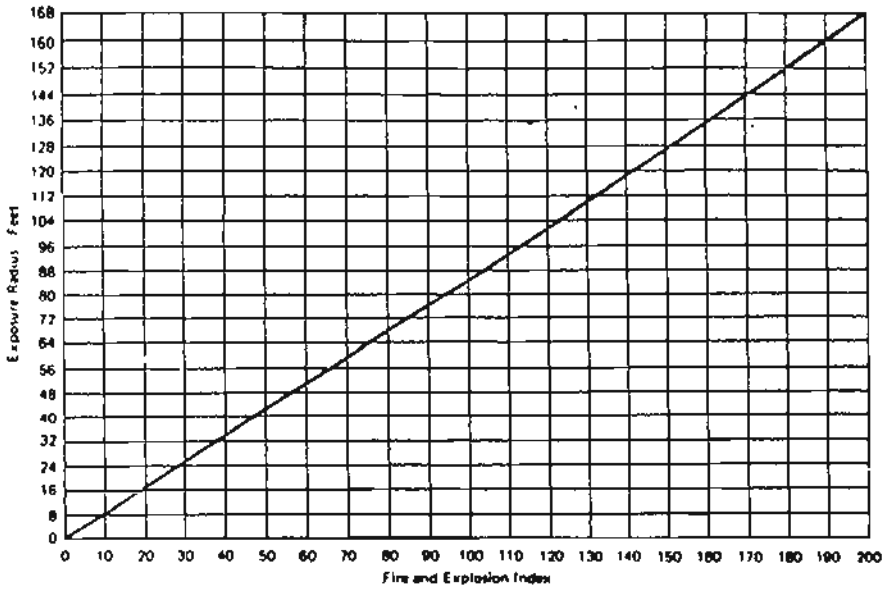


Fig. 4.11

damage probability from the Process Unit we are considering. In the same way we can see that the other Process Unit represents a 32% damage probability to an area of about 3633 square feet surrounding it.

The documentation provides a rough qualitative equivalent of the severity of the Fire and Explosion Index in a table similar to the one shown below:

Fire & Explosion Index Range	*Degree of Hazard*
1 - 60	Light
61 - 96	Moderate
97 - 127	Intermediate
128 - 158	Heavy
159 -	Severe

7. The Damage Factor represents the probable extent of any fire or explosion and so if we calculate the value of property within the Area of Exposure we could have some idea of what such an incident may cost. This is done by taking the replacement value of all such plant and multiplying it by the Damage Factor. Let us say that we had £280,000 in the Area then the Maximum Probable Property Damage, MPPD, would be £280,000 x 0.74 for the first Process Unit. This gives an expected damage figure, an MPPD, of £207,200.

8. The MPPD is very much a base figure as no account has been taken of any good features of the plant. So far we have concentrated on all those features which could increase the chance of loss. Now we can turn to the credit features and allow certain reductions in the MPPD as a result. A long list of credit points are given which include:
   - sprinklers
   - emergency shutdown systems
   - drainage
   - certain good operating procedures

   These various features produce a credit allowance of a figure between 0 and 1. All allowances to which the unit is entitled are multiplied together and the result is a Credit Factor. The MPPD is multiplied by the Credit Factor in order to reflect the good points in the plant or process and this provides us with the actual MPPD. In our example let us say we have a Credit Factor of 0.45. This would then produce an actual MPPD of £207,200 x 0.45 = £93,240. This is much lower than the base MPPD and reflects the good points for which credit has now been allowed.

9. We could pull all the work we have done on Process Unit One together in one summary table as follows:

   Fire and Explosion Index = 90
   Radius of exposure = 76ft
   Value inside Area of Exposure = £280,000
   Damage Factor = 0.74
   Base MPPD = £207,200
   Credit Factor = 0.45
   Actual MPPD = £93,240

   The steps are as we have outlined them above but clearly some experience in constructing the index is desirable and again the team approach may well be the best way to go about the whole job. The index does not identify individual risks but it does try to put some measure on the level of exposure arising out of the activities at the plant. With the index the risk manager can make comparisons across all the plant his company may operate and can monitor changes from year to year.

Chapter Five

STATISTICAL ANALYSIS OF RISK 1

5.0 In the previous chapters we have concentrated on the analysis of risk. This has involved both identification and then the formal analysis of the identified risk. We noted, as we did this, that some techniques were more concerned with identification than analysis, that others adopted a broad view of risk and that some concentrated on the qualitative rather than the quantitative analysis of risk.

In the end we agreed that the most important thing was that risk was brought to light and that a number of different techniques could be used to satisfy individual problems. Once risks have been identified we are often left with a large volume of information and can use this information in statistical analysis.

This chapter and the next are concerned with statistical risk analysis. The approach which has been adopted is, it is hoped, a practical one. Certain basic concepts in statistics will be introduced with the aid of a practical problem and the relevance of statistics should emerge.

5.1 Before introducing the example and beginning the chapter, let us make one or two points of a very general nature:

- What we will do in this chapter will be of an introductory nature. It is not possible, nor would it be worthwhile, in this study guide to provide a comprehensive text on statistics. This chapter and the one which follows will only be an introduction to a much wider topic.
- In addition to being introductory in nature these chapters are also selective. They will concentrate on techniques which seem to have some relevance for risk management. This means of necessity that certain topics will be left out.
- The chapters are written on the assumption that the reader has very little knowledge of statistics already. In fact a person with no previous knowledge should be able to read them without too much difficulty.
- Finally these chapters recognise the fact that statistics and quantitative risk analyses may not be the most popular topics among risk and insurance managers. However we live in an increasingly quantitative world and the risk manager must endeavour to keep abreast of developments.

5.2 Gathering Data

The first stage in statistical risk analysis is the gathering of data. Risk and insurance departments do generate large volumes of information on claims, policies, premiums etc. Consider for a moment all the data that is collected by and kept in your own department.

Very often this data is collected more as a matter of routine than by definite conscious decision to gather it. The risk manager has data and must look over it and see to what uses it can be put. In the unique case of a risk manager starting from scratch he can decide himself what he wants to collect but this will be unusual. However the data is gathered, one point still remains. The end result will only be as good as the data with which you start. For example you will not be able to analyse employee injuries by shift if you do not record the shift during which the accident occurred. By the same token there is as much danger in gathering unnecessary information. Why ask for the employee number when recording accident details if you do not intend to use this in some later analysis?

In deciding on what information to collect it is essential to have a clear picture of what the end result is likely to be. This is often best achieved by imagining what kind of statements you would like your analysis to produce at the end of the day. For example if you want to make statements such as;

> "Forty-three percent of all back injuries, many of which involved lifting sheet metal, occurred in the moulding plant during the early shift."

If this is the kind of statement, among many others, which you could imagine including in some report at a certain stage in the future then the correct data must be gathered. This simple statement implies that you must collect data on employee injuries which includes the type of injury, agent of injury, place of injury and time. It is therefore a useful discipline to consider all that you might want to say by way of a report at the end of your statistical analysis and then make sure you gather the appropriate data to begin with.

5.2.1 There are a number of techniques which can be employed in the gathering of data and most text books on statistics will include a comprehensive list of them. In the risk management situation it is likely that the data already exists. In only a few cases will the risk manager have to set about devising a system for collecting data.

What may be necessary is some adjustment in the format in which the data is gathered. If we take claims information as an example, it may well be that a report of, let us say, industrial injuries, is submitted to the insurance department. The risk manager will want to ensure that this report contains all the information he will need for his own analysis in addition to whatever is required by the insurers for their purposes.
The design of forms for gathering information then becomes quite important and this is an aspect of gathering data with which he may be concerned. The design of any form will depend on the nature of the data being collected, what the risk manager hopes to do with the data and any style which the company as a whole may normally adopt for reporting. There are however a number of general points we could make about documents which are intended to be used for gathering risk data.

5.2.2 The following points should be considered when designing any form for gathering information. This could be any form of information which the risk manager may want to collect e.g., information on accidents, fires, thefts, revisions to sums insured, staff lists, payroll and turnover figures etc., etc.

- The form should contain full instructions. The one thing which will ensure the non-return of a form is some problem which makes completion difficult. For example, if the questions are listed but the person completing the form is not clear as to whether he is to tick a correct statement or provide a full answer. Similarly if the person to whom the form is to be returned is not shown on the form then this could cause unnecessary delays.

  Included in the instructions should be some indication of why the form is required, what its objectives are and how it will be used.

- Ambiguities must be avoided. Each question must be clear, and this means clear from the point of view of the person asked to answer it. A question asking for details of any industrial accidents would have to clearly indicate what was meant by "details" and what constituted an "industrial accident". Where ambiguities creep in and different respondents answer according to their own interpretation then the data is of little value.

  It may also be in certain circumstances that pieces of machinery or processes have names which are in the form of a "slang" term or are shop floor names. In these cases it is best to use such terms and avoid ambiguities.

- Leading questions should be avoided. Street market researchers should not frame a question such as, "all reasonable people watch the ten o'clock news. Do you watch the ten o'clock news?" In the same way we cannot put on our form for gathering data a question such as, "All the plant managers within the company who are really anxious to minimise industrial accidents make a great deal of use of the recent safety posters sent. What use have you made of them?" Leading a person, by whatever means, is not acceptable.

- The form should be no more complicated than necessary. One possible disadvantage of the ease with which computers can handle large volumes of data is that people designing forms ask far more than they need. When forms were analysed manually there was a definite incentive to keep the form brief and we must try to stick to this idea. Long and complicated forms do tend to "put people off", there is no doubt about that. Our endeavour is to have the form completed quickly and accurately. Short crisp questions are more likely to achieve this end than lengthy, wordy forms.

- When designing the form, remember how the information is to be analysed. In many cases the data will be recorded on a computer and this will greatly speed up the eventual analysis. However, the designer of the form must remember this and gather data in a suitable way.

Generally speaking computers use numbers when analysing data gathering forms. The computer does not, for example, read in sentences and interpret them. It can read in a number which corresponds to a pre-defined sentence however. Take a simple question asking for the shift during which an accident took place. We could think of three ways among others, in which such a question could be put.

"During which shift did the accident occur?"

This will involve the respondent in answering, back shift, early shift, night shift or whatever shifts the company operates. The computer cannot however accept the words "back shift". It works with numbers and analyses them. What you can do is to provide alternatives for the person, from which he picks the correct one:

"Please indicate during which shift the accident occurred by ticking the appropriate box."

Early shift	☐	1
Day shift	☐	2
Back shift	☐	3
Night shift	☐	4

You can see a number against each box. When the computer is being fed the answer for this question it will be given the answer for each respondent in terms of the box number. The computer can then count the number of times the number '2' was given as the answer. If this turns out to be 55 times out of 100 forms completed then 55% of all incidents occurred during the day shift. In addition to this the computer has the capability to extract other information from those forms where the number '2' was answered in this question. For example it could calculate the location of the incidents. Another question may ask for location and provide five or six alternative locations. The computer can look at all the forms, where '2' was given as the answer in the shift question and work out where they had their incidents.

A third alternative would be to ask for the exact time, a question which would be asked in any case, and then have the computer programmed to interpret the time in terms of shifts. The starting and stopping times of shift would be entered into the computer programme and the machine would then convert each answer into a shift. This would be much more difficult where shifts, for example, overlap.

All that we have said here relates to those circumstances which forms are being designed. It may well be that in the majority of cases the data is already being sent to you by others. In these cases all you can do is take what is being sent and make sure that how it was gathered, by others, is satisfactory.

5.2.3 It may seem that we have spent a lot of time on the gathering of data. However, if the data is suspect before you begin the analysis then the whole exercise will be of little value.

Let us now create an example which we can follow through this chapter and the next. It is an example based on the claims register of an hotel company. Two pages from the register are shown in Fig. 5.1.

The hotel company has two hotels, one in Glasgow, the other in London and the register is only concerned with claims made by hotel guests. These claims are either claims for personal injury or loss or damage to personal effects. All claims paid to guests for injury or lost or damaged effects are met by the hotel itself without reference to any insurance.

This is a simple example and there are many other pieces of information which you may consider it to have been valuable to gather. However, the pages we have provided will be sufficient for our purposes at the moment.

5.3 Representation of Data

Now we have data! What will we do with it. In our example we have a settlement cost and four variables; location, type, age and sex. Our first task is to represent this in the most appropriate way for our own purposes. There are a number of different ways we can represent the data and we will look at each in turn. The main point is that the method chosen must match the need at the time. If your need is to provide an overall picture of fire losses in your company for some annual report then a different method would be selected from one which you may use when providing a technical report for your insurers.

5.3.1. The most elementary step is to prepare an un-ordered array concentrating on cost of settlement then we could provide a straight listing of settlement costs as shown in Fig. 5.2.

At least this gives some basic idea of how the costs are distributed. It still requires careful reading to be able to interpret how the costs are distributed and has severe limitation if large numbers are involved. We have sixty costs but you may be dealing with 500 claims and an ordered array would be of little or no value. Fig. 5.3 shows the ordered array.

5.3.2 To overcome these problems we can construct a frequency distribution. This simply condenses the array and is much easier to interpret. A frequency distribution of claims costs is shown in Fig. 5.4.

We can look at this distribution and see that claims costs are fairly evenly spread. The same "frequency distribution" simply describes what we have

CLAIMS REGISTER

NAME	SEX	AGE	NATURE	LOCATION	COST
[illegible] J	F	18	Ankle sprain	Glasgow	25
Smith F	M	38	Baggage	Glasgow	650
[illegible] P	F	50	Car	London	1900
[illegible] I	F	35	Clothes	Glasgow	1600
Bolton T	F	60	Broken leg	London	2550
[illegible]helmers S	F	21	Back injury	London	1400
[illegible]nderson B	M	60	Cut hand	London	250
Black Y	M	60	Personal Effects	Glasgow	850
[illegible]homson F	M	40	Baggage	Glasgow	350
Dickson S	M	30	clothing	Glasgow	425
[illegible]ccium T	F	35	Foot injury	London	2050
Boyce L	F	36	Car	London	2100
Davidson T	M	25	Camera lost	Glasgow	550
Elder B	M	38	Lost effects	Glasgow	610
Smith V	F	19	Cut hand	Glasgow	70
Brown G	F	23	Ankle strain	London	1500
Gray T	F	23	Stolen baggage	London	1600
Cox T	M	18	Hand injury	Glasgow	600
Young O	F	40	Car stolen	London	2425
Robertson S	M	15	Back injury	London	2000
Cowan V	F	50	Facial injury	London	2525
Cruickshank L	F	60	Car stolen	London	2750
Simms L	M	40	Foot injury	Glasgow	700
Ray B	M	19	Damaged clothing	Glasgow	900
Lamont J	F	65	Broken wrist	London	2850
Dean T	M	60	Cut finger	Glasgow	100
Cook B	F	40	Jewellery	London	1700
White L	F	39	Cut leg	London	2350
Williams T	F	57	leg injury	London	2600
[illegible] L	M	18	Bruise	Glasgow	100

Fig. 5.1

CLAIMS REGISTER

NAME	SEX	AGE	NATURE	LOCATION	COST
Morrison S	M	60	Jewellery	London	750
Merchant L	M	20	Sprained Ankle	London	1000
Watson R	F	62	Neck Injury	Glasgow	2900
Reid T	F	37	Car Damage	Glasgow	2100
Todd L	F	40	Clothes	London	2400
Russell V	M	55	leg Cut	Glasgow	1100
Sharples L	M	30	Cases lost	Glasgow	1200
[illegible]egan R	F	65	Broken leg	London	2875
[illegible]dder B	M	42	Hand Injury	Glasgow	1300
Stewart R	M	25	Cameras	Glasgow	1150
Rutherford L	F	31	Facial Injury	Glasgow	1650
[illegible]cKinnon D	F	55	Car Damage	London	2550
[illegible]cKenzie L	F	38	Jewellery	London	2200
[illegible]awards S	M	50	Personal Effects	Glasgow	540
King R	M	42	Sprain	Glasgow	400
Larson R	F	25	Cash Stolen	London	1850
Noble E	M	40	Hand Injury	Glasgow	1500
Winston D	F	32	Clothing lost	Glasgow	1675
Thatcher M	F	32	Cases Stolen	Glasgow	1700
Bean T	F	34	Leg Injury	Glasgow	2600
Foot M	M	50	Car Damage	London	1500
Steele D	M	20	Cut Finger	Glasgow	100
[illegible]ush D	M	19	Hand Cut	Glasgow	530
Jenkins R	M	55	Cut leg	Glasgow	800
Williams S	F	59	Broken Ankle	London	2700
Whitelaw W	F	41	Clothing	Glasgow	1900
Kinnock N	M	50	Car Damage	Glasgow	1175
[illegible]hore P	M	21	leg Cut	Glasgow	570
Russell E	M	50	Money	Glasgow	200
Willis N	M	60	Cases lost	Glasgow	300

Un-ordered Array of Data

25	1500	750	1850
650	1600	1000	1500
1900	600	2900	1675
1600	2425	2100	1700
2850	2000	2400	2000
1400	2525	1100	1300
250	2750	1200	100
850	700	2875	530
350	900	1300	500
425	2850	1150	2700
2050	100	1650	1900
2100	1700	2550	1175
550	2350	2200	870
610	2600	540	260
70	100	400	300

Fig. 5.2

Ordered Array of Data

25	600	1400	2100
70	610	1500	2100
100	650	1500	2200
100	700	1600	2350
100	750	1600	2400
250	850	1650	2425
260	870	1675	2525
300	900	1700	2550
350	1000	1700	2600
400	1100	1850	2700
425	1150	1900	2750
500	1175	1900	2850
530	1200	2000	2850
540	1300	2000	2875
550	1300	2050	2900

Fig. 5.3

CLAIMS COST £			Nº. f
0	<	600	15
600	<	1200	12
1200	<	1800	12
1800	<	2400	10
2400	<	3000	11
			60

Fig. 5.4

done. We have counted the number of times, or frequency with which each claim has occurred and how these claims are distributed. Compiling these frequency distributions is a fairly simple matter.

1. Find the highest and lowest value.
2. Divide by the number of groups or classes you want.
3. Create the classes.
4. Assign each value to a class or group.

In our example, if we do this we find:

1. The highest value of claim is £2,900 and the lowest is £25. £2,900 – £25 = £2,875.
2. The reason for constructing the distribution is so that it can represent the data easily. There is therefore no point in having 25 classes or, at the other end, 2 classes. There must be a reasonable spread and in most cases about five to six classes is adequate. We will settle on five classes £2,875 ÷ 5 = £575.
3. In creating the classes we have to think what the final distribution will look like. It would be very cumbersome if we started at the lowest value, £25 and rose in units of £575. By rounding the £575 up to £600 and starting at £0 then we created the classes shown in Fig. 5.4.

When creating the classes we use the symbol "<" which means "less than": And so the first class goes goes from £0 to less than £600. Theoretically this means £599.999999. In practice it means that a claim of £600 exactly will go in the second class and anything less than that will be put in the first class.

4. The sixty claims were then assigned to the classes, by methodically going down the register pages. It is wise to use a method like "five bar gates" or some other, to keep a note of how many values are being assigned to each class. Once we have done this we can count all the frequencies and they should add to 60.

The frequency distribution does give a clearer picture of what the data is telling us. It is also useful when comparing our data with someone elses or indeed when comparing sub-sets within the data. We could, for example, construct two distributions one for Glasgow and one for London claims.

We can tell right away that there are more Glasgow claims than London ones and that the distribution of these claims is quite different. The bulk of the Glasgow claims are small but the bulk of the London claims are at the higher end of the range of values.

CLAIMS COST £			ALL	GLASGOW	London
0	<	600	15 (25)	14 (40)	1 (4)
600	<	1200	12 (20)	10 (28)	2 (8)
1200	<	1800	12 (20)	7 (20)	5 (20)
1800	<	2400	10 (17)	3 (9)	7 (28)
2400	<	3000	11 (18)	1 (3)	10 (40)
			60	35	25

Fig. 5.5

5.3.3 These frequency distributions are much clearer than the un-ordered array but there is still more that we can do to make the data easier to interpret. Often it is necessary to express values as percentages. We can say that a certain percentage of all claims cost between £600 and £1,200 and another percentage cost between £1,200 and £1,800 and so on.

We can achieve this by constructing a relative frequency distribution. In the distribution in Fig. 5.5 we have shown the relative frequencies in brackets. From this we can say immediately that 18% of all claims cost more than £2,400. We can also see that the Glasgow hotel had 7 claims between £1,200 and £1,800 and the London hotel had two fewer. The relative frequencies however show that for both hotels, 20% of their claims were in the bracket £1,200 to £1,800.

We can also see, quite markedly, that while 40% of Glasgow claims cost less than £500 only 4% of London claims were in this range.

5.3.4 One other interpretation which is often made of data is to state the number of incidents up to a certain figure, or greater than some value etc. We could for example say that a certain percentage of all incidents cost more than £2,400, or a certain percentage cost at least £1,800.

These conclusions can be made if we construct a cumulative frequency distribution. Such a distribution is shown in Fig. 5.6.

CLAIMS COST £			f	CUMM. FREQ.	
0	<	600	15	15	60
600	<	1200	12	27	45
1200	<	1800	12	39	33
1800	<	2400	10	49	21
2400	<	3000	11	60	11
			60		

Fig. 5.6

You can see that we have two columns of cumulative frequencies. The first is in ascending order. There are 15 claims costing less than £600 and 12 incidents costing between £600 and less than £1,200. There are therefore 27 incidents costing less than £1,200. This procedure is carried out all through the distribution and so we have 39 claims costing less than £1,800, 49 costing less than £2,400 and 60 costing less than £3,000.

The second column shows the descending cumulative frequencies. We have 60, or all claims, costing more than £0. There are 45 claims costing more than £600 and so on.

We can also express these cumulative frequencies as relative cumulative frequencies if we want. We could say that 100% of all claims cost less than £3,000; 75% of claims [$\frac{45}{60}$] cost more than £600; 25% of claims [$\frac{15}{60}$] cost less than £600.

5.3.5 The techniques we have used so far are all based on the frequency distribution and simply re-arranged the data in an effort to make it clearer to the reader. Another method of representing data would be to draw it. There is a range of methods we could use for drawing our data, some based on the frequency distribution and others not.

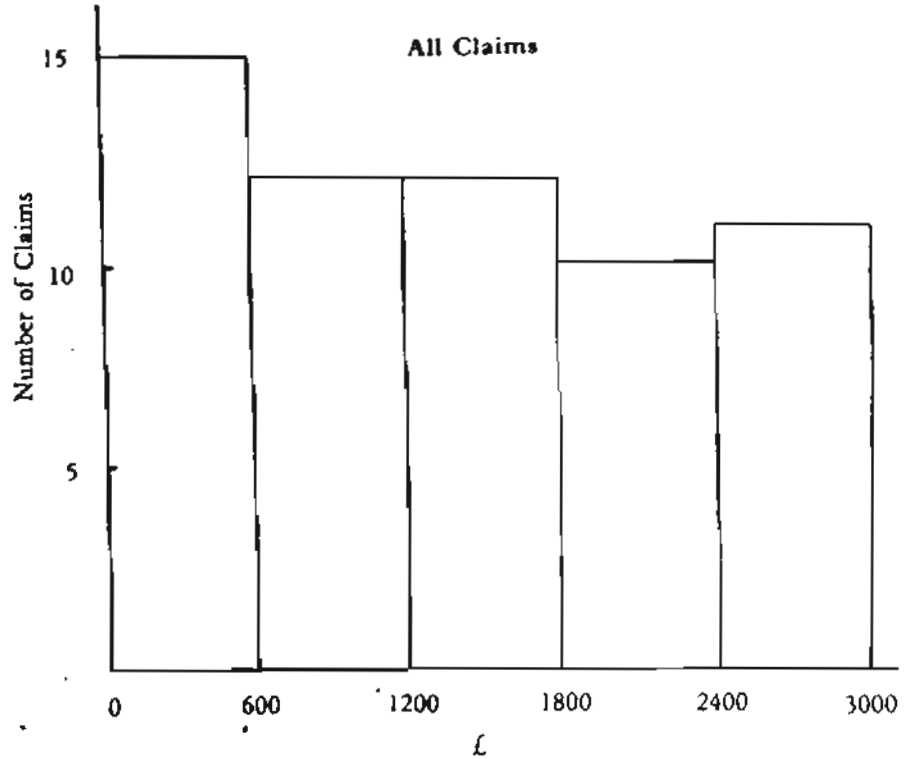


Fig. 5.7

Fig. 5.7. shows one drawing based on the frequency distribution of all claims. This method of representation is known as a histogram. The variable we are measuring is shown along the horizontal axis and the frequency with which the variable occurred is noted on the vertical axis. You can see that the horizontal axis shows the limits of the classes we used earlier in the frequency distribution shown in Fig. 5.4.

From this histogram we can see immediately that the frequency with which claims arise is fairly evenly spread over the range of values from £0 to £3,000. You may want to include this histogram in a report you are compiling on claims within the group or some note you are sending to hotel managers. You may, however, be more interested in comparing the two hotels in the group.

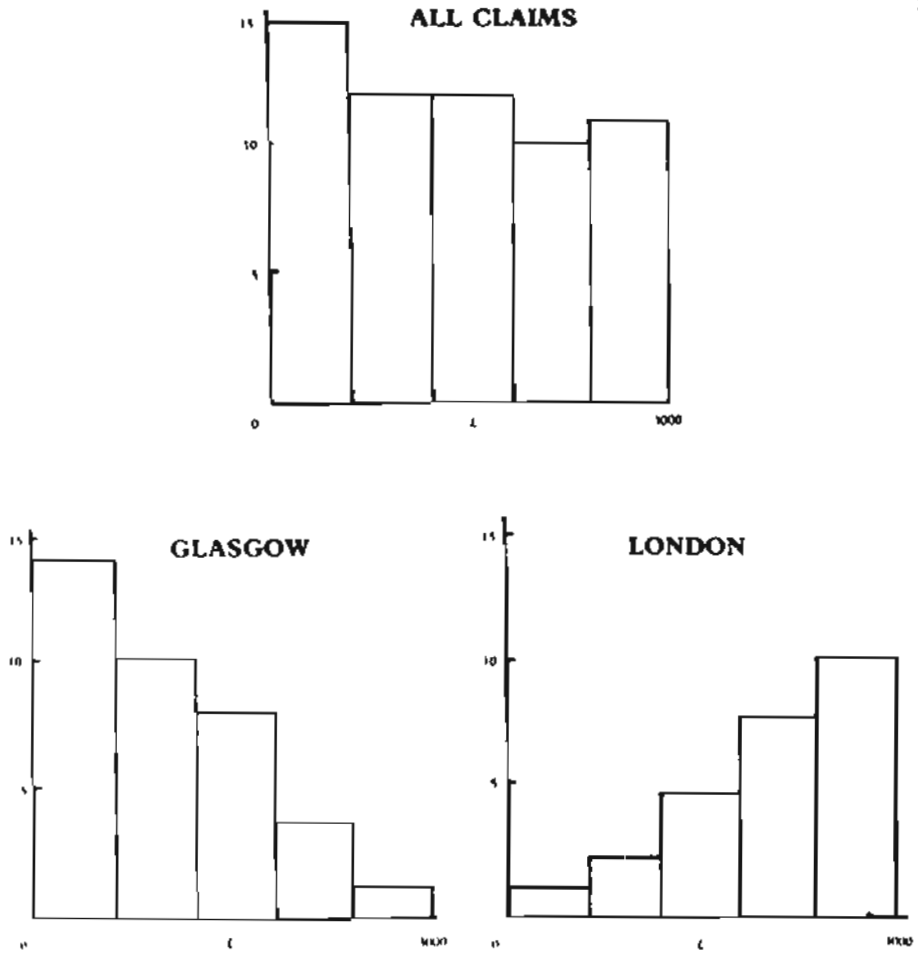


Fig. 5.8

The three drawings in Fig. 5.8 show the differences quite clearly. Neither hotel conforms with the pattern for all claims. When we contrast the two histograms for Glasgow and London we see that the bulk of Glasgow claims are small in value whereas the majority of London incidents are much more expensive. Histograms like this are a compelling way to make your point and will probably have much more impact than either of the frequency distributions shown in Fig. 5.5 or the pages from the register shown in Fig. 5.1.

The cumulative frequency distribution can also be drawn, when this is done the drawing is referred to as an ogive.

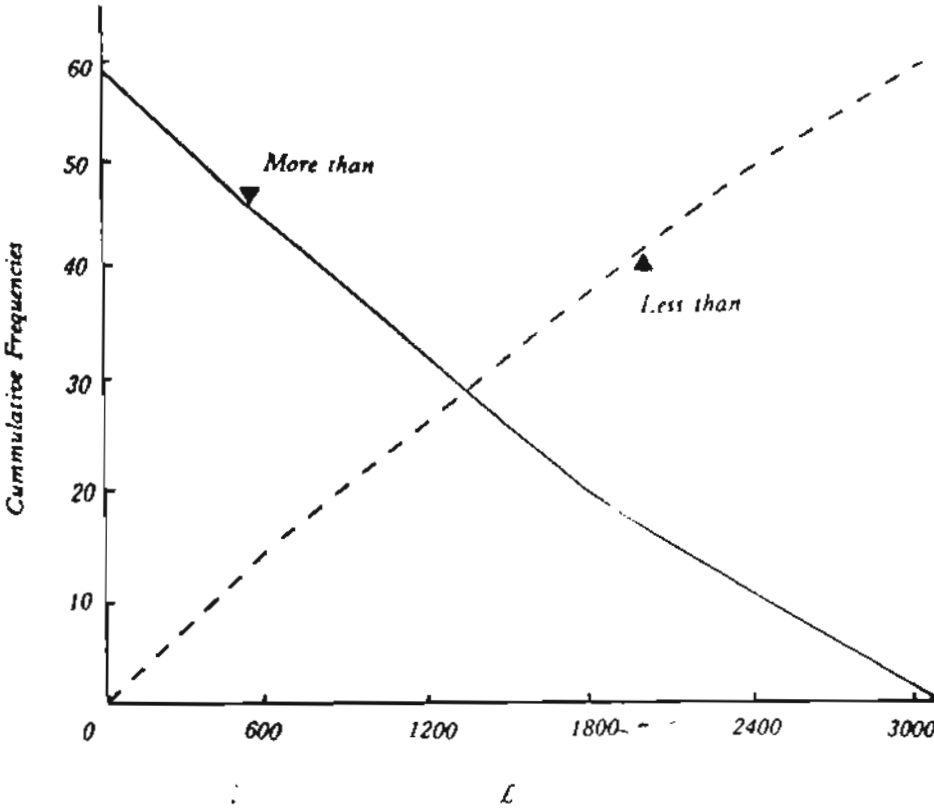


Fig. 5.9

An ogive of all claims costs is shown in Fig. 5.9. The horizontal axis i labelled with the cost of claims, in the same way as the histogram. Th vertical axis now shows the cumulative frequencies. The lines, "mor than" and "less than" are drawn using the limits of the various classes o

the frequency distribution. For the "less than" line we have plotted the figures from the first column of the cumulative frequency distribution shown in Fig. 5.6. When plotting the figures we use the top end of the classes i.e., the cumulative frequency of 15 is plotted against £600, 27 against £1,200 and so on. The "more than" curve uses the lower end of each class e.g., the 60 is plotted against £0, 45 against £600 and so on.

This means that when we use the lines to make interpretations for use we can say, using the "less than" line, that about 22 claims cost less than £1,000, or using the "more than" curve, that approximately 14 claims cost more than £2,000.

Notice that the lines cross at a point which has a monetary equivalent of £1,300 to £1,400. This value, let us say, for the moment that it is £1,350, must therefore be the value above and below which lie half of the claims. On the "more than" curve £1,350 corresponds to a cumulative frequency of 30 and on the "less than" curve it also corresponds to 30. Later we will see that this point, the point which splits the data itself, has a special significance for us.

5.3.6 All of the techniques used in 5.3.5 are based on the frequency distribution. There are also a range of more pictorial methods which we can use and which may be appropriate.

The choice of a particular method of representing data does depend on the use to which you want to put it and the following methods are often seen in company reports, newspapers, magazines and internal company reports. They are less 'scientific' in nature but nevertheless provide a means of representing the data and giving some other person an insight into what the data reveals.

The first of these more pictorial methods is the pie chart and an example is shown in Fig. 5.10.

This chart shows the division into male and female claimants. The entire circle represents all claimants and the segments are drawn to represent the proportions of the particular variables you want to show.

A second method is the bar chart. This is often seen in company publications. The one great advantage of the bar chart is that it is capable of illustrating more than one feature of the variable. The histogram, you will recall, illustrated the frequency with which a variable occurred. The bar chart can however show more than this. Fig. 5.11 shows a bar chart.

This chart shows the values of injury and property claims, sub-divided into the two hotels. On the one drawing we have value, type and location of claims. You can see how this could be valuable for displaying, for example, loss figures for various forms of risk and various plants.

Finally we could draw a graph representing the pattern of claims for the two hotels over the recent past. Fig. 5.12 shows such a graph. The years are on the horizontal axis and the cost on the vertical axis. The trend over these years is clearly shown for the two hotels.

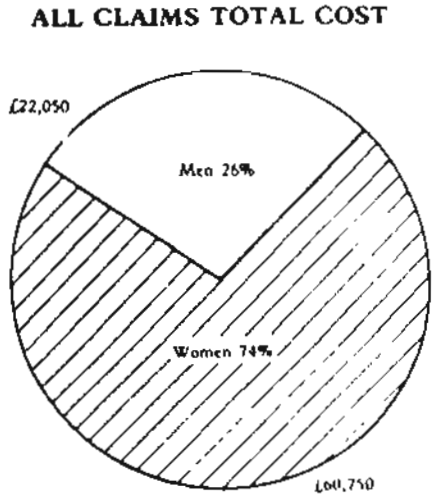


Fig. 5.10

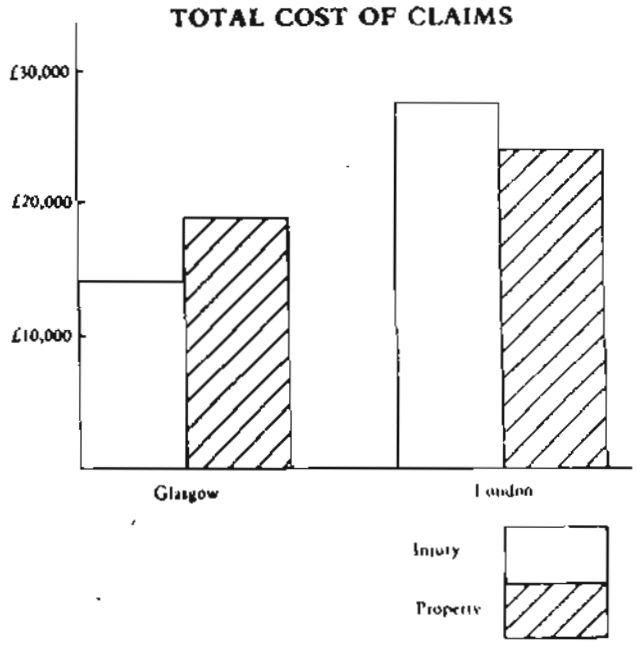


Fig. 5.11

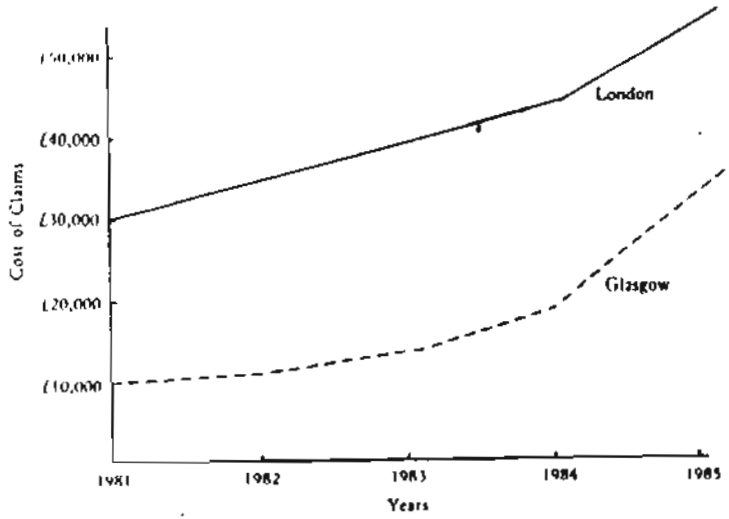


Fig. 5.12

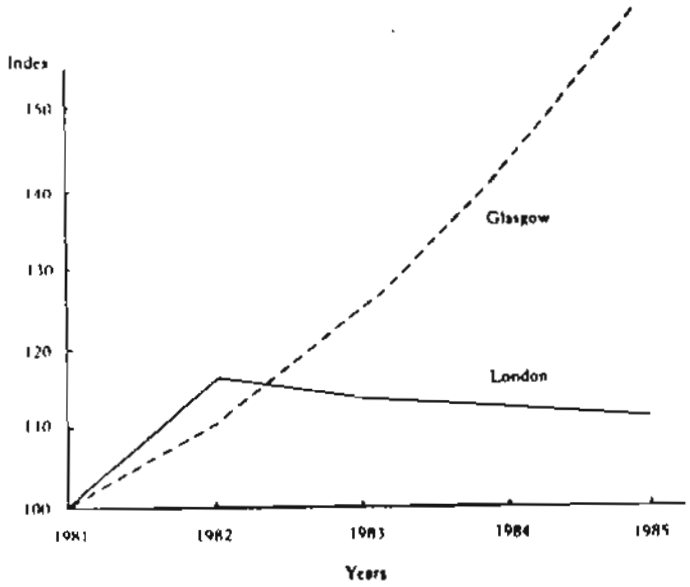


Fig. 5.13

What we must be careful to avoid is any misleading impressions which our graph may give. On looking at the graph we could easily come to the conclusion that the London hotel is really in a much worse position than the Glasgow hotel.

The claims are larger in size and increasing. The Glasgow claims are also increasing but they are much smaller. This is a fairly classic case of how simple it is to mislead the reader. The actual figures upon which the graph was drawn are:

	£ Glasgow	£ London
1981	10,000	30,000
1982	11,000	35,000
1983	14,000	40,000
1984	20,000	45,000
1985	32,000	50,000

These figures actually tell a different story from that implied by the graph. It would seem from the figures that the Glasgow claims are increasing at a much faster rate than London claims. If this is what you want to show then a slightly different graph would be more appropriate.

The graph in Fig. 5.13 shows the same data as 5.12 but this time the figures have been linked to a common base and expressed as increases year on year. For example, the 1981 figure for Glasgow was £10,000. By 1982 the figure was £11,000. If we let the 1981 figure equal 100 then the 1982 figure would be $\frac{11,000}{10,000}$ x 100 or 110. The 1983 figure of £14,000 would then become $\frac{14,000}{11,000}$ x 100 or 127 and so on.

When we change all of the figures for London and Glasgow we end up with:

	Glasgow	London
1981	100	100
1982	110	117
1983	127	114
1984	143	112
1985	160	111

When we plot these figures a quite different picture emerges as shown in Fig. 5.13. This time the Glasgow claims are seen to overtake the London claims. The graph is now reflective of the rate at which claims are increasing. It all depends on what you want to show and what objectives you wish to satisfy, as to how the data should be displayed.

Chapter Six

STATISTICAL ANALYSIS OF RISK 2

6.0 In the previous chapter we looked at the basic steps involved in gathering information and in representing that information in the best manner possible. We identified a number of techniques which can be used, each one having particular validity depending upon the objectives you had set for the exercise. When we represented data however, we were not carrying out any measurement of what we had found. All we were doing was drawing or representing the data in a suitable manner. In this chapter we turn to the business of taking measurements of the data in order that we can begin to make conclusions about what our data is telling us.

Taking measurements of the data is a little like taking a snapshot of the data. We will make some calculations which will be like a picture to us of what the data is like. We will for example want to know where in the whole spectrum of values our data lies. In other words where is our data located? If we are talking about claims then we will want to locate our claims in the spectrum of money values which claims could possibly assume. Is our data around the £200 mark or is it up around the £2,000 level? This is a basic statement which we would want to be able to make. In addition to locating the data we may also want to say something about how the data is spread out. It may be that we decide the data is located around the £500 mark and that it is tightly grouped around this figure. Alternatively we could find that our data is very widely dispersed around the general location of £500, with some claims down about £20 and others much higher at about £950. We will then need a way of measuring this dispersion, as saying where the data is located will not tell us the whole story. Even with location and dispersion described there is still one other aspect of the data which we may want to measure and that is the nature of the dispersion, that is, whether the data is grouped at one end of the scale of values or the other. Most insurance-type statistics on claims produce data which is grouped around lower monetary values. Most claims are reasonably small with only a very few extremely large ones in any year. What we will need then is some way of measuring this phenomenon. We will want to be able to say whether the data is grouped at the lower end or the higher end of the scale of values we are using. From what we have said it is clear that we require at least three measures of our data in order to have a picture of what the data is telling us. We will need a measure of location, of dispersion and of skew. This chapter takes each of these in turn and gives a brief introduction of them and illustrates their use by means of the hotel data from chapter five.

6.1 Measures of Location

What we want is some way to describe where the data is located. If someone asks what our claims at the hotel are like we want to be able to answer by giving a figure which locates our claims in the whole spectrum of money. The only one hundred percent accurate way of answering the question would be to list the entire number of claims and say to the

questioner that this is what the the claims costs are like. What we need is the snapshot idea which will capture the general thrust of the data and give the person a good idea of where the claims are located.

The normal method of measuring location is to express the data in the form of an average. There are however at least three forms of average and so we must be careful in our use of the term. We will look at each of these three averages in turn, the mean, median and mode.

6.1.1 The Mean

This is the form of average with which most people will be familiar. It is found by adding all the values of the variable under consideration and then dividing by the total number of variables. In the hotel data we have 60 claims which sum to £82,800. This gives an arithmetic mean of £1,380. And so we have now located the data in the sense that we can say the claims are around the £1,380 mark. This is the kind of snapshot we were looking for earlier, at least now the data can be described to anyone in a simple manner. What we have done can be described by a simple formula:

$$\bar{x} = \frac{\Sigma x}{n}$$

Where $\bar{x}$ is the arithmetic mean
Σx is the sum of all the values of the variable x
n is the number of values of x.

The simplicity of the arithmetic mean does bring some problems and we will look at them later.

Calculating the arithmetic mean is rarely as straightforward as adding all values and dividing by the total number. One of the main problems we will experience right at the start is that in many cases we do not have all the values of every variable. What we will most likely have is a grouped frequency distribution, as we have described in chapter five. The frequency distribution which we created for all claims was:

£			f
0	<	600	15
600	<	1200	12
1200	<	1800	12
1800	<	2400	10
2400	<	3000	11
			60

The difficulty which the frequency distribution causes is that we now do not have a value of each variable. The variable x was added and then divided by the total number. In the frequency distribution we have a class or an interval of values in place of an individual value.

What we require is a single number to insert in the formula for the arithmetic mean. The number would have to be representative of all the values in the class. A reasonable number to select is the mid-point and this is in fact the value most often chosen to represent the class. In ou example the mid-point of the first class would be £300, of the secon would be £900 and so on. What we have to remember is that the mid-poi only represents the values in the class. In other words, the value £30 represents the values in the first class. There are actually 15 numbers the first class and they are "each" being represented by the value £30 There are therefore 15 £300's in the first class, 12 £900's in the secon and so on. When we use these mid-points in the formula for the arithmet mean we will have to reflect this fact somehow.

The formula for the arithmetic mean of a grouped frequency distributi is:

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

Where Σfx is the sum of all values of x multiplied by the frequency wi which those values arose. In our example we will have to multiply all t mid-points by the frequencies with which they occur and then sum the products. The calculations are shown below.

£			x	f	fx
0	<	600	300	15	4500
600	<	1200	900	12	10800
1200	<	1800	1500	12	18000
1800	<	2400	2100	10	21000
2400	<	3000	2700	11	29700
				60	84000

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{84000}{60} = 1400$$

And so the arithmetic mean of the grouped frequency distribution £1400. You will see that this is slightly different from the arithmetic me which we calculated from the raw data. This is clearly due to the fact t we have lost some of the accuracy of the data by placing the claims groups. The value of x which we used was the mid-point which is o a representative value and not the actual value. The true mean and mean from the grouped distribution are not all that far apart and figure of £1400 will be good enough for most purposes.

There are a number of problems which flow from using the arithme mean and we could spend a great deal of time looking into them. Howe we are mainly concerned with the use of statistics, not the theory, and we mention two problems only of which users should be aware.

The first is that the arithmetic mean is not so suitable for certain ty of figures. Take for example the following set of figures.

Year	Claims	% Increase
1984	20	–
1985	30	150
1986	60	200

The final column shows the percentage increase in claims, and so in 1985 there was a 50% increase over the previous year. We could say that the average percentage increase has been 175%, i.e., $\frac{(150 + 200)}{2}$.

We can check this figure by applying it to the actual figures. When we do this we find:

Year	Claims
1984	20
1985	20 x 175% = 35
1986	35 x 175% = 61.25

These calculations however, produce an incorrect answer. The actual figure for 1986 is 60 not 61.25. The problem is that the arithmetic mean is not suitable for averaging out figures which are related to each other in the way that ours are i.e., one figure being a percentage of the one before. What we need for these situations is the "Geometric Mean" rather than the arithmetic mean. The geometric mean is found by the following formula:

$$\sqrt[n]{X_1 \times X_2 \times X_3 \times \ldots X_n}$$

In our small example we would have:

$$\sqrt[2]{150 \times 200}$$
$$= 173.21$$

If we test this on our figures we find:

Year	Claims
1984	20
1985	20 x 173.21 = 34.642
1986	34.642 x 173.21 = 60

This now corresponds exactly with what actually happened. The average annual percentage increase is therefore 73.21% and not 75%.

The second problem with the arithmetic mean is that it is easily distorted by extremely large or small values. The calculations for the arithmetic mean involve all values in the distribution and if there happens to be a very large value, for example, then this distorts the answer. In our hotel data we have 60 claims which total £82,800 giving an arithmetic mean of £1,380. If we had had 61 claims and the additional claim was £20,000 then the mean would have been £1,685. The mean has been pulled up by the one claim which was much higher than any other. If there are a few values in a distribution which are either much smaller or larger than the others then some mention should be made of this in any synopsis of findings.

6.1.2 The Median

The second form of average is the median. The median is the value which is exactly half way through the data, 50% of all values lie above it and 50% lie below. The median in the following list of figures is 10.

5, 7, 9, 10, 13, 15, 17

The median splits the data in half and so you are just as likely to have a value above it as below it. In our hotel data we had 60 claims and so there is no natural middle value. What we can do is to take the two values which straddle the middle i.e., the 30th and 31st values of £1,300 and £1,400. The mid-point between these is a good enough measure of the middle point of our entire distribution. The £1,300 and £1,400 were found from the ordered array in Fig. 5.3. The data must be ordered before finding the median. You cannot just pick the middle value of an unordered array as this would not give the value which is exceeded by half the values and which itself exceeds half.

The median for our data is therefore £1,350 which is fairly close to the arithmetic mean of £1,380. Notice, however, that if we had had the 61 claims mentioned earlier, with the additional claims being £20,000 the median would have been £1,400. This is a valuable quality of the median. It is not affected by extreme values in the distribution. It is always the value of the middle item, regardless of any extremes which there are.

We saw that the calculation of the arithmetic mean was slightly different when the data was grouped. Exactly the same is the case for the median. When we have a grouped distribution we cannot place the data in an ordered array? You might say, well why not unravel the data and just create the ordered array. We could certainly do this for our 60 claims but we would be much more reluctant to do it for 2,000 claims. What we need is a method of finding the median from the grouped frequency distribution. The grouped frequency distribution is:

£	f	Cumulative f
0 < 600	15	15
600 < 1200	12	27
1200 < 1800	12	39
1800 < 2400	10	49
2400 < 3000	11	60

We know that the median is the value associated with the middle claim. We have sixty claims, let us take 30 as the middle one for our purposes at the moment. The median is therefore the value associated with the 30th claim. Using the cumulative frequency distribution we know that the 30th claim will be in the class £1,200 < £1,800. There are 27 claims up to £1,200 and 39 up to £1,800, therefore the 30th must be in the class £1,200 - £1,800.

In fact the 30th claim is 3 claims into that class. There are 27 claims up to £1,200 and we want to move on another 3 claims to find the 30th. The class £1,200 < £1,800 has 12 claims in it and so we want to move 3/12ths of the way along the class. The width or interval of the class is £600 and so we want to go 3/12 x £600 i.e., £150 into the class. The class itself starts at £1,200 and £150 into it would bring us to £1,350.

This is exactly what we found earlier when we used the ordered array of data. The median found in this way will not always coincide with the true median but it will not be far out.

We can generalise what we have done in the following formula:

$$L_m + C_m \left[\frac{\frac{N}{2} - F_{m-1}}{f_m} \right]$$

Where:

L_m = lower limit of the class having the median in it i.e., the median class.
C_m = the width of the median class.
N = the number of values.
F_{m-1} = the cumulative frequency of the class immediately before the median class.
f_m = the frequency of the median class.

We can use the formula in our example to find:

$$1200 + 600 \left[\frac{\frac{60}{2} - 27}{12} \right]$$

$$= 1200 + 600 \left[\frac{3}{12} \right]$$

$$= 1200 + 150$$

$$= 1350$$

If you look back to Fig. 5.9 and what we said about it in paragraph 5.3.5 you will see that the ogive also gave us a median of approximately £1,350.

An extension of the thinking which went into the median can be used to find other interesting statistics. It may be useful for you to know the value which is exceeded by only 25% of claims rather than 50%. You could be considering some alternative risk financing mechaniams, new insurance covers, deductible levels etc. By altering our formula slightly we could find this value. What we are now looking for is the upper quartile, the value which splits your data so that 75% of all claims lie below it and 25% are above. The lower quartile would split your data the other way, it would give you the value above which lie 75% of all claims and below which lie 25%.

The upper quartile will be 3N/4 way through the data i.e., it will be the value associated with the 45th claim. In our previous formula we need only substitute "Q_u" for "m";

$$L_{Q_u} + C_{Q_u} \left[\frac{\frac{3N}{2} - F_{Q_u - 1}}{f_{Q_u}} \right]$$

$$= 1800 + 600 \left[\frac{45 - 39}{10} \right]$$

$$= 1800 + 600 \left[\frac{6}{10} \right]$$

$$= 2160$$

25% of all claims lie above £2,160.

We can do similar calculations to find deciles and percentiles i.e., to find the value which split the data into tenths or percentages. We might want to know the value below which lie only 10% of all claims. This would be the 1st decile. We can have the 2nd, 4th, 5th, 6th, 7th, 8th and 9th decile. In the same way we can have individual percentiles which split the data in more precise ways. We could find the 33rd percentile. This would be the point below which lie one third of your claims.

We mentioned earlier that the median had the advantage of not being affected by extreme values. There is one other useful feature which we should mention. The median and its off-shoots such as quartiles, deciles and percentiles can be used even where the data is incomplete.

Let us say that our grouped frequency distribution had a final class of:

£	f
greater then 2400	11

This is an open-ended class and such classes are often seen in real life. We know that there are eleven claims greater than £2,400 but we do not know what they are individually. Notice that in such circumstances the median will still be £1,350. All we have to know is the total number of claims and enough information to pinpoint the middle claim. You could not have calculated the arithmetic mean in such circumstances.

Depending on the circumstances therefore, it may be more appropriate to locate our data by describing the middle value rather than the arithmetic mean. But even the median is not appropriate in all circumstances.

Take the following numbers:

12,12,12,12,12,12,12,12,15,17,18,19,20,20,21,23,25.

The arithmetic mean is 16.12 and the median is 15. However, both of these statistics conceal one vital aspect of the data, there are eight values of 12. Almost half the data is made up of the one number.

6.1.3 The Mode

The solution to the above problem would be to use the mode. The mode is the most common number. It is quite common to use the mode in ordinary everyday language. When we talk of average holiday entitlement, average family size we are really expressing the modal holiday entitlement or the modal family size.

Many statistics text books give practical examples of when the mode should have been used in preference to the arithmetic mean. We could imagine an example in a simple planning decision. Let us say you are making provision for the supply of accident record forms at twenty different factories. You could work out the average number of forms used by twenty factories and send each one that number. This would be of no value if in fact most factories either have a very small number of accidents or a very large number. In such a case the distribution would be said to be bi-modal, it would have two modes. It would have been better to identify this first and despatch forms accordingly.

6.2 Measures of Dispersion

What we have done so far is to locate where our data is. In our case we were locating our claims in the whole spectrum of money. We saw that there were three main measures of location.

Location, however, is only one part of the story. We must add to it the question of dispersion. Take the following two sets of numbers:

	A	B
	10	1
	11	11
	12	21
$\bar{x}$ =	11	11
median =	11	11

Both have identical arithmetic means and medians but they are really quite different in the extent to which they are spread out around the measure of location. Series A is tightly grouped and is no more than 1 away from the mean. On the other hand series B is widely dispersed with a space of 10 between the mean and the other numbers.

We can think of this in terms of risk. If you had two factories and one produced claims according to series A and the other according to series B, then which is the riskier? It all depends on what you mean by "risk" but if you had to fund losses or charge them out, in advance, to the two factories then series A is much less risky. Claims will only be 1 away from the average.

Much the same problem faces the insurance underwriter. If he knew that claims would be tightly grouped around the mean then he could charge a premium and know that it would probably be sufficient. However, if claims are widely dispersed around the mean then it becomes much more difficult to know what to charge. Claims could be low but they could be

high and the underwriter does not know ahead of time whether you are one of those who will have low or high value claims, or whether this year the claims will produce a similar mean.

What we need is some measure of dispersion. The simplest measure is to calculate the range of values. The range is the space between the highest and lowest value. In our hotel data this is £2,900 - £25 = £2,875.

A much more valuable figure is the standard deviation. This measures the extent to which the values are dispersed around the arithmetic mean. It is possibly one of the best known of all statistical tools but probably no better understood.

We will illustrate how to calculate the standard deviation with a simple example and then return to our hotel data to show its application.

6.2.1 The Standard Deviation

Let us take the following figures:

x
4
7
11
12
15
23

The arithmetic mean is 12. Standard Deviation measures dispersion around the mean and so we can see that 4 is 8 from the mean, 15 is 3 from it and so on. If we compile a separate column of deviation from the mean we get:

x	$\bar{x}$	$(x - \bar{x})$
4	12	-8
7	12	-5
11	12	-1
12	12	0
15	12	3
23	12	11

We cannot simply add these dispersions as this would sum to zero. What we could do to avoid the zero sum is to square the values:

$(x - \bar{x})^2$
64
25
1
0
9
121

If we now add these squared deviations from the mean and divide by the number of values we have, we get an "average" squared dispersion:

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

$$= \frac{220}{6}$$

$$= 36.667$$

However, squared values are not much use to anyone i.e., squared accidents, squared fires etc. We must take the square root to return to normal values.

$$\sqrt{36.667}$$
$$= 6.05$$

The formula for all of this is:

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

If the data is in the form of a grouped frequency we then have:

$$s = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{\Sigma f}}$$

In this case we multiply each deviation by the frequency with which it occurred. An alternative formula for the standard deviation, and one which takes a little less time to calculate is:

$$s = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left[\frac{\Sigma fx}{\Sigma f}\right]^2}$$

It gives exactly the same answer as the other formula.

Today it is becoming less and less important to remember formula or even to know them. The use of computers, even business micros which are so common, mean that we can concentrate on the application of all the computational work, rather than the computations themselves. What we must understand are the conditions under which a particular statistic can be used and the interpretation of what we find.

Let us return now to the hotel data and see if we can illustrate the value of the standard deviation. The claims register shown in Fig. 5.1 shows the nature of the claims. There are two types recorded, either injury or property. What we could do is to examine each of these two types of claim to see what the data can tell us.

Taking the two types of claims, we have the following frequency distribution:

£	Injury	Property
0 < 600	9	6
600 < 1200	5	7
1200 < 1800	5	7
1800 < 2400	4	6
2400 < 3000	7	4

The standard deviation is found by one of the two formulas we showed above. Let us take the injury claims and use the formula:

$$s = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left[\frac{\Sigma fx}{\Sigma f}\right]^2}$$

we need to find Σfx, x^2 and fx^2.

x Mid-point	f	fx	x^2	fx^2
300	9	2,700	90,000	810,000
900	5	4,500	810,000	4,050,000
1,500	5	7,500	2,250,000	11,250,000
2,100	4	8,400	4,410,000	17,640,000
2,700	7	18,900	7,290,000	51,030,000
	30	42,000		84,780,000

We can put these figures into the formula:

$$s = \sqrt{\frac{84,780,000}{30} - \left[\frac{42,000}{30}\right]^2}$$

$$= \sqrt{2,826,000 - 1,400^2}$$

$$= \sqrt{866,000}$$

$$= 930$$

You can do exactly the same for the property claims and you should find a standard deviation of 791.

What do these two standard deviations actually tell us?

On its own the standard deviation is really very difficult to interpret. However it can be useful when comparing two different distributions and it is certainly useful in a whole range of statistical work far beyond what we might cover in this study guide.

In our example the means of both distributions, injury and property, are identical. The mean of the injury claims is seen in the standard deviation formula:

$$\frac{\Sigma fx}{\Sigma f} = 1400$$

If we calculate the mean property claim we find:

x Mid-point	f	fx
300	6	1,800
900	7	6,300
1,500	7	10,500
2,100	6	12,600
2,700	4	10,800
		42,000

$$\frac{\Sigma fx}{\Sigma f} = \frac{4,200}{30} = 1,400$$

Both types of claims are therefore located at the same place, but their dispersion is different. The injury claims are more widely dispersed than property claims. The difference is not very marked but it does exist. This takes us back to the point we made earlier. If we have two distributions with the same mean but one has a wider dispersion then it is the riskier of the two. The extent of the dispersion will determine just how risky the distribution is.

This direct comparison of the standard deviations was possible because the means were the same. Remember that the standard deviation measures dispersion around the mean and so if the mean for one distribution was substantially higher than another, the standard deviation would also be measured in larger figures. This may be due solely to the size of the numbers and not to a larger dispersion.

For example take the following two distributions:

A	B
4	40
7	70
9	90
10	100
$\bar{x} = 7.5$	$\bar{x} = 75$
$s = 2.29$	$s = 22.9$

Distribution B has a standard deviation of 22.9 which is much larger than the standard deviation for distribution A, but this is due solely to the fact that the size of the numbers in distribution B is 10 times that of A. The dispersion is in fact exactly the same. One way to compare this dispersion when the means are different is to express the standard deviation as a percentage of the mean. If we do this for the two distributions we get:

A	B
$\frac{s}{\bar{x}} \times 100$	$\frac{s}{\bar{x}} \times 100$
$\frac{2.29}{7.5} \times 100$	$\frac{22.9}{75} \times 100$
30.53%	30 53%

This figure is called the "coefficient of variation" and it allows us to compare different standard deviations even where the arithmetic means are quite different.

In our hotel data we could separate out the Glasgow injury claims from the London injury claims, when we do this we find:

Glasgow $\bar{x} = 850$
London $\bar{x} = 2069$

The Glasgow injury claims have a much lower mean than the London injury claims but what about their dispersion. The respective standard deviations are:

Glasgow $s = 804$
London $s = 816$

The standard deviations are not all that dissimilar but remember that the means are quite different. The coefficients of variation are:

Glasgow $\frac{804}{850} \times 100 = 94.59\%$

London $\frac{816}{2069} \times 100 = 39.44\%$

This shows quite clearly that the Glasgow injury claims, while having a lower mean cost than London, have a substantially higher dispersion. This implies that the range of claims in Glasgow is much wider than for London. The size of injury claims will cover a wide range, unlike London injury claims which are more tightly grouped around the mean, albeit a higher mean.

The coefficient of variation allowed us to make this comparison and can be useful whenever there are distributions of varying sizes. One such situation is comparison of costs for different currencies. When comparing sterling costs and dollar costs we have to take account of the differing values of the currencies themselves and avoid misleading conclusions. The coefficient of variation allows this to be done. Say that we have the following figures:

Britain $\bar{x} = £500$
$s = £350$
U.S.A. $\bar{x} = \$720$
$s = \$500$

It is difficult to compare these figures in a direct fashion. We could transfer all the values into either pounds or dollars but we could calculate the coefficient of variation and avoid that extra work:

Britain $\frac{£350}{£500} \times 100 = 70\%$

U.S.A. $\frac{\$500}{\$720} \times 100 = 69.4\%$

The dispersion within the two distributions is almost identical.

6.2.2 Skew

There is one final measure we must add, in order to describe our data fully. We have located our data and measured its dispersion around the mean. Look at these key statistics for all Glasgow and London Claims.

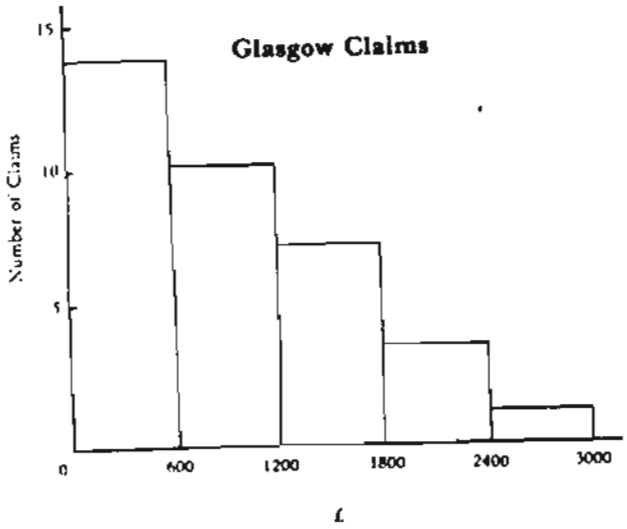


Fig. 6.1(a)

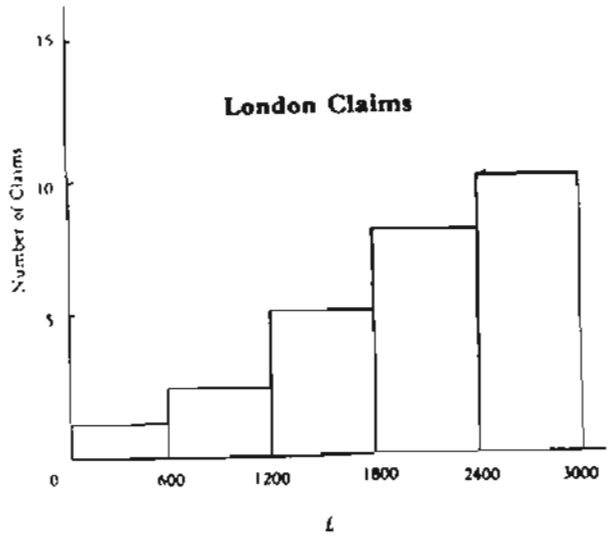


Fig. 6.1(b)

Glasgow $\bar{x}$ = £925 s = £696 Coefficient of variation = 75%
London $\bar{x}$ = £2019 s = £699 Coefficient of variation = 35%

Glasgow and London claims have roughly the same standard deviation but we know this is meaningless until we relate it to the arithmetic mean. When we do this we see that Glasgow claims are much more widely dispersed around a lower arithmetic mean. Going on our earlier comments we might say that Glasgow claims cost less but there is a high variation, while London claims cost more on average but are more predictable as they are more tightly grouped around the mean.

We have drawn both distributions in Fig. 6.1(a) and 6.1(b). From these two histograms you can see that the two distributions are really quite different. There is one feature we have not yet discussed, which can be observed from the two drawings. We have not yet measured the extent to which the data may be bunched or grouped at the lower or higher end of the distribution. We have not yet measured skew.

You can see from the drawings that the Glasgow claims are bunched at the lower end of the distribution. The frequency is highest over the lower values. This is the opposite of the London claims where the highest frequencies are measured over the high value claims. This bunching is what we refer to as skew and we must look now for some way of measuring it in a distribution.

In Fig. 6.2 we have drawn a distribution which has no skew, it is in fact symmetrical. The frequency distribution for this drawing is:

x	f
10	2
20	6
30	10
40	6
50	2
	$\Sigma f = 26$

The arithmetic mean is 30. The median is the value associated with the 13th number, and this is also 30. We can see here that when the distribution is symmetrical as we have shown in Fig. 6.2, the mean and the median will coincide.

When a distribution is skewed, either to the left or to the right, then the mean and median will not be the same values. The key to measuring skew lies with this fact. When the mean and the median are the same we have no skew, we could say we have zero skew. When the mean is greater than the median then the distribution will be bunched at the lower end of the values. Think about this, the mean is being pulled up by a few high values when in fact the majority of values are much lower.

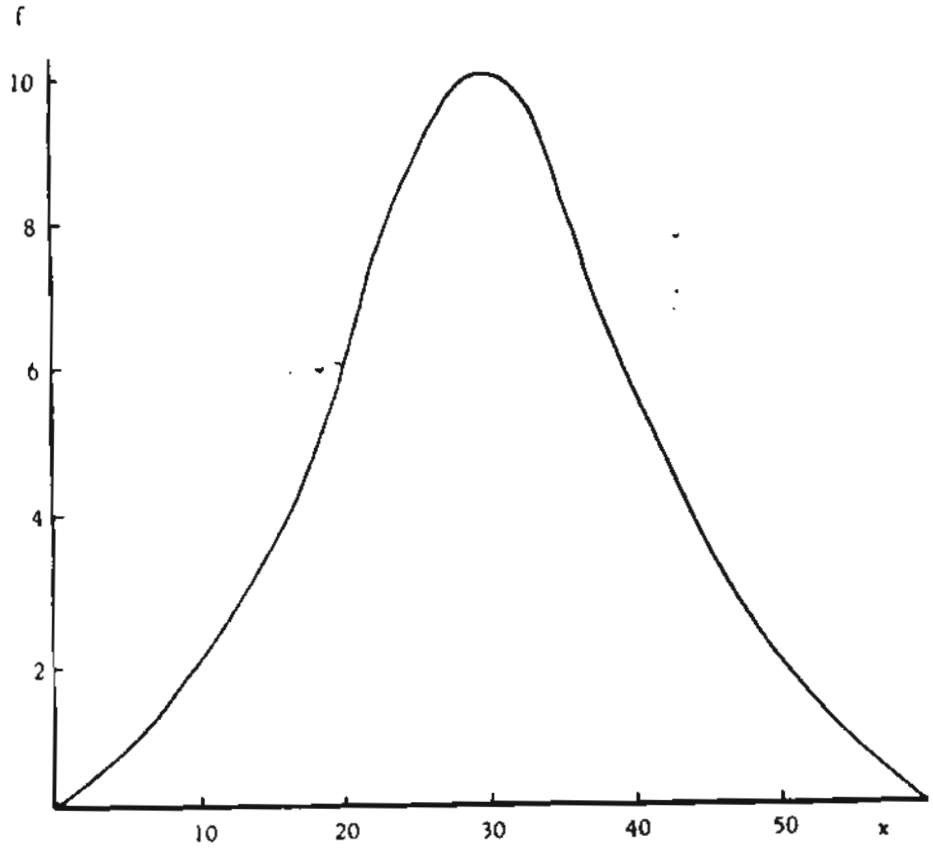


Fig. 6.2

The following distribution shows this:

x	f
10	10
20	7
30	5
40	3
50	1
	Σf = 26

The mean is $\frac{\Sigma fx}{\Sigma f} = \frac{560}{26} = 21.54$. The median is the value of the 13th number and this is 20. The mean is therefore greater than the median and the distribution is skewed to the righthand side of the distribution as we can see in Fig. 6.3. One formula for measuring skew is:

$$\frac{3(\text{mean} - \text{median})}{\text{St. Dev.}}$$

We are expressing the difference between the mean and the median in terms of the standard deviation. When the mean and median are the same then the formula produces "O" which we term zero skew.

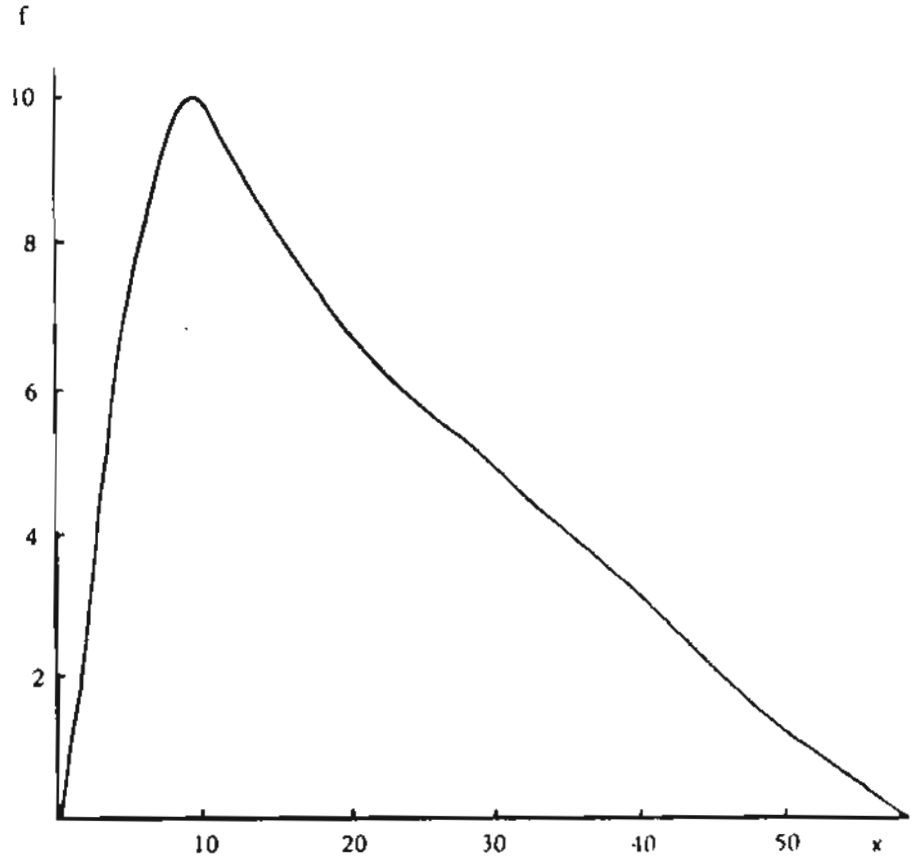


Fig. 6.3

In the distribution which is skewed to the right we find that an answer will always be positive. This is because the mean will be larger than the median in view of the high values, even although the bulk of values are bunched at the lower end. For the distribution in Fig. 6.3 we can calculate the standard deviation to be 11.67. Skew, known as Pearson's coefficient of skew is therefore:

$$\frac{3(21.54 - 20)}{11.67} = 0.4$$

This positive figure of 0.4 is the measure of skew. The most important thing is that the coefficient is positive, thus indicating that the distribution is bunched at the left and slopes down to the right.

The alternative to this "positive skew" would be negative skew. The following distribution shows this:

x	f
10	1
20	3
30	5
40	7
50	10
	$\Sigma f = 26$

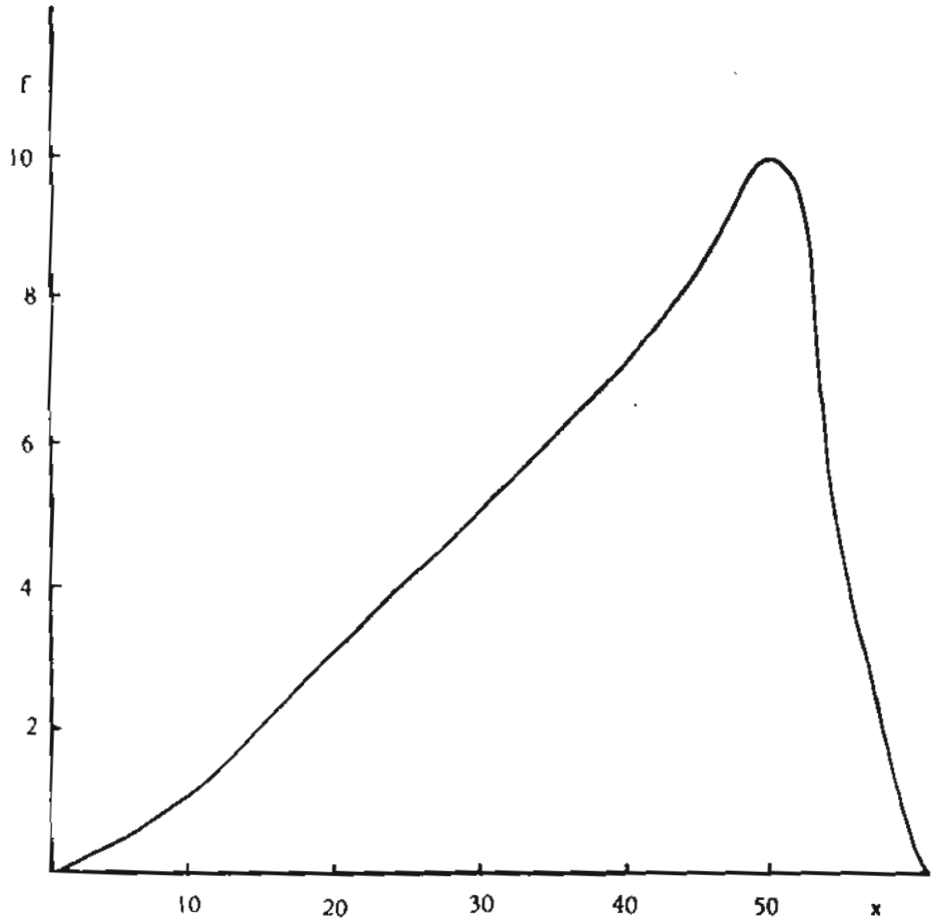


Fig. 6.4

The drawing in Fig. 6.4 illustrates this distribution and we can see that it is skewed the other way. The highest frequencies are associated with high values and this time the mean will be pulled down by the few low values.

The arithmetic mean is $\frac{\Sigma fx}{\Sigma f} = \frac{1000}{26} = 38.46$. The median is again the value of the 13th number which is 40. The standard deviation can be calculated to be 11.67. (We could have guessed that the standard deviation would have been the same because the actual dispersion is exactly the same for this distribution as it was for the distribution in Fig. 6.3. The only difference is the skew).

The coefficient of skew is therefore:

$$\frac{3(38.46 - 40)}{11.67} = -0.4$$

This is also the same, except that now it is negative which implies that the distribution is bunched at the higher end of the scale of values.

If we now return to our hotel data and try to apply what we have discovered, the figures necessary to calculate skew for the Glasgow and London claims distribution shown in Figs. 6.1(a) and 6.1(b) are:

Glasgow:	mean 925	median 700	standard deviation 696
London:	mean 2019	median 2100	standard deviation 699

Using Pearson's formula we have:

$$\text{Glasgow } \frac{(925 - 700)}{696} = +0.97$$

$$\text{London } \frac{(2019 - 2100)}{699} = -0.35$$

The Glasgow figures are positively skewed and the London claims negatively skewed.

We now have the three measures we need of our data, and in fact any data. We have measured location, dispersion and skew and it is only when all three have been found, as we saw, that we have a comprehensive picture of the data. In this chapter we have taken the time to draw distributions and indeed follow through formula. However, in the real world this is not what is done. In most cases there will be too many figures to start drawing distributions and there will be no need to carry out calculations by hand when computers can do all of the computational work for you. In fact a pocket calculator at little over £10 will compute standard deviations, for example, at the press of a single button.

What is important are two things. Firstly that we can understand what our data is like without having to take the extra time required in drawing it and secondly that we can interpret the statistics we compute.

Chapter Seven

PROBABILITY

7.0 There is one aspect of the statistical analysis of risk that we have not yet discussed in detail and that is probability. Measurement of the likelihood of loss is clearly an important aspect of the analysis of risk and is one with which risk managers should be familiar.

Discussion of likelihood inevitably leads on to measuring likelihood and measuring likelihood is what probability theory is all about. This chapter will introduce the whole idea of probability theory, the ways in which probabilities are derived, their uses and conclude with a discussion on probability distributions.

Those who are quite confident in their knowledge of probabilities can move on to the next chapter without running the risk of overlooking material they will require for later parts of the syllabus.

7.1 The Meaning of Probabilities

Probability theory is a topic which often causes confusion in the mind of those who are trying to get to grips with it for the first time. It is often the case that those people who are very good with statistical calculations, find probabilities a real problem and vice versa.

Probability theory sets out to attach a numerical value to our measurement of the likelihood of an event occurring. This probability figure must be between 0 and 1. A probability of 0 implies that the event is impossible, while a figure of 1 shows that it is certain to occur. Clearly there are few, if any, events which are either impossible or certain and most events therefore have a probability which lies between these two extremes.

A probability cannot exceed 1 or be a negative figure and if any calculations produce such a figure then the calculations are wrong. Where the probability lies on the range from 0 to 1 is an indication of how likely the event is. An event with a probability of 0.001 is quite unlikely, there is a one in a thousand chance of it occurring. An event with a probability of 0.95, on the other hand, is 95% certain to occur.

We can now use these statements to express the likelihood of different events occurring and will be able to rank different events according to how likely they are. We could say that the probability of fire at a particular plant is 0.2; of one of our vehicles being in a motor accident is 0.1; of theft from our shops is 0.01 and so on. Each statement expresses the likelihood of the event. In the above examples, fire seems the most likely event with theft being least likely.

If that was all there was to probability theory then people would not have the fear of it that they do. What we have to look at are the ways in which these probabilities are derived and then used.

7.2 Derivation of Probabilities

7.2.1 A Priori

The first method is the simplest to understand, but unfortunately is not very realistic. It had its origin in games of chance where for example a person wanted to know the likelihood of getting a particular playing card or a specific number on the roll of dice etc.

The a priori method expresses likelihood by taking the number of outcomes which you want and dividing this by the total number of all possible outcomes. And so if you wanted to know the probability of getting a red card from a pack of playing cards you would take the 26 red cards and divide by all cards to get 26/52 or 0.5. The total number of outcomes which you wanted is 26, there are 26 red cards, and the total number of all possible outcomes is 52 as there are 52 cards in the pack.

The same thinking can be used to find the probability of getting a 3 on the roll of a dice. There are six sides on a dice and each has a different number. Only one outcome corresponds to what you want and so the a priori probability is 1/6 or 0.1666.

In general terms we can say that we can calculate an a priori probability where all the events are equally likely to occur and all possible outcomes are known.

In the two examples we used above these rules certainly applied. You are just as likely to pick one card as another and just as likely to roll the dice and get a 3 as any other number. In the same way you knew in the playing card example that there were 52 possible outcomes when you selected a card and six possible outcomes when you rolled the dice.

These two assumptions are rather unrealistic in practical business situations. In most business problems the various events or outcomes to a problem are rarely equally likely and, in addition, it is often the case that the whole range of outcomes is not, or cannot be, known. In a very general sense we can see this in the problem of estimating losses. A risk manager may be trying to evaluate the cost of losses at a particular plant. In addition to the probability of various types of losses being unequal, there is also inequality within types. For example, the probability of having a large fire loss is not the same as the probability of there being a very small incident. We know that small incidents are really quite common whereas high fire losses are, fortunately, unlikely. It would not be possible therefore to use the a priori method of determining probabilities as the events with which we are concerned are not all equally likely.

The other condition which had to be satisfied before we could use the a priori method was that all possible outcomes had to be known. It is this condition which is really the most unrealistic in normal risk management problems. You will recall that in order to calculate the a priori probability we divided the desired number of outcomes by all possible outcomes. Using this approach to find the probability of there being three fires at your plant next year, you would divide 3, the number of desired outcomes, by the total number of all possible outcomes, which would be the total number of all possible fires. This figure just cannot be found!

At the end of any time period we will be able to calculate the number of fires we have had but we will never know the number of fires we did not have, and the total number of all possible fires is the sum of all those which took place and those which could have occurred but did not. It is similar to asking someone how many goals were not scored in a football match. We know how many were scored and that is all.

7.2.2 Relative Frequency

What was required was a method of determining probabilities which was based on the information we had. The relative frequency concept expresses likelihood in terms of the relative frequency with which similar events have occurred in the past.

A risk manager with 100 vehicles in a fleet would determine the probability of one of these vehicles being in an accident next year by looking back over the previous year and noting the number of vehicles that had been in an accident. If five vehicles had been involved in an accident in the previous year then we would say that the relative frequency with which vehicles were in accidents was 5/100 or 0.05.

We could do the same for fires. Let us say we had 50 fire incidents last year distributed as follows:

Cost	*Number*
$0 < 100$	25
$100 < 200$	15
$200 < 300$	6
≥ 300	4
	50

The relative frequency with which fires costing £300 or more occurred was 4 in 50 or 0.08. Assuming that there are no factors which alter the likelihood of loss then the probability of fires costing £300 or more next year is 0.08.

This method of determining probabilities also has its difficulties. There may be factors which change likelihood from year to year, or exact details of accidents in the past may not have been recorded. These two problems relate to cases where there was some previous record, albeit not suitable for deriving probabilities, but what of the case where no previous knowledge exists?

This could easily occur in the case of some new production process, chemical or building material. The process, chemical or material has not been used before and so the number of losses in the past from injury, disease or fire, respectively, cannot be calculated. The relative frequency concept is therefore not appropriate. In such cases we could turn to the third method of deriving probabilities.

7.2.3 Subjective

Where inaccurate or no historical information exists we can attempt to measure degrees of belief in the likelihood of an event occurring. This involves questioning yourself or others according to one of several methods which exist for the purpose of eliciting probabilities.

It is not necessary to go into the detail of the techniques here, what we might want to remember is that such techniques exist and can be useful.

7.3 Combining Probabilities

What we have discovered so far is the basic issue of how probabilities are derived. Having obtained a measure of the likelihood of an event or events occurring in the future, we will want to use this information in some way. There are certain rules for the manipulation of probabilities and we will look at the most important here.

7.3.1 Alternative Events

So far we have limited our examples to single events, the probability of fire, an accident, an injury etc. Often it is necessary or desirable to calculate the likelihood of either one or another event occurring. For example we may want to know the probability of there being a fire *or* a theft at a shop; of the London *or* Birmingham plant having a computer breakdown etc.

We call these probabilities, alternative probabilities and a rule exists to help us calculate them. The particular rule is known as the additions rule, for reasons which will become clear shortly.

Let us say that we operate a three shift system; early, late and night shifts. We want to know the probability of an accident occurring on either the early or late shifts. Without looking at probabilities at all we could imaginge that if 20% of all accidents took place during the early shift, 25% during the late shift and 55% during the night shift, we could say that 45% of all accidents occurred during either the early or late shifts. This should be obvious from the information we have available, 20% of all accidents were during the early shift and 25% during the late shifts and 45%, the total of these two, occurred during either the early or late shift.

If we reduce these percentages to probabilities then we would say that the probability of an accident during the early shift is 0.20, the late shift is 0.25 and the night shift is 0.55. We use a form of probability shorthand to avoid having to write "early shift", "day shift" etc., all the time. Let us say that the event that an accident occurs during the early shift is "A" during the late shift is "B" and during the night shift is "C", and so the probability of an accident occurring during the early shift would be written as P(A).

All the probabilities would then be:

$$P(A) = 0.20$$
$$P(B) = 0.25$$
$$P(C) = 0.55$$

In passing we can see that these probabilities sum to 1. This is because we only operate three shifts and accidents which occur will occur with certainty during one of these shifts.

We know already that the probability of (A) or (B) is 0.45. This was found by adding the individual probabilities:

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ &= 0.20 + 0.25 \\ &= 0.45 \end{aligned}$$

In the same way we could have calculated the probability of an accident occurring on either the late shift or night shift. This would be:

$$\begin{aligned} P(B \text{ or } C) &= P(B) + P(C) \\ &= 0.25 + 0.55 \\ &= 0.80 \end{aligned}$$

We must now notice one important feature of this example. The events we have discussed are *mutually exclusive*. By this we mean that the events cannot occur at the same time, in other words it is impossible for one accident to occur in both the early shift and the night shift. An accident can only be in one shift but not both.

This idea of mutual exclusivity could also apply for example to calculating the probability of an injured employee being male or female, of damages being above or below a certain figure, of employees being injured or killed etc. However, there are many cases where event will not be mutually exclusive.

Let us stay with the accidents to employees example. Out of 300 employees, you have, during the past year, recorded 25 accidents sustained by experienced employees i.e., those with more than 5 years relevant work experience, and 15 accidents sustained by those who had previously had an accident or accidents. Of the 25 experienced people, 5 had also sustained an accident in the past. We can represent this accident record in a diagram known as a Venn diagram. The square represents all employees and the circles, the two types of employees involved in accidents. This diagram is shown in Fig. 7.1

We can see in this diagram that out of the total number of employees there are the following categories:

- Those who had no accidents
- Those having an accident, who had work experience
- Those having an accident, who had had previous accidents.
- Those having an accident, who had both work experience and had had previous accidents.

These four groups can be identified in the diagram. One circle represents those having an accident who had work experience, another those having an accident who had had a previous accident and the overlap those with experience and a previous accident. The remainder of the square represents all those who had no accident.

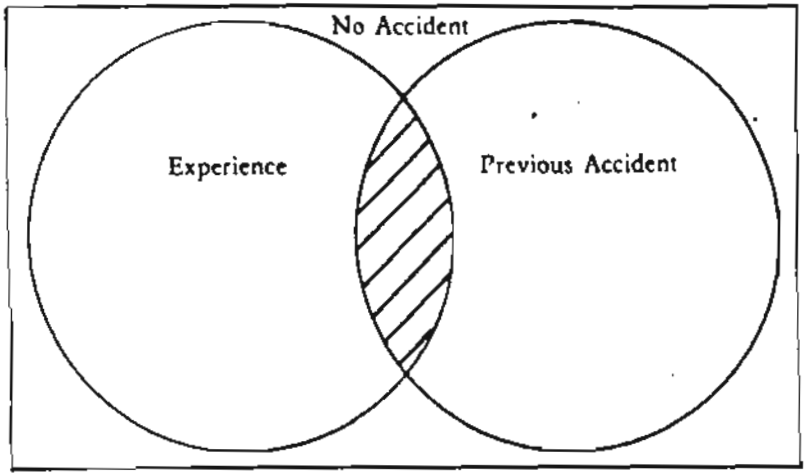


Fig. 7.1

When we insert the numbers we now we end up with the diagram in Fig. 7.2.

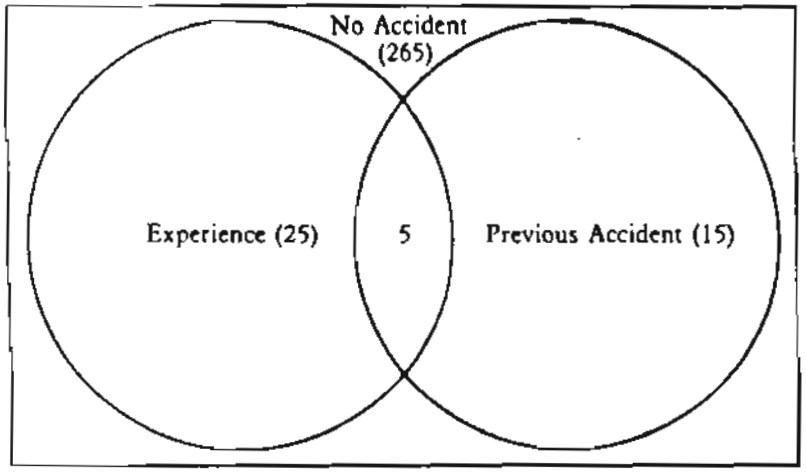


Fig. 7.2

Five people who had an accident had both experience and a previous accident. This leaves 20 who had experience but no previous accidents and 10 who had previous accidents but no work experience.

What then is the probability of a person involved in an accident being experienced or previously had an accident?

Using our notation of before we could let experience be (A) and previous accident be (B) and so we are looking for P(A or B). If we follow the rule we used earlier then we would say that P(A or B) = P(A) + P(B). Now P(A) is the event that a person in an accident had previous experience. We had 25 people and so the probability is 25/300 the probability of a person having an accident, who had previously had an accident is 15/200. Therefore:

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ &= \frac{25}{300} + \frac{15}{300} \\ &= \frac{40}{300} \end{aligned}$$

If this were correct it would mean that the probability of a person not having an accident would be $1 - \frac{40}{300}$ or $\frac{260}{300}$ i.e., it is certain that a person either has an accident or not. If the probability of having an accident is $\frac{40}{300}$ then when this is deducted from 1 which is the certainty we just mentioned we are left with $\frac{260}{300}$ as probability of not having an accident. We can see however from Fig. 7.2 that this figure of $\frac{260}{300}$ is wrong. The correct number of people not involved in an accident is 265 not 260. The formula P(A or B) = P(A) + P(B) does not seem to have worked in this case. The reason why the formula failed is the fact that the events are not mutually exclusive. When we added the $\frac{25}{300}$ to the $\frac{15}{300}$ we added the overlap of 5 twice. Of the 25 with experience there were 5 who had had a previous accident. In the same way the 15 with a previous accident included 5 with experience. It is the same 5 but we included it both in the 25 and the 15.

What we need to do is to deduct it once from the total. A formula for this would be:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \ \& \ B)$$

The expression P(A & B) is the probability of a person in an accident having experience *and* a previous accident. When we expand this formula we have:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$$
$$= \frac{25}{300} + \frac{15}{300} - \frac{5}{300}$$
$$= \frac{35}{300}$$

and so the probability of an accident victim having work experience or a previous accident is $\frac{35}{300}$ or 0.12. This means that the probability of a person not having an accident is $\frac{265}{300}$ or 0.88, which is correct according to the Venn diagram.

Fig. 7.3 shows the Venn diagram with all the relevant probabilities inserted.

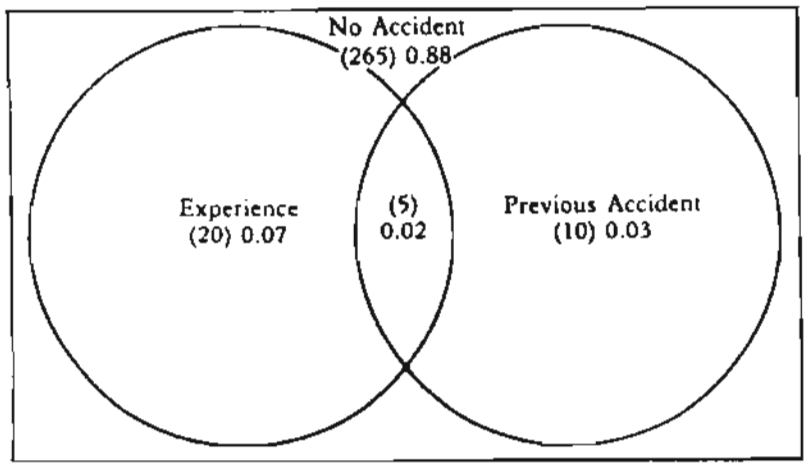


Fig. 7.3

We can now see from this that the probability of an accident victim having work experience but no previous accident is 0.07. The probability of a victim having a previous accident but not work experience is 0.03 and so on.

Apart altogether from the computation of probabilities it is often useful to illustrate complex problems in a simple diagram such as the Venn diagram in order to clarify some of the complex relationships which may exist in certain problems.

7.3.2 Joint Events

Another common way in which individual probabilities are combined i in the computation of probabilities for joint events. We might want t calculate the likelihood of there being a fire at our London factory *an* Birmingham plant during the next year. Based on past records and a element of subjective thinking, we believe that the probability of a fire i Birmingham, P(B), is estimated at 0.02, and in London, P(L), is 0.04.

What we now want is the probability of a fire in London and Birmingham P(L and B). When we stop to think about this we could imagine that th joint probability is likely to be smaller than the individual probabilities The likelihood of a fire in both plants is less likely than a fire in one o other. In fact we multiply the individual probabilities together in orde to arrive at the joint probability:

$$P(L \text{ and } B) = P(L).P(B)$$
$$= (0.04)(0.02)$$
$$= 0.0008$$

From there being a 1 in 25 chance of a fire in London and a 1 in 50 chanc of a fire in Birmingham there is now a 1 in 1250 chance of a fire at *bot* London and Birmingham. We call this rule for computing joint events the multiplication rule.

The simple formula of P(A and B) = P(A).P(B) holds good if events ar 'independent'. This word independent means that the occurrence of on event does not alter the likelihood of the other event occurring. This wa the case in the example we have used so far.

Consider, however, the case where we have two buildings within ver close proximity. The first building has a probability of going on fire o 0.05 and the other a probability of 0.02. The buildings are so close tha if one goes on fire then the other is almost certain to ignite, in fact if on of the buildings is on fire it has been estimated that there is an 85% chanc of the other catching fire.

If we now want to calculate the probability of both going on fire we wil have to take account of this new information. The simple formula P(A and B) = P(A).P(B) will not suffice as it assumes independence between the events, which is not the case.

By re-writing the formula we can reflect the fact that the events are no independent.

$$P(A \text{ and } B) = P(A).P(B/A)$$

The expression P(B/A) is the probability of 'B' given that 'A' has occurred. Using the figures we have above then:

$$P(A) = 0.05$$
$$P(B) = 0.02$$
$$P(B/A) = 0.85$$
$$P(A/B) = 0.85$$

When we calculate P(A and B) we now find:

$$\begin{aligned} P(A \text{ and } B) &= P(A).P(B/A) \\ &= (0.05)(0.85) \\ &= 0.0425 \end{aligned}$$

There is now approximately a 1 in 24 chance of both buildings igniting. This much more likely figure, than the 1 in 1250 calculated earlier, is due to the fact that the two events are not independent.

Notice that the probability of (A and B) is not the same as the probability of (B and A).

$$\begin{aligned} P(B \text{ and } A) &= P(B).P(A/B) \\ &= (0.02)(0.85) \\ &= 0.017 \end{aligned}$$

The figure of 0.017 is the probability that building (B) goes on fire and *then* building (A). Remember that building (A) was more likely to go on fire than building (B) and so it is important to remember which building goes on fire first.

7.3.3 Probability Trees

Before leaving the area of combining probabilities let us take a brief look at the use of probability trees. These trees are a useful way of illustrating the combination of events.

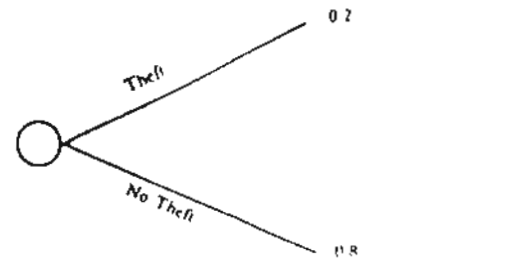
We can show the use of probability trees by means of a simple example. Let us say that at a particular site it is estimated that the likelihood of theft is 0.2. The likelihood of this theft then being of Fixtures and fittings, stock or plant has been put at (0.3), (0.5) and (0.2) respectively. Regardless of what kind of property is stolen the theft could be large or small. The probability of a large theft loss of fixtures and fittings is put at 0.7. The probabilities of large stock and plant thefts has been put at 0.5 and 0.1 respectively.

We can illustrate this problem by a probability tree and draw some interesting conclusions from it. Let us start with a simple tree of whether or not there will be a loss. This is shown in Fig. 7.4(a). In this tree we can see the two possibilities and the respective probabilities at the tips of the branches of the tree.

In Fig. 7.4(b) we have added on further branches showing the nature of the theft. The likelihood of theft was estimated at 0.2 and so the probability of no theft is 0.8. We know the probabilities of having fixtures and fittings, stock or plant stolen and these have been inserted in brackets above the relevant branches. At the tip of the tree we have now shown the combined probability of a theft being of one of the three types.

What we have calculated is the probability of having a theft *and* it being of Fixtures and Fittings etc. This is really just an application of the multiplication rule for joint events which we discussed earlier. We multiply the individual probabilities of the two events in order to arrive

Fig. 7.4(a)



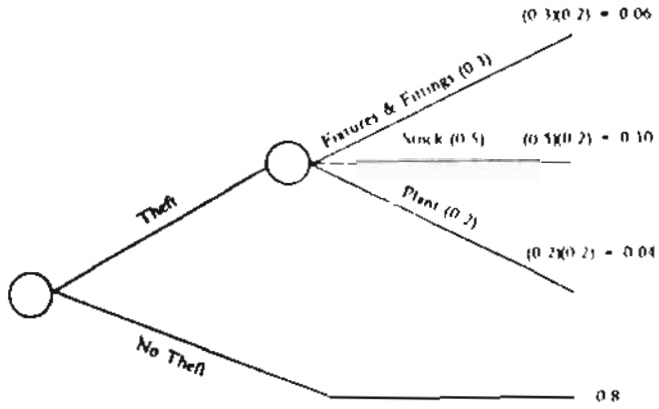


Fig. 7.4(b)

at the probability of the joint event. We can see now, for example, that the likelihood of having a theft of stock is 0.1 and of plant is 0.04.

In Fig. 7.5 we have added the final aspect of the problem, whether the theft was a large theft or small theft. The probability of having a theft of fixtures is calculated as 0.06, and we see this in Fig. 7.4(b). We now know that a large Fixtures theft is quite likely, probably due to the nature of the fixtures and fittings including micro computers, video machines etc. And so, if the probability of a theft of fixtures is 0.06 and the probability of any fixtures theft being large is 0.7 then the probability of having a fixtures theft *and* it being large is (0.06)(0.7) or 0.042.

We can do the same for all the types of property and the probabilities are shown at the tips of the branches of the tree.

We can now use the tree to calculate various other probabilities. For example, the probability of having a large theft of anything other than stock would be 0.046. We found this by saying that any large theft *other than* stock must be a large theft of either fixtures *or* plant. The probability of a large fixtures loss is 0.042 and of a large plant loss is 0.004. The probability of either of these happening is therefore (0.042) + (0.004) an example of the addition rule for alternative events.

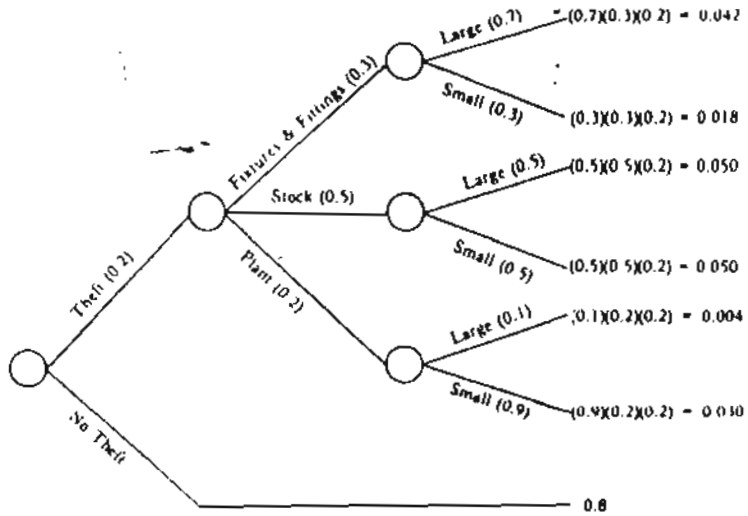


Fig. 7.5

Notice, just before we leave the tree, that the sum of all the probabilities at the tips of the branches is 1. We have a list of mutually exclusive events which are exhaustive of all possibilities and so one or other *must* happen, one or other is certain and so the addition of them all is 1.

This may have appeared to be a rather lengthy introduction to the idea of probability theory, and it has only been an introduction, but it is an important area. Whenever we speak of risk we imply the calculation or estimation of likelihood and we should endeavour to be familiar with the basic notion of probabilities.

7.4 Probability Distributions

We move on now to a particular application of probabilities. Probability distributions will help us to carry out many of the estimates of likelihood which we will want to make.

Let us firstly create a simple example in order to illustrate some fundamental points. A company has 100 vehicles in its motor fleet and over the past year it has kept a careful record of accidents. Of the 100

vehicles, 60 were not involved in any accidents, 20 were involved in one accident, 10 in two, 7 were in three and 3 were in four accidents during the year. This information is displayed in a frequency distribution as follows:

Number of accidents	*Number of vehicles*
0	60
1	20
2	10
3	7
4	3
	100

This is the kind of frequency we looked at in Chapter five. It might, however, be useful for the manager of the fleet to express the number of vehicles having accidents, in relative terms rather than absolute figures. A slight alteration to the distribution can bring this about as we see below:

Number of accidents	*Relative frequency*
0	60%
1	20%
2	10%
3	7%
4	3%
	100%

The manager can now see that 10% of his fleet had two accidents, 20% had one accident and so on. These relative frequencies can also be looked upon as probability statements. Imagine a pool of 100 vehicles and you were to pick one vehicle at random, what is the likelihood, the probability, that it has not been in an accident? Well 60% of all the vehicles were not involved in an accident and so the probability of selecting one is 0.6.

By re-writing the distribution we now get:

(x) *Number of Accidents*	*P(x)*
0	0.60
1	0.20
2	0.10
3	0.07
4	0.03
	1.00

This time we have defined the number of accidents as (x) and the frequency column has been altered to show the probability of x i.e., the probability of a particular number of accidents. We could now use this historical data as the base of our probability distribution. We have to assume, of course, that there are no changes envisaged over the period for which we are calculating probabilities.

We can also draw the probability distribution and we have done this in Fig. 7.6.

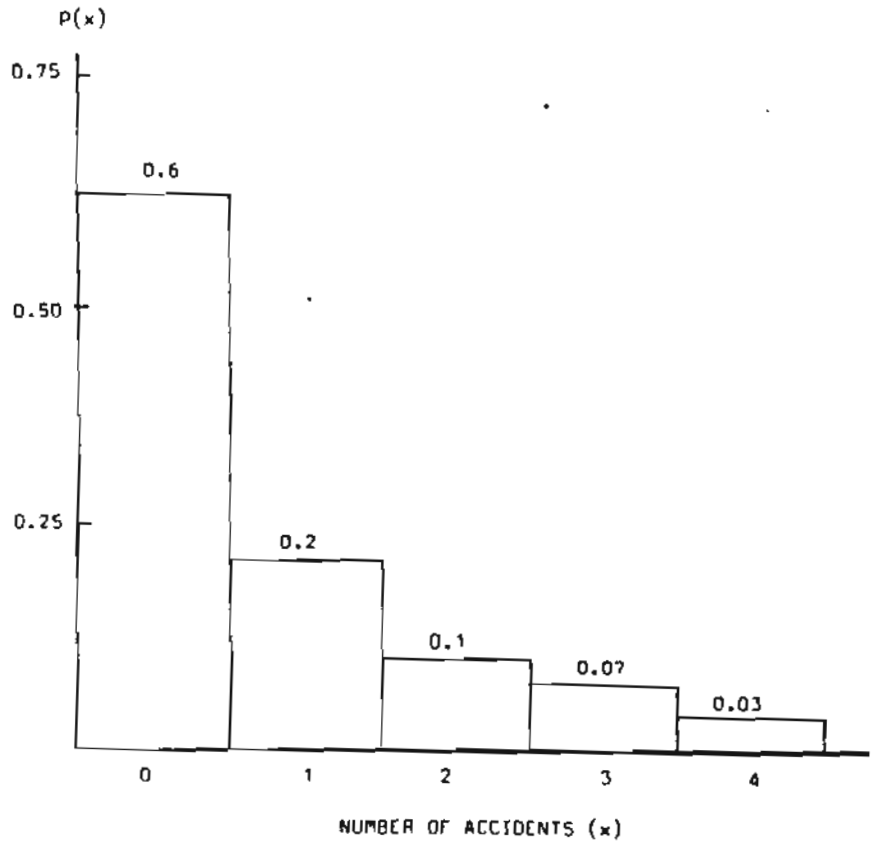


Fig. 7.6

This is similar to the histograms we saw earlier but this time the vertical axis shows the probability with which particular values of x occur. The variable (x) is the number of accidents each vehicle had. The use of the word variable simply denotes the fact that (x) does not have one particular value alone. The value can vary. In fact (x) is a random variable meaning that the exact value of (x) is not known at the outset of any time period. All we know is that there are a number of possible values which (x) could assume.

7.4.1 Discrete and Continuous Variables

One other important feature we should mention about our variable is that it can only be a whole number. It is only possible to have 0, 1, 2, 3 or 4 accidents. It is not possible for there to be 3.84 accidents. When a variable can only be a whole number like this we call it "discrete". The opposite of a discrete variable is one which can be expressed in fractions. Examples of this would include salaries, claim costs, premiums, weights, volumes, temperatures etc. Such variables are termed "continuous". This distinction may seem a little theoretical at this stage but its relevance should become clear as we move on.

In risk management terms we can see that, for example, the number of fires in a year is discrete but the cost of each fire is continuous.

We have looked at a probability distribution for a discrete variable, the number of accidents each vehicle in a motor fleet had. Let us look at a probability distribution for a continuous variable. The accident damage repair cost of the vehicle involved in the distribution we used earlier could be an example.

It would seem from the distribution that there were 73 individual accidents; 20 vehicles had 1 accident, 10 had 2, 7 had 3 and 3 had 4 accidents, in total this gives us:

$$\begin{aligned} 20 \times 1 &= 20 \\ 10 \times 2 &= 20 \\ 7 \times 3 &= 21 \\ 3 \times 4 &= \underline{12} \\ & \quad 73 \end{aligned}$$

Not all of these accidents will have involved accidental damage repair costs and those which did involve some repair would have had costs over a wide range of money. We could look at all the repair invoices and draw a frequency distribution. This we have done in the table below:

Cost of Repair	*Frequency*	*P(x)*
0 < 100	30	0.41
100 < 200	23	0.32
200 < 300	12	0.16
300 < 400	5	0.07
400 < 500	3	0.04
	73	1.00

The table also shows the probability of (x). In this case (x) refers to a range of money, and so the probability of accidental damage repair costs being between £200 and £300 is 0.16, while there is a 41% chance that the repairs will cost less than £100.

When this information is drawn it looks like the distribution in Fig. 7.7 (ignore the shaded area for the moment).

The interpretation of this distribution in Fig. 7.7 is different from the drawing in Fig. 7.6. In Fig. 7.6 we had a discrete variable. This meant that each value of (x), the variable, had an associated probability. We could read from the drawing and see that the probability of x being 3 was 0.07, and so on.

With the drawing in Fig. 7.7 it is not just as simple. In Fig. 7.7 we have a continuous variable. We are concerned with the cost of repairs and clearly that cost could be any fractional amount of money. It is not possible therefore to show a probability against each one of these

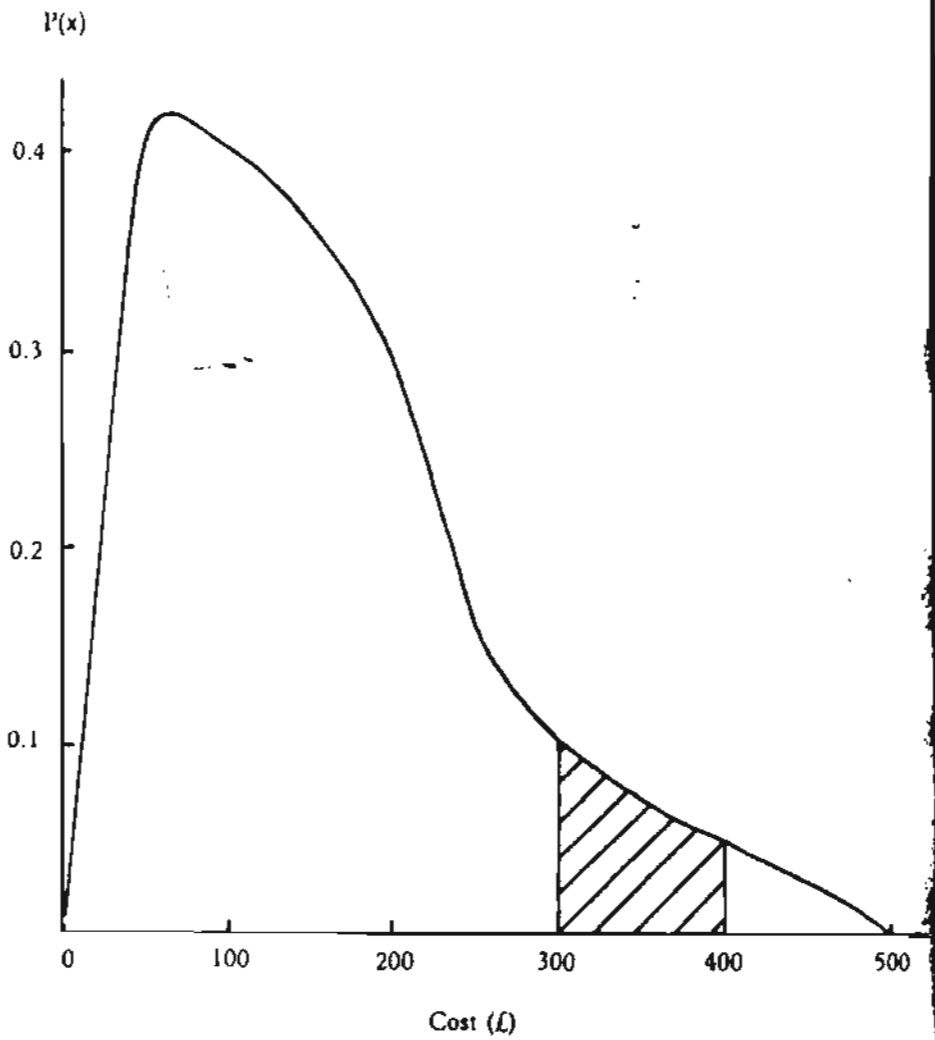


Fig. 7.7

potentially infinite number of different amounts of money. The best we can do is to assign probabilities to a range of values. In our example we grouped the different repair costs into five classes and could then express the likelihood of a claim costing between one amount and another.

Therefore, when we look at the probability distribution in Fig. 7.7 we cannot look upon the y-axis as representing the probability of any specific amount of money. The scale on the y-axis indicates only the height of the curve at any point. This, again, may seem rather abstract but think for a moment about what the curve actually represents, one hundred percent

of all claims costs are bounded by the curve. We know from the table we drew above that the probabilities of the various costs all summed to 1. We are sure that the cost of any one claim will be within the area bounded by the curve. We cannot however simply read off the drawing and find specific probabilities associated with certain amounts of money. For example on our drawing in Fig. 7.7 a cost of about £25 would seem to be associated with a probability of 0.2. However if we draw a horizontal line from 0.2 across the drawing we can see that it also cuts the curve at an amount equal to approximately £240. It just doesn't make sense to interpret the curve in the same way as we interpreted the drawing in Fig. 7.6.

Go back then for a minute to this idea of the height of curve. If 100% of costs are bounded by the curve then the total area under the curve has a probability of 1. We can then find the probability associated with any area under the curve. A specified area under the curve will be a proportion of the total area and this will give us our probability. In Fig. 7.7 we could find the probability of a claim being between £300 and £400 i.e., the area under the curve which is shaded. The actual procedure for calculating this probability can be quite complex, however as we will see later there are ways in which we can be helped.

7.4.2 Actual and Theoretical Distributions

The distributions we have used above could all be looked upon as 'actual' or 'empirical' distributions. They are actual or empirical in the sense that we obtained actual data and then drew the distributions which matched this data. There is no doubt, therefore, that these distributions are a good representation of our data, always assuming that our records and drawings are accurate.

You can imagine that this process will take some time. Each occasion when a probability has to be drawn someone will have to go to the sources of information and extract relevant data before drawing a probability distribution. Not only will the process take some time but we will find that many of the distributions we draw will have a general similarity.

This fact of the basic similarity of many probability distributions is not one which has escaped statisticians and over many years now, there has developed a number of theoretical distributions which closely approximate real world situations. This idea of theoretical distributions is one which many people find extremely difficult to follow, particularly if it is some time since they last studied mathematics. Let us see if we can introduce the idea with a simple non-mathematical analogy and then try to keep the maths to a minimum.

Imagine that you are in the process of buying a new suit. (I hope this is an example with which readers from both sexes can sympathise). One option open to you is to visit a tailor and have him take careful measurements of certain key areas such as waist, chest, arms, legs etc. He then sets to work and produces a suit. You may even go for a try-on of the suit and final adjustments can be made. The hope, of course, is that

the final suit will match your requirements exactly. Those who have one arm or leg shorter than the other will know exactly what I mean when I say how good it is to have a suit which matches these little peculiarities.

The one major drawback to this whole process is the cost. To have a suit made-to-measure will be expensive and will take some time. In the end, however, the finished product will be exactly what you need, assuming you have chosen a good tailor and you did not breathe in at the wrong time during the measurement session!

If the time and expense is too heavy you could always buy a suit off-the-peg. It may not be just as good a fit as the one made specially for you but it will be good enough. You could visit a large store and select a suit which matches, as close as possible, your own measurements. The store will have suits on racks all labelled with different sizes. Once you select the suit you will inevitably find that it doesn't fit exactly. Well, it can't, as the same suits with the same measurements are intended to fit hundreds or even thousands of customers. It will be good enough, one arm or leg may be a little short or long but on the whole it will look quite good.

This suit story is a useful analogy with our probability distribution story. Using an actual distribution is a bit like having a suit tailor-made. It is expensive to do and time consuming but will fit your data exactly. Using a theoretical distribution is like taking a suit off-the-peg. We select a distribution according to certain key measurements and use it. It will be good enough but will not, of course, be an exact fit.

We can imagine then that there are a range of theoretical distributions which will match real world situations. What we have to know is how to select a distribution for a particular set of data we may have.

In broad terms, these theoretical distributions are in two family groups. There is a family of continuous probability distributions and a family of discrete probability distributions. We will start with what is probably the best known of all such distributions, the normal distribution.

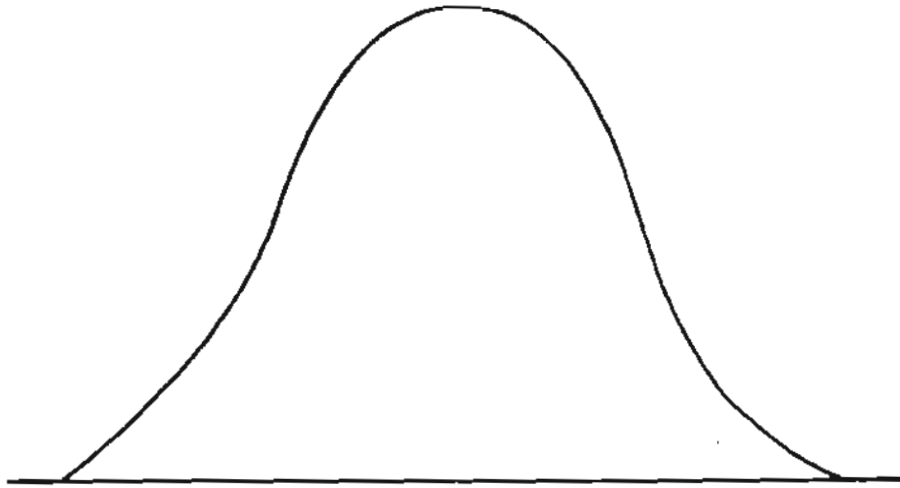
7.5 The Normal Distribution

The normal distribution is a member of the family of continuous probability distributions and we start with it as it is well known and is perhaps the simplest to introduce.

We said earlier that theoretical distributions match very closely what happens in many real life sets of data. If we take for example the heights of males in Great Britain we know that the vast majority are all of roughly the same height, or at least within six inches or so of each other. There are very few small people and very few large people.

In fact if we were to draw a distribution of the heights of men it would look something like the drawing in Fig. 7.8.

This shows that the bulk are of roughly the same height, with fewer people, or a lower probability of finding people, at either extreme.



Heights of Men

Fig. 7.8

If we think back now to the analogy with the suit, when we buy a suit off-the-peg we need to look for one which is roughly our size. We could say that there are some parameters which we work to. These parameters could be our height, chest size, arm length etc. In the case of many multiple stores the parameters are often reduced to one figure such as size 10 or 12, or small, medium or large. However it is done we will select our suit according to some key parameter or parameters.

In the same way our normal distribution is selected according to certain key parameters. It is still the same normal distribution but it will look slightly different to match different problems, just in the way that it will be the same suit which appears in size 36, 38, 42 and 44.

The basic shape of the normal distribution, in the same way as you can talk about the basic design and colour of a suit, is the bell shape we have shown in Fig. 7.8. It is symmetrical around the arithmetic mean and has a main characteristic, the bell shape. The mode and the median all coincide with the mean.

This idea of key parameters can be illustrated by thinking for a moment about temperatures. I have not calculated the annual mean temperature in London but would guess it to be about 55°F. There will be some days when it is extremely cold and some days when it is extremely hot (in those years when summer actually arrives) It is very similar to the heights data. There will be a few extremes with the most likely temperatures being evenly spread around the mean.

In fact the London tempertures will not vary all that much from the mean, the standard deviation will be relatively low when compared to some cities. Take, for example, Moscow. The mean temperature may be very

similar to London but the spread will be much wider. The distribution may still be normal in the sense that it is bell shaped with fewer at the extremes than in the middle but it will look different from the London distribution. The two distributions are shown in Fig. 7.9.

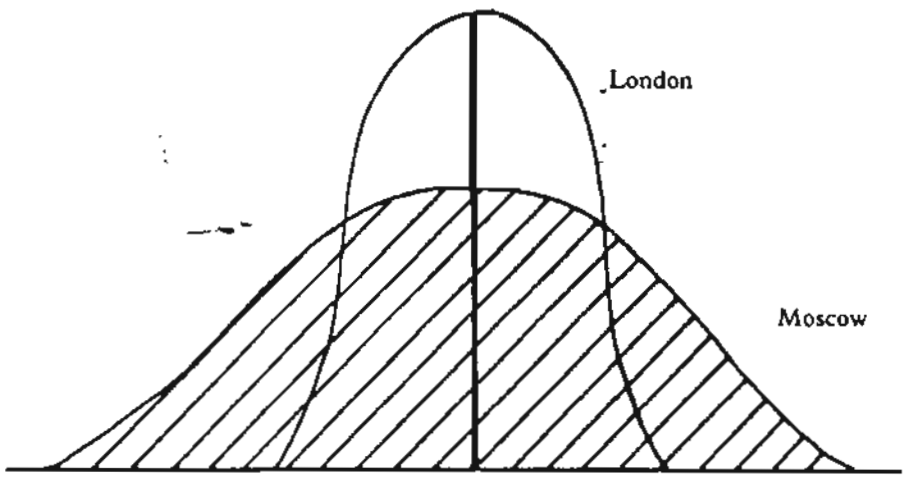


Fig. 7.9

Both distributions are bell shaped and while the mean are almost the same, the spread of temperatures is quite different.

If we compared Moscow with Miami we would get quite a different picture.

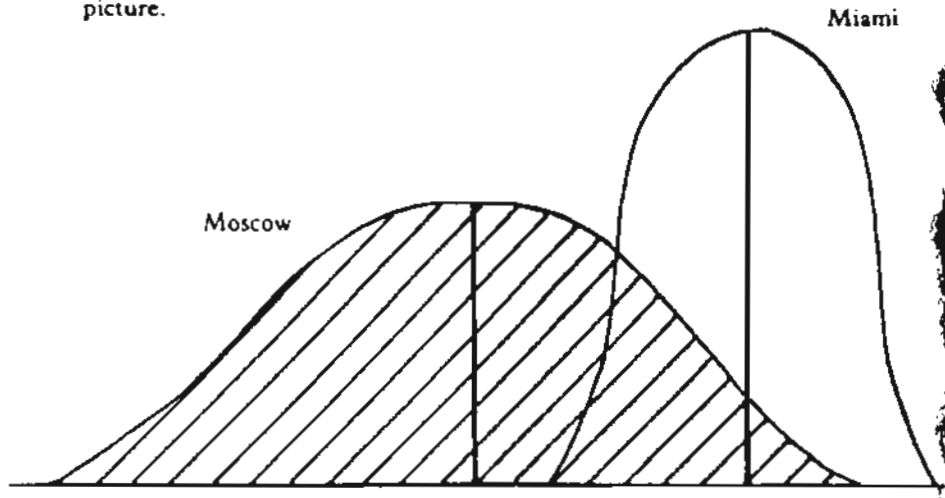


Fig. 7.10

In Fig. 7.10 we have the flatter pancake distribution this time compared with the very peaked distribution for Miami. The Miami temperatures will be tightly grouped around a much higher mean than Moscow.

Looking at the drawings in Figs. 7.9 and 7.10 we can begin to see the importance of the mean and Standard deviation, in determining the shape of the normal distribution. All we need in order to use a normal distribution is the mean and standard deviation of our data.

What then is the value of the normal distribution? Why go through all this theory? What benefit is to be derived?

The benefit lies in the fact that the normal distribution can be explained mathematically. There is an equation, making use of the mean and standard deviation of your data, which explains the curve and allows us to calculate areas under the curve between different points. This relates back to what we said earlier about being able to express the probability of an event being within a range, that range being an area under the curve.

We do not need to bother about the mathematics of the curve but what we may be interested in is the fact that by reading standard statistical tables we can find certain areas under the curve.

7.5.1 Using the Normal Distribution

You can see from the drawings in Figs. 7.8 to 7.10 that the normal distribution can be squat or peaked, depending on the parameters. In order to find areas under the curve then we would need an infinitive number of tables to match all the possible shapes of curves.

To overcome this problem we can re-write the x-axis in a standardised way. What we do in fact is to write the x-axis in terms of standard deviations around the mean. An illustration of this is shown in Fig. 7.11.

This shows a normal distribution with a mean of 30 and a standard deviation of 8. Let us say that this is a distribution of the time taken between the notification of an accidental damage claim in a motor fleet and the final settlement. On average the claim takes 30 days to clear up with a standard deviation of 8. You can see that the x-axis has two labels. The top label is the time measured in days. The lower one is the standardised measurement we mentioned earlier. This is referred to as "Z" and it measures the number of standard deviations we have moved from the mean. The value of z at the mean is obviously 0 and as the standard deviation is 8, then the value of z at 38 is +1. Similarly the value of z at 14 is −2 as we have moved two standard deviations down from the mean.

Well how will this help? Fortunately the normal distribution is symmetrical and so movements on one side of the mean are simply the mirror image of movements on the other side. In addition, the mathematics of the curve result in us being able to calculate the area under the curve between the mean and any number of standard deviations around the mean.

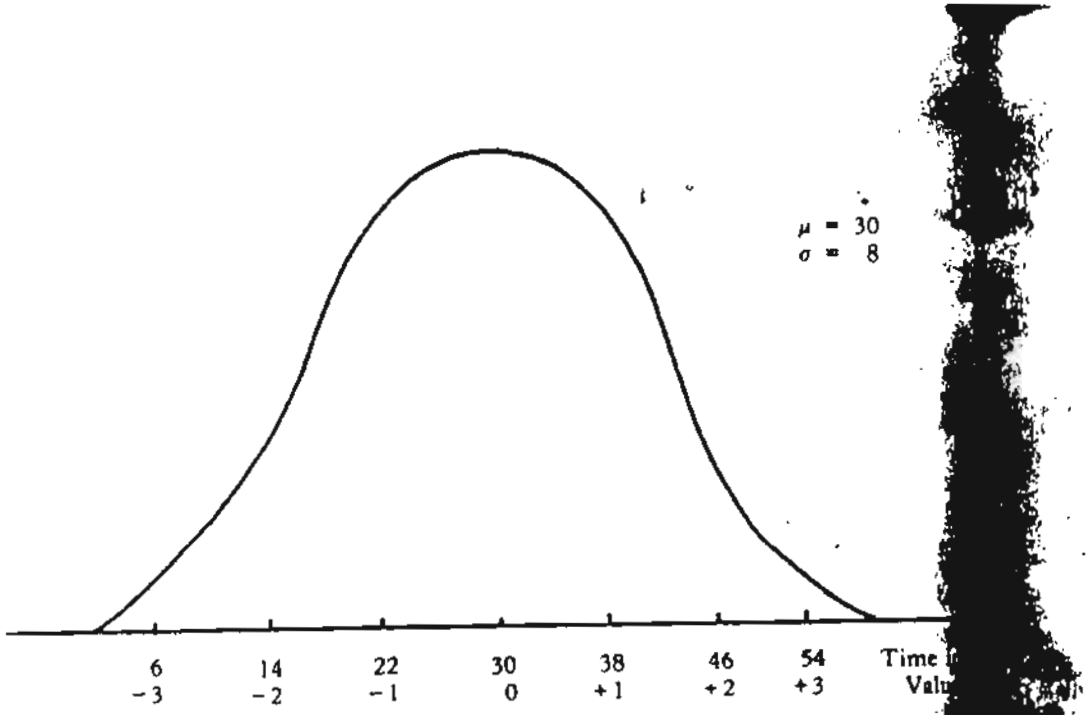


Fig. 7.11

Three well known areas are:

±1 standard deviation embraces 68.27% of all the values
±2 standard deviation embraces 95.45% of all the values
±3 standard deviation embraces 99.73% of all the values

And so in our simple example we can say that 95.45% of all claims took between 14 and 46 days to settle. We found this by moving two standard deviations above and below the mean.

These areas will be so, regardless of whether the curve is peaked or squat and so you can begin to see the great value in being able to carry out these calculations.

The standardised values come into their own when you do not want to move in neat numbers of standard deviations. Say we wanted to know the chance of a claim taking between 30 and 35 days to settle, we want the area under the curve bounded by 30 and 35. What we do is to express this range in a standardised way so that we end up with a z value which we then use to consult statistical tables.

We have covered a distance of 5, 35 - 30, and in terms of standard deviations this would be ⅝th of a standard deviation because a full standard deviation is 8. In other words what we have done is:

$$\frac{x - \mu}{\sigma}$$

where x is the value
μ is the mean
σ is the standard deviation

In our example this is:

$$\frac{35 - 30}{8} = \frac{5}{8} = 0.625$$

What this means is that the value 35 is 0.625 standard deviations away from the mean. Fortunately we have standard statistical tables which tell us the area under a normal curve for any distance around the mean. Most of these standard tables follow the same pattern.

A sample extract would be:

Areas under the normal curve

z	.00	.01	.02	.03	.04	.05	.06
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515

We use these tables by reading down the z column to find the first number after the decimal place and then look along the columns to find the second number. In our case we had z = 0.625 which we could round to 0.63. We look down to find 0.6 and then along to .03 and find a value of 0.2357. This tells us that 23.57% of all values under the curve lie between 30 and 35 days. The probability of a claim taking between 30 and 35 days is therefore 0.2357.

Let us say that for some reason we want an estimate of the chance that claims will take longer than 42 days. It may be that our investments or access to funds changes after such a period. What is the probability of a claim taking longer than 42 days to settle?

In Fig. 7.12 we have shaded the appropriate area.

We know that 50% of all claims take longer than 30 days, this is because the distribution is symmetrical around the mean. What proportion of the area is greater than 42? Our formula, $\frac{x - \mu}{\sigma}$ measures distances around the mean and so we cannot find a z score for distances beyond 42. What we can do is find the area under the curve between 30 and 42 and then subtract it from the 50% which we know lies above 30. In formula terms this would be:

$$\frac{x - \mu}{\sigma} = \frac{42 - 30}{8} = 1.5$$

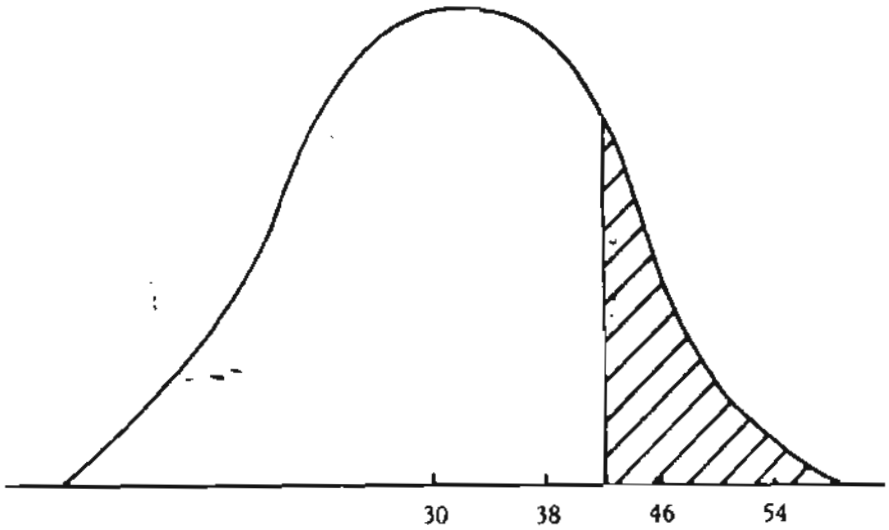


Fig. 7.12

We have a z value of 1.5 which according to our brief extract, represents 43.32% of the area under the curve. And so if the area between 30 and 42 is 43.32% then the area beyond 42 must be 50% − 43.32% or 6.68%. There is only a 6.68% chance of a claim taking more than 42 days to settle. The probability is 0.0668.

The normal distribution is only one of the family of theoretical distributions which are suitable for use with continuous variables. Other distributions include the exponential.

7.6 Binomial Distribution

We must now turn to discrete variables and look at theoretical distributions suitable for these cases. Discrete distributions are very common in risk management. One particular application is in calculating the number of accidents we may expect. Accidents can only assume a whole number as we have said before, and are therefore discrete. Any distribution of discrete variables would conform more to the drawing in Fig. 7.6 and would not be suitable for treatment by the normal distribution which was confined to continuous variables.

Rather than build an actual distribution every time we want to calculate probabilities for a discrete variable, we can turn to the family of theoretical probability distributions for discrete variables.

The binomial distribution is one such distribution. The basic philosophy underpinning the distribution is the same as for the normal distribution; we will have a mathematical equation which will describe a distribution which will be a good match for many real world problems involving discrete variables. Let us begin with a simple problem.

Let us say that you operate a fleet and each year half of the fleet is involved in some kind of accident. One particular branch has three vehicles and you want to calculate the probability of one of the vehicles being in an accident. Why you would want to do this is rather obscure but it will act as a suitable example for us in the meantime.

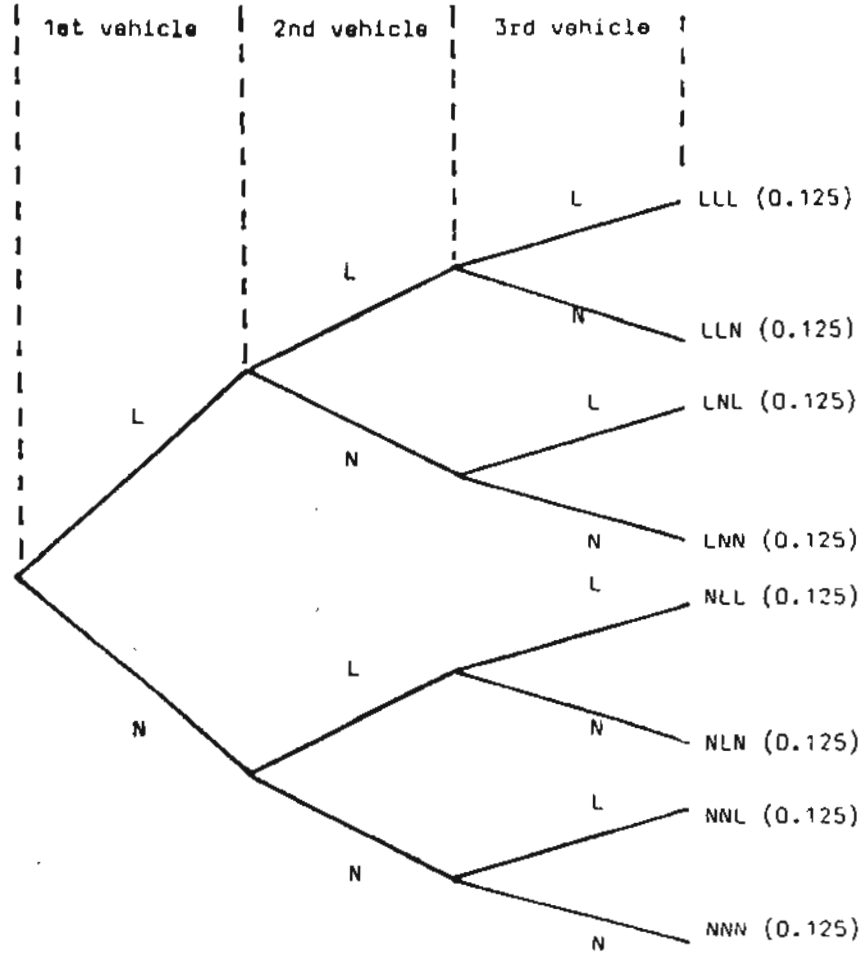


Fig. 7.13

The basics of our problem are that the probability of a vehicle being in an accident is 0.5, half of all vehicles are usually involved in accidents. There are three vehicles and we want to know the probability of one vehicle only, having an accident.

We know from our earlier work on probability that the probability of three vehicles having an accident would be (0.5)(0.5)(0.5) i.e., P(1st *and* 2nd *and* 3rd having accidents) is 0.125. But what is the probability of one vehicle, only, being in an accident. We cannot simply say that it is P(1st in an accident *and* 2nd not *and* 3rd not) as we do not know that the first will be in an accident. It could well be the second or the third vehicle which has the accident. In other words we have a number of possible ways in which we could have one accident.

1st is involved 2nd is not involved 3rd is not involved
1st is not involved 2nd is involved 3rd is not involved
1st is not involved 2nd is not involved 3rd is involved

There are three ways in which one of the three vehicles could be involved in an accident, but we cannot forget that there may be no accidents at all, or two vehicles may be involved etc. In the tree in Fig. 7.13 we have drawn all the possibilities.

Of the eight possible ways our accident record may end up we can identify the three which involve only one accident.

LNN
NLN
NNL

The chance of one vehicle being in an accident is therefore ⅜ or 0.375. We could also have arrived at this figure by adding up the 0.125's at the tips of the tree in Fig. 7.13 which involved only one accident.

You can imagine that this process would be very cumbersome if we had 25 vehicles and wanted to calculate the chance of 5 being in accidents. This is where the theoretical binomial distribution comes in.

If we can say that:

- we can only have two outcomes, success/failure, accident/no accident, fire/no fire.
- the individual trials are independent i.e., whether one vehicle is involved in an accident does not alter the probability of the second being involved in one etc.
- the probabilities do not change from trial to trial i.e., the likelihood of fire does not increase over time etc.

then we can make use of the theoretical, binomial probability distribution.

You will recall from the discussion on the normal distribution that it was described by two parameters, the mean and the standard deviation. The binomial distribution is explained by n and p where n is the number of trials i.e., the number of cars, number of defective parts etc., and p is the

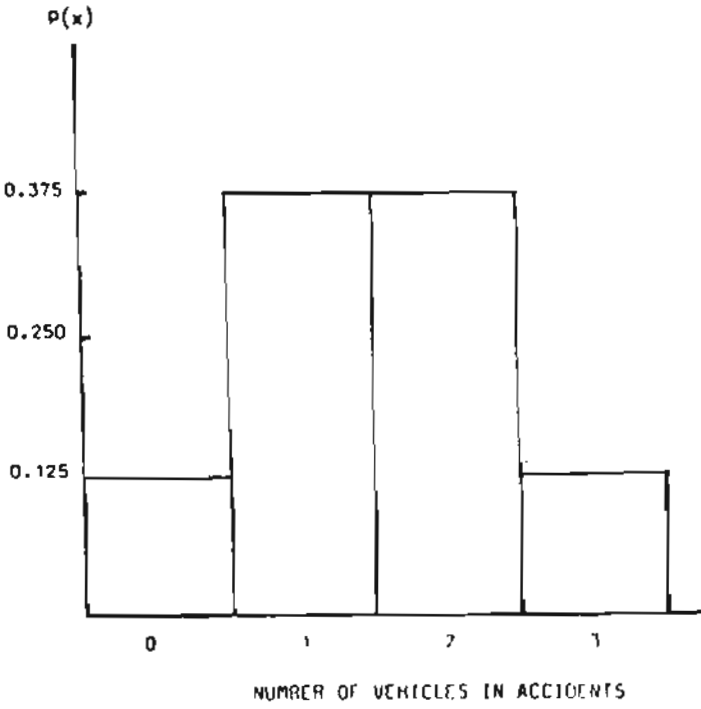


Fig. 7.14

probability of success i.e., the probability of having a specific num defective parts etc. Just as with the normal distribution we will t to consult standard tables to obtain relevant probabilities. Before this, let us examine the two parameters. You will recall th parameters of a theoretical distribution act in the same way as yo height, weight, chest size, arm length act in selecting an off-the-pe You choose that suit which most closely matches your own measure

In the example we started with we had n = 3 and p = 0.5. This g eight possible ways in which we could have one vehicle involve accident. The eight possibilities are shown at the tips of tree in Fi If we draw a histogram of these ways we would end up with the d in Fig. 7.14. We had one way in which we could have no accident ways we could have one accident, three ways we could have two ac and one in which we could have three accidents.

We can see from the drawing that the probability of one vehicle b an accident is 0.375. This is the figure we calculated earlier. As th of n increases, the number of vehicles increases, then the shape distribution will change. As n gets larger the distribution become

peaked. For example if we now have five vehicles we find that there are 32 possible ways we could end up at the end of a year. You could draw the tree for this yourself and see if you agree. Of these 32 ways there are five which involve only one vehicle in an accident. The histogram is in Fig. 7.15.

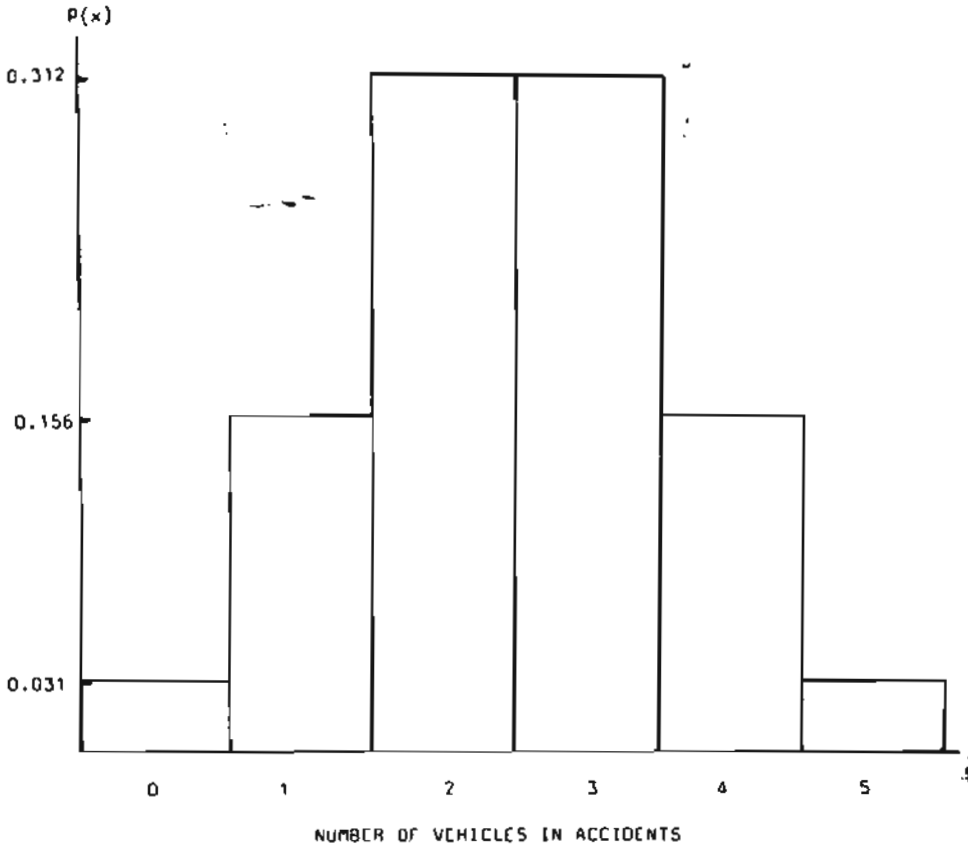


Fig. 7.15

The probability of one of the five vehicles being in an accident is now 0.156. The parameter, n, has altered the shape of the distribution. As the number of trials, in this case the number of vehicles, increases then the distribution becomes more and more symmetrical, in fact it eventually assumes the characteristic bell shape of the normal distribution. However at the moment we can see that the parameter, n, has some effect on the distribution.

The other parameter is p, the probability of the event. In our examples so far we have kept this probability at 0.5. Let us see what effect a change to 0.1 would have.

If you go back to the drawing in Fig. 7.13 you will see that each of the eight combinations had the same likelihood of occurring. This is due of course to the fact that the likelihood of a vehicle being involved in an accident was the same as not being involved in an accident. Let us change the probability to 0.1. It is now much less likely to have a vehicle in an accident. What effect will this have on the shape of the distribution.

Well we can list the eight possible outcomes and show their relative likelihoods:

L = 0.1
N = 0.9

LLL = 0.001
LLN = 0.009
LNL = 0.009
LNN = 0.081
NLL = 0.009
NLN = 0.081
NNL = 0.081
NNN = 0.729

The eight probabilities sum to one just as they did in Fig. 7.13 but they are not all equal. The likelihood of one vehicle being in an accident is now:

LNN = 0.081
NLN = 0.081
NNL = 0.081
0.243

This compares to a probability of 0.375 which we found when the probability of an accident was 0.5. We can draw the histogram of this new distribution and see what has happened to the shape of it.

In Fig. 7.16 we can see that the shape is now quite skewed. The change in probability means that the likelihood of no accidents is fairly high and the chance of all vehicles being in accidents is very low. Notice that the change in the probability simply alters the shape of the distribution. A low probability of accident will give the positive skew in Fig. 7.16 whereas a high probability of accident will give a negative skew, i.e., the peak would be to the right of the distribution.

What we need then are tables which will help us to find probabilities. An extract of one such table is shown on the next page:

To find a probability you locate the value of n, the value of r and the probability. And so we would find the probability of having one accident from three vehicles where the probability is 0.5 by looking down the n column to 3, and then moving along the row corresponding to an r value and 1 to reach the column with a probability of 0.5. When we do this we find the probability to be 0.375. Use the table to find the probability of having one vehicle involved in an accident. You have three vehicles and the probability is 0.1 of a vehicle being in an accident. The figure should be the same as we found earlier, 0.243.

		p			
n	r	0.05	0.1	0.25	0.5
3	0	857	729	422.	125
	1	135	243	422	375
	2	007	027	141	375
	3	*	001	016	125
5	0	774	590	237	031
	1	204	328	396	156
	2	021	073	264	312
	3	001	008	088	312
	4	*	*	015	156
	5	*	*	001	031

n = number of trials e.g., number of vehicles
r = number of events e.g., number of accidents
P = the probability

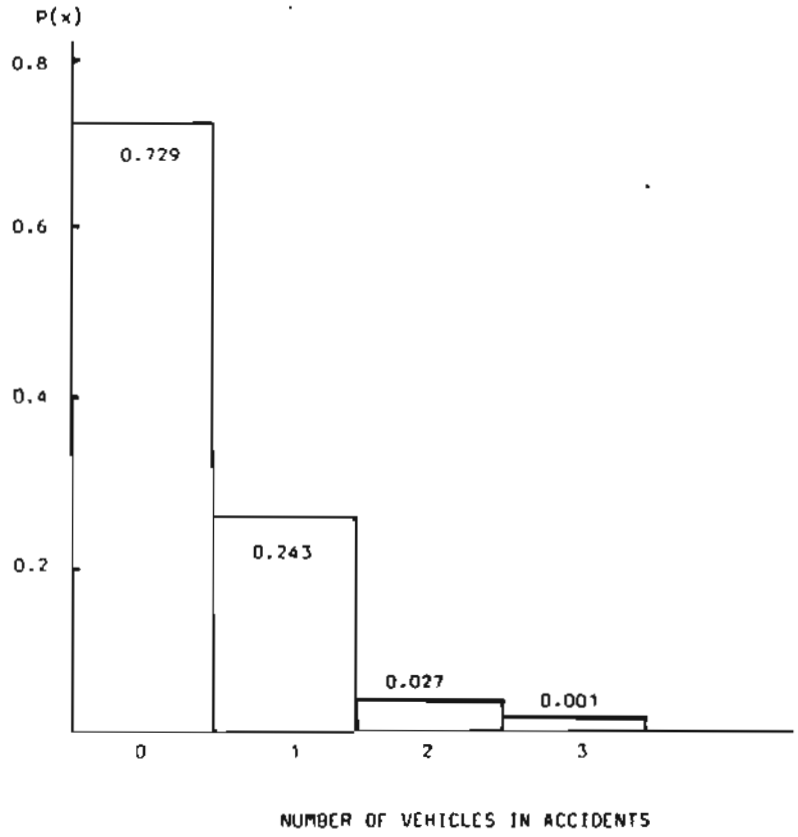


Fig. 7.16

8.0

Chapter Eight

REPORT WRITING

A wide range of topics has been covered by this text so far. We have looked at human behaviour and risk; we have examined a number of detailed risk identification and analysis techniques; we have looked at the role of statistical risk analysis and spent some time on probabilities.

All of this work is not, however, an end in itself. The whole reason for studying these concepts was in the hope that our eventual analysis of risk may be carried out more effectively. But what will happen to the findings of your risk analysis? Even assuming that all the techniques we have touched on in this text are applied expertly to each problem, what will happen then?

Clearly, the results of your analysis have to be acted upon. Just as the techniques in this text are not an end in themselves, so the analysis of risk is not an end in itself. There may be occasions when you yourself may take some action on the results of your analysis but there will also be many times when the results of your work will have to be communicated to others.

These risk analysis reports are extremely important but will only be one form of report which may be produced by the risk management department. Other reports which the department may produce include:

- the annual risk management report. This may be distributed widely within the company or may be for senior management consumption only. Much more on these annual reports is included in the course text, "Corporate Risk Management".
- Reports may be generated on major items of capital expenditure, you may wish to install a new security system, sprinkler system or build some fire defences. These items of expenditure will have to be justified and your justification included in a brief report.
- An annual report on losses within the company may be prepared. This could be separate from the annual report which has a far wider scope, and would concentrate on reporting the range of incidents which occurred. It may also point our or highlight lessons to be learned.
- The result of important decisions may be communicated in the form of a report. For example, the Safety Committee may have made some decision which will affect the duties of safety representatives and so a report may be prepared and distributed.

The list could go on and on. Whatever the report is about it must clearly reflect your intentions whether it is a report intended to persuade people to your point of view, or to justify expenditure or simply to report facts, the report must perform the function you intended for it. In this course book we are primarily concerned with risk analysis but the comments we will make about report writing will hold true for many different forms of report.

We will concentrate on three aspects of report writing:

- Preparation
- Format
- Writing

Having done that we can make some brief comments about presenting oral reports.

8.1 Preparation for Report Writing

Very few people will be able to sit down and write a good report without any preparation. Those who do, probably end up producing a report that looks "quickly done". Preparation is essential. It may be a bit frustrating to be preparing and apparently not actually writing anything but time spent in preparation is a wise investment.

A number of points should be borne in mind when preparing for report writing.

8.1.1 Know Your Reader

You must stand a far better chance of your report having the desired effect if you know who your reader is likely to be. Remember that all your reader knows will be contained on the printed pages of your report. You will not be there to amplify points, to give examples to answer questions. The reader will read your report and form his opinion almost exclusively on the strength of it.

Viewed this way it is clear that a knowledge of the reader is essential if you hope to win him or her to your point of view. At least three aspects should be remembered.

a) Language style. The wrong style of language could be a disaster. Writing a "chatty" report to someone who prefers a far more formal approach can be just as bad as writing a terribly formal report to someone who responds to brief, chatty reports.

The choice is almost between a formal, impersonal style or a conversational personal approach. Clearly there are many stages between these two extremes and the idea is to gauge where your reader is on that continuum.

As a general rule we could say that reports going up the way, to higher levels, tend to be more formal and impersonal than reports travelling along or downward. In addition, reports concentrating on financial matters tend to be less conversational than others. Obviously, a report about some critical incident or serious loss would not be written in an amusing, personal style of language.

Here are two different examples of ways in which the same thing could be reported. The first is much more formal than the second:

"The current problem, highlighted by the finance committee Convener, associated with time delays in small claims settlement

could be resolved by allowing each subsidiary to create a fund out of which losses would be met, without reference to the centralised claims function".

"John Cumming has raised the problem of delays when subsidiary companies have to have small claims settled by the insurance department. How about letting each subsidiary settle their own claims up to £500 and reporting settlements to the insurance department twice a year?"

b) Reader Bias. The second point to bear in mind when thinking about the potential reader is to gauge whether or not he or she has any particular bias.

Your reader may be well known for disliking modern technology, he may be highly profit motivated or safety conscious. It is important to gauge these things before framing the report.

It would be unfortunate if you were trying to acquire funds for the purchase of a computerised record keeping system and prepared a report assuming the reader was enthusiastic about computers, when in fact he or she was quite antagonistic towards modern technology. Knowing that he or she disliked computers would certainly condition how you should frame the report.

c) Familiarity with the Subject Matter. Try also to ascertain how much your reader knows of the subject matter. You can avoid wasting space, and consequently time, if you find out how much he or she already knows about the issue under consideration. For example, if you are preparing a report on the possible purchase of a sprinkler system and find that the person who ultimately will make the decision is a mechanical engineer, then this will influence the report.

If on the other hand the decision maker was strictly a finance person then some of the technical terms and functions may have to be worded differently.

Try also to find out how much the reader knows about the background to the report. There is no point in including a lengthy introduction, painting the full background if the reader is well acquainted with all of this. These points, and indeed many which will follow, may seem common sense. They are just that, but it is interesting to note the number of times that common sense seems to take a back seat. It does no harm to have these things brought to the forefront of our thinking.

8.1.2 Purpose

The second point to remember under the general heading of preparation is to establish the purpose of the report. The last thing you want is for a person to receive the report, read it and then say, "why did I get this report, what am I expected to do now, what was that all about?".

It is useful, at the very outset, to have a statement of the purpose of the report. You can keep this for your own use or can include it later at some point in the report. Having a purpose statement does help to focus attention on the business in hand and the job of preparing a statement of purpose can, itself, be a useful discipline.

The presence of a purpose statement should help avoid some of the reactions we mentioned above. In addition it may be possible to refer to previous correspondence or reports, previous meetings, the decisions at other meetings etc. All of these should locate the purpose of the report in the mind of the reader and avoid confusion.

8.1.3 The Parameters of the Report

It is just as important to establish what the report does not cover as to establish what is covered. The purpose statement will give the broad purpose but the more particular parameters must also be defined.

For example, the purpose may be to prepare a report on industrial injury claims within the company. This may be the purpose but in addition you may define certain parameters such as, industrial injury claims made during the last financial year, industrial injury claims which were also the subject of civil action etc. The purpose is the broad brush and the parameters more closely define the remit of the report.

It might then be useful to state in the report, possibly close to the purpose statement, that the report does not include the following, or that the report is limited to the following.

8.1.4 Management Support

For those reports where you are trying to persuade someone or some committee to your point of view then it is important that you have support. This touches on the very difficult area of "corporate politics". There is no denying that every organisation has its own form of politics. You cannot learn how to identify this or deal with it, from a text book. It is very much something which you learn by experience, often hard experience.

Where you want to persuade someone to your point of view, or sell an idea or obtain permission for some action, then you will have to work out who is important in the decision making process. You would not want the report to be the first notice they had of your intentions.

In other words you may have to do some prior lobbying or selling.

8.1.5 Timing

Remember that someone has to read your report, or at least you hope they will read it. If this is the case then timing is important. Try not to send the report to someone when you know they are busy, when they are out of office, during heavy holiday periods or at the run up to year ends or other crucial dates in the corporate calendar.

Timing could be crucial in obtaining the desired response to your report and so think carefully about it when preparing the report.

8.2 Format of the Report

Let us turn now to the report itself and consider the various formats it could take. All the preparation has been done and you are now ready to begin writing the report, but firstly you must work out what format or structure the report is to have.

A fairly straight forward format would be:

Title
Introduction
Body of the Report
Summary
Conclusions and recommendations.

This is a simple structure and will not satisfy the needs of every kind of report but it is reasonably representative of what you might expect to find in many reports. We can look at each part in turn.

a) **Title.** This would be a one page cover to the report and the title would explain the general purpose of the report. Often it is necessary to have an explanatory sub-title in addition to the main title, and so we have:

INDUSTRIAL INJURY CLAIMS

An evaluation of the cost of Industrial Injury Claims

or,

GROUP INSURANCE COST

An analysis of the cost of insurance, for all group companies.

In addition to the title you would also want to include your own department or name and the date the report is submitted.

b) **Introduction.** The introduction to a report can be fairly brief and should cover topics such as:
- The purpose
- The method
- The scope and parameters
- The definition of any technical terms used.

We have touched on most of these topics before, when we were discussing the preparatory work necessary. The list of definitions is important. In the case of a report on industrial injury claims you can imagine using words such as "claim", "absence", "injury" etc. Each of these words has a lay meaning and a more technical meaning and so you will want to define each one clearly so that there is no ambiguity.

c) **Body of the Report.** We are going to look at the actual writing of the report later but there are some general points we could make

about the body of the report at this stage. This is the heart of the report and it is obviously important that it captures the real purpose of the whole report.

Each report will of course be quite different but there are possibly some general pointers which could apply across a wide range of reports:

- try and lead the reader to your conclusion. This is particularly the case if the report is intended to persuade someone to your point of view or is asking for permission to use funds etc. You must try to frame the body of the report so that statements which are supportive of your case are made. Leading the reader to your conclusion is not easy but with a little practice you can develop the ability to leave the reader believing that there is little alternative to your recommendation.

- Support your major statements. It is very easy to get carried away with the first point and to produce statements which cannot be supported. You can imagine a report which is attempting to persuade management to install a new security system at a particular plant. You want to lead the reader to the conclusion that money should be spent on buying the system for that plant.

  Carried away with this desire you might make statements like, "this particular plant has the worst theft record in the group" or "more thefts take place at this plant than any other". These statements may well be accurate and could well convince the reader of your case. You will however need to support these statements with findings of some kind or another. Perhaps an appendix could be included which shows a comparison of theft losses for all plants. You cannnot make such sweeping statements without providing supporting evidence.

- Leave calculations to an appendix. It is often very difficult to maintain the impetus of the body of a report if it is continually interrupted by calculations. Statistics which support findings or conclusions, financial calculations and any other forms of numbers can be assigned to an appendix without damaging the integrity of the report.

  At the end of the day it is probable that only a few people will read through any calculations and those who do want to do this can still do so by consulting the appendix.

  In the main we could generalise and say that most people are not terribly numerate. The average reader may well appreciate being relieved of the necessity to 'wade' through some 'heavy' calculations on the way to a conclusion.

- Explain the significance and implications of your findings. Where important points are made then you should take time to explain their significance. This is not to imply that the reader could not work it out for him or herself but it is a useful reinforcement.

You could have come to the conclusion that the majority of claims from employees involve skin complaints. The report could then briefly outline the implfcations or significance of this finding:

"as the bulk of employee claims involve skin complaints it is essential that an immediate investigation takes place to establish

i) the nature of all chemicals and other substances handled by employees,

ii) the system of work for handling chemicals and hazardous substances,

iii)the current safety equipment,

iv) the extent to which current safety equipment is used and the extent to which employees are encouraged to use safety gear".

These implications of the major finding could have been worked out by any competent reader but if the report highlights them then they are reinforced in the mind of the reader and there can be no ambiguity about what should now be done.

- Calculate the cost of any recommendations. Following on from the previous point it is essential that all proposals are costed. There is little point in suggesting that a particular kind of glove be purchased or new washhand basins installed etc., if costs are not given. Costs of all proposals should be included and properly stated. By properly stated we mean that they should be couched in terms that those involved in the finance of the company will appreciate. Much more on this is said in the course on "Business Finance".

  Risk Management departments must submit their costings in just the same way as any other corporate function applies for funding. There is no point in relying on emotion or mystery alone.

- Predict objections. Almost every report will be read by at least one person who objects to the findings or proposals suggested. (If only one person objects you are probably lucky.) It often seems as if people get a report and then feel that they must find at least one objection.

  Try to anticipate these objections. For example, you may be suggesting that the company cancels all accidental damage cover on motor vehicles. You could reasonably anticipate that people will object on the grounds of the possible costs of such an idea. Either people will say that a large number of expensive incidents in the year could occur or that the company will be inundated with small "scrapes and dents". Try to anticipate these objections, for example, "some may suggest that a number of large, expensive incidents may occur during any one year and wipe out any cost savings by way of premium reductions. Our historical data does not support this view however. Over the last eighteen years we have never had more than three incidents per annum and on current day values the aggregate cost of claims has never exceeded £14,500. This is well within the expected savings."

Another point of view is that we may be inundated with a number of small claims on the basis that the "company will meet the bill". "This could have been a valid objection had it not been for the fact that we have been carrying a £1,000 accidental damage excess for the past five years and have experienced no increase in small claims over that period."

By anticipating these objections you may take some of the heat out of subsequent discussion. You may also prove to others that you have given sufficient thought to the issues.

- Try to avoid your proposals being rejected. Everyone wants their proposals accepted but outright acceptance may not be too likely. What you want to do is to avoid an outright 'no'. For example, in the above example of cancelling the accidental damage cover you may think that the answer is either yes or no. You want to avoid the 'no' and so you may include an alternative to a straight 'yes' which is not 'no'.

  For example you may suggest that if it is decided not to cancel the cover, that an alternative may be either to reduce the accidental damage cover to catastrophe cover or postpone a decision and monitor the claims for a year.

  Either way the report does not end up on the shelf. It is likely that at least one of the alternatives will be selected, and this still keeps the issue alive.

d) Summary. A brief but concise summary should be placed at a suitable point in the report. Remember that many people may only read the summary and so it should cover all the essential points and make mention of the main conclusions.

c) Conclusions and Recommendations. This is the "bottom" line of the report. Many people will read the introduction, the summary and skip over the body of the report to get to the conclusions.

  State the main conclusions clearly and list all recommendations leaving no ambiguity in the mind of the reader.

  These five points all referred to the format of the actual report itself. We now turn our attention to the business of actually writing the report.

8.3 Writing the Report

It is very difficult to know what to say about writing the report without going back to the basic rules we learned at school. There are possibly one or two points which should be remembered.

a) Avoid using jargon. Plain English is usually good enough for most reports. We have all read reports where people had to,
   "operate within fixed cost envelopes"
   "access funds from an ever decreasing reserve"
   "prioritise capital projects"
   "rank products profit-wise"

   These may look good to you but are probably not highly regarded by others. One area where jargon is rampant is that of the computer and those who have begun using the computer often have a zeal for it which is reflected in their language.

   The request to, "purchase a Winchester in order to bump up the RAM to 20 megabytes in order to dispense with the double density floppies, now that the new generation DB package is operational" will most probably leave people rather cold, ... acceptance - wise!

b) Use simple words and phrases. There is a great temptation when writing a report to get out the "Rogets Thesaurus" and use a "fancy" word when a simple word would do.

   This usually impresses nobody but yourself.

c) Avoid padding out the report with unnecessary words and phrases. The quality of a report is often in inverse proportion to its quantity. As someone once said, "verbosity often conceals a paucity of ideas".

   There are certain phrases and words which creep into many reports and are really not necessary.

   "At this point in time the value of the premiums in monetary terms is continuing to maintain a steady increase. The reason for this may be due to the fact that the number of claims has not followed a downward trend.

   Try and reduce the above paragraph. We could easily put a line through:
   - at this point in time
   - the value of
   - in monetary terms
   - continuing to maintain a steady
   - the reason for this
   - due to the fact
   - not followed a downward trend.

   By editing out these unnecessary phrases we could end up with:
   "Premiums have continued to increase due to a corresponding rise in claims".

   It is a useful discipline to read over a report you wrote a number of weeks or years ago and see how you would re-write it today.

d) Use short paragraphs. You may often have picked up a report and it looked like one long paragraph from start to finish. People like to read short paragraphs.

They can only retain a certain amount in their mind at one time and while reading a long paragraph will almost certainly lose track of what it was about.

Split long paragraphs in order to make the report easier to read. You can make use of headings, numbering, underlining etc. All of these ideas help to split up a page and help the reader.

These are very general points about report writing itself and you should try to look over a number of reports in your department looking for useful help and being critical where appropriate.

8.4 Oral Presentations.

The final topic we will look at is the business of making oral presentations. Much of what we have said about preparing to write the report will apply to the preparatory work necessary for making an oral presentation.

One or two specific points could be made:

i) try and have a look at the room in which the presentation is to be made beforehand. This will ensure that there is somewhere to place your notes, the seating is as you would like it and any visual aid equipment is ready for use.

ii) Speak slowly and distinctly. There is little point in preparing an excellent report and then muff your way through it to the extent that no one understands a word.

iii) Use visual aids if you can. Listening to the one voice for any length of time can be monotonous and it is quite valuable to have slides or a flip chart to use. People can then, at least, have the diversion of looking at something.

Make sure however that the visual aid can be seen. There is nothing more annoying to an audience than a slide being put up and taken away before it can be read. In addition you must ensure that the print on any slide is of a size that it can be read, and of course you should not stand in front of the screen.

iv) Try to look at the audience as much as possible. It is very easy for people in a group, especially a large group, to feel left out and eventually disinterested in the proceedings. Try to look at the people. It used to be the case that public speakers were advised to fix their eye on something at the back of the room so that they would not be distracted. This may help the speaker but doesn't do much for the audience.

v) If you have to read notes, try to make sure that your notes are in order. Try also to avoid it sounding as if you are reading notes. You can certainly add to this list from your own experience of listening to oral presentations.

To conclude, let us say that the report, either oral or written, is often the end result of some risk analysis or other project and as such assumes quite an importance. Do not spend hours on a risk analysis project and spoil all chances of having your findings recognised by rushing at the writing of the report.

INDEX