope angle frequency of the 'scree slope'?
2. What is the **modal** slope angle frequency (or frequencies) of the 'gully in scree'?
3. How should the **modal form** of slope angles in the 'gully in scree' be described?
4. It is not possible to work out the medians or means for the data in Table 1. Why not?

Table 1 : Frequency of slope angle in 'scree' and 'gully in scree'

Slope angles in degrees	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 & over
Scree slope	3	4	7	40	14	1
Gully in scree	3	8	5	22	8	0

Fig. 1 : Frequency of slope angle in 'scree' and 'gully in scree'



Examine Table 2, which gives the discharge and rainfall for the River Thurso and Halkirk respectively over a 15 day period :

5. Draw up a table with the discharge data in rank order as in the column labelled 'Rank X' (it has been started for you). What is the **median** value of the discharge over the 15 day period?
6. Draw up a second table this time with the rainfall data in rank order as in the column labelled 'Rank Y' (it has been started for you). What is the **median** value of the rainfall over the 15 day period?
7. Using the raw data, calculate the **mean** values of both the discharge and rainfall.
8. Explain why there is such a large difference between the **median** and **mean** values in the case of the discharge data, yet not with regard to the rainfall data.

Table 2 :

Day	Discharge in cumecs X	Rainfall Y	Discharge rank ordered Rank X	Rainfall rank ordered Rank Y
1	6.6	9.6	284.2	41.6
2	62.4	21.8	142.2	23.4
3	95.1	9.6	120.2	21.8
4	94.3	17.4	95.1	
5	83.8	4.6		
6	68.7	2.0		
7	32.9	7.6		
8	32.8	6.6		
9	120.2	41.6		
10	284.2	23.4		

11	142.2	2.4		
12	41.4	2.6		
13	31.2	11.4		
14	46.6	13.0		
15	60.5	2.8		

EXTENSION

Read page 9 to 11 of Skills & Techniques for Geography A-level (EPICS).
Do the Review on page 11.

Measures of Dispersion :

Knowing the measure of central tendency is useful, but it helps more to know how the rest of the data is spread or **dispersed** around the mode, the median and the mean. For example, the average temperature in July in the north of Scotland is 15°C ; but that does not tell what the hottest or coldest temperatures were during the month. If, on the other hand, we knew the hottest temperature was 27°C and the coldest 9°C, then we could say that the temperature range was 18°C. Compare two sets of data :

Data Set 1	Data Set 2
1	6
4	7
6	8
9	9
12	10
14	11
<u>21</u>	<u>12</u>
Mean of Set 1 ($\bar{x}_1$) = <u>9</u>	Mean of Set ($\bar{x}_2$) = <u>9</u>

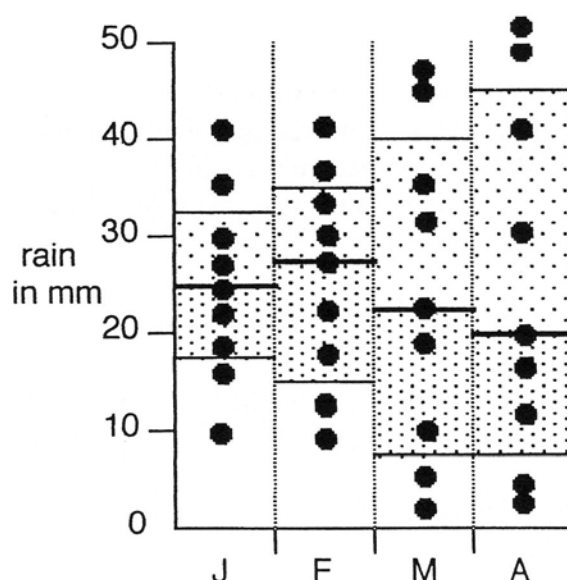
Both data sets share the **same mean**, but clearly **differ** in their **dispersion around the mean**. The maximum for Set 1 is 21 and for Set 2 is only 12 ; the minimum is 1 for Set 1 and 6 for Set 2. But even between the mean and max./min. values there is a spread of data.

Measures of dispersion can be procured for two of the three measures of central tendency – the **mean** and the **median**. In the case of the median, **quartiles** are employed ; in the case of the mean, **standard deviation** is calculated. Let's start with **medians & quartiles** :

Median & Quartiles :

The **median** is the middle value in a rank-ordered data set. **Quartiles** are a 'middle value' also, but this time they lie between the median and the lowest and highest ranked values. This gives the **lower quartile** and the **upper quartile**. The range between the lower and upper quartiles is called the **inter-quartile range** (see Fig. 2 below).

Fig. 2



In this example, 9 years of monthly rainfall totals have been graphed and the medians and quartiles inserted. Note how the highlighting of the **inter-quartile range** brings out the **variability** of the rainfall over the four months. Quartiles are particularly useful in identifying **patterns of variability** in rainfall or river discharge or in any data which is not normally distributed, otherwise 'skewed'.

ACTIVITIES 4

Examine Fig. 2 :

1. What is the median value for January?
2. What **changing trend in rainfall** do the median values for the four months highlight?
3. What are the **inter-quartile ranges** for January and April?
4. What can you learn from the graph about the variability of the rainfall over the 4 month period?

EXTENSION 4

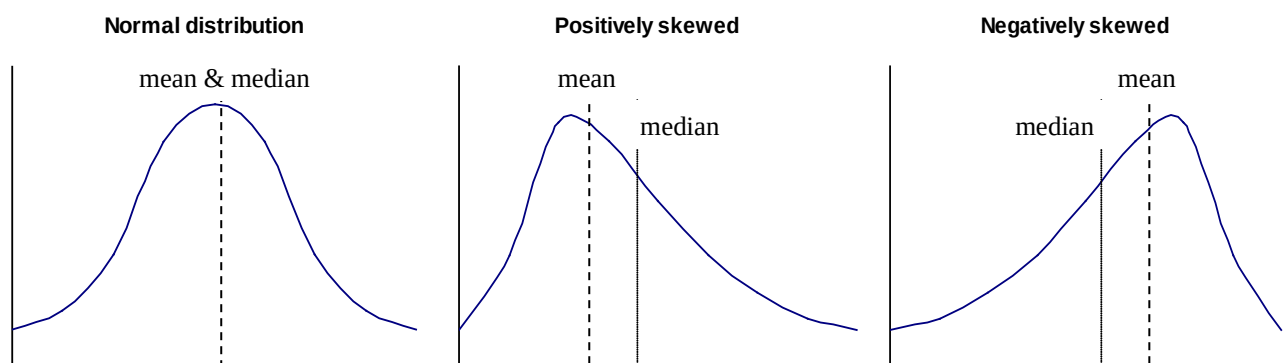
Read pages 11 to 14 of Skills & Techniques for Geography A-level (EPICS) on medians and quartiles. Do the Review on p. 12.

Skewness & Normal Distribution :

When choosing whether to use **median & quartiles** or **mean & standard deviation**, it is essential to examine the data for **skewness**. Skewness is caused by either the highest or lowest values being too dominant in a data set. Skewness pulls the **mean** towards the dominant values and results in statistical unreliability.

Where a data set is **normally distributed** (see fig. 3), the mean and median are similar in value, so the **mean** and **standard deviation** should be applied. They are mathematically more robust measures. Indeed, the '**normal distribution**' of data underlies all more rigorous statistical techniques used. Where the data is skewed however, the mean and median differ in value considerably (see fig. 3). In such a situation it is necessary to apply **median & quartiles**.

Fig. 3



Standard Deviation :

Standard deviation is the measure of dispersion around the **mean**. The importance of standard deviation becomes clearer when we move from descriptive to inferential statistics. This is because dispersion around the mean lies at the heart of the **probability** problem.

To work out the standard deviation (S or SD) of a data set :

- draw up a table similar to that shown in Table 3 - "Barley production on selected farms";
- add up the total in column 'x' (in this case 'hectares of barley') ;
- calculate the mean i.e. the sum of 'x' divided by the total number of values 'n' : $\Sigma x/n$
- subtract the mean from each value, i.e. '(x - bar x)'
or $(x - \bar{x})$
- there will be negatives ; so, to get rid of the negatives, square $(x - \bar{x})$
i.e. $(x - \bar{x})^2$
- sum the total of $(x - \bar{x})^2$
- insert the values for 'n' and $\Sigma (x - \bar{x})^2$ into the equation for standard deviation ('S' or 'SD')
- and calculate it.

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$$

ACTIVITIES 5

1. Complete the following example on 'barley production' (Table 3) in your own folder.
2. What is the **mean** and **standard deviation** for the 'hectares of barley' grown?
3. Identify the farm(s) lying more than 1SD above the mean and the farm(s) lying more than 1 SD below the mean.
4. Suggest further research that might be undertaken in order to identify the reasons for the wide variation in barley production across the farms.

Table 3 : Barley production on selected farms

Farm	barley (in hectares) x	difference from the mean (x - $\bar{x}$)	difference squared (x - $\bar{x}$) ²
Dubbie's	7		
Clartie's	3		
Puddock Law	15		
Hillie's	11		
Steenie's	19		
Coo' Toon	6		
The Laird's	21		
The Hame Farm	27		
Badenyeuchers	9		
Broadlands	16		
$\Sigma x =$		$\Sigma (x - \bar{x})^2 =$	
$\bar{x} =$		SD =	

EXTENSION :

Read pages 14 to 16 of Skills & Techniques for Geography A-level (EPICS) on medians and quartiles.

Sampling :

A complete set of data is called the '**population**'. **Census data**, for example, is essentially **population data** in more ways than one. Another example could be where all 1st year pupils are surveyed to test whether they prefer tarantulas to gerbils as pets. Gathering every scrap of data, however, is time-consuming. So you might select a group of 1st years for your survey, e.g. all the 1st year pupils in a class of 28 out of a total S1 roll of 214 in the school. This is called sampling and any data gathered is a **sample**. Sampling saves time and energy and, if properly done, should give very reliable results.

But it is so easy to **bias the sample**! The stones picked up may be those easiest to lift or they may be orientated the way you think they ought to point. You may avoid asking 'auld wifies' about their shopping habits for fear of getting a whack from their brollie or handbag (which is usually large and filled with lead!). Either way you are consciously or unconsciously introducing **bias** into your sample. So it is necessary to understand the nature and conduct of sampling.

Sampling may be spatial or temporal. Sampling a map or an area of sand dunes or a quarry face are forms of **spatial sampling**. **Temporal sampling** would include recording changes over time, e.g. daily weather recordings, stream discharge or traffic flows. The type of sampling undertaken within both, however, can be **random**, **systematic** or **stratified**. If sampling is spatial, it could be done using sample **points**, **lines** or **areas** (see fig. 4).

ACTIVITIES 6

1. Examine a 1:50 000 O.S. map of your local area. How would you sample in order to calculate the percentage of the total map area covered in water bodies (rivers, lochs, ponds, sea)? Justify your choice of sampling method.
2. You are sampling a Solid Geology map in which there are 3 main rock types present. What sampling technique are you likely to employ and why?
3. You are sampling a map area in which the valleys run roughly east-west. Explain why systematic sampling might be inappropriate.
4. You are sampling vegetation along a transect using a quadrat. What type of sampling technique is involved.
5. You have to sample the sediments in the face of the stream bank in Fig. 5. Which sampling method would you choose and why?

Fig. 4

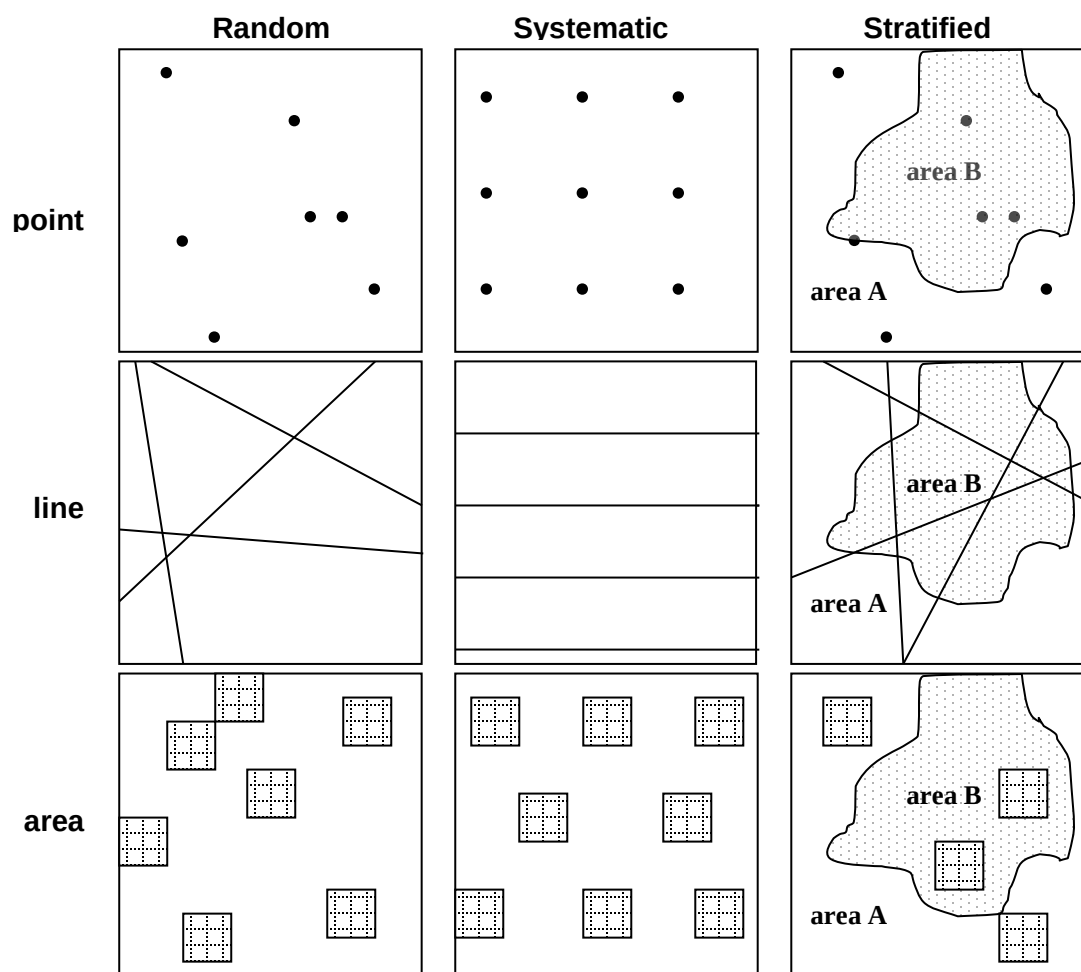


Fig. 5



Testing for sampling errors :

Any **mean** calculated from population data is the **true mean**. Sampled data only gives a **sample mean**. There is inevitably **sampling error**. This is where testing for sampling error comes in. There are a number of methods available, but the ones you need to know are **variance**, the **standard error of the mean** and the **Chi²-test** (note that Chi² is also used to test relationships).

Variance :

The larger the sample, the smaller the deviation around the mean. Where samples are small in number, a check should be carried out to verify whether the sample is '**representative**' of the population. Variance is one such check. It is calculated as :-

$$\text{variance } (\sigma^2) = \frac{\sum (x - \bar{x})^2}{n}$$

or, put in words, the **variance** is the sum of the 'difference from the mean of individual items' divided by the 'number of items in the data set'.

N. B. : **variance** is the method used where there is 'more than one sample of a population', in order to check whether the **sample means** of all samples coincide (i.e. come close to the **true mean**).

Coefficient of Variation (V) :

A very useful measure of variability is the **coefficient of variation (V)**. It is simply the standard deviation expressed as a percentage of the mean.

$$V = \frac{SD}{\bar{x}} \times \frac{100}{1}$$

The value of the coefficient of variation is that it permits the easier comparison of data sets being in percentage form. It is also very useful when isopleth or choropleth mapping variabilities, e.g. of rainfall.

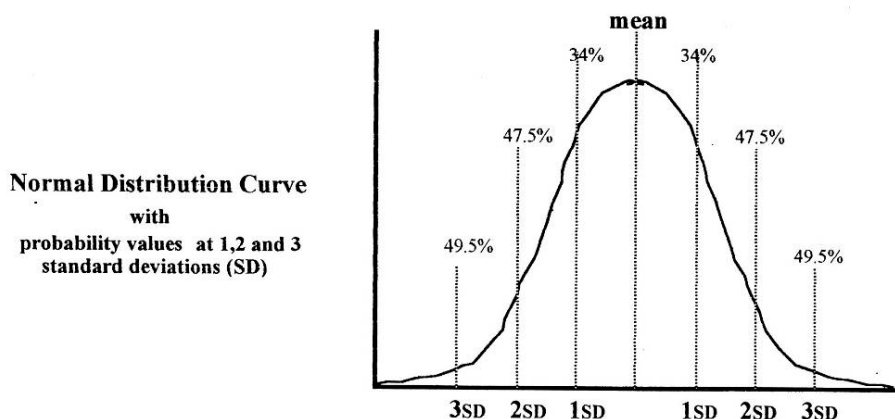
Standard error of the mean :

The second method used to check the validity of one's sample is to calculate the **standard error of the mean**. It seeks to measure how close the **sample mean** lies to the **population mean** by examining the size of the deviation around the mean. In a normal distribution the probability is that :-

- it is unlikely that more than 68% of values will exceed 1 SD from the mean and . . .
- it is very unlikely that more than 95% will differ by more than 2 SD's from the mean and . . .
- it is highly improbable that any value within the sample will lie outside 3 SD's

Standard error of the mean is based on the assumption of **normal distribution** of data, as seen in Fig. 6. This is vital to what comes next, when dealing with inferential statistics.

Fig. 6



Calculation of the standard error of the mean = $\frac{\sigma}{\sqrt{n}}$

$$\frac{s}{\sqrt{n}}$$

i.e. the sample's standard deviation divided by the square root of the number of items in the sample.

Probability :

Probability is encountered frequently, mainly due to the complexities of the real world. Statistics provide the best method for testing probabilities. We cannot say for certain that **all** 1st year pupils 'go bananas' on the full moon ; there may be a **probability**, however, that most do! We expect that most commuters travel the shortest route to and from home, but we cannot say that for definite. Some may travel a longer route occasionally e.g. to pick up a message or just to enjoy the change of scenery. Knowing the **level of probability** helps us to predict the outcome in many situations. This is where the application of standard error of the mean to normally distributed data comes in.

When **standard error of the mean** is applied to normally distributed data on a scale of 0 to 1, the probability that I will swim the English Channel is 0 ; that is absolutely certain, because I can't swim! The probability that I will die is 1 ; we all do! If I tossed a coin, the probability is that it would land heads down 50% of the time or, on the scale 0 to 1, a 0.5 probability.

It is generally better to express any probability by its percentage significance as it makes it easier to understand. So any result lying within the 0.05 or **95%** level is '**significant**' ; any within the 0.01 or **99%** level is '**very significant**' ; and any within the 0.001 or **99.9%** level is '**highly significant**'.

In testing probabilities there are three terms you need to be clear on :

- **probability** - the likelihood of something being 'true' or 'false'.
- **significance** - the strength of the probability.
- **degrees of freedom** - the limits set on any significance values related to the total number of items within the sample tested (n) or (n – 1) even (n – 2), depending on the statistical test and the number of data sets used.

Inferential Statistics

So far we have only considered **descriptive statistical techniques**. We now move on to the area of **inferential statistics**.

As the name suggests, inferential statistics **infer** some relationship. It hints at the presence or absence of a relationship between two sets of data. This is useful in that it allows for **hypothesis testing** in which 'cause and effect' **relationships** are sought.

The main inferential tests used are :

- **chi-squared test (χ^2)** - with **nominal data** in frequency format
- **Spearman's Rank correlation coefficient** - with **ordinal data** or interval data ranked in ordinal form (good with skewed data)
- **Pearson's Product correlation coefficient** - with **interval data** (normally distributed)
- **linear regression** - with **interval data** (normally distributed) can provide a predictive tool

Chi-squared Test - χ^2 :

So often the data collected in the 'field' is **nominal data**, e.g. rock type, angularity, traffic types, farm crops. The problem then arises as to how best to process that data. The **Chi-squared test (χ^2)** provides one of the best solutions, particularly where the data is :

- in frequency format
- has a total number of frequencies exceeding 20
- and no category with less than 5 frequencies

An example of the kind of data involved is shown in Table 4. Here, data was collected on the presence of water in gullies developed on scree slopes. Some of these gullies were directly linked to stone chutes above, others were simply developed on scree. A total of 64 gullies were observed, so the total number of frequencies well exceeds the minimum total number of frequencies i.e. 20. Two categories, however, have less than 5 observed frequencies – streamflow/chute at 4 and streamflow/scree at 1. Any Chi² result in this instance would be unreliable. While the omission of the streamflow categories is the best solution in this instance, it is sometimes possible to consolidate categories.

Table 4

Type of water presence	chute	scree
streamflow	4	1
seepage at head	8	6
seepage in gully	9	5
no water present	7	24

The equation for Chi-squared (χ^2) is :

$$\text{Chi-squared} - \chi^2 = \sum (O - E)^2 / E$$

where :

O = what has been **observed**

E = what would be **expected**, if no difference/relationship existed.

Worked example of Chi² for one data set :

Coire orientations are set out as shown in Table 5(a). Note that 3 extra columns have been added to enable the calculation of Chi². The first step is to set up a **null hypothesis (H₀)** in order to test whether there is a significant difference in coire orientations or not. The null hypothesis is written as follows :

H₀ : that there is **no** significant difference in the orientation of coires.

If the null hypothesis (H₀) proves untenable, then an **alternative or working hypothesis (H₁)** will replace the null hypothesis and is written as :

H₁ : that there is **a** significant difference in the orientation of coires.

The calculation is simple where one data set is concerned. The **expected values (E)** are the sum of the **occurrences (Σ(O))** which are 84 here, divided by the number of categories (**n**) in this case 4, i.e. 84/4. So the 'E' value (**or expected value**) is 21 for each category (see Table 5(a)).

Table 5(a)

orientation	no. of coires		(O - E)	(O - E) ²	(O - E) ² /E
	O (observed)	E (expected)			
NE	37	21			
SE	22	21			
SW	6	21			
NW	19	21			

n = 4 Σ (O) = 84

Σ(O - E)²/E =

Having obtained the **expected values (E)**, the calculations are then completed for **(O – E)**, **(O – E)²** and **(O – E)²/E**. Lastly, the **(O – E)²/E** values are summed to give the final result for Chi² in this worked example of 23.14 (see Table 5(b))

The **Chi² (χ²)** result is then compared to those in a table of critical values at '**n - 1**' degrees of freedom, i.e. the 'total no. of categories (**n**) minus one'. So the **degrees of freedom (df)** equals **3** (i.e. (4 – 1)).

Table 5(b)

orientation	no. of coires		(O – E)	(O – E) ²	(O – E) ² /E
	O (observed)	E (expected)			
NE	37	21	16	256	12.19
SE	22	21	1	1	0.05
SW	6	21	-15	225	10.71
NW	19	21	-2	4	0.19
n = 4	Σ (O) = 84		Σ(O – E) ² /E = 23.14		

Because the result **exceeds** the critical value in the table for the given degrees of freedom at the required significance level, then the **null hypothesis** must be rejected and replaced by the **alternative hypothesis**. The final result in this example is summarised as follows (the same pattern should be applied elsewhere when summarising any Chi² result) :

"The Chi² test result of 23.14 exceeds the 0.001 or 99.9% significance level at 3 degrees of freedom of 16.27. The null hypothesis must be rejected and replaced by the alternative hypothesis – that there is a highly significant difference in coire orientations"

If the result had fallen below the critical value, then the null hypothesis would have to be accepted. The description of the result would be as follows :

"The Chi² test result of 4.12 (e.g.) lies below the 0.01 or 90% significance level at 3 degrees of freedom of 6.25. The null hypothesis stands – there is no significant difference in coire orientations"

You are not finished, however, at that point! You need to go on to explain why a highly significant difference in coire orientations occurs ; is it due to prevailing winds drifting snow into sheltered hollows, or does shade from the sun play a part, or is there another factor involved? Even if there proved to be no significant difference, the question still remains as to why.

Note at this point that the expected (E) values for one data set need not always be the same for each data item, as it is here with E = 21. It may be that the data set is being compared to **population ratios**, in which case the ratios are entered in the **E** column.

Testing 2 or more variables using χ² :

Chi² opens the possibility of testing for relationships when more than one variable is involved. The expected values are calculated differently however. Assume you want to test whether rock type affects the angularity of the clasts produced : there are now 2 variables – **rock type** and **angularity**. The first step is to write the **null hypothesis** as follows :

H₀ : that there is **no** relationship between rock type and the angularity of the clasts produced.

If the null hypothesis (**H₀**) proves untenable, then the alternative hypothesis (**H₁**) replaces the null hypothesis and is written as :

H₁ : that there is **a** relationship between rock type and the angularity of the clasts produced.

The data is laid out in a table as in Table 6(a) below (omitting those parts in italics) :

Table 6(a)

	angular		rounded		rows
	O	E	O	E	R
granite	12		38		$\sum R$ (granite)
sandstone	14		16		$\sum R$ (sstn)
schist	34		29		$\sum R$ (schist)
columns K	$\sum K$ (angular)		$\sum K$ (rounded)		(total)

Procedure :

- sum the rows (**R**) for 'granite', 'sandstone' and 'schist'
- sum the columns (**K**) for the 'angular' and the 'rounded' (see Table 6(b))

Table 6(b)

	angular		rounded		rows
	O	E	O	E	R
granite	12		38		50
sandstone	14		16		30
schist	34		29		63
columns K	60		83		143

- for each **expected (E)** value, multiply the $\sum R$ by the $\sum K$ and divide by the **total sum of R and K** ; so the expected value for granite/angular is (50x60)/143 which equals 20.98, for sandstone/rounded it is (30x83)/143 which is 17.41. The full **expected (E)** values are shown in Table 6(c) :

Table 6(c)

	angular		rounded		rows
	O	E	O	E	R
granite	12	20.98	38	29.02	50
sandstone	14	12.59	16	17.41	30
schist	34	26.43	29	36.57	63
columns K	60		83		143

- list the rock types/angularity as follows and work out the following equation for each :

$$\begin{array}{llll}
 \text{granite / angular} & = & (O - E)^2/E & = 3.84 \\
 \text{sandstone / angular} & = & (O - E)^2/E & = 0.16 \\
 \text{schist / angular} & = & (O - E)^2/E & = 2.17 \\
 \text{granite / rounded} & = & (O - E)^2/E & = 2.78 \\
 \text{sandstone / rounded} & = & (O - E)^2/E & = 0.11 \\
 \text{schist / rounded} & = & (O - E)^2/E & = 1.57 \\
 \hline
 \chi^2 & = & \sum (O - E)^2/E & = 10.63 \quad (\text{total})
 \end{array}$$

The χ^2 value of **10.63** must then be compared to those given in the table of critical values for Chi-squared at the respective **degrees of freedom (df)** for the data used. In this case **df = 2** , since **df = (R - 1)(K - 1)** i.e. (the number of rows (R) minus 1) times (the number of columns (K) minus 1). The final result is described as follows :

"The χ^2 value of **10.63** exceeds the 0.01 or 99% significance level at 2 degrees of freedom of 9.21. The null hypothesis must therefore be rejected and replaced by the alternative hypothesis - : that there is **a** relationship between rock type and the angularity of the clasts produced".

Again you need to go further and explain the reasons why a relationship exists between rock type and angularity – is one rock type more prone to attrition than another ; if so which?

NOTE : the two main applications of the **Chi-squared Test (χ^2)**

- i. it can be used to test whether a sample (in frequency/or nominal data form) is representative of the 'population'. If the H_0 proves correct, the sample can be accepted as representative. In this case the H_0 is written as :-

H_0 = "there is no significant difference between the sample and the total population"

- ii. it can be used to test whether a relationship exists between variables, eg. the no. of shoppers and the types of shop. In this case the H_0 is written as :-

H_0 = "there is no relationship between the no. of shoppers and the types of shops visited.

If the **null hypothesis (H_0)** must be rejected, the **alternative** or **working hypothesis (H_1)** is given instead.

ACTIVITIES 7

1. Wind directions were recorded over the period of one year and the results are presented in Table 7. Test whether there is a significant difference in wind directions using χ^2 .

Table 7

	Frequency of wind directions				
	O (observed)	E (expected)	(O - E)	(O - E) ²	(O - E) ² /E
NE	67				
SE	74				
SW	102				
NW	98				

n = Σ (O) = Σ (O - E)²/E =

2. Observations were made on the flatness or roundness of three different rock types. The results are shown in Table 8. Draw up Table 8, then :

(a) Set up a null hypothesis to test the data.

(b) Calculate the χ^2 value using the formula :

$$\chi^2 = \Sigma(O - E)^2/E \quad df = (R - 1)(K - 1)$$

(c) Describe the result in terms of the null hypothesis (use the χ^2 significance tables in the Appendix of the booklet).

(d) Explain the reasons for your results.

Table 8

	flat		oblong		square		(rows)
Rock type	O	E	O	E	O	E	R
schist	27		19		16		
quartzite	18		20		35		
granite	7		26		31		
K (columns)							

schist/flat	= (O - E) ² /E	=
schist/oblong	= (O - E) ² /E	=
schist/square	= (O - E) ² /E	=
quartzite/flat	= (O - E) ² /E	=
quartzite/oblong	= (O - E) ² /E	=
quartzite/square	= (O - E) ² /E	=
granite/flat	= (O - E) ² /E	=
granite/oblong	= (O - E) ² /E	=
granite/square	= (O - E) ² /E	=
		$\Sigma(O - E)^2/E =$ _____

EXTENSION

Read p.31 to 35 of Skills & Techniques for Geography A-level (EPICS). Do the Review on p.33.

Correlation and Correlation Coefficients

Correlation is simply '**co-** relation' ('**co-**' is derived from the Latin '**cum**' meaning '**with**') or a relationship 'with'. In correlation a meaningful relationship is sought between one variable and another, e.g. between the levels of poverty and the % unemployed.

But a meaningful relationship depends firstly on the presence of a statistical one. Yet a statistical relationship may occur where no meaningful relationship exists. For example, you might find a strong positive statistical relationship between the length of the queue at the Tuck Shop and the height of the 6th years on duty! But such relationships may be spurious (you could always test it though!)

Prior to undertaking any statistical calculations, it is worth drawing a **scatter graph** of the data involved. The graph should be drawn with the **independent variable** along the **x-axis** and the **dependent variable** along the **y-axis**. A scatter graph saves a lot of trouble, especially if ultimately little or no relationship exists. It may also highlight a **curvilinear relationship** which would require logarithmic conversion prior to the application of more robust statistical tests. Much geographical data does require logarithmic conversion - the best examples are those of **river discharge measurements** and **sediment size analysis** (see fig. 7).

Fig. 7

direct relationship



inverse relationship



no relationship



Two methods are outlined here for checking for a statistical relationship – Spearman's Rank and Pearson's Product Moment correlation coefficients.

Spearman's Rank Correlation Coefficient (Rs) :

Spearman's Rank correlation coefficient (Rs) is a very useful tool particularly when dealing with skewed data, e.g. GDP and GNP are heavily skewed and are frequently used when testing relationships related to development status indices.

But **Spearman's Rank** can only be used with **ordinal data** and with **no less than 7 observed pairs of samples**. It is easy, however, to convert interval data to ordinal format.

Worked example of Spearman's Rank (Rs) :

In this example the relationship between life expectancy and levels of adult obesity in MEDCs is examined. The data is set out as shown in Table 9.

- Set up a null hypothesis – that there is no relationship between life expectancy and % adult obesity.
- The first procedure involves **rank ordering both sets of data**. The data can be ranked either from lowest to highest or vice versa BUT both sets must be ranked in the same way. In this case the data sets are ranked from lowest to highest. Note that in the case of equal data values, the ranking is given as 'n=' for each item of the same value ; so 1st, 2nd, 3rd =, 3rd =, then to 5th. The last ranking must equal the total number of items ; if not, there is an error in the ranking.
- Having done that, **rank Y** is subtracted from **rank X** to give the **difference (d)**. The difference is then squared (**d²**) in order to get rid of the negatives ; (**d²**) is then added to give $\sum d^2$.
- The values for 'n' and the $\sum d^2$ are inserted into the equation and the calculation completed.

$$Rs = 1 - \frac{6\sum d^2}{n^3 - n}$$

Table 9

MEDCs	% adult obesity X	life expectancy Y	Rank order of X	Rank order of Y	d	d ²
Japan	3.1	83	1	15	-14	196
Norway	6.3	82	2	14	-12	144
Italy	9	81	3	12=	-9	81
Denmark	9.4	78	4	3=	1	1
Sweden	9.9	81	5	12=	-7	49
Netherlands	11.1	79	6	5=	1	1
Belgium	12.7	79	7	5=	2	4
Germany	12.9	79	8	5=	3	9
Ireland	13	79	9	5=	4	16
Spain	13.2	80	10	10=	0	0
Canada	14.9	80	11	10=	1	1
Czech Rep.	15	76	12	1	11	121
USA	32.1	77	15	2	13	169
Greece	22.1	79	13	5=	8	64
Finland	22.4	78	14	3=	11	121
n = 15					$\sum d^2 =$	977

$$Rs = 1 - \frac{(6 \times 977)}{(3375 - 15)}$$

$$Rs = 1 - \frac{(5862)}{(3360)}$$

$$Rs = 1 - 1.745$$

$$Rs = \underline{\underline{-0.745}}$$

- Finally the result for **R_s** , in this case – **0.745**, is compared to the significance values for Spearman's Rank at the correct degrees of freedom ($df = (n - 2)$, so $df = 13$) and presented as follows :

"The result for Spearman's Rank (R_s) of -0.745 lies just below the 99% significance level of 0.746 but above the 95% level of 0.567 at 13 degrees of freedom. The null hypothesis must therefore be rejected and replaced by the alternative hypothesis 'that there is a significant relationship between life expectancy and % adult obesity and that it is an inverse relationship'. In other words, as the % adult obesity increases so life expectancy decreases".

Again, as in χ^2 , the result (**R_s**) is simply the statistical relationship. It is necessary to go on to explain how that relationship exists or otherwise. In the example here we probably were expecting an inverse relationship between life expectancy and the % adult obesity (*them 'wot' over-eat, get fat and suffer heart attacks!*). The relationship, however, is not as clearcut as that. In the Activities that follow you are required to carry out a similar test on Less Economically Developed Countries (LEDCs).

Note in the passing that you may be given degrees of freedom at 'n', $(n - 1)$ or $(n - 2)$. Use accordingly.

ACTIVITIES 8

1.
 - (a) Set up a null hypothesis to test for a relationship between life expectancy and adult obesity in LEDCs (data in Table 10).
 - (b) Calculate the Spearman's Rank correlation coefficient (R_s) value for the data and describe the significance of the relationship in terms of the null hypothesis.
 - (c) Account for the relationship.

Table 10

LEDCs	Life expectancy X	Adult obesity % Y	Rank order of X	Rank order of Y	d	d ²
Ethiopia	51	0.3				
India	63	0.5				
Nepal	61	0.5				
Malawi	46	1.1				
Zambia	40	1.5				
Indonesia	67	2.3				
Burkina Faso	49	2.4				
China	73	2.9				
Nigeria	48	2.9				
Kenya	45	3.2				
Ghana	57	4				
Philippines	67	4.6				
Brazil	72	11				
Bolivia	65	15.1				
Peru	72	15.7				
Egypt	68	22.8				
Mexico	74	23.4				

n =

$\sum d^2 =$

$$R_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

2. Set up a null hypothesis and complete the calculation of **Spearman's Rank correlation coefficient (Rs)** for the data in Table 11. Then describe and analyse your findings.

$$R_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

Table 11

Sample sites	Distance downstream X	Discharge Y	Rank X	Rank Y	d	d ²
1	5.7	0.205				
2	6.2	0.399				
3	6.6	0.246				
4	7.0	0.422				
5	7.2	0.404				
6	7.5	0.727				
7	7.9	1.324				
8	8.1	0.813				
9	8.3	1.368				
10	8.5	1.358				
11	10.5	1.917				
12	10.7	1.160				
13	11.1	1.240				
14	11.5	1.895				
15	12.3	5.691				

EXTENSION

Read p.25 to 27 of Skills & Techniques for Geography A-level (EPICS).

Analysis of correlation coefficient results :

The answers for Spearman's Rank and Pearson's Product correlation coefficients can never exceed 1 or -1. So, for example, answers of 1.356 or -1.462 are impossible. ; your calculations must be wrong! Also, in both cases, answers must be checked against their respective significance tables at the correct degrees of freedom. Generally, the closer the answer is to **1** or **-1** the stronger the correlation. Where the answer is in the '**negative**', the relationship is **inverse** ; in other words as X increases, so Y decreases. Be very careful, therefore, not to ignore the minus!

You have calculated your **correlation coefficient** and found a strong relationship exists. **So what?** All you have at this point is clear proof of a **statistical relationship** ; you now need to justify the existence of a '**cause and effect**' relationship. If you have already thought out your initial hypothesis well, there should be little problem in justifying it given your statistical evidence.

Do note, however, that the relationship between 'two chosen variables' may conceal a third variable unknown to you. For example, a strong relationship exists between river discharge and channel width/depth ratio ; but there would be, since the calculation of discharge is 'velocity times the area of the channel cross-section (i.e width x depth)'. Vice versa, where little or no relationship exists, it does not necessarily indicate the lack of one ; the presence of hidden variable(s) can cancel the one you are measuring.

"Can I trust my results then?" **Yes!** You have proved a statistical relationship, which no one can argue against. It is now up to you to analyse your result well.

Pearson's Product Moment Correlation Coefficient (r) :

Pearson's Product Moment correlation coefficient (r) is more statistically robust than Spearman's Rank ; but it involves much more maths, and also assumes the presence of normally distributed data. There are at least three equations that can be used and in each case the way the data is tabulated differs :

1.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}}$$

2. where 'dx' and 'dy' are $(x - \bar{x})$ and $(y - \bar{y})$ respectively

$$r = \frac{\sum dx \cdot dy}{\sqrt{(\sum dx^2 \cdot \sum dy^2)}}$$

3. where capital 'X' and 'Y' are $(x - \bar{x})$ and $(y - \bar{y})$ respectively

$$r = \frac{\sum XY}{\sqrt{(\sum X^2 \cdot \sum Y^2)}}$$

My personal preference is **equation 3** and it is that which is applied to the tabulation of data shown in the worked example that follows. (**N.B.** : capital 'X' and 'Y', used in equations, are not small case 'x' and 'y').

Worked example of Pearson's Product Moment correlation coefficient :

The example used here (see Table 12) seeks to test whether a relationship exists between water temperature (°C) and the pH of the water in a mountain stream.

- (a) Draw up a working hypothesis (H_1) : *"that there is a relationship between variable 'x' (water temperature) and variable 'y'" (pH of water)"*.
- (b) Set out the data in a table similar to that below.
- (c) Work out the mean of 'x' and the mean of 'y'.
- (d) To find the values for 'X' and 'Y' subtract the **mean** from the value of 'x' and 'y' respectively.
- (e) Square all values of 'X' and 'Y' (to get rid of the minuses) for columns 'X²' and 'Y²'
- (f) Multiply all values of 'X' by 'Y' to complete column 'XY'
- (g) Sum all the values required along the foot of the table.
- (h) Insert the values as appropriate in the equation and calculate it.
- (j) Check the result against a table of significance for Pearson's at '(n – 1)' degrees of freedom (see p. 31).

Table 12

	Water temp.(°C) x	pH of water y	X = (x – bar x)	Y = (y – bar y)	X ²	Y ²	XY
1	1.5	5.50	-1.16	-1.216	1.346	1.479	1.411
2	2.0	5.96	-0.66	-0.756	0.436	0.571	0.499
3	2.5	7.27	-0.16	0.554	0.026	0.307	-0.089
4	2.5	7.38	-0.16	0.664	0.026	0.441	-0.106
5	2.8	7.25	0.14	0.534	0.019	0.285	0.075
6	3.0	7.03	0.34	0.314	0.115	0.098	0.107
7	3.0	6.89	0.34	0.174	0.115	0.030	0.059
8	3.0	6.68	0.34	-0.036	0.115	0.001	-0.012
9	3.2	6.57	0.54	-0.146	0.292	0.021	-0.079
10	3.1	6.63	0.44	-0.086	0.194	0.007	-0.038
n = 10 $\sum x = 26.6$ $\sum y = 67.16$ $\bar{x} = 2.66$ $\bar{y} = 6.716$ $\sum X^2 = 2.684$ $\sum Y^2 = 3.240$ $\sum XY = 1.826$							

$$r = \frac{\sum XY}{\sqrt{(\sum X^2 \cdot \sum Y^2)}}$$

$$r = \frac{1.826}{\sqrt{(2.684 \times 3.240)}}$$

$$r = \underline{\underline{0.619}}$$

A word of caution when using Pearson's Product – it is very easy to make a simple error in the calculation of X^2 , Y^2 and XY and also in the summing up of the same. Don't rush it ; take time to double check as you go. If an error does occur, scan each column for a value that seems inordinately large or small and recalculate it. There is no easy way round this, which is why it is good policy to graph the data first as a **scatter graph** before setting out on the computation of Pearson's.

ACTIVITIES 9

The figures for **distance downstream** and **stream channel width** are given in Table 13 following :

1. (a) What might you expect the relationship to be between distance downstream and the width of the stream channel?
- (b) State the relationship in terms of a working hypothesis (H_1).
- (c) Using **Pearson's Product Moment corr. coeff.**, test the **working hypothesis (H_1)**.
- (d) Describe the relationship in terms of its significance - check the result against the table of **significance** at **(n – 1) degrees of freedom** in the tables at the end.
- (e) Draw a **scatter graph** of the data and insert a 'best-fit' line. Why might it have been better to have done the scatter graph first?

Table 13

Sample sites	downstream (km) x	channel width y	X (x - $\bar{x}$)	Y (y - $\bar{y}$)	X ²	Y ²	XY
1	0.5	0.97					
2	2.7	0.92					
3	3	1.4					
4	3.9	1.61					
5	0.7	0.56					
6	1.5	0.63					
7	1.8	1.1					
8	4.3	1.36					
n =	$\sum x =$ $\bar{x} =$	$\sum y =$ $\bar{y} =$		$\sum X^2 =$	$\sum Y^2 =$	$\sum XY =$	

$$r = \frac{\sum XY}{\sqrt{(\sum X^2 \times \sum Y^2)}} \quad \text{degrees of freedom } (n - 1)$$

2. Examine the data below in Table 14 on life expectancy and the number of deaths from malaria per 1000 of the population in several African countries.

Table 14

Country	No. of malarial deaths per 1000/pop. x	Life expectancy y	X	Y	X ²	Y ²	XY
Angola	0.938	41	0.632	-5.77			
D.R.of Congo	0.336	44	0.030	-2.77			
Eritrea	0.022	55	-0.284	8.23			
Ghana	0.169	57	-0.137	10.23			
Kenya	0.005	48	-0.301	1.23			
Malawi	0.676	40	0.370	-6.77			
Mozambique	0.189	42	-0.117	-4.77			
Nigeria	0.050	44	-0.256	-2.77			
Somalia	0.001	47	-0.305	0.23			
Sudan	0.088	57					
Tanzania	0.441	46					
Uganda	0.411	49					
Zambia	0.656	38					
n = 13	$\sum x = 3.982$ $\bar{x} = 0.306$	$\sum y = 608$ $\bar{y} = 46.77$			$\sum X^2 =$	$\sum Y^2 =$	$\sum XY =$

$$r = \frac{\sum XY}{\sqrt{(\sum X^2 \times \sum Y^2)}} \quad \text{degrees of freedom } = (n - 1)$$

- (a) Set up a null hypothesis to test the data for a relationship. State also your choice of probability (90%, 95%, 99% or 99.9%).

- (b) Calculate Pearson's Product Correlation Coefficient for the data using the equation given below.
- (c) Describe your result in terms of the null hypothesis.
- (d) Malaria is only one factor in life expectancy in Africa. Identify and explain some of the other factors that should be taken into consideration when examining the life expectancies of the countries in Africa (use an Atlas for further reference).

Linear Regression :

The value of correlation is that it indicates a relationship between two variables. **Linear regression** goes a step further, in that it permits the **prediction** of what is likely to happen at requested values of either variable. It is based on working out a mathematically accurate '**best-fit line**'. The **regression equation** takes the following form :-

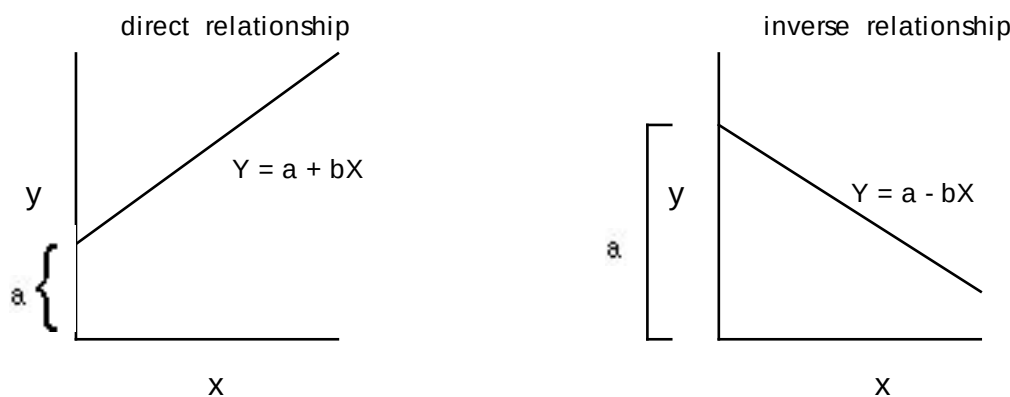
$$Y = a + b.X$$

or

$$Y = a - b.X$$

where '**a**' is the point of intersection on the '**Y**' axis, and the slope is '**b**'. The graphs in Fig. 8 show this.

Fig. 8



As the values for **X** and **Y** are present in the data, it only requires to find the values for '**a**' and '**b**'. The method used for calculating linear regression is as follows :-

Equation for calculation of 'b' in regression :

Again three versions of the equation for calculating the value of '**b**' are shown ; all come to the same thing.

$$1. \quad b = \frac{\sum XY - \frac{\sum(X) \cdot \sum(Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$2. \quad b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$3. \quad b = \frac{\sum(dX \cdot dY)}{\sum(dX)^2} \quad \text{where 'd' is the deviation from the mean}$$

To find the value of 'a', the calculated value for 'b' is inserted into the equation below along with the mean values for 'X' and 'Y' :-

$$a = \bar{Y} - b.\bar{X} = \frac{\sum Y}{n} - b.\frac{\sum X}{n}$$

A worked example of linear regression :

Pearson's Product Moment Correlation Coefficient suggests a significant relationship of 95% probability between stream channel gradient and pebble mean size. It has been decided that it is worth carrying on with the calculation of **regression line** (see Table 15). The process is as follows :

Table 15

reach	gradient X	pebble mean size Y	X ²	XY
1	1.1	247	1.21	271.7
2	1.3	273	1.69	354.9
3	1.5	639	2.25	958.5
4	1.8	328	3.24	590.4
5	1.4	958	1.96	1341.2
6	1.3	589	1.69	765.7
7	3.1	884	9.61	2740.4
8	3.0	1105	9.00	3315.0
n = 8	$\sum X = 14.5$ $\bar{X} = 1.81$	$\sum Y = 5023$ $\bar{Y} = 627.87$	$\sum X^2 = 30.65$	$\sum XY = 10337.8$

- The data is set out in a table with two extra columns, one for X² and the other for XY
- Complete the calculations for X², XY, $\sum X$, $\sum Y$, $\sum X^2$ and $\sum XY$; also $\bar{X}$ and $\bar{Y}$
- Insert the required values into the equation for 'b'. Be very careful that you do not confuse the $\sum X^2$ with $(\sum X)^2$

$$b = \frac{\sum XY - \frac{\sum(X) \cdot \sum(Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

so :

$$b = \frac{10337.8 - \frac{(14.5 \times 5023)}{8}}{30.65 - \frac{(14.5)^2}{8}}$$

$$b = \frac{10337.8 - 9104.19}{30.65 - 26.28}$$

$$b = \frac{1233.61}{4.37}$$

$$b = \underline{\underline{282.29}}$$

- Having found the value of 'b', insert the values for 'b', $\bar{X}$ and $\bar{Y}$ into the equation below to find the value of 'a' :-

$$a = \bar{Y} - b.\bar{X}$$

$$= 627.87 - (282.29 \times 1.81)$$

$$a = \underline{116.92}$$

- Insert the computed values for 'a' and 'b' into the equation for linear regression below :

$$Y = a + b.X$$

so the linear regression equation for stream gradient/pebble size is :

$$Y = \underline{116.92 + (282.29 \times X)}$$

The main value of undertaking **linear regression** in statistical analysis is to permit the prediction of **variable Y** given a known value of **variable X**. So in this example it would now be possible to calculate the mean size of pebbles at a given stream channel gradient. So, for example, for a stream channel gradient of 3.0, the expected pebble mean size would be :

$$Y = 116.92 + (282.29 \times 3.0) = \underline{963.97}$$

ACTIVITIES 10

- The data in Table 16 is taken from a study into the soil properties of a West Highland mountain.
 - What relationship might you expect between **soil pH** and the **the amount of organic matter** in the soil? Why might you expect that relationship?
 - Using the following data (Table 16), calculate the **linear regression** estimate for soil pH against organic matter content in the soil.
 - Predict the soil pH that would be expected if the amount of organic matter was 2gm.
 - What other factors could affect soil pH?

Table 16

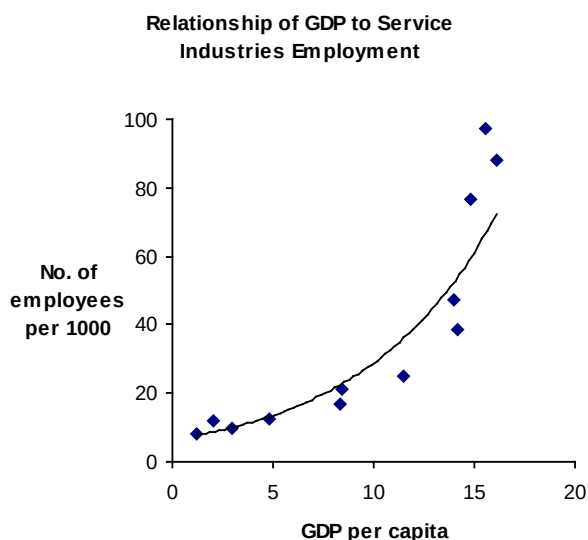
Site	Organic matter (gm) X	Soil pH Y	X ²	XY
1	0.43	6.8		
2	0.99	6.7		
3	0.56	6.3		
4	1.2	5.6		
5	1.23	5.9		
6	1.04	5.8		
7	0.92	6		
8	1.94	5.5		
9	1.11	6.2		
10	1.13	6.3		
n =	$\sum X =$ $\bar{X} =$	$\sum Y =$ $\bar{Y} =$	$\sum X^2 =$	$\sum XY =$

Non-linear Relationships :

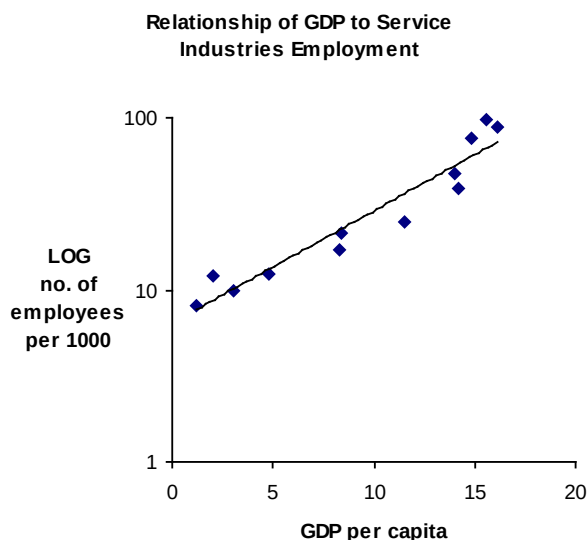
Quite often in geography relationships are **non-linear** or **curvilinear**. The use of simple **linear regression**, in such cases, gives poor results - remember, that more advanced statistical techniques assume **normal distribution of data** and in curvilinear relationships **skewing** is present.

All is not lost however. Such data can be transform using **logarithms**. This can be done either by the use of **semi-log** or **double-log** graph paper if plotting a **scatter graph** or by the conversion of the 'y' and/or 'x' variables to '**log y**' and/or '**log x**'. In the example below (Fig. 9(a) and (b)) the data for the number of employees per 1000 in service industries is quite heavily skewed ; the logarithmic conversion straightens out the curve, so making it possible to apply Pearson's Product and Linear Regression.

Fig. 9 : (a)



(b)



Example of how non-linear data is set out in a table for Pearson's Product (Table 17):

Note that an extra column is added for the data set to be logged, in this case it is the 'y' variable that needs logarithmically converted. If both 'x' and 'y' variables require log conversion, then two extra columns are added. Variables 'x' and/or 'y' are then log converted and used from then on instead of the raw data.

Table 17

GDP per capita (in '00's \$) x	No. of employees per 1000 in service industries y	(Log y)	X (x - $\bar{x}$)	Log Y (Log y - Log $\bar{y}$)	X ²	Log Y ²	X.LogY
2.0	12.0						
1.2	8.0						
14.8	76.4						
8.3	17.0						
11.5	25.0						
14.2	38.6						
14.0	47.3						
$\sum x =$	$\sum \log y =$			$\sum X^2 =$		$\sum X \cdot \log Y =$	
$\bar{x} =$	$\log \bar{y} =$			$\sum \log Y^2 =$			
n = 12							

The equation remains the same apart from the insertion of the log values in the appropriate places.

$$r = \frac{\sum X \cdot \log Y}{\sqrt{(\sum X^2 \times \sum \log Y^2)}}$$

Example of how non-linear data is tabulated and calculated for Linear Regression (Table 18) :

Note that extra columns are added for **log x** and **log y**. If only 'x' or 'y' variables require log conversion, then one extra column is added.

Table 18

Day	Streamflow (cumecs) x	Sed. discharge (tonnes/day) y	log x	log y	logx ²	logxy
1	125.4	12698.0				
2	70.0	1287.9				
3	68.9	2575.9				
n =	$\frac{\sum \log x}{\log \bar{x}} =$		$\frac{\sum \log y}{\log \bar{y}} =$		$\sum \log x^2 =$	$\sum \log xy =$

The calculation of '**log b**' uses the equation :

$$\log b = \frac{\frac{\sum \log xy - (\sum \log x \cdot \sum \log y)}{n}}{\frac{\sum \log x^2 - \frac{(\sum \log x)^2}{n}}{n}}$$

The calculation of '**log a**' uses the equation :

$$\log a = \frac{\frac{\sum \log y}{n} - \log b \left(\frac{\sum \log x}{n} \right)}{1}$$

And finally the values of '**log a**' and '**log b**' are inserted into the linear regression equation :

$$\log y = \log a + \log b \cdot (\log x)$$

N.B. : while you are not expected to deal with logarithmic calculations in the exam, you do need to know about them. It provides a way round the problems of skewed data and permits the use of the more robust statistical methods of Pearson's Product and Linear Regression. You may also wish to attempt it in your Geographical Study, particularly if you have data with some very high values that skew the distribution.

On **EXCEL**, the **function (f)** button gives you the option of finding out how skewed a set of data is under **skewness** in the statistical menu.

ACTIVITIES 11 :

The data in Table 19(a) refers to the streamflow in m³/secs (cumecs) and sediment discharge in metric tons/day over a 7 day period on a river in the Mid West of America.

Use of Spearman's Rank correlation coefficient (Rs) :

1. (a) Set up a null hypothesis to test the data.
- (b) Calculate Spearman's Rank Corr. Coeff. (Rs) to test the null hypothesis, using the equation below.

Table 19(a)

Day	Streamflow (cumecs) x	Sed. discharge (tonnes) y	Rank X	Rank Y	d	d ²
1	125.44	12698.0				
2	70.00	1287.9				
3	68.86	2575.9				
4	27.86	197.7				
5	41.44	1623.5				
6	13.38	25.4				
7	24.92	87.1				
n =			Σd ² =			

df = n

$$R_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

Use of Pearson's Product Moment correlation coefficient (r) on the same data :

2. (a) For the same data in Table 19(b), complete the calculation for Pearson's Product Moment Corr. Coeff. (r), using the equation below.

Table 19(b)

Day	Streamflow (cumecs) x	Sed. discharge (tonnes/day) y	X	Y	X ²	X ²	XY
1	125.4	12698.0	72.3	10055.8	5233.1	101119113.6	727436.6
2	70.0	1287.9	16.9	-1354.3	285.6	1834128.5	-22887.7
3	68.9	2575.9					
4	27.9	197.7	-25.2	-2444.5	637.1	5975580.3	61699.2
5	41.4	1623.5	-11.7	-1018.7	136.0	1037749.7	11878.0
6	13.4	25.4	-39.7	-2616.8	1577.7	6847642.2	103939.3
7	24.9	87.1					
n = 7	x = 371.9	y = 18495.5	X ² =			XY =	
	bar x =	bar y =	Y ² =				

df = (n - 1)

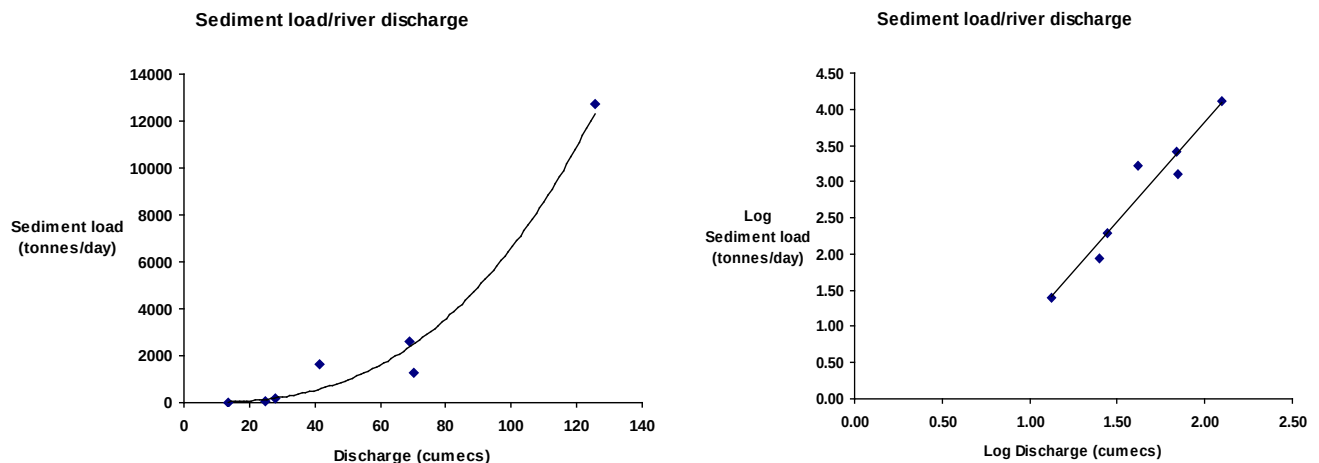
$$r = \frac{\sum XY}{\sqrt{\sum X^2 \times \sum Y^2}}$$

- (c) Describe both results (Rs and r) in terms of the null hypothesis.
- (d) Evaluate the validity of the statistical techniques used and the results obtained.

Use of scatter graphs on the same data :

3. Examine the two scatter graphs in Fig. 10.
 - (a) Explain why logarithmic conversion of the data is needed before applying to it both Pearson's Product and Linear Regression statistical techniques.
 - (b) Why is the drawing of a scatter graph a useful first step in the testing for relationships?
 - (c) What is the main purpose of using linear regression?

Fig. 10



Use of linear regression on the same data :

4. Now using the same data in Table 19(c), complete the calculation of the linear regression equation for the relationship of streamflow against sediment discharge.

Table 19(c)

Day	Streamflow (cumecs) x	Sed. discharge (tonnes/day) y	log x	log y	logx ²	logxy
1	125.4	12698.0	2.10	4.10	4.40	8.61
2	70.0	1287.9	1.85	3.11	3.40	5.74
3	68.9	2575.9	1.84	3.41	3.38	6.27
4	27.9	197.7	1.44	2.30	2.09	3.32
5	41.4	1623.5	1.62	3.21	2.62	5.19
6	13.4	25.4	1.13	1.40	1.27	1.58
7	24.9	87.1	1.40	1.94	1.95	2.71
n = 7			$\sum \log x = 11.37$		$\sum \log x^2 = 19.11$	
			$\sum \log y = 19.48$			$\sum \log xy = 33.42$

- (a) First calculate the value of 'log b' using the equation given below.

$$\log b = \frac{\frac{\sum \log xy - (\sum \log x \cdot \sum \log y)}{n}}{\frac{\sum \log x^2 - \frac{(\sum \log x)^2}{n}}{n}}$$

- (b) Now calculate the value of '**log a**' using the equation given below.

$$\log a = \frac{\sum \log y}{n} - \log b \left(\frac{\sum \log x}{n} \right)$$

- (c) Insert the values of '**a**' and '**b**' into the linear regression equation :

$$\log y = \log a + \log b.(\log x)$$

- (d) Predict, using the regression equation, the sediment discharge that would occur if the streamflow was at 130 cumecs.

Note : *x = 130 so you need to find the value of **log 130** before you apply it to the equation and then at the end convert back the result for **log y** to '**y**' using antilogs.*

EXCEL-ling Statistics

"Do I need to 'number-crunch' every data set I want to examine statistically, e.g. in my Geographical Study?" No! You do need to know how each statistical technique is worked out and you will be required to do a 'number-crunching' question in the exam. But for the Geographical Study (or Geog. Issue even) you can and should make free use of a good spreadsheet application such as **EXCEL**. Use the '**function**' tool '**f**', to be found in the **INSERT** draw-down menu. Click on '**Statistical**' and it gives a large menu of statistical options. Among those of most use are :

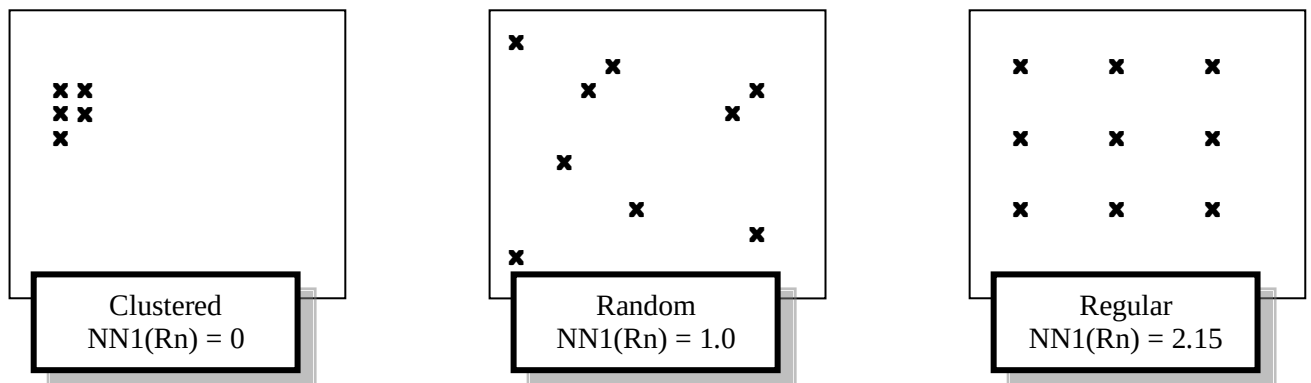
AVERAGE (mean) **STDEV** (standard deviation) **PEARSON** (Pearson's Product)
QUARTILE (median, upper and lower quartiles, maximum and minimum)

Nearest Neighbour Analysis :

Spatial analysis may sound like something out of Star-trek, but is in fact a posh geographical term for the study of patterns and distribution over a large area or space. Geographers are very much involved in the study of such patterns. A pattern of settlement, for example, may be described as **dispersed**, **nucleated** or **linear**; trees may be clustered, regularly spaced or randomly spaced. One technique used to bring some form of objectivity to the analysis of pattern is that of **nearest neighbour analysis**.

The **nearest-neighbour index (Rn)** is derived by averaging the distance between each point and its nearest neighbour. The computed value can range between 0 and 2.15, where 0 indicates a highly clustered pattern and 2.15 a highly regular pattern (see fig.11).

Fig. 11



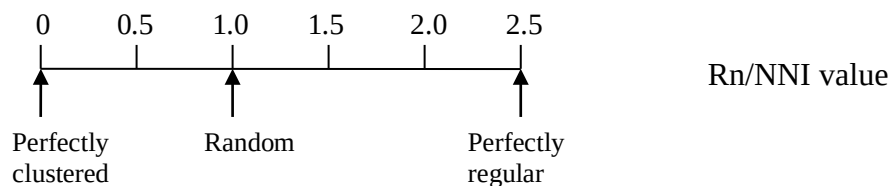
The equation for working out the nearest-neighbour index (**Rn**) is :-

$$Rn = 2\bar{D} \sqrt{(n/a)}$$

where $\bar{D}$ is the average distance between each point and its nearest neighbour ($\Sigma D/n$),
 n the number of points under study
and a the size of the area in metres² or kilometres².

The diagram for nearest-neighbour index (**Rn**) significance is :-

Diagram of linear NNI significance



While **nearest-neighbour** provides one method of spatial analysis, it does suffer from several problems which must be kept in mind when interpreting the results :-

- should measurements be straight-line distances or should they follow the line of the road or railway?
- why nearest neighbour, why not 2nd nearest or 3rd nearest?
- does the choice and size of the area affect the end result?
- does the ONE overall index (**Rn**) hide important sub-patterns?
- what is the effect of paired distributions, or linear distributions?
- even though the **Rn** may suggest a random pattern, is the pattern far from random but purely the function of soil type or relief?

Significance Tables for Chi², 'r' and 'Rs'

Chi-square

(significance values at 'n' degrees of freedom)

degrees of freedom (df)	0.10 90%	0.05 95%	0.01 99%	0.001 99.9%
1	2.71	3.84	6.63	10.83
2	4.61	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.51
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.09	26.12
9	14.68	16.92	21.67	27.88
10	15.99	18.31	23.21	29.59
11	17.28	19.68	24.73	31.26
12	18.55	21.03	26.22	32.91
13	19.81	22.36	27.69	34.53
14	21.06	23.68	29.14	36.12
15	22.31	25.00	30.58	37.70
16	23.54	26.30	32.00	39.25
17	24.77	27.59	33.41	40.79
18	25.99	28.87	34.81	42.31
19	27.20	30.14	36.19	43.82
20	28.41	31.41	37.57	45.31
21	29.62	32.67	38.93	46.80
22	30.81	33.92	40.29	48.27
23	32.01	35.17	41.64	49.73
24	33.20	36.42	42.98	51.18
25	34.38	37.65	44.31	52.62
26	35.56	38.89	45.64	54.05
27	36.74	40.11	46.96	55.48
28	37.92	41.34	48.28	56.89
29	39.09	42.56	49.59	58.30
30	40.26	43.77	50.89	59.70

Pearson's Product (r)

(significance values at 'n' degrees of freedom)

degrees of freedom (df)	0.05 95%	0.01 99%
1	0.997	1.000
2	0.950	0.990
3	0.878	0.959
4	0.811	0.917
5	0.754	0.875
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606
16	0.468	0.590
17	0.456	0.575
18	0.444	0.561
19	0.433	0.549
20	0.423	0.537
21	0.413	0.526
22	0.404	0.515
23	0.396	0.505
24	0.388	0.496
25	0.381	0.487
26	0.374	0.479
27	0.367	0.471
28	0.361	0.463
29	0.355	0.456
30	0.349	0.449

n	0.05 95%	0.01 99%
5	1.000	-
6	0.886	1.000
7	0.786	0.929
8	0.738	0.881
9	0.683	0.833
10	0.648	0.794
12	0.591	0.777
14	0.544	0.715
16	0.506	0.665
18	0.475	0.625
20	0.450	0.591
22	0.428	0.562
24	0.409	0.537
26	0.392	0.515
28	0.377	0.496
30	0.364	0.478

Spearman's Rank (Rs)

(significance values at 'n' degrees of freedom)

Note :

if the degrees of freedom are missing from the table e.g. 15 or 17, then take the significance value 1-df below it. So 15-df becomes 14-df.

ANSWERS

Activities 1

1. extract 2 - gives clear and precise evidence to back up the argument for funding

Activities 2

1. inferential
2. descriptive
3. (a) interval (but ranked ordered) (b) nominal (c) interval
(d) nominal (e) ordinal (f) interval (g) ratio
(h) nominal

EXTENSION

- (a) interval (b) interval (c) nominal (d) ordinal (e) nominal

Activities 3

1. angle 30-39
2. angle 10-19 and 30-39
3. bimodal
4. it is nominal data
5. median of discharge = 62.4
6. median of rainfall = 9.6
7. mean of discharge = 80.2 mean of rainfall = 11.8
8. The discharge figures are generally larger, so the difference in the median and mean would be larger as a matter of course. But the discharge data is much more skewed than the rainfall data.

EXTENSION

- (a) (teacher check) (b) average dist. = 6-10km
modal dist. = 3-5km
median dist. = 0-2km
- (c) i. the modal as it includes most shoppers ii. the average (mean)
- (d) skewed data such as this are not best dealt with using the mean ; this is because the mean is best suited to data with a normal distribution.

Activities 4

1. (a) 25mm (b) January = 15mm (c) April = 36mm
(c) as time progresses from Jan. to April the rainfall becomes increasingly variable.

EXTENSION

- (teacher check)

Activities 5

- 1.

Farm	barley (in hectares) x	difference from the mean (x - $\bar{X}$)	difference squared (x - $\bar{X}$) ²
Dubbie's	7	-6.4	40.96
Clartie's	3	-10.4	108.16
Puddock Law	15	1.6	2.56
Hillie's	11	-2.4	5.76
Steenie's	19	5.6	31.36
Coo' Toon	6	-7.4	54.76
The Laird's	21	7.6	57.76
The Hame Farm	27	13.6	184.96
Badenyeuchers	9	-4.4	19.36
Broadlands	16		
$\Sigma x = 134$		$\Sigma (x - \bar{X})^2 = 512.4$	

2. $\bar{X} = 13.4$ SD = 7.16
3. Farms below 1SD from the mean = Clarties and Coo Toon ; farms above 1SD from the mean = The

Laird's and The Hame Farm

- Soil analysis, pH levels, altitude above sea level of farms, aspect, slope, distance from market outlets e.g. distilleries are among some areas worth following up order to identify the reasons for the wide variation in barley production across the farms.

EXTENSION

(teacher check)

Activities 6

- The total map area is calculated by counting the no. of grid squares across the map and multiplying them by the no. of grid squares up the side. How the OS map is best sampled then for water bodies depends on the map area concerned. Random sampling by generating grid refs from random number tables is good but time-consuming. Systematic sampling using the easting/northing intersections works well, but not in a map area with distinct linear lochs or east/west rivers.
- Stratified sampling would be used where the map clearly divides into 3 areas linked to rock type. It allows comparison of the areas and the effects rock type has on whatever.
- Systematic sampling may not be appropriate in a map area with distinct linear features. There is a strong possibility of over- or under-estimating the variable being sampled.
- It is linear and systematic sampling as regards the transect, and areal and systematic within the confines of the quadrat.
- Stratified sampling would be required first, as there are two very distinct layers in the face of the bank. Random sampling of each layer would then be appropriate.

Activities 7

- Null hypothesis is 'that there is no significant difference in wind directions over the year'.

	Frequency of wind directions				
	O (observed)	E (expected)	(O - E)	(O - E) ²	(O - E) ² /E
NE	67	85.25	-18.25	333.06	3.91
SE	74	85.25	-11.25	126.56	1.48
SW	102	85.25	16.75	280.56	3.29
NW	98	85.25	12.75	162.56	1.91

n = 4 $\sum (O) = 341$

$\sum (O - E)^2/E = 10.59$

$\chi^2 = 10.59$. The χ^2 value of 10.59 at 3 degrees of freedom (n - 1) lies above the 95% significance value of 7.82 but below the 99% value of 11.34. The null hypothesis must be rejected. There is therefore a significant difference in wind directions over the year, though not highly significant. (The fact that there are only 341 wind observations over the year is due to 24 calm days).

- Null hypothesis is that there is no relationship between rock type and clast shape
 - The initial workings for the expected values are as shown. The final result of $\chi^2 = 20.02$

	flat		oblong		square		(rows)
Rock type	O	E	O	E	O	E	R
schist	27	16.2	19	20.25	16	25.55	62
quartzite	18	19.07	20	23.84	35	30.08	73
granite	7	16.72	26	20.9	31	26.37	64
K (columns)	52		65		82		199

- The χ^2 result of 20.02 exceeds the 99.9% value of 18.46 at 4 degrees of freedom. The null hypothesis must be rejected ; there is a highly significant relationship between rock type and clast shape.
- Schist tends to split along mica rich layers due to its foliation, while granite and quartzite are more massive rocks.

EXTENSION

- strong possibility of a relationship between the nos. of cirques and the altitude.
- null hypothesis - there is **no relationship** between the no. of cirques and the altitude.
- expected frequency = 17.5
- Chi-squared (χ^2) = 32.32
- The χ^2 value even at 3 degrees of freedom has less than 1 : 1000 chance of occurring so the null hypothesis must be rejected. There is therefore a strong relationship between the occurrence of cirques and the altitude.

Activities 8

1. (a) There is no relationship between life expectancy and % adult obesity in LEDCs
- (b) $R_s = 0.578$ (see table below for calculation of d^2 ; $n^3 - n = 4896$). At 17 degrees of freedom the R_s value of 0.578 exceeds the 95% significance value of 0.492 but not the 99% value of 0.645. There is therefore a significant relationship between life expectancy and % adult obesity in LEDCs. The null hypothesis must be rejected.
- (c) The direct relationship between life expectancy and % adult obesity in LEDCs is surprising, as it is the complete opposite of that for MEDCs. However the reason may be due to the high levels of malnutrition and poverty in LEDCs. The poor die young, those that are 'well-fed' (obese) live longer.

LEDCs	Life expectancy X	Adult obesity % Y	Rank order of X	Rank order of Y	d	d ²
Ethiopia	51	0.3	6	1	5	25
India	63	0.5	9	2	7	49
Nepal	61	0.5	8	2	6	36
Malawi	46	1.1	3	4	-1	1
Zambia	40	1.5	1	5	-4	16
Indonesia	67	2.3	11	6	5	25
Burkina Faso	49	2.4	5	7	-2	4
China	73	2.9	16	8	8	64
Nigeria	48	2.9	4	8	-4	16
Kenya	45	3.2	2	10	-8	64
Ghana	57	4	7	11	-4	16
Philippines	67	4.6	11	12	-1	1
Brazil	72	11	14	13	1	1
Bolivia	65	15.1	10	14	-4	16
Peru	72	15.7	14	15	-1	1
Egypt	68	22.8	13	16	-3	9
Mexico	74	23.4	17	17	0	0

2. There is no relationship between distance downstream and discharge. $R_s = 0.882$ ($n = 15$, $d^2 = 66$). Because the R_s value of 0.882 far exceeds the 99% significance value of 0.689 at 15 degrees of freedom, the null hypothesis must be rejected and replaced by the alternative hypothesis which states that there is a very significant relationship between distance downstream and discharge. With increasing distance downstream a water course is joined by tributaries which add more water to the discharge.

Activities 9

1. (a) One would expect channel width would increase with increasing distance downstream.
 - (b) that there is a significant relationship between channel width and distance downstream.
 - (c) $r = 0.802$ ($n = 8$, $\sum x = 18.4$, $\sum y = 8.55$, $\bar{x} = 2.3$, $\bar{y} = 1.07$, $\sum X^2 = 13.9$, $\sum Y^2 = 0.971$, $\sum XY = 2.948$)
 - (d) At 7 degrees of freedom the Pearson's (r) result of **0.802** exceeds the 99% significance level of 0.798. There is therefore a very significant relationship between channel width and distance downstream.
 - (e) Drawing a scatter graph can save a lot of unnecessary computation if little or no relationship exists.
2. (a) That there is no relationship between the no. of deaths from malaria per 1000 people and life expectancy for the selected African countries. Choice of probability is at 95%.
 - (b) Calculations given below.
 - (c) The Pearson's Product result of -0.625 well exceeds the 95% level of significance of 0.532 at 12 degrees of freedom. The null hypothesis must be rejected. There is therefore a significant relationship between the no. of deaths per 1000 people and life expectancy.
 - (d) The incidence of other major killers will also play a part, e.g. AIDS/HIV, cholera, bilharzia, sleeping sickness. So too will levels of medical care e.g. natal care, no. of doctors, hospital beds available, vaccination. A country's water and sewerage services can affect levels of general health. Literacy too and levels of individual prosperity. In a question such as this it is important to work through each factor mentioned and to back it up with evidence from the Atlas (i.e. data).

Country	from malaria per 1000/pop. x	Life expectancy y	X	Y	X ²	Y ²	XY
Angola	0.938	41	0.632	-5.77	0.400	33.29	-3.647
D.R.of Congo	0.336	44	0.030	-2.77	0.001	7.67	-0.082
Eritrea	0.022	55	-0.284	8.23	0.081	67.73	-2.339
Ghana	0.169	57	-0.137	10.23	0.019	104.65	-1.398
Kenya	0.005	48	-0.301	1.23	0.091	1.51	-0.371
Malawi	0.676	40	0.370	-6.77	0.137	45.83	-2.504
Mozambique	0.189	42	-0.117	-4.77	0.014	22.75	0.558
Nigeria	0.050	44	-0.256	-2.77	0.065	7.67	0.709
Somalia	0.001	47	-0.305	0.23	0.093	0.05	-0.070
Sudan	0.088	57	-0.218	10.23	0.048	104.65	-2.234
Tanzania	0.441	46	0.135	-0.77	0.018	0.59	-0.104
Uganda	0.411	49	0.105	2.23	0.011	4.97	0.234
Zambia	0.656	38	0.350	-8.77	0.123	76.91	-3.072

n = 13, $\sum x = 3.982$, $\sum y = 608$, $\bar{x} = 0.306$, $\bar{y} = 46.77$, $\sum X^2 = 1.099$, $\sum Y^2 = 478.31$, $\sum XY = -14.321$,
df = (n - 1) = 12

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \times \sum Y^2}} = \frac{-14.321}{\sqrt{(1.099 \times 478.31)}} = \frac{-14.321}{\sqrt{525.663}} = \frac{-14.321}{22.927} = \underline{\underline{-0.625}}$$

Activities 10

- The higher the organic content in the soil the lower the pH. This is due to the soil becoming more acidic with the build up of mor or peat within the soil profile.
 - $\sum X = 10.55$, $\sum Y = 61.1$, $\sum XY = 63.285$, $\sum X^2 = 12.633$
 $b = -0.782$ $a = 6.935$
 the estimated relationship is : $Y = 6.935 + (-0.782 \times 2)$, so $Y = \underline{\underline{5.317}}$
 - With an organic content in the soil of 2gm the pH would be 5.3
 - Factors such as moisture content, drainage, muir burn, hill liming, grazing levels could all affect the pH of a hill soil

Activities 11

- that there is no relationship between streamflow and sediment discharge at 95% significance.
 - Calculation of Spearman's Rank Corr. Coeff. (Rs) given below.

Day	Streamflow (cumecs) x	Sed. discharge (tonnes) y	Rank X	Rank Y	d	d ²
1	125.44	12698.0	7	7	0	0
2	70.00	1287.9	6	4	2	4
3	68.86	2575.9	5	6	-1	1
4	27.86	197.7	3	3	0	0
5	41.44	1623.5	4	5	-1	1
6	13.38	25.4	1	1	0	0
7	24.92	87.1	2	2	0	0
n = 7						$\sum d^2 = 6$

df = n = 7

$$R_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

$$R_s = 1 - \frac{(6 \times 6)}{343 - 7} = 1 - \frac{36}{336} = 1 - 0.107 \quad \quad \quad R_s = \underline{\underline{0.893}}$$

2. (a) Calculation of Pearson's Product Moment Corr. Coeff. (r) given below.

Day	Streamflow (cumecs) x	Sed. discharge (tonnes/day) y	X	Y	X ²	X ²	XY
1	125.4	12698.0	72.3	10055.8	5233.1	101119113.6	727436.6
2	70.0	1287.9	16.9	-1354.3	285.6	1834128.5	-22887.7
3	68.9	2575.9	15.8	-66.3	248.4	4395.7	-1044.9
4	27.9	197.7	-25.2	-2444.5	637.1	5975580.3	61699.2
5	41.4	1623.5	-11.7	-1018.7	136.0	1037749.7	11878.0
6	13.4	25.4	-39.7	-2616.8	1577.7	6847642.2	103939.3
7	24.9	87.1	-28.2	-2555.1	794.1	6528536.0	72002.7
n = 7	x = 371.9 bar x = 53.1	y = 18495.5 bar y = 2642.2	ΣX ² = 8911.9		ΣXY = 953023.3		
			ΣY ² = 123347146				

df = (n - 1)

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \times \Sigma Y^2}} \quad r = \frac{953023.3}{\sqrt{(8911.9 \times 123347146)}} = \frac{953023.3}{1048454.8} \quad r = 0.909$$

- (c) Both results show a highly significant relationship between streamflow and sediment discharge. Spearman's value of 0.893 exceeds the 95% significance of 0.786 at 7 df ; Pearson's value of 0.909 at 6 df even exceeds the 99% significance level of 0.834. The null hypothesis must therefore be rejected.
- (d) There are serious flaws in using both methods with the data as it stands, therefore the results are not reliable. Seven pairs of data for Spearman's is insufficient for reliability ; however the fact that the data is skewed does substantiate its use. Pearson's requires normal distribution of the data, which is lacking here. The alternatives are to use Spearman's or to logarithmically convert the data. The latter option is the only feasible one here and does allow for a more rigorous testing of the data.
3. (a) The data is heavily skewed as seen in the curvilinear graph. When logged, the graph displays a linear normal distribution curve.
- (b) drawing a scatter graph may save complicated calculations if no relationship exists.
- (c) Linear regression provides a mathematical model for the prediction of the Y-value at any given X-value. So in this case it is possible to compute the sediment load for a given streamflow e.g. 130 cumecs.

4. Calculation of the Linear Regression equation below.

Day	Streamflow (cumecs) x	Sed. discharge (tonnes/day) y	log x	log y	logx ²	logxy
1	125.4	12698.0	2.10	4.10	4.40	8.61
2	70.0	1287.9	1.85	3.11	3.40	5.74
3	68.9	2575.9	1.84	3.41	3.38	6.27
4	27.9	197.7	1.44	2.30	2.09	3.32
5	41.4	1623.5	1.62	3.21	2.62	5.19
6	13.4	25.4	1.13	1.40	1.27	1.58
7	24.9	87.1	1.40	1.94	1.95	2.71
n = 7		Σlogx = 11.37	Σlogx ² = 19.11			
		Σlogy = 19.48			Σlogxy = 33.42	

- (a) Value of 'log b' :

$$\log b = \frac{33.42 - \frac{(11.37 \times 19.48)}{7}}{19.11 - \frac{(129.28)}{7}} = \frac{33.42 - 31.64}{19.11 - 18.47} = \frac{1.78}{0.64} \quad \log b = 2.78$$

(b) Value of '**log a**' :

$$\log a = \frac{19.48}{7} - 2.78 \cdot \frac{(11.37)}{7} = 2.78 - 4.5 \quad \log a = \underline{\underline{-1.72}}$$

(c) The linear regression equation for streamflow against sediment discharge :

$$\log y = -1.72 + 2.78(\log x)$$

(d) The sediment discharge predicted at 130 cumecs streamflow :

$$x = 130, \log x = 2.11 : \log y = -1.72 + 2.78(2.11) = y = \underline{\underline{13803.8 \text{ cumecs}}}$$