

UNIT III

UNIVARIATE ANALYSIS

SYLLABUS

Introduction to Single variable: Distributions and Variables - Numerical Summaries of Level and Spread - Scaling and Standardizing – Inequality - Smoothing Time Series.

- ❖ Introduction to Single Variable
- ❖ Distributions and Variables
- ❖ Numerical Summaries of Level and Spread
- ❖ Scaling and Standardizing
- ❖ Inequality
- ❖ Smoothing Time Series

UNIT III

UNIVARIATE ANALYSIS

3.1. INTRODUCTION TO SINGLE VARIABLE

3.1.1. DISTRIBUTIONS AND VARIABLES

How many households have no access to a car? What is a typical household income in Britain? Which country in Europe has the longest working hours? To answer these kinds of questions we need to collect information from a large number of people, and we need to ensure that the people questioned are broadly representative of the population we are interested in.

Conducting large-scale surveys is a time-consuming and costly business. However, increasingly information or data from survey research in the social sciences are available free of charge to researchers and students. The development of the worldwide web and the ubiquity and power of computers makes accessing these types of data quick and easy.

The aim is to explore data. We can use the 'Statistical Package for the Social Sciences' (SPSS) package to start analysing data and answering the questions posed above.

Preliminaries

Two organizing concepts have become the basis of the language of data analysis: cases and variables. The cases are the basic units of analysis, the things about which information is collected. The word variable expresses the fact that this feature varies across different cases.

We will look at some useful techniques for displaying information about the values of single variables, and will also introduce the differences between interval level and ordinal level variables.

Variables on Household Survey

It is a multipurpose survey carried out by the social survey division of the Office for National Statistics (ONS). The main aim of the survey is to collect data on a

range of core topics, covering household, family and individual information. Government departments and other organizations use this information for planning, policy and monitoring purposes, and to present a picture of households, family and people in Great Britain.

Person-id	Age	Sex	Units of alcohol per week	Drinking Classification	NC-SEC5
1	27	1	24	4	5
2	27	2	8	3	1
3	27	2	27	3	4
4	6	1	.	.	-6
5	5	1	.	.	-6
6	77	1	8	2	1
7	65	2	14	3	2
8	51	2	3	2	2
9	33	1	9	3	5
10	25	1	9	2	5
11	49	1	352	5	3
12	16	1	2	1	97
13	66	1	0	1	4
14	65	2	0	1	5
15	47	2	0	-9	2
16	42	1	6	1	2
17	15	1	.	.	-6
18	13	2	.	.	-6
19	47	2	0	1	1
20	44	1	5	1	1

Fig. 3.1. Specimen data from the 2005GHS (Individual file)

Column 5 contains a variable that indicates individuals classification of themselves in terms of the amount of alcohol they usually drink. It has five ranked categories:

1. hardly drink at all
2. drink a little
3. drink a moderate amount
4. drink quite a lot
5. drink heavily

Column 5 indicates the social class of individual based on the occupation.

1. Managerial and professional occupations
2. Intermediate occupations
3. Small employers and own account workers
4. Lower supervisory and technical occupations
5. Semi-routine occupations

Bar Charts and Pie Charts

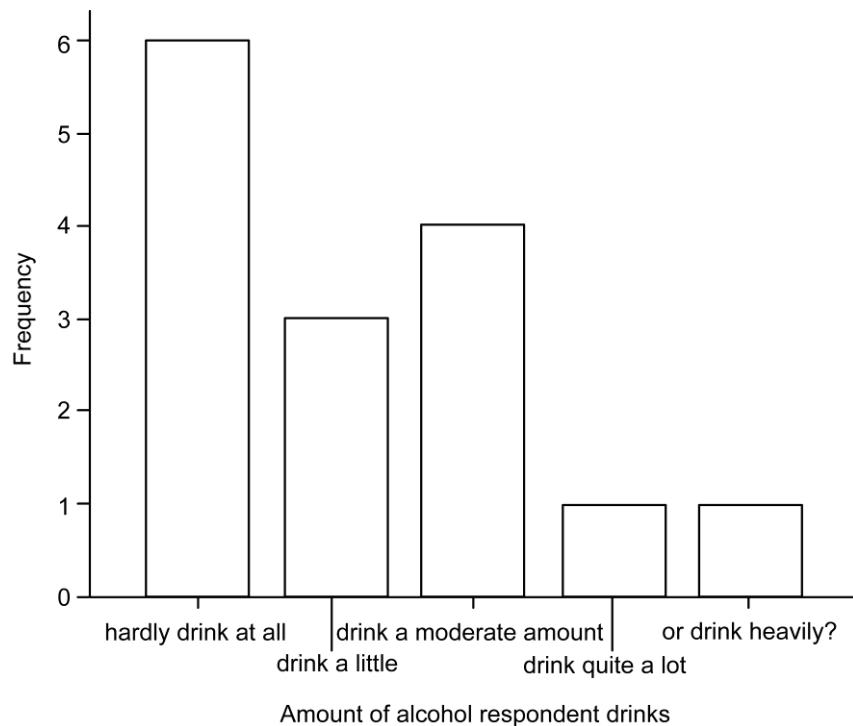


Fig. 3.2. Amount of alcohol respondent drinks

One simple device is the bar chart, a visual display in which bars are drawn to represent each category of a variable such that the length of the bar is proportional to the number of cases in the category.

A pie chart can also be used to display the same information. It is largely a matter of taste whether data from a categorical variable are displayed in a bar chart or a pie chart. In general, pie charts are to be preferred when there are only a few categories and when the sizes of the categories are very different.

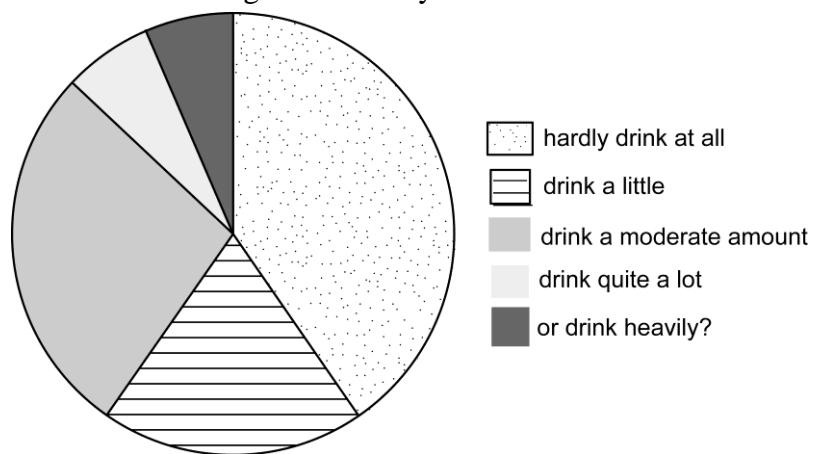


Fig. 3.3.

Bar charts and pie charts can be an effective medium of communication if they are well drawn.

Histograms

Charts that are somewhat similar to bar charts can be used to display interval level variables grouped into categories and these are called histograms. They are constructed in exactly the same way as bar charts except that the ordering of the categories is fixed, and care has to be taken to show exactly how the data were grouped.

3.2. NUMERICAL SUMMARIES OF LEVEL AND SPREAD

Let focus on the topic of working hours to demonstrate how simple descriptive statistics can be used to provide numerical summaries of level and spread. The chapter will begin by examining data on working hours in Britain taken from the

General Household Survey discussed in the previous chapter. These data are used to illustrate measures of level such as the mean and the median and measures of spread or variability such as the standard deviation and the midspread.

Working hours of couples in Britain

The histograms of the working hours distributions of men and women in the 2005 General Household Survey are shown in figures 3.1 and 3.2. We can compare these two distributions in terms of the four features introduced in the previous chapter, namely level, spread, shape and outliers. We can then see that:

- ❖ The male batch is at a higher level than the female batch
- ❖ The two distributions are somewhat similarly spread out
- ❖ The female batch is bimodal suggesting there are two rather different underlying populations
- ❖ The male batch is uni-modal

Summaries of level

The level expresses where on the scale of numbers found in the dataset the distribution is concentrated

Residuals

Another way of expressing this is to say that the residual is the observed data value minus the predicted value and in this case $45 - 40 = 5$. Any data value such as a measurement of hours worked or income earned can be thought of as being composed of two components: a fitted part and a residual part. This can be expressed as an equation:

$$\text{Data} = \text{Fit} + \text{Residual}$$

The median

The value of the case at the middle of an ordered distribution would seem to have an intuitive claim to typicality. Finding such a number is easy when there are very few cases. In the example of hours worked by a small random sample of 15 men (figure 3.4), the value of 48 hours per week fits the bill. There are six men who work fewer hours and seven men who work more hours while two men work exactly 48

hours per week. Similarly, in the female data, the value of the middle case is 37 hours. The data value that meets this criterion is called the median: the value of the case that has equal numbers of data points above and below it. The median hours worked by women in this very small sample is 11 hours less than the median for men. This numeric summary of the level of the data therefore confirms our first impressions from simply looking at the histograms in figures 3.1 and 3.2 that women generally work shorter hours than men.

Men's working hours (ranked)
30
37
39
40
45
47
48
Median value 48
50
54
55
55
67
70
80

Fig. 3.4. Men's working hours ranked to show the median

The arithmetic mean

Another commonly used measure of the centre of a distribution is the arithmetic mean. Indeed, it is so commonly used that it has even become known as the average. It is conventionally written as A (pronounced 'A bar'). To calculate it, first all of the values are summed, and then the total is divided by the number of data points. In more mathematical terms:

$$\frac{1}{N} \sum_{i=1}^N \gamma_i$$

We have come across N before. The symbol Y is conventionally used to refer to an actual variable. The subscript i is an index to tell us which case is being referred to. So, in this case, Y_i refers to all the values of the hours variable. The Greek letter Σ , pronounced 'sigma', is the mathematician's way of saying 'the sum of'.

Summaries of Spread

The second feature of a distribution visible in a histogram is the degree of variation or spread in the variable.

Once again, there are many candidates we could think of to summarize the spread. One might be the distance between the two extreme values (the range). Or we might work out what was the most likely difference between any two cases drawn at random from the dataset.

The midspread

The range of the middle 50 per cent of the distribution is a commonly used measure of spread because it concentrates on the middle cases. It is quite stable from sample to sample. The points which divide the distribution into quarters are called the quartiles (or sometimes 'hinges' or 'fourths'). The lower quartile is usually denoted QL and the upper quartile $Q0$. (The middle quartile is of course the median.) The distance between QL and $Q0$ is called the midspread (sometimes the 'interquartile range'), or the dQ for short.

Men's working hours (ranked)	
30	
37	
39	
40	
	$Q_L = 42.5$
45	
47	
48	
48	
50	
54	
55	
	$Q_U = 55$
55	
67	
70	
80	

Fig. 3.5. Men's working hours ranked and showing the upper and lower quartiles

There is a measure of spread which can be calculated from these squared distances from the mean. The standard deviation essentially calculates a typical value of these distances from the mean. It is conventionally denoted s , and defined as:

$$s = \sqrt{\left[\frac{\sum (Y_i - \bar{Y})^2}{(N-1)} \right]}$$

The deviations from the mean are squared, summed and divided by the sample size and then the square root is taken to return to the original units. The order in which the calculations are performed is very important. As always, calculations within brackets are performed first, then multiplication and division, then addition (including summation) and subtraction. Without the square root, the measure is called the variance, s^2 . The layout for a worksheet to calculate the standard deviation of the hours worked by this small sample of men is shown in figure 3.6.

Y	Y – Ȳ	(Y – Ȳ)²
54	3	9
30	-21	441
47	-4	16
39	-12	144
50	-1	1
48	-3	9
45	-6	36
40	-11	121
37	-14	196
48	-3	9
67	16	256
55	4	16
55	4	16
80	29	841
70	19	361
Sum = 765		Sum of squared residuals = 2472

Fig. 3.6. Worksheet for standard deviation of men's weekly working hours

$$s = \sqrt{\left[\frac{\sum (Y_i - \bar{Y})^2}{(N-1)} \right]} = \sqrt{\frac{2472}{14}} = 13.29$$

Interpreting Locational Summaries

In the examples discussed above the locational statistics for only a very small subsample of data of 15 cases from the GHS 2005 have been calculated by hand. It is useful to experiment with calculating locational statistics in this way in order to reach a better understanding of the meaning of these summary statistics. However, with larger batches of data the median, quartiles (and deciles) can be calculated very easily using a package such as Excel or SPSS.

Total Work Hours (Men)

N	Valid	6392
	Missing	8188
Median		39.000
Minimum		.00
Maximum		97.00
Percentiles	25	37.0000
	50	39.0000
	75	42.8750

Total Work Hours (Women)

N	Valid	6127
	Missing	9362
Median		35.0000
Minimum		.00
Maximum		97.00
Percentiles	25	20.0000
	50	35.0000
	75	37.5000

We can see that on average men tend to work more hours per week than women (39.2 hours vs 29.6 hours) and also the higher standard deviation for women, 12.3 vs 11.6 for men indicates that there is more variation among women in terms of the hours they usually work per week. It should also be noted that the figures for the means and standard deviations are pasted directly from the SPSS output. We can see that in each case the number of decimal places provided is four for the mean and five for the standard deviation.

Total Work Hours (Men)

N	Valid	6392
	Missing	8188
Mean		39.2268
Std. Deviation		11.64234

Total Work Hours (Women)

N	Valid	6127
	Missing	9362
Mean		29.5977
Std. Deviation		12.31122

3.3. SCALING AND STANDARDIZING

Data are produced not given

The word 'data' must be treated with caution. Literally translated, it means 'things that are given'.

There are often problems with using official statistics, especially those which are the by-products of some administrative process like, for example, reporting deaths to the Registrar-General or police forces recording reported crimes. Data analysts have to learn to be critical of the measures available to them, but in a constructive manner. As well as asking 'Are there any errors in this measure?' we also have to ask 'Is there anything better available?' and, if not, 'How can I improve what I've got?'

Improvements can often be made to the material at hand without resorting to the expense of collecting new data.

We must feel entirely free to rework the numbers in a variety of ways to achieve the following goals:

- ❖ to make them more amenable to analysis
- ❖ to promote comparability
- ❖ to focus attention on differences.

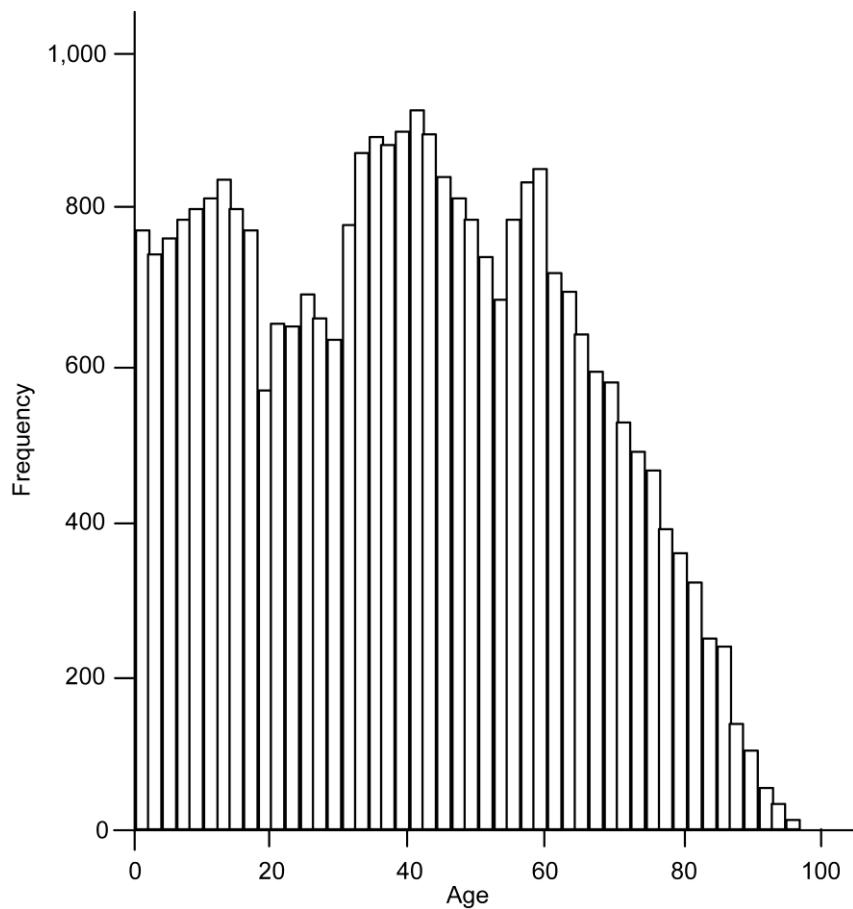


Fig. 3.7. Histogram

Consider various manipulations that can be applied to the data to achieve the above goals:

(i) Adding or subtracting a constant

One way of focusing attention on a particular feature of a dataset is to add or subtract a constant from every data value.

For example, in a set of data on weekly family incomes, it would be possible to subtract the median from each of the data values, thus drawing attention to which families had incomes below or above a hypothetical typical family.

The change made to the data by adding or subtracting a constant is fairly trivial. Only the level is affected; spread, shape and outliers remain unaltered. The reason for doing it is usually to force the eye to make a division above and below a particular point. A negative sign would be attached to all those incomes which were below the median in the example above. However, we sometimes add or subtract a constant to bring the data within a particular range.

(ii) Multiplying or dividing by a constant

Instead of adding a constant, we could change each data point by multiplying or dividing it by a constant.

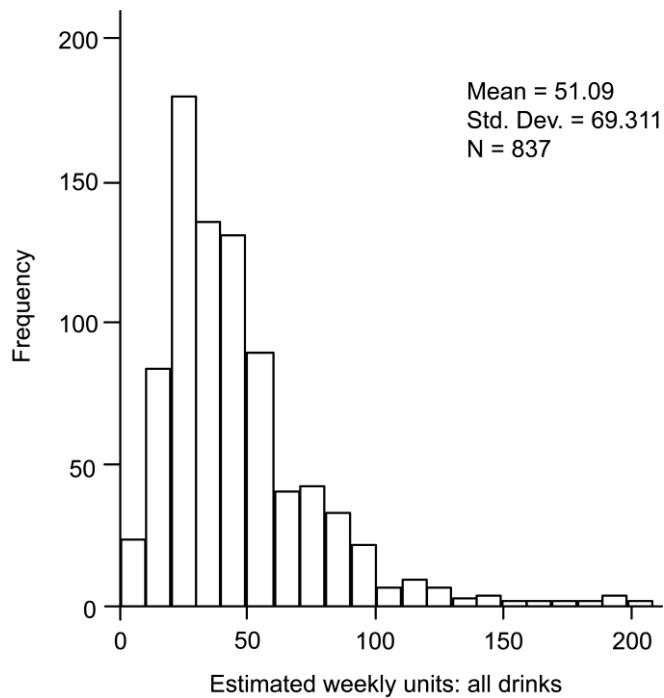


Fig. 3.8.(a) Histogram of weekly alcohol consumption of men who describe themselves as 'drinking quite a lot' or 'heavy drinkers'

A common example of this is the re-expression of one currency in terms of another. For example, in order to convert pounds to US dollars, the pounds are multiplied by the current exchange rate. Multiplying or dividing each of the values

has a more powerful effect than adding or subtracting. The result of multiplying or dividing by a constant is to scale the entire variable by a factor, evenly stretching or shrinking the axis like a piece of elastic. To illustrate this, let us see what happens if data from the General Household Survey on the weekly alcohol consumption of men who classify themselves as moderate or heavy drinkers are divided by seven to give the average daily alcohol consumption.

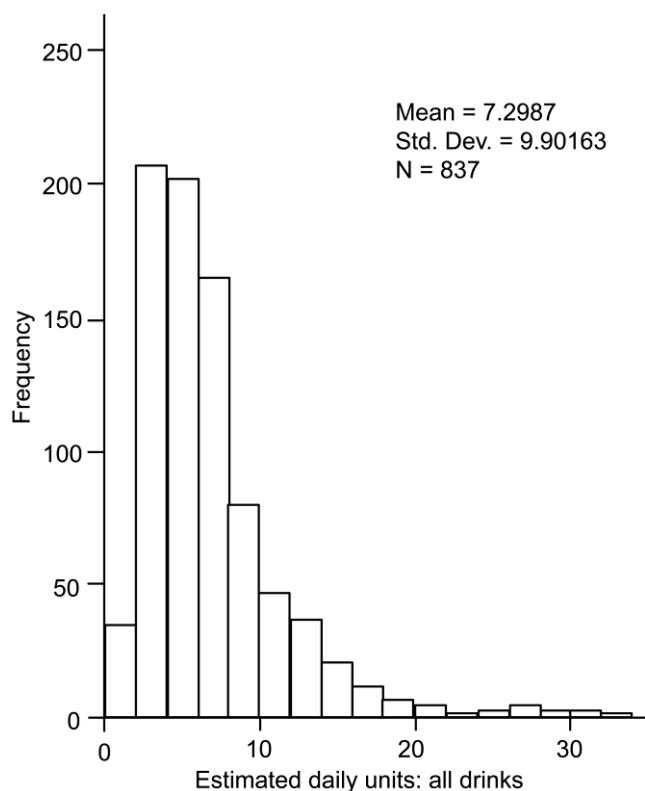


Fig. 3.8. (b) Histogram of daily alcohol consumption of men who describe themselves as 'drinking quite a lot' or 'heavy drinkers'

The overall shape of the distributions in figures 3.8 (a) and 3.8 (b) are the same. The data points are all in the same order, and the relative distances between them have not been altered apart from the effects of rounding. The whole distribution has simply been scaled by a constant factor.

In SPSS it is very straightforward to multiply or divide a set of data by a constant value. For example, using syntax, the command to create the variable drday ‘Average

daily alcohol consumption' from the variable drating 'Average weekly alcohol consumption' is as follows:

COMPUTE DRDAY — DRATING/7.

Alternatively, to create a new variable 'NEWVAR' by multiplying an existing variable 'OLDVAR' by seven the syntax would be:

COMPUTE NEWVAR = OLDVAR*7.

The 'Compute' command can also be used to add or subtract a constant,

for example:

COMPUTE NEWVAR = OLDVAR + 100.

COMPUTE NEWVAR = OLDVAR - 60.

The value of multiplying or dividing by a constant is often to promote comparability between datasets where the absolute scale values are different. For example, one way to compare the cost of a loaf of bread in Britain and the United States is to express the British price in dollars. Percentages are the result of dividing frequencies by one particular constant - the total number of cases.

(iii) Standardized Variables

In sections 3.2 and 3.3, we saw that subtracting a constant from every data value altered the level of the distribution and dividing by a constant scaled the values by a factor. In this section we will look at how these two ideas may be combined to produce a very powerful tool which can render any variable into a form where it can be compared with any other. The result is called a standardized variable.

To standardize a variable, a typical value is first subtracted from each data point, and then each point is divided by a measure of spread. It is not crucial which numerical summaries of level and spread are picked. The mean and standard deviation could be used, or the median and midspread:

$$\frac{Y_i - \bar{Y}}{s} \text{ or } \frac{Y_i - M(Y)}{dQ}$$

A variable which has been standardized in this way is forced to have a mean or median of 0 and a standard deviation or midspread of 1.

Two different uses of variable standardization are found in social science literature. The first is in building causal models, where it is convenient to be able to compare the effect that two different variables have on a third on the same scale.

The second use which is more immediately intelligible: standardized variables are useful in the process of building complex measures based on more than one indicator. In order to illustrate this, we will use some data drawn from the National Child Development Study (NCDS). This is a longitudinal survey of all children born in a single week of 1958.

There is a great deal of information about children's education in this survey. Information was sought from the children's schools about their performance at state examinations, but the researchers also decided to administer their own tests of attainment.

Rather than attempt to assess knowledge and abilities across the whole range of school subjects, the researchers narrowed their concern down to verbal and mathematical abilities. Each child was given a reading comprehension test which was constructed by the National Foundation for Educational Research for use in the study, and a test of mathematics devised at the University of Manchester.

The two tests were administered at the child's school and had very different methods of scoring. As a result they differed in both level and spread. As can be seen from the descriptive statistics in figure 3.4, the sixteen-year-olds in the National Child Development Study apparently found the mathematics test rather more difficult than the reading comprehension test. The reading comprehension was scored out of a total of 35 and sixteen-year-olds gained a mean score of 25.37, whereas the mathematics test was scored out of a possible maximum of 31, but the 16-year-olds only gained a mean score of 12.75.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.Deviation
Age 16 Test 1–reading Comprehension	11920	0	35	25.37	7.024
Age 16 Test 2–Mathematics Comprehension	11920	0	31	12.75	6.997
Valid N (listwise)	11920				

Fig. 3.9. Descriptive statistics for reading comprehension and mathematics test scores from NCDS age 16

The first two columns of figure 3.11 show the scores obtained on the reading and mathematics test by fifteen respondents in this study. There is nothing inherently interesting or intelligible about the raw numbers. The first score of 31 for the reading test can only be assessed in comparison with what other children obtained. Both tests can be thought of as indicators of the child's general attainment at school. It might be useful to try to turn them into a single measure of that construct.

1 Raw reading score	2 Raw maths score	3 Standardized reading score	4 Standardized maths score	5 Composite score of attainment
31	17	0.8	0.61	1.41
33	20	1.09	1.04	2.12
31	21	0.8	1.18	1.98
30	14	0.66	0.18	0.84
28	14	0.37	0.18	0.55
31	11	0.8	-0.25	0.55
29	8	0.52	-0.68	-0.16
28	17	0.37	0.61	0.98
23	8	-0.34	-0.68	-1.02
25	13	-0.05	0.04	-0.02
19	8	-0.91	-0.68	-1.59
32	25	0.94	1.75	2.69
31	22	0.80	1.32	2.12
29	8	0.52	-0.68	-0.16
30	17	0.66	0.61	1.27

Fig. 3.10. Scores of reading and mathematics tests at age 16

In order to create such a summary measure of attainment at age 16, we want to add the two scores together. But this cannot be done as they stand, because as we saw before, the scales of measurement of these two tests are different. If this is not immediately obvious try the following thought experiment. A 16-year-old who is average at reading but terrible at mathematics will perhaps score 25.4 (i.e. the mean score) on the reading comprehension test and 0 on the mathematics test. If these were summed the total is 25.4. However, a 16-year-old who is average at mathematics but can't read is likely to score 12.7 (i.e. the mean score) on the maths score and 0 on the reading comprehension. If these are summed the total would only be 12.7. If the two tests can be forced to take the same scale, then they can be summed.

This is achieved by standardizing each score. One common way of standardizing is to first subtract the mean from each data value, and then divide the result by the standard deviation. This process is summarized by the following formula, where the original variable 'Y' becomes the standardized variable 'Z'

$$Z = \frac{(Y_i - \bar{Y})}{\text{St.Dev.}}$$

For example, the first value of 31 in the reading test becomes:

$$(31 - 25.37) / 7 \text{ or } 0.8$$

The same individual's mathematics score becomes $(17 - 12.75) / 7$, or 0.61. This first respondent is therefore above average in both reading and maths. To summarize, we can add these two together and arrive at a score of 1.41 for attainment in general.

Similar calculations for the whole batch are shown in columns 3 and 4 of figure 3.11. We can see that the sixth person in this extract of data is above average in reading but slightly below average (by a quarter of a standard deviation) in mathematics. It should also be noted that any individual scoring close to the mean for both their reading comprehension and their mathematics test will have a total score close to zero. For example, the tenth case in figure 3.11 has a total score of -0.02.

The final column of figure 3.11 now gives a set of summary scores of school attainment, created by standardizing two component scores and summing them, so attainment in reading and maths have effectively been given equal weight.

It is very straightforward to create standardized variables using SPSS. by using the Descriptives command, the SPSS package will automatically save a standardized version of any variable.

First select the menus.

Analyze > Descriptive Statistics > Descriptives

The next stage is to select the variables that you wish to standardize, in this case N2928 and N2930, and check the box next to ‘Save standardized values as variables.’ The SPSS package will then automatically save new standardized variables with the suffix Z. In this example, two new variables ZN2928 and ZN2930 are created.

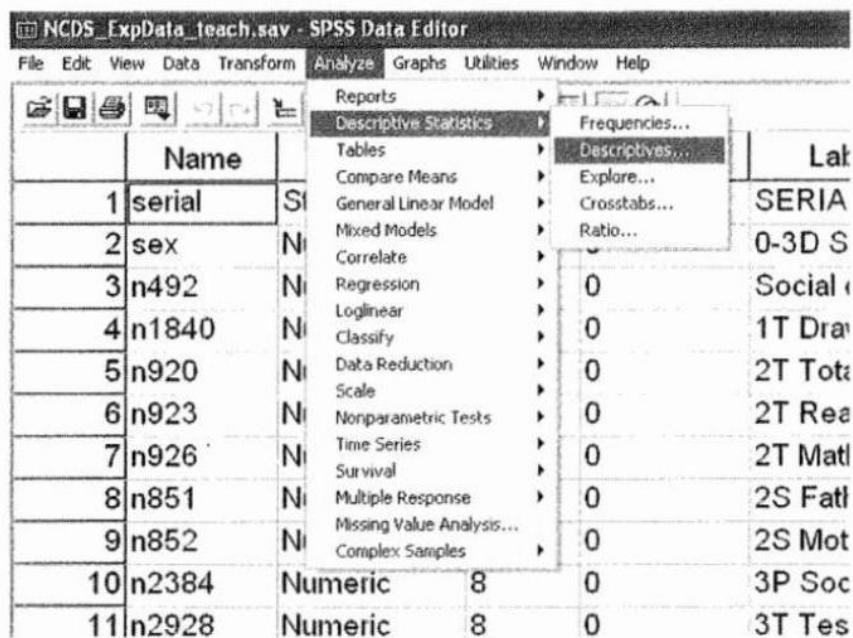


Fig. 3.11. Creating standardized variables using SPSS

The syntax to achieve this is as follows:

DESCRIPTIVES

VARIABLES = n2928 n2930 /

SAVE /STATISTICS = MEAN

STDDEV MIN MAX.

Standardizing the variables was a necessary, but not a sufficient condition for creating a simple summary score. It is also important to have confidence that the components are both valid indicators of the underlying construct of interest.

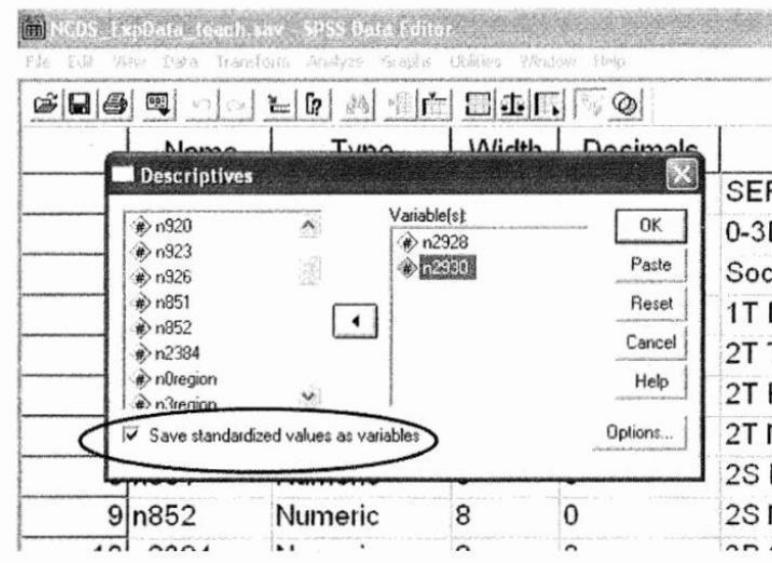


Fig. 3.12. Selecting variables to standardize

(iv) The Gaussian distribution

We are now ready to turn to the third feature of distributions, their shape. With level and spread taken care of, the shape of the distribution refers to everything that is left. In order to summarize the shape of a distribution, it would need to be simple enough to be able to specify how it should be drawn in a very few statements. For example, if the distribution were completely flat (a uniform distribution), this would be possible. We would only need to specify the value of the extremes and the number of cases for it to be reproduced accurately, and it would be possible to say exactly what proportion of the cases fell above and below a certain level.

However, many distributions do have a characteristic shape — a lump in the middle and tails straggling out at both ends. How convenient it would be if there was an easy way to define a more complex shape like this and to know what proportion of the distribution would lie above and below different levels.

One such shape, investigated in the early nineteenth century by the German mathematician and astronomer, Gauss, and therefore referred to as the Gaussian

distribution, is commonly used. It is possible to define a symmetrical, bell-shaped curve which looks like those in figure 3.14, and which contains fixed proportions of the distribution at different distances from the centre. The two curves in figure 3.14 look different — (a) has a smaller spread than (b) — but in fact they only differ by a scaling factor.

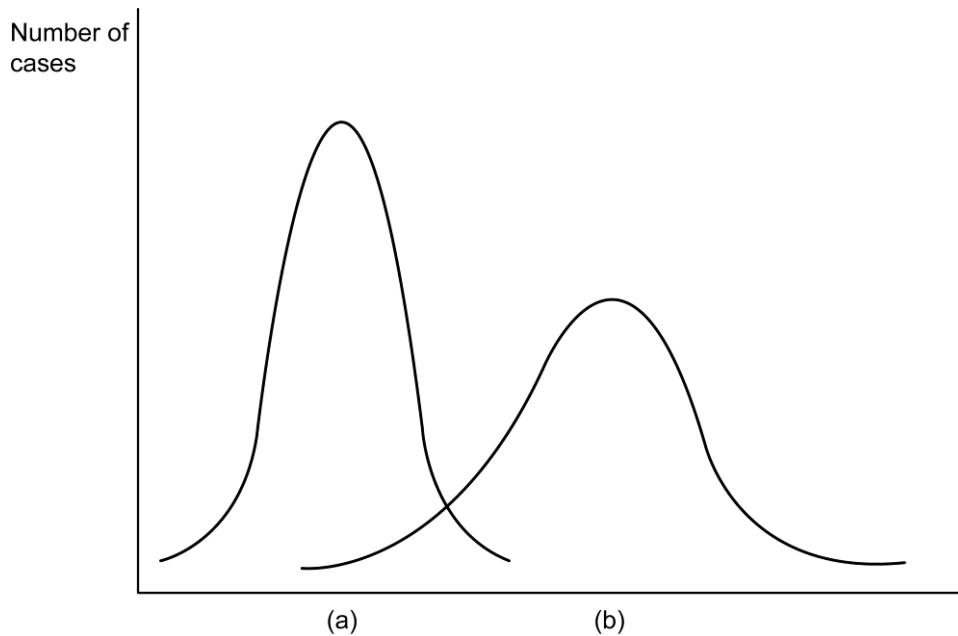


Fig. 3.13. The Gaussian distribution

Any Gaussian distribution has a very useful property: it can be defined uniquely by its mean and standard deviation. Given these two pieces of information, the exact shape of the curve can be reconstructed, and the proportion of the area under the curve falling between various points can be calculated.

This bell-shaped curve is often called ‘the normal distribution’. Its discovery was associated with the observation of errors of measurement. If sufficient repeated measurements were made of the same object, it was discovered that most of them centred around one value (assumed to be the true measurement), quite a few were fairly near the centre, and measurements fairly wide of the mark were unusual but did occur. The distribution of these errors of measurement often approximated to the bell-shape in figure 3.14.

(v)) Standardizing with respect to an appropriate base

In the scaling and standardizing techniques considered up to now, the same numerical adjustment has been made to each of the values in a batch of data. Sometimes, however, it can be useful to make the same conceptual adjustment to each data value, which may involve a different number in each case.

A batch of numbers may be reworked in several different ways in order to reveal different aspects of the story they contain. A dataset which can be viewed from several angles is shown in figure 3.15: the value of the lower quartile, the median and the upper quartile of male and female earnings in the period between 1990 and 2000. The data are drawn from the New Earnings Survey that collects information about earnings in a fixed period each year from the employers of a large sample of employees.

Year	Male Earnings			Female Earnings		
	QL	M	Qu	QL	M	Qu
1990	193.4	258.2	347.5	136.2	177.5	244.7
1991	206.9	277.5	376.5	150.6	195.7	271.6
1992	219.3	295.9	401.9	161.4	211.3	295.9
1993	226.0	304.6	417.3	168.2	221.6	309.1
1994	231.1	312.8	427.3	174.6	229.4	320.1
1995	237.1	323.2	442.7	179.5	237.2	332.5
1996	245.2	334.9	460.7	186.8	248.1	347.3
1997	256.4	349.7	480.0	196.1	260.5	364.7
1998	265.3	362.8	499.0	203.6	270.0	379.1
1999	274.5	374.3	517.3	213.3	284.0	398.2
2000	284.7	389.7	537.7	223.6	296.7	417.6

Fig. 3.14. Male and female earnings 1990-2000 gross earnings in pounds per week for full-time workers on adult rates whose pay was not affected by absence

As the figures stand, the most dominant feature of the dataset is a rather uninteresting one: the change in the value of the pound. While the median and mid-spreads of the money incomes each year have increased substantially in this period, real incomes and differentials almost certainly have not. How could we present the data in order to focus on the trend in real income differentials over time?

One approach would be to treat the distribution of incomes for each sex in each year as a separate distribution, and express each of the quartiles relative to the median. The result of doing this is given in figure 3.16.

Year	Male Earnings			Female Earnings		
	QL	M	Qu	QL	M	Qu
1990	75	100	135	77	100	138
1991	75	100	136	77	100	139
1992	74	100	136	76	100	140
1993	74	100	137	76	100	139
1994	74	100	137	76	100	140
1995	73	100	137	76	100	140
1996	73	100	138	75	100	140
1997	73	100	137	75	100	140
1998	73	100	138	75	100	140
1999	73	100	138	75	100	140
2000	73	100	138	75	100	141

Fig. 3.15. Male and female earnings relative to medians for each sex

3.4. INEQUALITY

Prosperity and Inequality :

There are a number of reasons why we might want to reduce inequality in society. For example, as Layard (2005) argues, if we accept that extra income has a bigger impact on increasing the happiness of the poor than the rich, this means that if some

money is transferred from the rich to the poor this will increase the happiness of the poor more than it diminishes the happiness of the rich. This in turn suggests that the overall happiness rating of a country will go up if income is distributed more equally. Of course, as Layard acknowledges, the problem with this argument is that it only works if it is possible to reduce inequality without raising taxes to such an extent that there is no longer an incentive for individuals to strive to make money so that the total income is reduced as a result of policies aimed at redistribution. It is clearly important to understand the principal ways of measuring inequality if we are to monitor the consequences of changing levels of inequality in society. This chapter will focus on how we can measure inequality in such a way as to make it possible to compare levels of inequality in different societies and to look at changes in levels of inequality over time.

Income and Wealth :

Considered at the most abstract level, income and wealth are two different ways of looking at the same thing. Both concepts try to capture ways in which members of society have different access to the goods and services that are valued in that society. Wealth is measured simply in pounds, and is a snapshot of the stock of such valued goods that any person owns, regardless of whether this is growing or declining. Income is measured in pounds per given period, and gives a moving picture, telling us about the flow of revenue over time.

For the sake of simplicity, we restrict our focus to the distribution of income. We will look in detail at the problems of measuring income and then consider some of the distinctive techniques for describing and summarizing inequality that have evolved in the literature on economic inequality.

There are four major methodological problems encountered when studying the distribution of income:

1. How should income be defined?
2. What should be the unit of measurement?
3. What should be the time period considered?
4. What sources of data are available?

Definition of Income

To say that income is a flow of revenue is fine in theory, but we have to choose between two approaches to making this operational. One is to follow accounting and tax practices, and make a clear distinction between income and additions to wealth. With this approach, capital gains in a given period, even though they might be used in the same way as income, would be excluded from the definition. This is the approach of the Inland Revenue, which has separate taxes for income and capital gains. In this context a capital gain is defined as the profit obtained by selling an asset that has increased in value since it was obtained. However, interestingly, in most cases this definition (for the purposes of taxation) does not include any profit made when you sell your main home.

The second approach is to treat income as the value of goods and services consumed in a given period plus net changes in personal wealth during that period. This approach involves constantly monitoring the value of assets even when they do not come to the market. That is a very hard task.

So, although the second approach is theoretically superior, it is not very practical and the first is usually adopted.

The definition of income usually only includes money spent on goods and services that are consumed privately. But many things of great value to different people are organized at a collective level: health services, education, libraries, parks, museums, even nuclear warheads.

The benefits which accrue from these are not spread evenly across all members of society. If education were not provided free, only families with children would need to use their money income to buy schooling.

Sources of income are often grouped into three types:

- ❖ earned income, from either employment or self-employment;
- ❖ unearned income which increases from ownership of investments, property, rent and so on;
- ❖ transfer income, that is benefits and pensions transferred on the basis of entitlement, not on the basis of work or ownership, mainly by the government but occasionally by individuals .

Lower boundary of group (£ per week gross income)	2003/4
2 nd decile	£ 124
3 rd decile	£ 193
4 th decile	£ 263
5 th decile	£ 351
6 th decile	£ 445
7 th decile	£ 558
8 th decile	£ 673
9 th decile	£ 828
10 th decile	£ 1092

Fig. 3.16. Lower boundaries of each gross income decile group

Measuring inequality: quantiles and quantileshares :

Figure 3.17 illustrates one method for summarizing data on the income received by households. It displays the gross income of different deciles of the distribution (gross income is defined as income from employment, self-employment, investments, pensions, etc. plus any cash benefits or tax credits). For example, figure 3.17 shows that in 2003/4 the poorest ten per cent of households had a gross income of less than 124 pounds per week, while the richest ten per cent of households had a gross income of over 1,092 pounds per week. The median gross income is 445 pounds per week.

An alternative technique for examining the distribution of incomes is to adopt the quantile shares approach. This is illustrated in figure 3.18, which is a modified version of a table produced as part of the annual report from the Office for National Statistics 'The effects of taxes and benefits on household income'. The income of all units falling in a particular quantile group — for example, all those with income above the top decile, is summed and expressed as a proportion of the total income received by everyone.

	Percentage shares of equivalized income for ALL households			
	Original income	Gross income	Disposable Income	Post-tax income
Quintile group				
Bottom	3	7	8	7
2 nd	7	11	12	12
3 rd	15	16	17	16
4 th	24	22	22	22
Top	51	44	42	44
All households	100	100	100	100
Decile group				
Bottom	1	3	3	2
Top	33	29	27	29

Fig. 3.17. Percentage shares of household income, 2003-4

Cumulative income shares and Lorenz curves :

Neither quantiles nor quantile shares lend themselves to an appealing way of presenting the distribution of income in a graphical form. This is usually achieved by making use of cumulative distributions. The income distribution is displayed by plotting cumulative income shares against the cumulative percentage of the population.

The cumulative distribution is obtained by counting in from one end only. Income distributions are traditionally cumulated from the lowest to the highest incomes. To see how this is done, consider the worksheet in figure 3.19. The bottom 5 percent receive 0.47 percent of the total original income, and the next 5 percent receive 0.51 percent. In summing these, we can say that the bottom 10 per cent receive 0.98 per cent of the total original income. We work our way up through the incomes in this

fashion. It can be noted that the first two columns of this table are simply a more detailed version of the data presented in figure 3.18. For example, from figure 3.18 we can see that the top quintile group receives 51 per cent of original income; this figure is also obtained if you sum the first three numbers in the first column of figure 3.19.

The cumulative percentage of the population is then plotted against the cumulative share of total income. The resulting graphical display is known as a Lorenz curve. It was first introduced in 1905 and has been repeatedly used for visual communication of income and wealth inequality. The Lorenz curve for pre-tax income in 2003/4 in the UK is shown in figure 3.20.

Cumulative share of population	Percentage of total income received by the quantile		Cumulative share of total income	
	Original income	Post-tax income	Original income	Post-tax income
100	21.6	18.9	100	100
95	11.8	9.8	78.4	81.1
90	17.6	15	66.6	71.3
80	13.5	12.1	49	56.3
70	10.5	10	35.5	44.2
60	8.5	8.6	25	34.2
50	6.3	7.4	16.5	25.6
40	4.6	6.3	10.2	18.2
30	2.9	5.3	5.6	11.9
20	1.72	4.3	2.7	6.6
10	0.51	1.76	0.98	2.3
5	0.47	0.54	0.47	0.54

Fig. 3.18. Cumulative income shares : 2003-4

Lorenz curves have visual appeal because they portray how near total equality or total inequality a particular distribution falls. If everyone in society had the same income, then the share received by each decile group, for example, would be 10 per cent, and the Lorenz curve would be completely straight, described by the diagonal line.

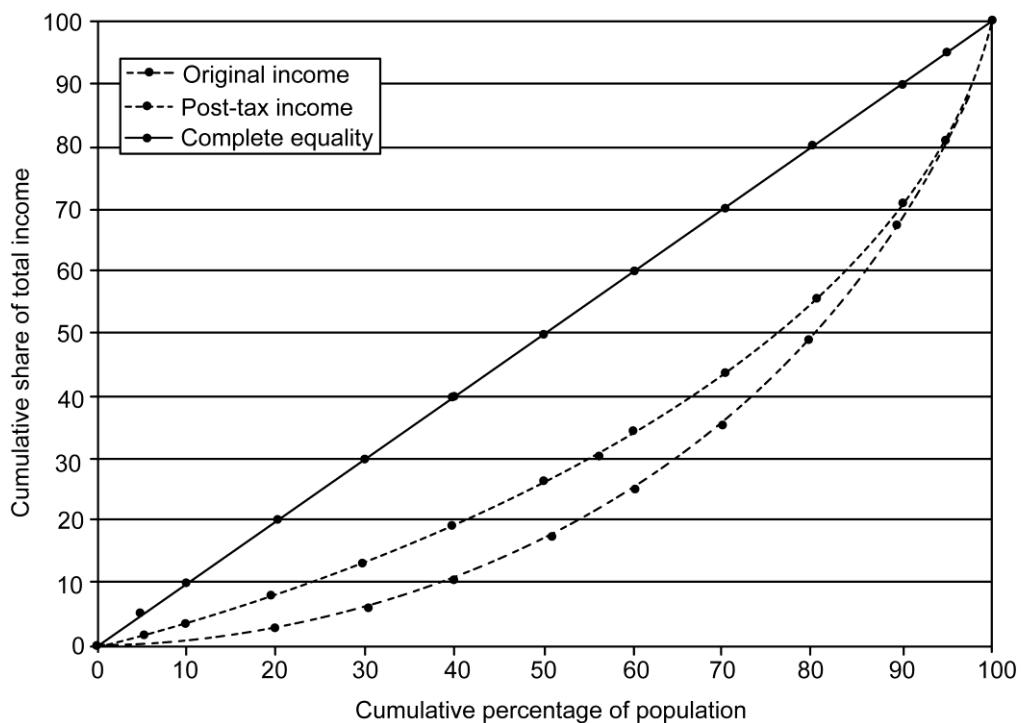


Fig. 3.19. Lorenz curves of income: 2003-4

Desirable properties in a summary measure of inequality

Scale independence

However, it is important that the measure be sensitive to the level of the distribution. Imagine a hypothetical society containing three individuals who earned 5,000, 10,000 and 15,000 pounds respectively. If they all had an increase in their incomes of 1 million pound, we would expect a measure of inequality to decline, since the differences between these individuals would have become trivial. The standard deviation and midspread would be unaffected. A popular approach is to log income data before calculating the numerical summaries of spread. If two distributions differ by a scaling factor, the logged distributions will differ only in

level. However, if they differ by an arithmetic constant, they will have different spreads when logged. The existence of units with zero incomes leads to problems, since the log of zero cannot be defined mathematically. An easy technical solution to this problem is to add a very small number to each of the zeros. If a numerical summary of spread in a logged distribution met the other desirable features of a measure of inequality, we could stop here. Unfortunately, it does not.

The principle of transfers

It makes intuitive sense to require that a numerical summary of inequality should decline whenever money is given by a rich person to a poor person, regardless of how poor or how rich, and regardless of how much money is transferred (provided of course that the amount is not so big that the previously poor person becomes even richer than the previously rich person). One numerical summary — the income share of a selected quantile group — fails to meet this principle. By focusing on one part of the distribution only, perhaps the top 5 per cent, it would fail to record a change if a transfer occurred elsewhere in the distribution. Similar objections apply to another commonly used summary, the decile ratio, which simply expresses the ratio of the upper decile to the lower decile. Other inequality measures meet this principle, and so are to be preferred.

However, they unfortunately still fail to agree on an unambiguous ranking of different societies in terms of income inequality, because they are sensitive in different ways to transfers of varying amounts and at different points in the income scale. Cowell (1977) argues that the principle of transfers should be strengthened to specify that the measure of inequality should be sensitive only to the distance on the income scale over which the transfer is made, not to the amount transferred.

He also adds a third principle to the two considered here, that of decomposition: a decline in inequality in part of a distribution should lead to a decline in inequality overall. We shall return to these more stringent criteria below.

3.5. SMOOTHING

Time series such as that shown in the second column of figure 3.21 are displayed by plotting them against time, as shown in figure 3.22. When such trend lines are smoothed, the jagged edges are sown off. A smoothed version of the total numbers of

recorded crimes over the thirty years from the mid 1960s to the mid 1990s is displayed in figure 3.23.

Year	Total Recorded Crimes	Year	Total Recorded Crimes
1965	1,133,882	1980	2,688,235
1966	1,199,859	1981	2,963,764
1967	1,207,354	1982	3,262,422
1968	1,289,090	1983	3,247,030
1969	1,488,638	1984	3,499,107
1970	1,555,995	1985	3,611,883
1971	1,646,081	1986	3,847,410
1972	1,690,219	1987	3,892,201
1973	1,657,669	1988	3,715,767
1974	1,963,360	1989	3,870,748
1975	2,105,631	1990	4,543,611
1976	2,135,713	1991	5,276,173
1977	2,636,517	1992	5,591,717
1978	2,561,499	1993	5,526,255
1979	2,536,737	1994	5,252,980

Fig. 3.20. Total numbers of recorded crimes: 1965-94

Most people, if asked to smooth the data by eye, would probably produce a curve similar to that in figure 3.23, which has been derived using a well-defined arithmetic procedure described later in the chapter. However, smoothing by an arithmetic procedure can sometimes reveal patterns not immediately obvious to the naked eye.

The aim of smoothing

Figure 3.22 was constructed by joining points together with straight lines. Only the points contain real information of course. The lines merely help the reader to see

the points. The result has a somewhat jagged appearance. The sharp edges do not occur because very sudden changes really occur in numbers of recorded crimes.

They are an artefact of the method of constructing the plot, and it is justifiable to want to remove them. According to Tukey (1977, p. 205), the value of smoothing is 'the clearer view of the general, once it is unencumbered by detail'. The aim of smoothing is to remove any upward or downward movement in the series that is not part of a sustained trend.

Sharp variations in a time series can occur for many reasons. Part of the variation across time may be error. For example, it could be sampling error. The opinion-poll data used later in this chapter were collected in monthly sample surveys, each of which aimed to interview a cross-section of the general public, but each of which will have deviated from the parent population to some extent. Similarly, repeated measures may each contain a degree of measurement error. In such situations, smoothing aims to remove the error component and reveal the underlying true trend. But the variable of interest may of course genuinely swing around abruptly. For example, the monthly count of unemployed people rises very sharply when school-leavers come on to the register. In these cases, we may want to smooth to remove the effect of events which are unique, or which are simply not the main trend in which we are interested. It is good practice to plot the rough as well as the smooth values, to inspect exactly what has been discarded.

In engineering terms we want to recover the signal from a message by filtering out the noise. The process of smoothing time series also produces such a decomposition of the data. In other words, what we might understand in engineering as

$$\text{Message} = \text{Signal + Noise}$$

becomes

$$\text{Data} = \text{Smooth + Rough}$$

This choice of words helps to emphasize that we impose no a priori structure on the form of the fit. The smoothing procedure may be determined in advance, but this is not the case for the shape and form of the final result: the data are allowed to speak for themselves. Put in another way, the same smoothing recipe applied to different

time series will produce different resulting shapes for the smooth, which, as we will see in, is not the case when fitting straight lines.

As so often, this greater freedom brings with it increased responsibility. The choice of how much to smooth will depend on judgement and needs. If we smooth too much, the resulting rough will itself exhibit a trend. Of course, more work is required to obtain smoother results, and this is an important consideration when doing calculations by hand. The smoothing recipe described later in the chapter generally gives satisfactory results and involves only a limited amount of computational effort.

Most time series have a past, a present and a future. For example, the rising crime figures plotted in figure 3.22 and figure 3.23 are part of a story that begins well before the 1960s and continues to the present day. However, the goal of the smoothing recipes explained in this chapter is not the extrapolation of a given series into the future. The following section provides the next instalment in this story and discusses what happened after the very dramatic increases in total recorded crime in the early 1990s.

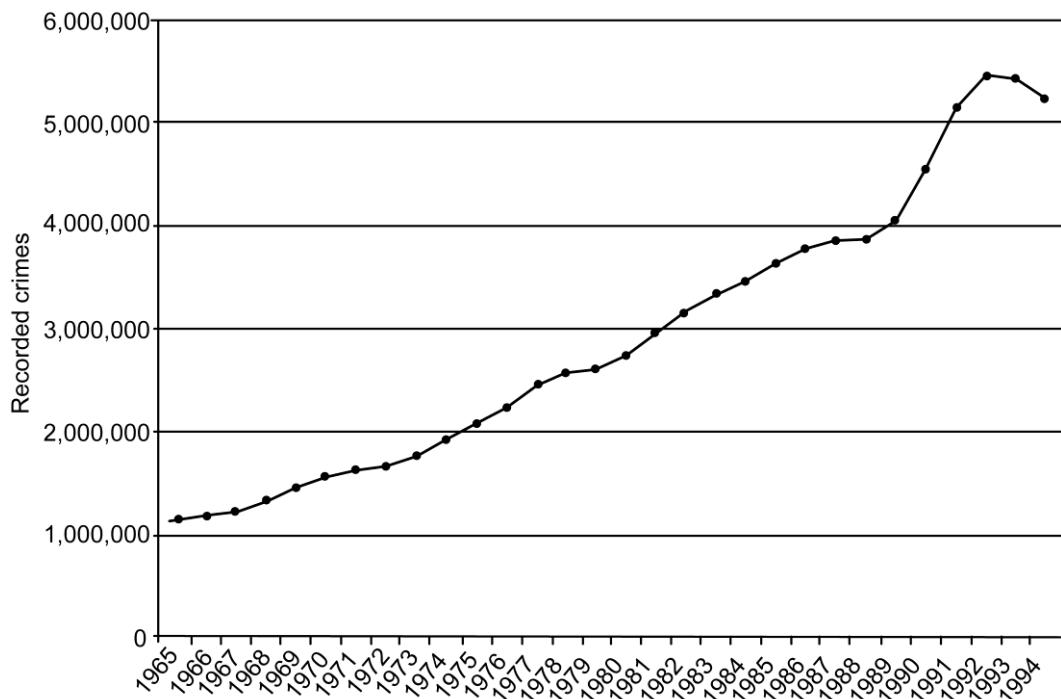


Fig. 3.21. Total number of recorded crimes: unsmoothed

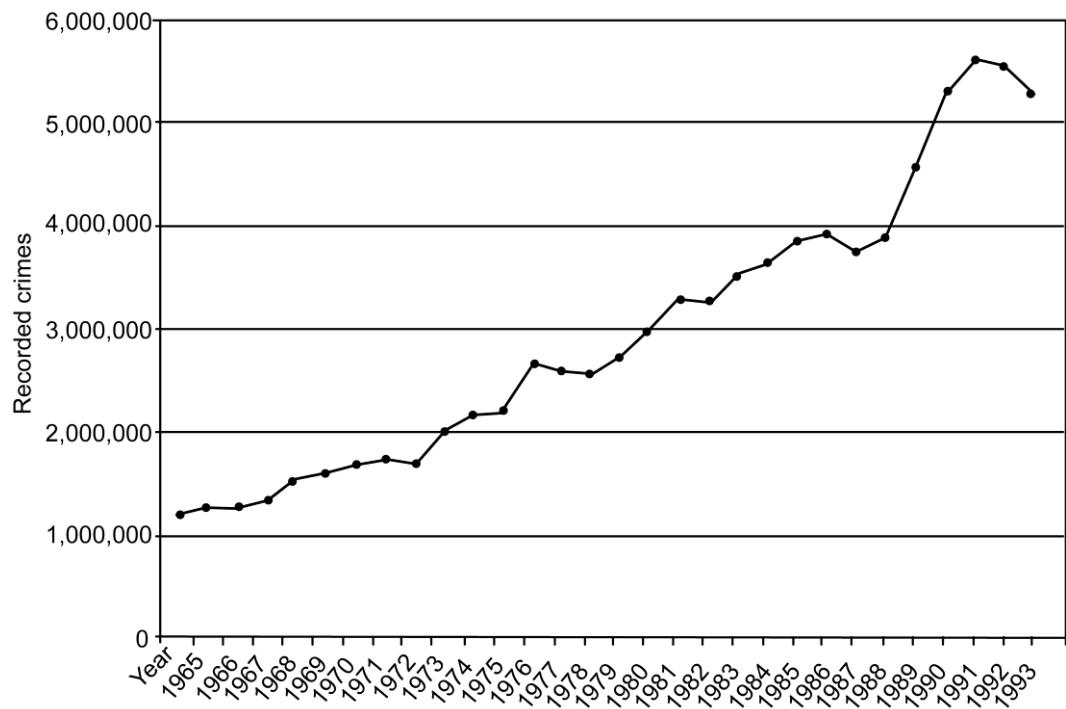


Fig. 3.22. Total recorded crimes 1965-94: smoothed

TWO MARKS QUESTION AND ANSWERS (PART- A)

1. *What is a Univariate analysis?*

Among all the forms of analytical methods that data analysts practice, univariate analysis is considered one of the basic forms of analysis. It is typically the first step to understanding a dataset. The idea of univariate analysis is to first understand the variables individually. Then, you move into analyzing two or more variables simultaneously.

2. *What is the basis for data analysis?*

Two organizing concepts have become the basis of the language of data analysis: cases and variables. The cases are the basic units of analysis, the things about which information is collected. The word variable expresses the fact that this feature varies across different cases.

3. *Distinguish Bar charts and pie charts*

One simple device is the bar chart, a visual display in which bars are drawn to represent each category of a variable such that the length of the bar is proportional to the number of cases in the category.

A pie chart can be used to display the above said information but in a different perspect as whether data from a categorical variable are displayed in a bar chart or a pie chart. In general, pie charts are to be preferred when there are only a few categories and when the sizes of the categories are very different.

4. *Define level and spread in data exploration*

We will focus on the working hours to demonstrate how simple descriptive statistics can be used to provide numerical summaries of level and spread. We begin by examining data on working hours in Britain taken from the General Household Survey. These data are used to illustrate measures of level such as the mean and the median and measures of spread or variability such as the standard deviation and the midspread.

5. *What is a mid-spread?*

The points which divide the distribution into quarters are called the quartiles (or sometimes 'hinges' or 'fourths'). The lower quartile is usually denoted Q_L and the upper quartile Q_U . (The middle quartile is of course the median.) The distance between Q_L and Q_U is called the midspread (sometimes the 'interquartile range'), or the d_Q for short.

6. *Differentiate scaling and standardizing.*

Subtracting a constant from every data value altered the level of the distribution and dividing by a constant scaled the values by a factor. These two ideas may be combined to produce a very powerful tool which can render any variable into a form where it can be compared with any other. The result is called a standardized variable.

7. *Define a Gaussian Distribution.*

Many distributions do have a characteristic shape a lump in the middle and tails straggling out at both ends. One such shape, investigated in the early

nineteenth century by the German mathematician, Gauss, and therefore referred to as the Gaussian distribution, is commonly used. It is possible to define a symmetrical, bell-shaped curve which contains fixed proportions of the distribution at different distances from the centre.

8. *Why there is a need to reduce in-equality?*

If we accept that extra income has a bigger impact on increasing the happiness of the poor than the rich, this means that if some money is transferred from the rich to the poor this will increase the happiness of the poor more than it diminishes the happiness of the rich. This in turn suggests that the overall happiness rating of a country will go up if income is distributed more equally.

9. *What Is a Lorenz Curve?*

A Lorenz curve, developed by American economist Max Lorenz in 1905, is a graphical representation of income inequality or wealth inequality. The graph plots percentiles of the population on the horizontal axis according to income or wealth and plots cumulative income or wealth on the vertical axis.

10. *Define smoothing time series*

Smoothing is usually done to help us better see patterns, trends for example, in time series. Generally smooth out the irregular roughness to see a clearer signal. For seasonal data, we might smooth out the seasonality so that we can identify the trend. Smoothing doesn't provide us with a model, but it can be a good first step in describing various components of the series. The term **filter** is sometimes used to describe a smoothing procedure. For instance, if the smoothed value for a particular time is calculated as a linear combination of observations for surrounding times, it might be said that we've applied a linear filter to the data

11. *What is the Gini Coefficient?*

The Gini coefficient (Gini index or Gini ratio) is a statistical measure of economic inequality in a population. The coefficient measures the dispersion of income or distribution of wealth among the members of a population. What is the Gini Coefficient?

The Gini coefficient (Gini index or Gini ratio) is a statistical measure of economic inequality in a population. The coefficient measures the dispersion of income or distribution of wealth among the members of a population.

PART - B & C

1. Categorize variables and explain its distribution.
2. Define level and summarize in detail.
3. Summarize spread depicting an example.
4. Elaborate scaling and standardizing with an example.
5. Clarify in-equality with a proper example.
6. Explain the concept smoothing time series with an example.
