FIELDING and GILBERT (2006c)

MEASURES OF CENTRAL TENDENCY AND DISPERSION

5

CONTENTS

MEASURES OF CENTRAL TENDENCY

A frequency distribution allows us to see the overall pattern in the distribut of respondents on a variable. It is often also useful to summarize the distribut with a single number, a statistic called a **measure of central tendency**.

There are three main measures of central tendency, the **mode**, the **median** a the **mean**, each appropriate for a different level of measurement. Remember t the different levels of measurement describe the relationships between categories of a variable (see Chapter 1).

THE MODE

The mode is the simplest measure of central tendency. It is simply the value the observation that occurs most frequently in the distribution. Sometime distribution has more than one value with similarly large numbers of obser tions. This is called a **bimodal** distribution if there are two modal values **multimodal** if there are more.

The mode can be calculated for nominal, ordinal or interval level variabl but it is the *only* measure of central tendency applicable to nominal variables.

In Exhibit 5.1, the modal value of the nominal variable, marital status, *married*. Simply, there are more married people than of any other marital stat Note that the mode is always the value of the variable and *not* the number cases. So the mode is the value 1 or its label, married.

For grouped data, the mode is the midpoint of the interval which has t largest frequency. Exhibit 5.2 displays the grouped variable, AGEGROUP. V see that the modal class interval is 30–39 so the mode is 34.5, the midpoint that interval.

MARSTAT MARITAL STATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Married	2653	57.3	57.3	57.3
	2 Cohabiting	346	7.5	7.5	64.7
	3 Single	934	20.2	20.2	84.9
	4 Widowed	378	8.2	8.2	93.0
	5 Divorced	226	4.9	4.9	97.9
	6 Seperated	94	2.0	2.0	100.0
	7 Same sex cohab	2	.0	.0	100.0
	Total	4633	100.0	100.0	

Exhibit 5.1 *Frequency listing for marital status illustrating 'married' as the mode Source: ONS, 1995*

AGEGROUP

	code:	label:	midpoint	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	10–19	14.5	255	5.5	5.5	5.5
	2.00	20–29	24.5	758	16.4	16.4	21.9
	3.00	30–39	34.5	938	20.2	20.2	42.1
	4.00	40–49	44.5	806	17.4	17.4	59.5
	5.00	50–59	54.5	689	14.9	14.9	74.4
	6.00	60–69	64.5	570	12.3	12.3	86.7
	7.00	70–79	74.5	427	9.2	9.2	95.9
	8.00	80–89	84.5	170	3.7	3.7	99.6
	9.00	90–99	94.5	20	.4	.4	100.0
	Total			4633	100.0	100.0	

Exhibit 5.2 *Frequency listing for age group demonstrating the mode is age group 30–39 with midpoint 34.5*

USING SPSS TO CALCULATE THE MODE

Measures of central tendency can be obtained in SPSS with an extension of the **Frequencies** command described previously. As explained in Chapter 2, you select **Analyze|Descriptive Statistics ▸|Frequencies**... and then the variable for analysis. However, this time, before you click on **OK**, click the button marked **Statistics** to bring up the dialog box in Exhibit 5.3.

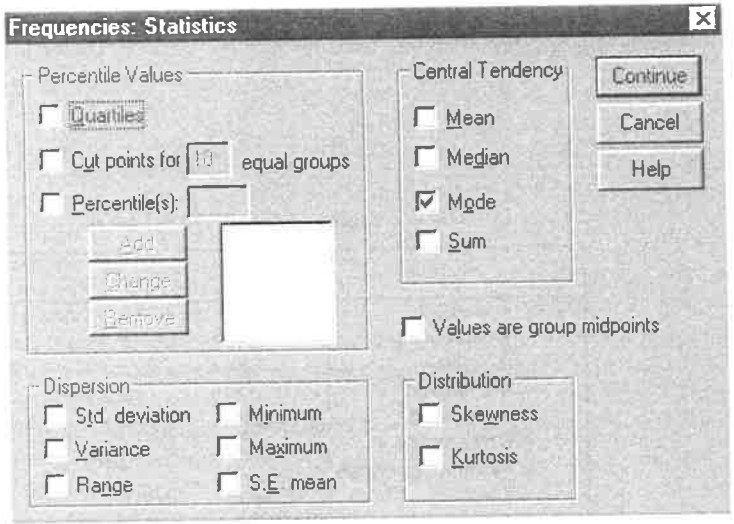

Exhibit 5.3 *The **Frequencies: Statistics** dialog box showing how to select the mode*

To calculate the mode simply select **Mode** from the **Central Tendency** box. Note, however, that if there are multiple modes, SPSS takes the lowest value.

PROPERTIES OF THE MODE

The mode, as the simplest measure of central tendency, has no mathematical properties; you cannot add, subtract, multiply or divide the mode. And one limitation with the mode is that it is possible to have more than one mode for a variable when two categories have exactly the same number of cases.

Another problem with the mode is illustrated by the following example. Imagine a hypothetical data set (A) of six cases with the following scores on a variable:

2 3 3 3 4 6

And here is a second data set (B), also with six cases and the following scores on another variable:

3 3 9 10 12 15

Both variables have a mode of 3, yet while a mode of 3 is clearly in the middle of A and describes its central tendency, this is not the case with B. The most frequent value, in this case, 3, is not necessarily the most typical value in data set B.

THE MEDIAN

The second measure of central tendency, used primarily for ordinal variables, but also appropriate for interval/ratio variables, is called the median. It is the middle value when the observations are arranged in order of magnitude. This means that there are as many scores or cases above the median as below it. The median identifies the *position* of an observation.

Consider the test scores from eleven students in Exhibit 5.4. To work out which is the middle value, we first have to rearrange them into ascending order and then mark the one which has as many values below as above (see Exhibit 5.5). Value 7 is positioned in the middle of this set of data. This median value allows you to summarize a set of numbers, the data set, with just one number or statistic.

8	2	12	4	5	2	7	15	10	7	5

Exhibit 5.4 *Test scores for eleven students*

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
2	2	4	5	5	7	7	8	10	12	15

▲

Exhibit 5.5 *The test scores rearranged in ascending order showing the 'middle' or median value*

The example above had an odd number of cases, so the middle data value has as many cases above as below. But what happens if we have an even number of observations? In Exhibit 5.6 there are two 'middle' observations. In this instance the median value lies half way between the 4th and 5th observations. The median value of this distribution is the average of the values of the two middle observations:

$$\text{median} = \frac{5+7}{2} = 6 \tag{5.1}$$

A formula that can be applied to calculate the position of the median for any number of observations is:

$$\text{median} = \frac{(n+1)\text{th}}{2} \text{ observation} \tag{5.2}$$

So

$$\text{median of 11 observations} = \frac{(11+1)\text{th}}{2} \text{ observation} = 6\text{th observation} \tag{5.3}$$

and

$$\text{median of 8 observations} = \frac{(8+1)\text{th}}{2} \text{ observation} = 4.5\text{th observation} \tag{5.4}$$

which we interpret as half-way between the 4th and the 5th observations.

2	3	5	5	7	7	9	12

▲ ▲

Exhibit 5.6 *Eight scores in order showing two middle values*

PERCENTILES

As we have seen, the median value is the middle value of a set of observat arranged in order of magnitude. If the median value is 7, as with the el observations in Exhibit 5.5, 50 per cent of the observations have values *less* 7. For this reason, it is also called the 50th percentile. Sometimes newspa report on the number of people living below the 5th percentile in income, th the number in the bottom 5 per cent by income. The definition of **a percenti** the smallest score below which a given percentage of cases fall. So if a salar £26,000 is reported as the 95th percentile, it means that 95 per cent of respondents have salaries *lower* than £26,000.

QUARTILES

Some percentiles are of particular importance. We have seen that the media the 50th percentile. Each half of a distribution can be further divided to **quartiles**. The **lower quartile** is half-way between the lowest value and median and the **upper quartile** is half-way between the highest value and median. We often write the lower quartile as Q_1, the median as Q_2 and the up quartile as Q_3.

Equation (5.2), the formula for the position of the median, can also be wri as

$$\text{median} = 0.5(\text{n} + 1)\text{th observation} \tag{5}$$

In the same way, the formula for the 25th percentile or lower quartile can written

$$\text{lower quartile} = 0.25(n+1)\text{th observation} \tag{5}$$

and the formula for the 75th percentile, or upper quartile can be written as

$$\text{upper quartile} = 0.75(n+1)\text{th observation} \tag{5}$$

Exhibit 5.7 demonstrates the calculation of the upper and lower quartiles the set of 11 marks seen previously.

USING SPSS TO CALCULATE THE MEDIAN

While SPSS will calculate the median (by selecting **Analyze|Descript Statistics▸|Frequencies**... and changing the **Statistics**), it is easy to find median category just by looking at the frequency table produced by SP Consider the ordinal variable, SOCLASE, social class, from the 1995 Gene Household Survey. There are 4417 valid responses which could be any one

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
2	2	4	5	5	7	7	8	10	12	15

▲ ▲ ▲

Q_1	Q_2	Q_3
Lower quartile	Median	Upper quartile
0.25 (11+1)th obs.	0.5 (11+1)th obs.	0.75 (11+1)th obs.
= 3rd obs.	= 6th obs.	= 9th obs.
= 4	= 7	= 10

Exhibit 5.7 *Calculation of the median and the upper and lower quartiles*

seven values. See Exhibit 5.8 for a frequency listing of the seven valid responses, ranging from code 1 for those in class I to code 7 for those in the armed forces. In order to find the median value, look at the column headed cumulative percentages. We know that 50 per cent of the cases lies above the median and 50 per cent below, so we need to look in which category the middle observation falls. From Exhibit 5.8, we can see that 29.4 per cent are in the two highest classes. If we include the next category, Class IIIN (non-manual) we reach a cumulative percentage of 53.2. We crossed the 50 per cent point, or middle observation, somewhere in the Class IIIN group. Therefore, this category is the median category.

SOCLASE SOCIAL CLASS OF INDIVIDUAL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 SOCIAL CLASS I	189	4.1	4.3	4.3
	2 SOCIAL CLASS II	1110	24.0	25.1	29.4
	3 SOC CLASS IIIN	1050	22.7	23.8	53.2 ←
	4 SOC CLASS IIIM	956	20.6	21.6	74.8
	5 SOCIAL CLASS IV	796	17.2	18.0	92.8
	6 SOCIAL CLASS V	298	6.4	6.7	99.6
	7 ARMED FORCES	18	.4	.4	100.0
	Total	4417	95.3	100.0	
Missing	-9 DNA:NEVER WORKED	216	4.7		
Total		4633	100.0		

Exhibit 5.8 *Calculating the median from the frequency output of the ordinal variable,* SOCLASE
Source: ONS, 1995

PROPERTIES OF THE MEDIAN

As with the mode, the median has no mathematical properties. You cannot add subtract, multiply or divide the median. It only informs us of the point of the distribution at which there are 50 per cent of the cases above and 50 per cent below. We can say that at least half the respondents summarized in Exhibit 5.8 are of social class IIIN or below.

An advantage of the median is that it is unaffected by extreme values. For example, if the largest score in a distribution is made even larger, the median will not change. Consider the two sets of data in Exhibit 5.9. In both, the median remains at value 7, the 5th observation, regardless of the fact that the first set of numbers ranges from 2 to 15 and the second set from 2 to 99. For this reason the median is known as a **resistant measure** of central tendency.

2	3	5	5	7	7	9	12	15

▲

2	3	5	5	7	7	9	12	99

▲

Exhibit 5.9 *Two sets of nine values demonstrating the resistant properties of the median*

THE MEAN

The third measure of central tendency, appropriate for interval/ratio variables but not for nominal or ordinal, is the arithmetic average or the mean. Unlike the median, which is based on the rank or position of a value, it is based on the actual values of scores. It allows us to compare groups on the basis of the amounts of a characteristic possessed by the group, relative to their size.

The arithmetic mean is calculated by summing all the values of the observations and then dividing by the number of observations. Consider five people with ages:

19, 25, 20, 21, 17

To calculate the mean age, you add up the ages and then divide by 5, the number of people. This would give a mean age of 20.4.

This procedure for calculating the mean can be expressed as a formula:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.8}$$

Let us break this formula down to its separate components. First of all, $\overline{X}$ (X bar) is the **mean**. The capital Greek letter, Σ, stands for 'the sum of'. Therefore, the expression

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

means the sum of the value of all the xs from the first case, where $n = 1$ and therefore x is x_1, to the last case, n, where x is x_n.

Exhibit 5.10 shows the ages of the same five people, but with age called x, so person 1's age is called x_1 and person 2's age is called x_2 and so on. Substituting into the equation and, since $n = 5$, we get:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 19 + 25 + 20 + 21 + 17 = 102 \tag{5.9}$$

$$\therefore \overline{X} = \frac{102}{5} = 20.4 \tag{5.10}$$

confirming that the mean, $\overline{X}$, is 20.4 years.

THE MEAN OF A GROUPED VARIABLE

Although it is quite a familiar situation in a survey to be asked 'In which age or salary group do you belong?', then the researcher has a dilemma if they want to calculate the mean age or salary. They did not collect individual age data but collected grouped data.

Case number	AGE	Let's call AGE X
1	19	x_1
2	25	x_2
3	20	x_3
4	21	x_4
5	17	x_5

Exhibit 5.10 *Calculating the mean of* AGE

If SPSS is asked to give a frequency listing of the grouped vari AGEGROUP, the result is the output seen in Exhibit 5.11. Asking SPS calculate the mean does not help – it gives you a mean of the codes, 1– mean of 4.1448. Clearly, this is not the mean age but the mean of the c given to the age groups.

To calculate the mean you need to assess the midpoint for each interval then multiply that midpoint by the number of cases in that interval. midpoint can be seen as the best estimate of all those in that category – in c words, the average age or salary of all those in that particular interval. If then add up all the cumulative ages for each interval you arrive at a very l number which, if divided by the number of people, will give you a mean based on midpoint estimates. This procedure is demonstrated in Exhibit 5.12

A short-hand way of describing this procedure to calculate the mean grouped data is

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i m_i}{n} \tag{5}$$

AGEGROUP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	255	5.5	5.5	5.5
	2.00	758	16.4	16.4	21.9
	3.00	938	20.2	20.2	42.1
	4.00	806	17.4	17.4	59.5
	5.00	689	14.9	14.9	74.4
	6.00	570	12.3	12.3	86.7
	7.00	427	9.2	9.2	95.9
	8.00	170	3.7	3.7	99.6
	9.00	20	.4	.4	100.0
	Total	4633	100.0	100.0	

Statistics

AGEGROUP

N	Valid	4633
	Missing	0
Mean		4.1448

Exhibit 5.11 *Inappropriate calculation of the mean of a grouped variable usin SPSS*

Stated class intervals	Real class intervals	Midpoint (m_i)	Frequency (f_i)	$f_i m_i$
10–19	9.5–19.5	14.5	255	3697.5
20–29	19.5–29.5	24.5	758	18571
30–39	29.5–39.5	34.5	938	32361
40–49	39.5–49.5	44.5	806	35867
50–59	49.5–59.5	54.5	689	37550.5
60–69	59.5–69.5	64.5	570	36765
70–79	69.5–79.5	74.5	427	31811.5
80–89	79.5–89.5	84.5	170	14365
90–99	89.5–99.5	94.5	20	1890
		Total	4633	212878.5
			212878.5/4633 = 45.95	

Exhibit 5.12 *Calculation of the mean of the grouped variable* AGEGROUP

where f_i is the number of cases in an interval, and m_i is the midpoint of the interval and n is the total number of cases.

The sum of the product of the frequencies (f_i) and the midpoint (m_i) of each interval equals 212,878.5. It is a simple task to divide this number by the *total* number of valid cases, 4633, to give a grouped mean of 45.95:

$$\text{grouped mean} = \bar{X} = \frac{212{,}878.5}{4633} = 45.95 \qquad (5.12)$$

PROPERTIES OF THE MEAN

The mean is adversely affected by extreme values. Using the data from the example of the calculation of the median, consider what happens to the value of the mean (Exhibit 5.13).

As we saw, the median is not affected by changing the last score from 15 to 99. The median is called a resistant statistic because it is not affected by extreme values. However, the mean is not a resistant measure, and is greatly affected by the one extreme value of 99, changing from 7.2 to 16.6. In different circumstances, this can either be an advantage (the value of the mean summarizes the values of all the scores) or a disadvantage (its value is sensitive to the precise value of the scores).

Exhibit 5.14 summarizes the values of the three measures of central tendency – the mode, median and mean – for the variables AGE and AGEGROUP. The statistics for AGE were provided by the computer but the statistics for AGEGROUP, except the median which is just included for completeness, are those that we have calculated by hand. Notice that the values for the grouped variable are remarkably similar to those of the ungrouped variable. This demonstrates the validity of using grouped data where ungrouped data are not available. Notice also that the modes, using either variable, are much lower than

2	3	5	5	7	7	9	12	15

▲

Median =7
Mean =7.2

2	3	5	5	7	7	9	12	99

▲

Median =7
Mean =16.6

Exhibit 5.13 *Two sets of values demonstrating how the mean is affected by extreme values*

Statistic	Ungrouped value Calculated by SPSS	Grouped value Calculated by hand
Mode	31	34.5
Median	45	(44.5)**
Mean	45.97	45.95
Valid cases: 4633 Source: 1995 General Household Survey		

Exhibit 5.14 *Comparing the mode, the median and the mean for the grouped and ungrouped variable age*
*** It is possible to calculate the median for grouped variables bu formula is quite complicated. Here the result is provided for comparative purposes only.*

either the means or the medians. This is because there are more people in lower age groups; the distribution is said to be **skewed**. In the next section will be investigating the shape of the distribution, but this table demonstrates dangers of relying on any one measure of central tendency which may not good summary of the overall distribution. If the distribution is skewed in direction then a good resistant measure to use is the median.

Another distribution which is often skewed is income (see Exhibit 5.15). M cases are in the lower-income brackets, yet a few individuals may have l incomes and these will raise the mean, making it unrepresentative of the sar as a whole. In Exhibit 5.15 the mean weekly earnings are £272, yet the me is much lower at £219. This is a clear indication of a skewed distribution.

Statistics

EARNINGS USUAL GROSS WEEKLY EARNINGS: INDIVIDUAL

N	Valid	2164
	Missing	2469
Mean		271.5209
Median		218.5450
Mode		230.76

Exhibit 5.15 *The mode, median and mean for usual gross weekly earnings Source: ONS, 1995*

SUMMARY OF MEASURES OF CENTRAL TENDENCY

Three measures of central tendency, the mode, median and the mean, were introduced as ways of summarizing the distribution of a variable. Exhibit 5.16 shows the appropriate measures of central tendency for different levels of measurement. It will be noted that although a measure may be primarily intended for one level of measurement, it may in certain circumstances, for example a skewed distribution, be the most suitable measure for another.

	Level of measurement of the variable			
Level of measurement of the statistic used	Interval/Ratio	Nominal	Ordinal	Interval/Ratio
	Nominal *mode*	✔	✔ but may not be the best measure	✔ but may not be the best measure
	Ordinal *median*	!! Logical errors	✔	✔, in fact as a resistant measure may be the best if distribution is skewed
	Interval/Ratio *mean*	!! Logical errors	!! Logical errors	✔

Exhibit 5.16 *Summary of levels of measurement and appropriate measures of central tendency*

THE SHAPE OF A DISTRIBUTION

In the previous chapter we saw how it was possible to create a histogram continuous variable by grouping the categories and making the areas in the of the chart proportional to the frequencies of each group. Exhibit 5.17 histogram of the variable AGE, grouped into 10-year intervals. It displa uneven stepped outline.

Exhibit 5.18 and 5.19 also display histograms of this variable, but decreasing intervals: first 5 years, then 1 year. Because age was only meas using one-year intervals we cannot reduce the intervals size below 1.

As we reduce the interval size the steps get smaller and the outline o distribution becomes smoother. If we continued to reduce the interval size increased the sample size we could further 'smooth' the steps of the histo until it formed a perfectly smooth outline. However, because age was meas on a discrete, yearly, interval, Exhibit 5.19 is as 'smooth' as we can get these data. A line has been drawn on Exhibit 5.20 to show a possible outli silhouette.

As Exhibit 5.20 demonstrates, the distribution of a continuous variable have a shape, in addition to a measure of central tendency which summa the distribution. The shape can be described in terms of several quantities number of modes, skewness, kurtosis and spread.

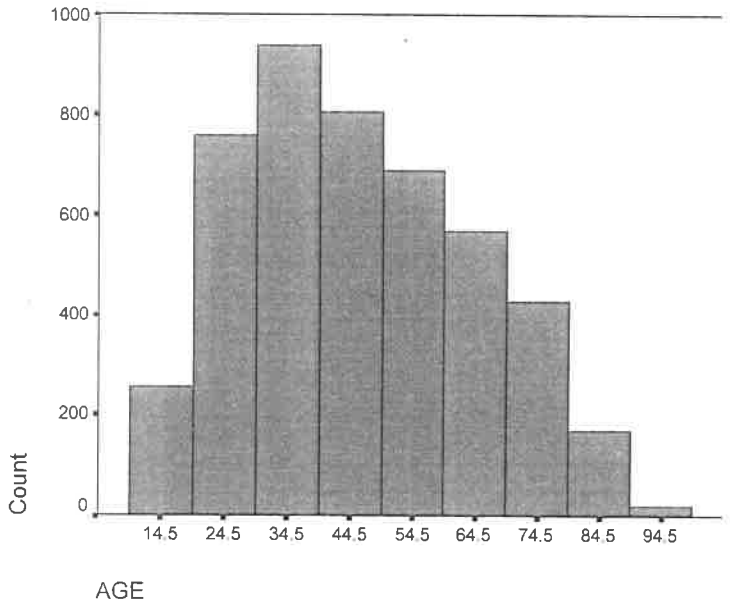

Exhibit 5.17 *Histogram of age in ten-yearly intervals*

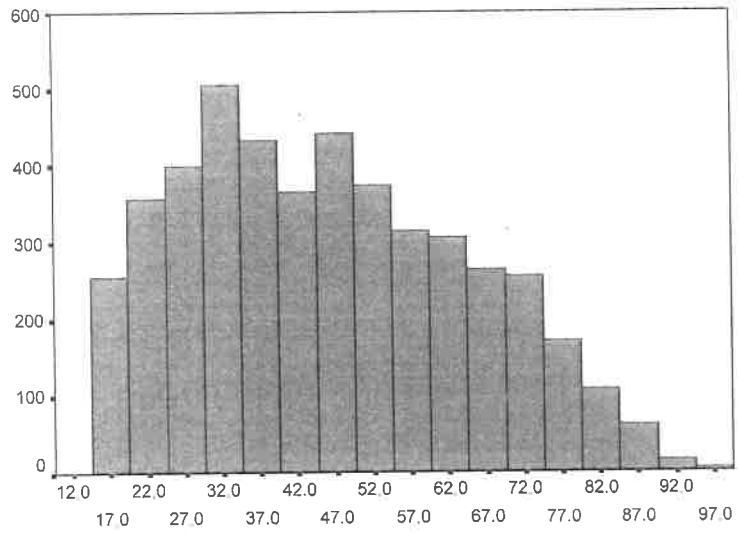

Exhibit 5.18 *Histogram of age in five-yearly intervals*

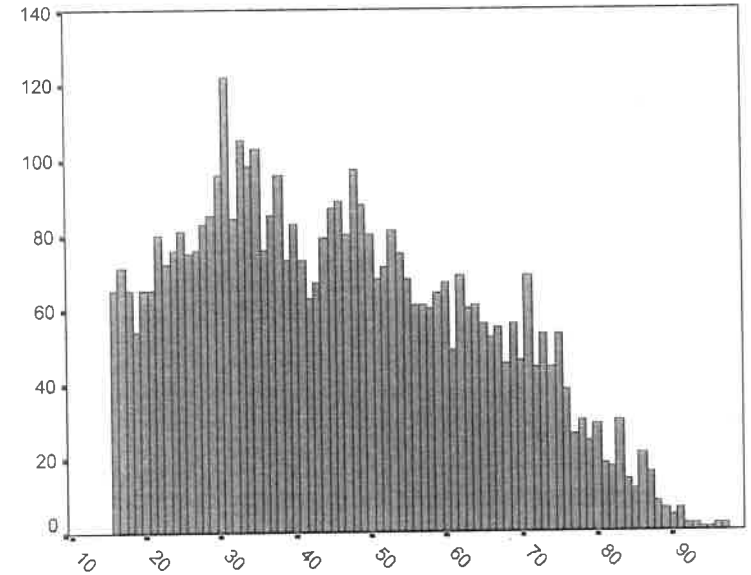

Exhibit 5.19 *Histogram of age in one-yearly intervals*

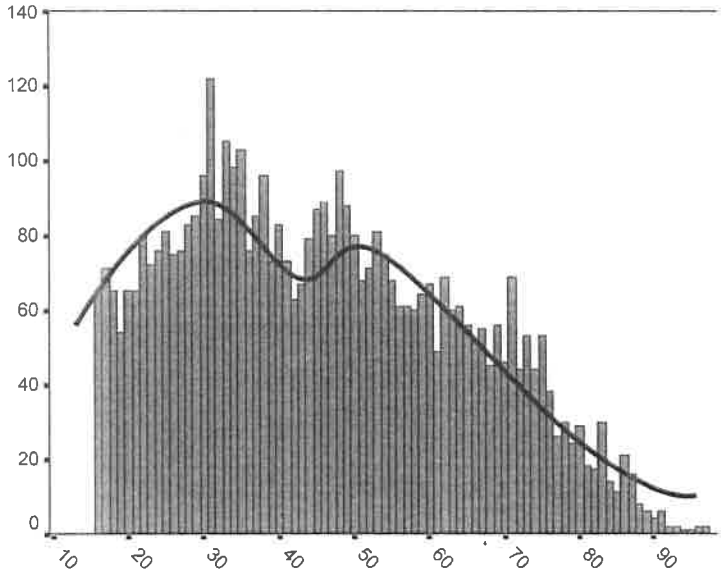

Exhibit 5.20 *Histogram of age with a curve superimposed to demonstrate the 'shape' of the distribution*

NUMBER OF MODES

The distribution can be unimodal (have one peak), bimodal (have two peak: it can be multimodal (with several peaks).

SKEWNESS

The tail of the peak (or peaks) can be stretched out to the left (the lower val in which case we say that the distribution has **negative skew**, or it ma stretched out to the right, the higher values, when it has **positive skew**. Or it lean neither to the left or right and be **symmetrical** (see Exhibit 5.21).

KURTOSIS

If the peak is pointed, which indicates that many observations are clc clustered around the mode, then the distribution is called a **leptokurtic** distı

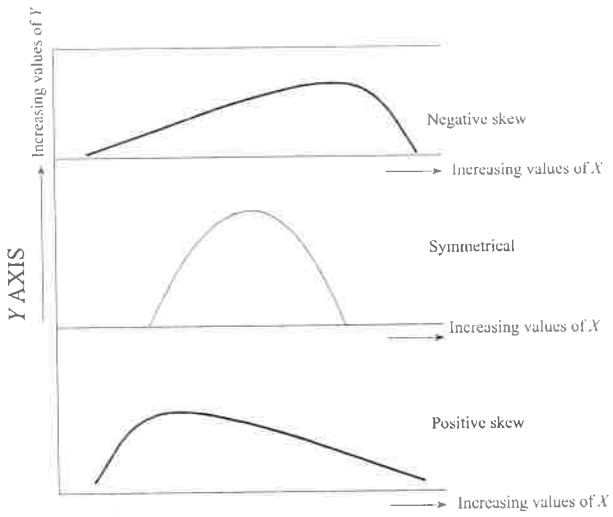

Exhibit 5.21 *Shape of distributions: skewness*

tion and to have **positive kurtosis**. If, on the other hand, the distribution is widely spread out, it is called **platykurtic** and is said to have a **negative kurtosis**. However, if it has an even spread then it is **mesokurtic** (see Exhibit 5.22).

MEASURES OF DISPERSION OR SPREAD

We have seen so far that we can describe variables in terms of their measure of central tendency (mode, median or mean) and, for continuous variables, the shape of the distribution. There is a further, and very important, feature of distributions for continuous variables and that is their spread or dispersion.

If all cases within a sample have the same value for a variable, for example all respondents in a survey earned £20,000, that variable, salary, would have no spread. There would be no variation as far as salary is concerned. The highest value would equal the lowest value and in this sample it would seem that everyone earned the same, £20,000. There would be no point in looking at differences between men's and women's earnings. In fact, there would be no point in trying to explain salary differential between any subgroups because there is none. In this hypothetical situation, salary is not a *variable*, it is a *constant*. However, in reality there is usually wide variation between individuals' salaries. This variation is what makes social research interesting. For these

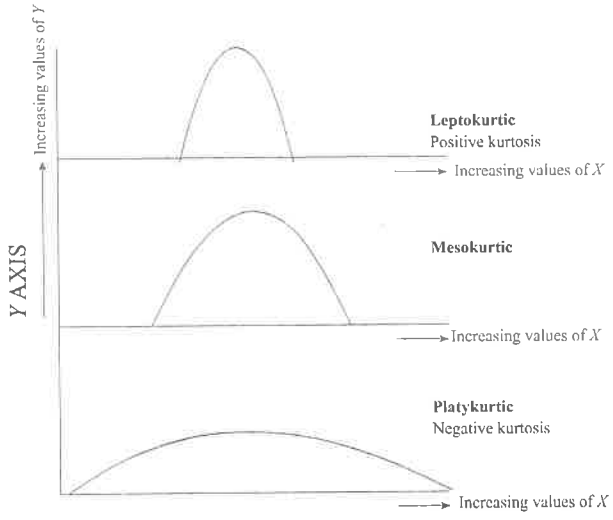

Exhibit 5.22 *Shape of distributions: kurtosis*

reasons we need to find ways of describing the spread of values caused by tl variation. We can then measure the spread, compare different subpopulatio (e.g. males and females) and possibly explain the variation.

We will start with the simplest way of describing variation, the **range**, a then go on to more sophisticated measures, the **variance** and **standard devi tion**.

THE RANGE

The range is the highest value of a distribution minus the lowest value. This one of the weaknesses of using the range as a measure of variation: it relies only two values. Consider:

2, 5, 6, 7, 9, 99 range $= 99 - 2 = 97$

2, 5, 6, 7, 9, 11 range $= 11 - 2 = 9$

Clearly, the value of 99 in the first set of numbers is at odds with the rest of t numbers and the range of 97 gives a false impression of the actual values.

In addition, the range can depend on the size of the population or samp Generally, you get a larger range for larger populations where you have a grea chance of extreme values.

INTERQUARTILE RANGE

One way round the problem of relying on just two values for the range is to calculate the **interquartile range** (IQR). The IQR is simply the value for the lower quartile subtracted from the value for the upper quartile:

$$\text{IQR} = \text{upper quartile } (Q_3) - \text{lower quartile } (Q_1) \qquad (5.13)$$

Whereas the range relies on the values of the highest and the lowest values, the IQR ignores the values of the highest 25 per cent and the lowest 25 per cent of cases. Exhibit 5.23 summarizes these two measures of spread in a symmetrical distribution.

USING SPSS TO COMPUTE THE RANGE AND INTERQUARTILE RANGE

To illustrate how SPSS provides a value for the range for a set of values, consider the data in Exhibit 5.24 for two sets of marks for ten students, one for a sociology exam and one for a psychology exam. This is the `MARKS.SAV` data set which you may have created in Chapter 2.

We select **Analyze|Descriptive Statistics ▸|Frequencies**.... After selecting the variable of interest, we select **Statistics**... in the **Frequencies** dialog box. In the **Statistics** dialog box (Exhibit 5.25) we would select **Range** in the box marked **Dispersion**. In addition, we could select **Mean**, **Mode** and **Median** in the **Central Tendency** box and **Quartiles** in the **Percentile Value** box.

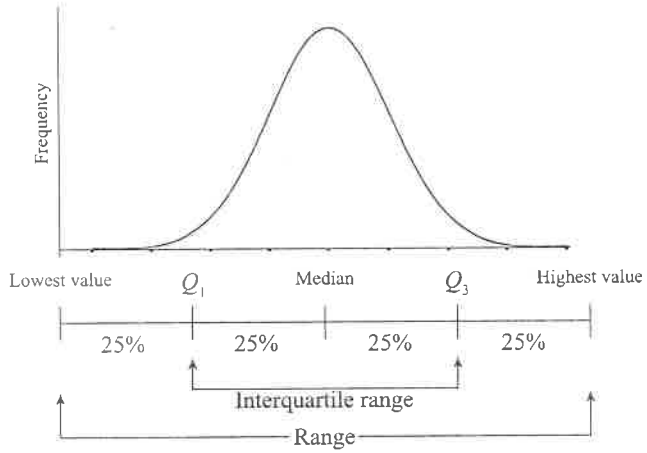

Exhibit 5.23 *The relationship between the range and the interquartile range*

	sociolog	psycholo
1	55.00	49.00
2	45.00	55.00
3	66.00	45.00
4	34.00	56.00
5	67.00	73.00
6	53.00	37.00
7	78.00	12.00
8	45.00	85.00
9	57.00	58.00
10	60.00	63.00

Exhibit 5.24 *Sociology and psychology marks for ten students*
Source: `MARKS.SAV`

Before clicking **OK** in the main **Frequencies** dialog box, we need to des **Display frequency tables**. This ensures that we do not get a long listin individual frequency counts, just the statistic(s) we asked for. The result SPSS is shown in Exhibit 5.26.

The **Frequencies** procedure does not calculate the IQR for you but it is easy to find it for yourself using the values of the 25th and 75th percentiles.[1]

$Q_3 \qquad Q_1$

Sociology: $66.25 - 45 = 21.25$

Psychology: $65.5 - 43 = 22.5$

So the sociology marks have a range of 44 and an IQR of 21.25, while psychology marks have a range of 73 and an IQR of 22.5. The psycho marks have a lower minimum and a higher maximum than the sociology m

[1]The **Explore** procedure in SPSS for Windows will give you an IQR statistic.

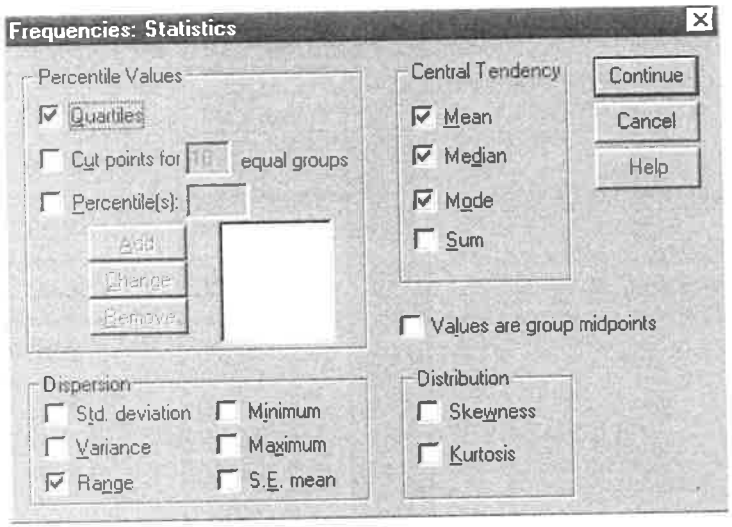

Exhibit 5.25 *The **Frequencies: Statistics** dialog box. Calculation of the quartiles and range in SPSS*

Statistics

		PSYCHOLO	SOCIOLOG
N	Valid	10	10
	Missing	0	0
Mean		53.3000	56.0000
Median		55.5000	56.0000
Mode		12.00[a]	45.00
Range		73.00	44.00
Percentiles	25	43.0000	45.0000
	50	55.5000	56.0000
	75	65.5000	66.2500

[a] Multiple modes exist. The smallest value is shown

Exhibit 5.26 ***Frequencies: Statistics** output in SPSS showing the range and quartiles*

These extreme values, however, are ignored in the calculation of the interquartile range, resulting in similar values for the sociology and psychology IQRs.

The IQR and the range are based on the *position* of the observations, not on their actual values. If we want to assess the spread of a variable in a way which

takes into account all the observations, we need to use the actual value data and consider how they are distributed around the average value. We know *how far* each mark is above or below the mean. A measure of v that does consider the values of each observation is the **variance**.

THE VARIANCE

The variance (denoted by s^2) is based on the deviations or distances observation from the central or average observation. In Exhibit 5.2 sociology mark and the mean value have been placed on a line. In addit have calculated how far each mark is from the mean and made a note of i

If we then add up all these deviations each side of the mean we find tl sum to identical amounts; +48 above the mean and −48 below the mea we added these deviations together we would end up with zero. Clea measure is not going to get us far. However, there is one way we can end a positive deviation, and that is to calculate the *square of each deviat* sum these squared deviations. Exhibit 5.28 charts these calculations sociology marks.

However, we have not quite arrived at a measure of spread because ou for the sum of the squared deviations will obviously get bigger as w greater numbers of observations and therefore deviations to sum. So we divide by the number of observations to control for the number of cases.

However, although this gives us a good measure when we have g information from the whole population of interest, it tends to *underestim* value of the variance when we have only collected information from a s In other words, we need to divide by slightly less than the number of resp in order to get a slightly larger value for the variance. This *fudge facto* out to be the number of cases less than one. So, since we have only cc

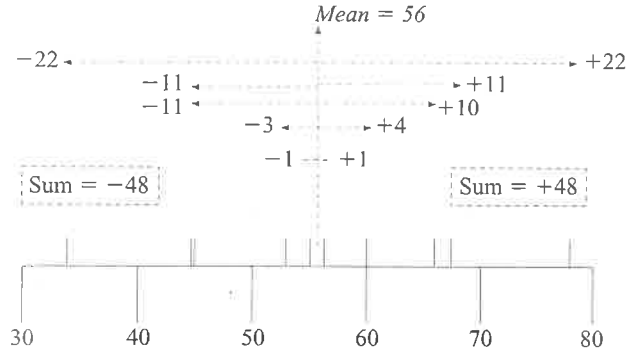

Exhibit 5.27 *Distances from the mean for each sociology mark*

Sociology Mark X	Mean $(\overline{X})$	Mark−Mean $(X-\overline{X})$	$(X-\overline{X})^2$
55	56	−1	1
45	56	−11	121
66	56	10	100
34	56	−22	484
67	56	11	121
53	56	−3	9
78	56	22	484
45	56	−11	121
57	56	1	1
60	56	4	16
SUM (Σ) =		0	1458

Exhibit 5.28 *Calculation of the sum of the deviations from the mean for each mark*

marks from 10 students we need to divide the sum of the squared deviations by $N-1$ or $10-1$, or 9.

$$s^2 = \frac{1458}{10-1} = \frac{1458}{9} = 162 \qquad (5.14)$$

The short-hand way of writing these instructions for the calculation of the variance for a sample is

$$s^2 = \frac{\sum(X-\overline{X})^2}{N-1} \qquad (5.15)$$

where s^2 is the sample variance. And the formula for a population is

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \qquad (5.16)$$

where σ^2 is the population variance and μ is the population mean.

Both Equations (5.15) and (5.16) both involve the subtraction of the mean from each value. An alternative formula which provides a short-cut for the calculation of the variance by hand is

$$s^2 = \frac{\sum X^2 - (\sum X)^2/N}{N-1} \qquad (5.17)$$

Exhibit 5.29 demonstrates the steps necessary to calculate the variance psychology marks using equation (5.17). Using this formula you only nee columns in the table of calculations; one for the mark and one for the squared. Then it is relatively quick and simple to sum each column to pr the sum of all the marks and the sum of all the squared marks.

USING SPSS TO CALCULATE THE VARIANCE

As with the calculation of the range, we select **Analyze|Descriptive Sta ▸|Frequencies...**. This time, however, we select **Variance** in the **Statist**

Psychology Mark x	x^2
49	2401
55	3025
45	2025
56	3136
73	5329
37	1369
12	144
85	7225
58	3364
63	3969
Sum = Σ = 533 $\therefore(\Sigma x)^2$ = 284,089	Sum = Σ = 31,987 $\therefore\Sigma x^2$ = 31,987

$$s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/N}{N-1}$$

Equation 1.1

$$s^2 = \frac{31987 - 284089/10}{10-1} = \frac{31987 - 28408.9}{9} = 397.57$$

Exhibit 5.29 *Calculation of the variance using the alternative formula*

dialog box (see Exhibit 5.3). The result is shown in Exhibit 5.30. The variance of the psychology marks (397.57) is greater than that for the sociology marks (162). This confirms what we have already seen when we calculated the range. The psychology marks are more spread out and show greater variation.

THE STANDARD DEVIATION

The variance is calculated by summing *squared* deviations. This means that the units of variance are on a squared scale. For example, if the data were measured in pounds, the variance would be in pounds squared. To create a measure of variation on the same scale as the original marks data we need to take the square root of the variance which is called the **standard deviation**.

The standard deviation (denoted by s or SD) is the square root of the summed deviations divided by the number of cases. For a sample, the denominator is the number of cases minus one. The formula for the standard deviation is

$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{n - 1}} \qquad (5.18)$$

The larger the standard deviation, the more spread out are the data. Conversely, the smaller the standard deviation, the less spread out and the more similar are the data. A standard deviation of 0 occurs when all scores are the same so there is no deviation around the mean.

To see what it means to have two distributions with different standard deviations, imagine two ambulance crews, crew A in North London and crew B in South London, both with an average response time of 10 minutes. However, the standard deviation for the response times for crew A is 9 minutes and the standard deviation for crew B is 5 minutes. Where would you rather live?

If you live in North London and called the ambulance out enough times you would wait 10 minutes on average for it to arrive. However, since North London has the larger standard deviation you may have to wait considerably longer on any one occasion or you may be lucky and it would arrive in just a few minutes. In South London, where the response time has a smaller standard deviation, the times will be closer to the average of 10 minutes, and they will deviate from the

Statistics

		PSYCHOLO	SOCIOLOG
N	Valid	10	10
	Missing	0	0
Variance		397.57	162.00

Exhibit 5.30 *The variance output in SPSS using the* ***Frequencies*** *procedure*

mean to a lesser degree than those in North London. So your chances of ha to wait a long time for an ambulance crew are greater if you live in N London than if you live in the South.

What is the likelihood of making a wrong guess about the chances of get a slow ambulance? If certain conditions are satisfied the chance of gues wrongly can be quantified. These conditions are to do with the shape of distribution. If the distribution of the variable you are interested in, be it e marks or ambulance response times, forms a specific shape of curve, you use the properties of this special sort of curve to quantify the chances of get any particular range of scores or times. This special shape is called a **nor curve** and is the subject of the next chapter.

SUMMARY OF MEASURES OF SPREAD

In this chapter we have considered the range, the interquartile range, variance and the standard deviation as alternative measures of spread. Exl 5.31 summarizes the notation we have used for these measures of spread.

EXERCISE ON THE MEAN, MODE AND MEDIAN

Exhibits 5.32–5.34 show the outputs from SPSS when asked to dis frequencies for a number of variables from the 1995 General Household Sur For each variable do the following:

(a) Find the mode.
(b) *If applicable*, find the median.
(c) *If applicable*, find the mean.
(d) Write brief comments about your conclusions from considering the va of the mode, median and the mean.

Type of Statistic	Sample	Population
Variance	s^2	σ^2 (sigma)
Standard deviation	s	σ
Mean	$\overline{X}$	μ (mu)
Number of observations	N	N

Exhibit 5.31 *Notation for the variance and standard deviation*

NADULTS NUMBER OF ADULTS IN HOUSEHOLD

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	812	32.8	32.8	32.8
	2	1286	52.0	52.0	84.9
	3	267	10.8	10.8	95.7
	4	91	3.7	3.7	99.4
	5	13	.5	.5	99.9
	6	2	.1	.1	100.0
	7	1	.0	.0	100.0
	Total	2472	100.0	100.0	

Exhibit 5.32 *Frequency output for number of adults in household*

HHTYPF1 HOUSEHOLD TYPE F

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 1 PERSON ONLY	703	28.4	28.5	28.5
	2 2+ UNREL ADULTS	60	2.4	2.4	30.9
	3 M.CPLE, DEP CH	595	24.1	24.1	55.0
	4 M.CPLE, INDEP CH	159	6.4	6.4	61.5
	5 M.CPLE, NO CH	712	28.8	28.8	90.3
	6 LONE P, DEP CH	152	6.1	6.2	96.5
	7 LONE P, INDEP CH	72	2.9	2.9	99.4
	8 2+ FAMILIES	14	.6	.6	100.0
	9 SAME SEX COHAB	1	.0	.0	100.0
	Total	2468	99.8	100.0	
Missing	-9	4	.2		
Total		2472	100.0		

Exhibit 5.33 *Frequency output for household type*
Note. M.CPLE -married couple
CH -children
Lone P -lone parent
UNREL -unrelated

EXERCISE ON MEASURES OF SPREAD

NB. In all the following exercises, assume you have a population and use N rather than $N - 1$ in the calculation of the standard deviation.

1. Guess which of the following two lists has the larger standard deviation. Check your guess by computing the standard deviation for both lists.

EDLEV HIGHEST EDUCATIONAL QUALIFICATION

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 HIGHER DEGREE	53	1.1	1.5	1.5
	2 FIRST DEGREE	335	7.2	9.3	10.7
	3 TEACHING QUAL	65	1.4	1.8	12.5
	4 OTH HIGHER QUAL	239	5.2	6.6	19.2
	5 NURSING QUAL	73	1.6	2.0	21.2
	6 GCE A LEVEL 2+	161	3.5	4.5	25.6
	7 GCE A LEVEL 1	284	6.1	7.9	33.5
	8 GCSE&OLV5+, SG1–2	461	10.0	12.8	46.3
	9 GCSE&OLV1–4, NOCQ	73	1.6	2.0	48.3
	10 GCSE&OLV1–4, NOCQ	316	6.8	8.8	57.0
	11 COMM Q, NO O LEVS	148	3.2	4.1	61.1
	12 CSE GRD 2–5	117	2.5	3.2	64.4
	13 APPRENTICESHIP	92	2.0	2.5	66.9
	15 FOREIGN QUALS	61	1.3	1.7	68.6
	16 OTHER QUALS	31	.7	.9	69.5
	17 NO QUALS	1102	23.8	30.5	100.0
	Total	3611	77.9	100.0	
Missing	-9 NEV WENT TO SCH	1022	22.1		
Total		4633	100.0		

Exhibit 5.34 *Frequency output for educational level*
GCE A level -General Certificate of Education Advanced Level
GCSE -General Certificate of Secondary Education
OLV -General Certificate of Education Ordinary Level
CSE -Certificate of Secondary Education
CM Q -Commercial qualifications
NOCQ -No commercial qualifications

List A	7	8	8	8	10	13
List B	8	8	9	9	9	11

2. (a) For each list of numbers work out the mean and the standard deviation

List A	4	3	6	4	4	3
List B	11	10	13	11	11	10

(b) How is list A related to list B? How does this relationship carry over the average? The standard deviation?

3. Repeat Exercise 2 for the following two lists:

List A	4	3	6	4	4	3
List B	8	6	12	8	8	6

4. Repeat Exercise 2 for the following two lists:

List A	0	−3	5	−4	−4	3
List B	0	3	−5	4	4	−3

GRAPHICS FOR ANALYSIS

CONTENTS

EXPLORATORY DATA ANALYSIS

Exploratory methods are designed to give you an idea about the range, and structure of the data collected. They include simple graphical method are quick to carry out and interpret. Interpretation therefore tends to be and impressionistic. These methods tend to be used as a first step in any an to give you ideas about further areas to explore and as the prelude to a rigorous analysis where ideas you have formed about the data can be tested

STEM AND LEAF DIAGRAMS

Stem and leaf diagrams are one of the simplest exploratory techniques, ar used to display either discrete or continuous data which have at leas significant places, for instance tens and units. The first significant place bec the stem, and the second the leaf.

To illustrate this technique, consider the following twelve numbers...