PART V: EMPIRICAL GENERALISATION

Successful prediction and scientific knowledge depend on results which are known to generalise. In Chapter 19 we briefly consider the different methods of data collection that can lead to empirical generalisations, and in Chapter 20 the ways in which descriptive generalisations lead to explanation and deeper understanding.

CHAPTER 19

Observation and Experimentation

The different approaches to collecting empirical data can broadly be classified as:

(i) censuses or sample surveys which are statistically representative;
(ii) observational studies, in which the observer selects things to observe or measure; and
(iii) experiments, in which the observer deliberately controls or varies some factors.

The boundaries can be a little fuzzy, such as where selection stops and deliberate variation starts, but an artificial laboratory experiment clearly differs from the mere accumulation of observational records. The crucial element is the amount of control exercised by the observer.

The characteristic feature of a scientific result is that it must be generalisable, i.e. repeatable. Any extended study of empirical phenomena must involve more than one set of data and repetition therefore becomes a key element in data collection.

19.1 Repetition

To report that children's heights and weights have followed the equation $\log w = .02h + .76$ in a single study is merely to present an isolated finding. The initial study must be repeated if the result is to gain any scientific meaning.

The first repetition is the most dramatic. If the initial result does not hold again, then we have learned that it cannot generalise, at least not in any simple way. But if the same result *does* hold a second time, we know it could generalise even further.

Establishing the mere possibility that the result is repeatable is clearly only the beginning. But even a few successful repetitions can greatly progress the study if the new conditions of observation are different enough. For example, we needed only three or four studies in Part II to see that the

height/weight relationship, $\log w = .02h + .76$, held despite differences in race, sex, age, nation, time, or observers.

No study can be repeated under literally identical conditions. But what matters is that the *result* can be repeated, not whether all the conditions of observation can be duplicated. Indeed, repetition in various *different* situations is required, to determine the range of different conditions or factors under which the result holds. What factors to vary therefore becomes a paramount question.

19.2 **Factors to Vary**

With a completely new result one usually aims to repeat the initial study with as few changes as possible: the same observer, the same apparatus, the same source of material, etc. (although *some* things must change, e.g. time). The purpose is to see easily and quickly whether the same result can be repeated at all.

After this one tries to extend the degree of generalisation. This is best done by changing the conditions of observation as far as possible without making a substantial difference to the results. But one may even go beyond this expected point, since it will either set a limit on the possible range of generalisation or lead to an unexpected, and hence even more dramatic, extension.

For example, one would not follow up a result on the height/weight relationship for white boys with more data on white boys, but with data on white girls or on black boys. Usually more than one factor is varied at a time, so one might go from a study of white boys in Birmingham in 1947 to a study of black girls in Ghana in 1970. If the result holds again, then a single study will have achieved a major breakthrough. One will have found that neither the change of race, sex, nation, time or observer, nor a lot of other non-explicit factors like temperature or seasons, affects the relationship. (There might be compensating factors, e.g. a difference due to sex cancelling a difference due to race; but such details would be covered in subsequent work, where white girls and black boys would also be covered.)

If the ambitious extension of conditions to black girls in Ghana did not work, one would have established one of the limits of the earlier result. To determine why the breakdown occurred, one would have to backtrack on the separate factors. Was the breakdown caused by the difference in time, sex, race, nation, or some other factor like age or nutrition? These further studies would now concentrate on one major factor at a time (but other more exploratory ones could still be varied as well). Varying one factor at a time is only important when the factor actually affects the results.

But even when only one factor is varied, the outcome must still be confirmed by repetition. No single study is all-important. (The "crucial experiment"

of science is largely a myth since a study is regarded as crucial only if it has become repeatable.)

19.3 Statistical Surveys

Sample surveys or censuses are, at their simplest, a method of collecting data where no factors are varied at all. Such surveys lead to statistical averages whose validity depends entirely on the data being statistically representative. Thus if 22% of the population lives in Town A, then a nationally representative survey must take 22% of its data from Town A.

Simple statistical surveys are a "last resort" method, used when there is no more structured way of collecting the information. One problem is that, as a technique, the simple representative sample survey cannot cope with the basic requirement of repeating a study under different conditions. For example, taking two samples from the same statistical population is not independent repetition-the results *must* be the same, except for differences due to random sampling errors. On the other hand, taking two samples under different conditions, e.g. sampling from the same set of people on two different days, is no longer mere sampling. One has deliberately changed the empirical conditions of observation, i.e. the populations, and non-statistical control has been introduced.

More generally, control can be introduced into surveys and censuses by stratification. A national opinion survey might show that 40 % of the people like the Government. Instead of merely reporting that fact, the results can be broken down to show that it is 40% in the North and in the South, and that it is 60% in large towns and only 20 % in small towns and rural areas.

This set of results is more meaningful than the simple national average. By exercising more control over the data, the observer in effect has a number of *different* surveys. Statistical representation is only needed to deal with the uncontrolled variability within each of these smaller surveys. The purely *statistical* element in most structured surveys and censuses is to ensure that little or no systematic error or "bias" has crept into the results.

19.4 Observational Studies

In most observational studies the observer exercises a good deal of control. He selects the variables he will measure and the various conditions under which he will measure them. There is even some interference with the material which may affect the results, since getting people to answer questions or to stand up straight in order to measure their heights is not their normal behaviour. But in an observational study the observer generally does not try to change his material substantially.

Many observational studies are designed so that different factors vary together, instead of just one at a time. For example, a group of patients treated with a certain drug may have recovered better than an untreated group. But aside from the drug treatment the two groups may have differed in the intensity or nature of their illness, their ages, their previous medical histories, the amount of nursing care they received, etc. This makes it difficult to assign causes.

To identify the effects of the medical treatment more closely one has to try to eliminate other possible factors by comparing patients of the same degree of illness, by imposing the same amount of nursing care, etc. The process of eliminating possible factors is piecemeal and certainly not foolproof. It does not positively establish any form of cause-and-effect. This is what makes scientific progress slow and laborious.

But the strength of the observational approach lies in its facility for producing *negative* results: that some factor does *not* matter. A single observational study showing that the equation $\log w = .02h + .76$ holds for both boys and girls shows that in general sex cannot affect the result.

It may seem naive to base a conclusion on a single observational result, but one is not claiming that all boys are like all girls. One is merely saying that the relationship has been found to be unaffected by sex, the factor did not matter in this study. If in the next study we find that girls *are* different, the cause must be some *additional* factor, like the ages of the girls. Or there may have been something special about the *first* study, if that result never repeats. One does not usually have to rely on a single study for long since the first result will be either confirmed or limited by information from later data.

19.5 **Controlled Experimentation**

Experimentation can often speed-up the observational approach. Here the investigator deliberately varies or adjusts some of the factors in the situation. If temperature is a relevant factor, he does not wait for the desired temperature level to occur naturally, but does something to make it warmer or colder.

Experimentation has two major roles: deliberate control, to keep specific factors the same in different studies, and deliberate variation, to see what will happen. Thus the typical laboratory experiment is a way of creating *unnatural* observational conditions outside the normal range of variation. Much of the success of science has been due to this power to explore hypotheses by creating artificial situations.

But an experimenter cannot necessarily control all of the factors in a situation. For example, he might select one group of patients and deliberately treat them with a drug and select another "control" group whom he does *not* treat with the drug. He can control the way the drug is administered, try to

keep certain background factors like nursing care the same, and generally try to eliminate many of the varying factors that might cause confusion in a purely observational study, where the observer does not even decide which patients are to be treated. But even the deliberate experimenter will usually not wholly succeed in eliminating other factors. His two groups of patients may still not be fully comparable. Consciously or unconsciously, the untreated group may have been selected to have a different chance of recovery, or the drug may have side-effects which required additional treatment and *that* might have caused the difference in recovery rates. When faced with material that is variable and unpredictable, the experimenter's judgment alone is not enough to avoid the possibility of bias.

19.6 **The Randomised Experiment**

About 50 years ago Sir Ronald Fisher invented a way to control the variability of experimental material through the randomised experiment.

Suppose we have a total of 32 available patients. If we randomly divide them into two equal groups of 16, A and B, we leave the possibility of systematic bias, that the patients in Group A are inherently more likely to improve, literally to chance. (With large enough samples there will only be a very small possibility that the two groups are substantially different. Even with small samples we can estimate the probability of this happening with statistical tests of significance and confidence limits.) If Group A is now treated with the drug but not Group B, we exclude the effects of the variability of the material, subject to the known and usually small degree of statistical risk.

This kind of randomised experiment can greatly cut the time and effort required to eliminate uncontrolled variability in an experimental situation. But the approach still has its limitations.

Firstly, randomised experimentation, and in many cases experimentation of any kind, is often impossible, e.g. in many parts of biology, medicine, sociology, economics, geography, and, until recently, all of astronomy. For example, one cannot randomly allocate children to either "nuclear families" or "communes" and observe the difference. While randomised experiments are helpful, they are in any case not crucial; many scientific results were obtained before Fisher made his discovery in the 1920's.

Secondly, randomisation provides no safeguard that the treatment, the factor deliberately. varied, was the only difference between the experimental and control groups. Randomisation only eliminates the inherent variability of the test material itself. The "true" effect of the treatment might still have been due to an impurity and not to the drug itself. Or the shear fact of administering the treatment might have caused some patients to improve: a "placebo" effect.

Thirdly, a randomised experiment will show the effects associated with the treatment (within probability limits), but it does not follow that the result will be repeatable. Randomisation cannot eliminate specific local factors, e.g. that the 32 patients in the clinical trial may have some mineral deficiency due to where they all live, or that the local strain of the virus is perhaps more susceptible to treatment.

As always, the study must be repeated to determine whether the result occurs again under other conditions. But two randomised experiments no longer constitute a controlled experiment. Many of the differences between two studies cannot be controlled, let alone be "randomised away". Thus any sequence of separate experiments degenerates into an observational type of situation. It is therefore the observational approach, with all its complications, that is central to scientific methodology.

19.7 **The Design of Experiments**

Sometimes a more complex experimental design enables a series of separate experiments to be carried out under more comparable conditions, thereby eliminating much of the ambiguity in interpreting an isolated result. If randomisation can be used in the larger design, the gain in eliminating extraneous forms of variation is even greater.

With the proper design, many different experiments can be conducted simultaneously, whilst still varying only one factor at a time. This is another major advance pioneered by Fisher. Such experiments do not necessarily require increased resources, e.g. increased numbers of readings. Indeed, these "factorial" types of design often reduce the amount of statistical error in the experiment,

Table 19.1 illustrates how a more complex experimental design for a clinical drug test can clarify some of the complicating factors mentioned earlier, like the level of nursing care, the level of dosage, and the possible "placebo" effect of the treatment.

TABLE 19.1 A More Complex Clinical Trial

(4 patients in each "cell")

	Dosage				
	High	Low	Placebo	Control	Total
Nursing Care					
High	4	4	4	4	16
Normal	4	4	4	4	16
Total	8	8	8	8	32

We still have a total of 32 patients, but instead of merely dividing them into two sub-groups of 16, one to be treated and one untreated, we have broken them into eight cells of 4 each.

Since level of nursing care was felt to be a possible cause of patients' responses, the patients are evenly divided with two different levels of nursing. Then within each group of 16, 4 patients are given a "high" dosage level of the drug, 4 a "low" level, 4 are a plain untreated "control" group, and 4 are given an inert tablet to simulate the effects of being "treated". The low number in each cell does not necessarily reduce the effective sensitivity of the experiment in establishing the results of the treatment.

Suppose first that (i) the variation in the dosage level makes no difference, (ii) the placebo has no effect (the results being the same as for the control group), and (iii) the variation in the two levels of nursing care makes no difference. We then have several major results from the one experiment: a comparison of 16 "treated" and 16 "untreated" patients just as before, *plus* the knowledge that dosage levels, placebos, and nursing care do not affect this comparison.

Next, suppose that nursing care matters but the medical drug is irrelevant. Then we have a comparison between high and normal levels of nursing, again based on total samples of 16 patients each.

Thirdly, suppose that both the drug and nursing care have an effect but that they do not "interact", i.e. the drug has the same effect at both levels of nursing, and the higher level of nursing produces better results irrespective of the improvement due to the drug. Then we have the result for the drug based still on a comparison of two groups of 16, and the result for nursing based on two groups of 16, two experiments for the price of one. We also have the knowledge that there is no "interaction", i.e. that the drug works in the same way at the different levels of nursing. This is the beginning of empirical generalisation.

Finally, suppose that several of the factors operate and "interact", e.g. that the placebo has a positive effect at the normal level of nursing but none at the high nursing level (where patients are getting enough "attention" anyway). In such a complex case, one needs to compare results based on the smaller sub-samples in the design. These will be statistically less accurate than results based on groups of 16, but usually this loss in statistical sensitivity will be outweighed by the additional information gained about the relative complexity of the situation. Any comparison based on single experimental and control groups of 16 each would have been superficial and potentially misleading and the sooner one knows about this, the better.

Factorial designs are generally easier to interpret if they are randomised, e.g. if patients are allocated to each sub-group at random. Then the initial statistical analysis can often follow the lines of Fisher's Analysis of Variance procedures. (See Section 18.6 and Exercise 19G.) The theory of the design

and analysis of such experiments has been vastly elaborated in the last few decades, as described in various specialist texts (e.g. Fisher, 1935; Cochran and Cox, 1957; Cox, 1958).

19.8 Theoretical Norms

Controlled experimentation is largely concerned with comparing a treated group with an untreated control group to determine what would have happened without the treatment. But such a direct empirical check is only required when relatively little is known about the subject-matter. Often enough is known to *predict* the normal response levels.

For example, when we analysed the half-yearly purchases of Corn Flakes and other breakfast cereals earlier, the data had been collected under somewhat unnatural conditions (e.g. only *five* established brands could be bought from the retail outlet in question). One basic question in the initial study was whether these artificial conditions affected the observed purchasing behaviour (see Charlton *et al.*, 1972). But it was not necessary to collect data on normal purchasing behaviour under conditions somehow "matched" with the experimental ones (e.g. same product, same part of the country, same types of household, same season, etc.). It was already known that under everyday conditions the incidence of light and heavy buyers of a brand generally follows the Negative Binomial Distribution (NBD) within close limits of approximation: i.e. for a large variety of different product-fields (food and non-food), for large and small brands, in the U.K. and the U.S., for different lengths of time period, etc. All that was needed was to compare the experimental results with such validated *theory*, as illustrated in Table 19.2.

TABLE 19.2 Corn Flake Purchases in 24 Weeks

(From Table 12.9a)

	Number of Purchases											
	0	1	2	3	4	5	6	7	8	9	10	11+
% households buying												
Observed %	39	14	10	6	4	4	3	3	2	2	2	10
NBD %	35	16	10	7	6	5	3	3	2	2	2	9

More generally, a doctor dealing with patients diagnosed as having acute appendicitis does not run a controlled experiment ("the operation was successful because half the patients died"). Instead, he already knows from past experience what would happen most of the time if he did not operate.

Again, the physicist measuring temperatures by looking at the-length of a column of mercury in a glass tube does not check every reading against "control readings" obtained from beakers of boiling water and crushed ice at the end of his laboratory bench. Instead, his past experience enables him to predict successfully that the mercury would reach the 100 mark for boiling water and 0 for crushed ice. All he need do is check very occasionally that nothing has gone wrong. Similarly, in analysing the height/weight data for children earlier in this book we did not compare the numerical data for one group with that for another group. Instead, we only compared the data with the theoretical abstraction log w $= .02h + .76$ which had held for all the previous data.

One of the most immediate and powerful uses of empirically based theory is to provide such norms based on prior knowledge. It largely replaces the use of statistically designed experiments, because it is easier and more effective than collecting and analysing new empirical data for control or comparison every time. It is the usual procedure in more mature subject areas.

19.9 Summary

The planned experiment, expecially one using randomisation, can greatly reduce the ambiguities in interpreting any finding. More than one factor can be varied in a suitably designed experiment, so that a wide range of generalisations can be established in a single study.

But repeating a controlled experiment under different conditions becomes a form of observational study because the differing conditions cannot be experimentally controlled. Thus the "observational" approach rather than experimentation remains the most basic form of data collection in science.

CHAPTER 19 EXERCISES

Exercise 19A. The Conditions of Observation

An empirical observation is made under numerous conditions. Can these be classified?

Discussion.

Take the measurement of the boiling-point of water as an example. The conditions of observation can be classified as:

(i) conditions used in the analysis, e.g. a correction for the atmospheric pressure ;
(ii) conditions recorded but not used in the analysis, e.g. how much heat was applied;
(iii) conditions observed but not recorded, e.g. the time of day;
(iv) conditions that could have been observed but were not, e.g. the humidity ;

(v) conditions that could not have been observed, e.g. the phenomenon of super-heating when the relevant concept or measuring procedure was still unknown.

One could record a myriad of factors in any experiment, but one generally ignores those which previous experience has shown to be more or less irrelevant (such as what was going on in the next room, what the observer ate for breakfast, how many measurements he had already made that day, etc.). Only when an unexplained discrepancy occurs may one start digging into such other factors for a possible explanation.

Exercise 19B. Repeating an Experiment

Should we aim to repeat an experiment under identical conditions?

Discussion.

No, for two reasons. Firstly, repeating an observation under identical conditions is impossible *(something* must have changed, e.g. time or place, etc.). Secondly, if identical repetition were possible it would be pointless since we would know beforehand that we must get the same result.

It also follows that a large number of repetitions is pointless if the conditions observed and/or recorded are all very similar. Variety of conditions is what matters.

Exercise 19C. Explaining a Discrepancy

What does a chemistry teacher do when litmus paper turns blue instead of red on being exposed to an acid?

Discussion.

The teacher does not suppose that he has disproved a law of nature, but says "Sorry, something has gone wrong!" and tries again. If the paper still turns blue, he checks whether he has used the right bottle.

All scientific laws only hold under a specified range of conditions and he knows that discrepancies will occur if these conditions are not fulfilled. Occasionally a discrepant observation arises which is outside the "normal" type of exception (Fleming's discovery of penicillin has already been mentioned as a popular example). But most of the time we know (or guess) that discrepant results are due to types of error which are already known about.

In the early stages of studying a topic, exceptions will not yet be well understood, so they require following-up. The various conditions of observation have to be checked and further observations made. Sometimes a discrepancy is non-repeatable. It then remains an isolated exception, with no sort of explanation. But just because it was not repeatable the exception usually becomes increasingly unimportant, a once-only event.

Exercise 19D. The Uncertainty Principle

Is sociology a science?

Discussion.

Qualms are often raised about the scientific study of sociological phenomena, since the act of observation will affect those who are observed. But this also occurs elsewhere. For example, 50 years ago in subatomic physics Heisenberg's *Uncertainty Principle* said that the "quanta" of energy used in trying to measure the position and the speed of an electron are so large relative to the electron itself that they interfere with it too much to measure both position and speed simultaneously.

In fact, everything affects everything else. Lifting one's hand to write affects the gravitational pull of every body in the universe, but mostly only to a trivial extent. The aim of science is to isolate and codify those phenomena which are effectively related to only a few other variables.

There is plenty of experience, both everyday and scientific, to show that there are generalisable regularities in the social sciences, i.e. phenomena which are *not* affected by a myriad of other factors. These phenomena may not always be the important problems which the practical minded person wants to solve immediately, but then the physicist still cannot measure both the position and speed of an electron, nor readily transmute lead into gold.

Exercise 19E. Time as a Condition of Observation

"Forecasting is always difficult, especially when it concerns the future." Discuss.

Discussion.

If the same result has been observed at several different points in time, we know that time as such cannot affect it. We can therefore predict that the result will hold again in the future within the range of conditions already covered, e.g. for different observers, different places, etc.

If such a prediction fails, we know that time itself cannot be the cause. The failure must be due to some more specific factor in the new conditions of observation, e.g. because the temperature or the lighting was different.

Forecasting problems arise because we do not know how factors *other* than time will change, or what their effect on the result will be.

Exercise 19F. The Analysis of a Factorial Experiment

Table *19.3* sets out the average readings of a clinical trial of a drug, designed along the lines of Table 19.1. Discuss the main steps of the analysis.

Discussion.

The marginal averages show three main effects:

(i) treated patients score substantially higher than non-treated patients;

(ii) dosage level makes a marked difference, but there is no placebo effect;

(iii) there is a small difference in favour of the normal level of nursing.

Table 19.3 Mean Values of A Diagnostic Measure

	Treatment Dosage				
	High	Low	Placebo	Control	Average
Nursing Care					
High	82	62	42	50	59
Normal	78	68	62	52	65
Average	80	65	52	51	62

But there is also some indication of "interaction" between the treatment and nursing factors. The results for the two dosage levels differ more at the higher level of nursing than at the normal level, and the placebo group has a substantially lower score at the higher nursing level. Perhaps the normal level of nursing was better in some ways. (The artificially high level of nursing might have been supervised by a tyrannical head-nurse.)

If this was the first experiment of its kind, the interpretation of the observed effects cannot be clear-cut. But the relatively elaborate experimental design allows stronger conclusions about the treatment than would have been possible with just a simple comparison between a treated and a control group of patients.

However, we still do not know to what degree the results have been affected by the allocation of different patients to the various design "cells". This question could have been eliminated by allocating the 32 patients randomly to the 8 "cells".

Exercise 19G. The Analysis of Variance

How would one test the significance of the results in the preceding experiment if the patients had been randomly allocated to the different design cells?

Discussion.

The usual statistical procedure used to establish whether random sampling errors caused the differences in Table 19.3 is the Analysis of Variance (see Section 18.6). This process uses the sampling distribution of the F-ratio and is based on breaking the total variance of all the readings about their overall mean into different components.

The figures below divide the total variance of the data in the clinical drug trial into different components for the four treatment levels, the two nursing levels, the interaction of treatment and nursing, and the pooled residual variance (based on the variances of the four readings in each cell).

	Degrees of Freedom	Variance Estimate	F-ratio
Treatment	3	1,477	10.6
Nursing	1	288	2.1
Interaction	3	208	1.5
Residual	24	140	
Total	31	281	—

The residual variance is the basic figure used to test the statistical significance of the differences in the observed readings for treatment, nursing and interaction. We are essentially seeking to establish whether the mean values in the different cells differ only because of random sampling errors, or because either the drug or nursing had effects.

We calculate the residual variance by first looking at the variance of the four readings from the mean in each cell. If these eight cell variances are approximately equal, we can pool them into an overall residual variance, giving a value of 140 in our case. (If the cell variances differ markedly, the analysis becomes more complex.) Thus the pooled residual variance is an estimate of σ^2, the variance of the individual patients' readings about the population mean of 62.

We next look at the variances of various sub-group means. Starting with the treatment itself, we have four levels of dosage and therefore four mean values in Table 19.3. The variance of this set of mean values is

$$\{(80-62)^2+(65-62)^2+(52-62)^2+(51-62)^2\}/3 = 184.7.$$

On the null hypothesis that the treatment has had no effect, this figure should be an estimate of the sampling variance of the means in each cell. It should therefore equal σ^2 divided by the sample size of each mean, i.e. σ^2/n. Since in our case $n = 2 \times 4 = 8$, the treatment variance multiplied by 8, i.e. $184.7 \times 8 = 1{,}477$, should give an estimate of σ^2 just as the pooled residual variance of 140 did.

The two values in our case, 1,477 and 140, clearly differ markedly. Whether the difference is real (reflecting a treatment effect) or probably only due to random sampling errors in allocating particular patients to the different types of treatment is what one seeks to determine by calculating the ratio of the two variance estimates, $1{,}477/140 = 10.6$.

If the null hypothesis is correct and the treatment has not been effective, this "F-ratio" should follow an F-distribution with 3 and 24 degrees of freedom. (The degrees of freedom reflect the number of independent comparisons in the data. With four treatment levels there are only three such comparisons, one less than the number of readings. There are $8 \times 3 = 24$ degrees of freedom for the residual variance.)

Referring to tables of the appropriate F-distribution we see that an F-ratio of 10.6 is statistically significant. Thus we must reject the null hypothesis and conclude that the treatment has been effective; the *apparent* differences in the last row of Table 19.3 reflect something real.

The same steps can be carried out for the nursing effect and for the interaction between different dosage levels and different nursing levels. (The interaction should normally be tested first.) We will find that neither is statistically significant. This indicates that the large difference between the two nursing levels in the placebo group was due to random allocation of patients to the different design cells. The implication is that a similar difference would occur by chance in most experiments like this with only 32 readings (though not always in the placebo group).

Once we have established that the treatment effect is significant, we still have to estimate and interpret the effects in detail. We tested the treatment on an overall basis. The next step is to determine which specific comparisons matter, i.e. the low level of the drug versus no treatment, the two dosage levels against each other, and so on. As mentioned in Section 18.6, selecting the largest differences by inspection and then testing them for

statistical significance leads to technical difficulties. But in our example prior expectations provide specific hypotheses; i.e. whether the treatment is *generally* effective, whether the higher dosage is *more* effective, and whether the placebo has an effect compared with no treatment. Such prior hypothesis can then be tested for statistical significance by standard t-tests or the like.

Exercise 19H. HowManyFactorsinaFactorialDesign?

Could the clinical trial discussed in the two preceding exercises have had more than two factors (treatment and nursing)?,

Discussion.

A fully balanced experimental design needs at least one reading for every possible combination of factors. Thus with a total of 32 patients we could have had two additional factors of two levels each, e.g.

four dosage levels = 4
two nursing levels = 2
younger vs. older patients = 2
severe vs. mild cases = 2

This would lead to a 4 x 2 x 2 x 2 design with 32 test combinations.

By exercising more control on the allocation of different patients to each cell (e.g. two old, two young; two mild cases, two severe cases), the experimenter derives more information from the one study. Normally it is not advisable for *every* factor in an experiment to be expected to produce a striking effect (or at least not a striking interaction with other factors). This would make the results too complex to interpret. Instead it is generally better to include some factors which one merely wants to establish as having *no* effect,

Such fully utilised controlled experiments are obviously beneficial in gaining greater information and saving time, effort and money. In practice they are, however, often difficult to organise.

Exercise 191. LatinSquareDesigns

What is lost if there is less than one observation for each combination of factors?

Discussion,

Consider the following "Latin Square" design in an agricultural trial, using nine plots of land.

		Fertiliser level 10 lbs	5 lbs	0
Fertility of Plot	High	A	B	C
	Med	B	C	A
	Low	C	A	B

A certain fertiliser is applied at 10 lbs, 5 lbs, and zero rates to plots classified as being of high, medium and low fertility. Three different types of seed, A, B, and C, are also used.

Thus we have three factors, each at 3 levels, making a total of 27 possible combinations, but only 9 plots of land. All possible combinations cannot be measured, e.g. there is no reading for 10 lbs of fertiliser applied to seed B in a high-fertility plot.

However, the average reading for the 10 lbs fertiliser level comes from plots of high, medium, and low fertility using all three seeds, A, B, and C. In the same way, every level of every factor is balanced on all the levels of the other factors. For example, the three *low*-fertility plots are given seeds A, B, and C, and the three fertiliser levels. It is therefore possible to make balanced comparisons of the "main effects", e.g. the effect of fertiliser level as an average across the different plot levels and seeds.

The limitation of the Latin Square design is that it is generally not possible to establish *interactions*. For example, we cannot tell whether 10 and 5 lbs of fertiliser have different effects on high- and medium-fertility plots because different types of seed were used, A and B in one case, and B and C in the other.

This type of experimental design is therefore useful mainly when previous work has shown that such interactions are unlikely, or when one is merely "fishing" to see what *major* effects might occur.

The "Graeco-Latin" Square below is a still more ambitious design.

	10 lbs	5 lbs	0
High	$A\alpha$	$B\beta$	$C\gamma$
Medium	$B\gamma$	$C\alpha$	$A\beta$
Low	$C\beta$	$A\gamma$	$B\alpha$

Here a fourth factor, the use of different fungicidal "dressings" of the seed, say, has been introduced at three levels α, β, and γ. Each level of each factor is still fully balanced against every level of the other factors. Each column and row of the design has one α, one β, and one γ, and one A, one B, and one C.

The names of these two designs stem from the Roman and Greek letters used in them. They exemplify more advanced cases in the statistical theory of experimental design.

Exercise **19J.** The **Important** Factors

Why do scientists so often concentrate on "academic" questions instead of practical problems?

Discussion.

One reason is that before any problem can be solved it is first necessary to establish the effect of the *major* factors (the ones that most affect the variable in question) even if these factors are uninteresting from the practical point-of-view. For example, we may be interested in the effects of race on children's stature and growth, but first we have to establish the

effects of *age*. If we do not know how age affects stature, we will not reach any valid conclusions about the relatively minor differences (if any) related to race.

Exercise 19K. TheChoiceoftheResearchDesign

There has been a great deal of discussion about the association between smoking and lung cancer. The evidence has largely been based on surveys which examined smoking and lung cancer in human populations.

Discuss the prccedures which may be used in such surveys to give rise to the strong presumption of a link. (Adapted from a specimen "Question-and-Answer" prepared for the Market Research Society by Mr. Colin Greenhalgh.)

Discussion.

One possible procedure is as follows:

(a) A representative sample of the young adult (or even child) population is recruited. This sample should be large enough for small differences in a rare characteristic, the contraction of lung cancer, to be significantly demonstrated between smokers and non-smokers, probably within sub-groups thought to be relevant (e.g. by sex, occupation, parents' smoking and health history, etc.).

(b) The smoking habits of this sample are recorded over time (e.g. by self-completion diaries or by regular personal interviews).

(c) Their medical history over time is also recorded; specifically, of course, the apparent cause of death for any who are unfortunate enough to die in the course of the survey. Alternatively, the survey could be continued until *all* the informants are dead and age and cause of death are recorded.

(d) A simple breakdown analysis is tabulated of mortality rates, cause of death and/or ailments suffered, between smokers and non-smokers. If the incidence of lung cancer among the smokers is shown to be higher than among the non-smokers, then this is taken as strong evidence to support the suggested association.

Such a simple-minded analysis can be refined not only by *type* of smoking (filter or non-filter cigarettes, cigars, pipe, etc.) but also by *mount* of smoking.

However, this kind of a survey would require a long run of continuous data and take many years before any conclusive association became apparent. Instead, a survey could base its evidence on (probably less accurate) data of past smoking behaviour collected by an interview with each recruited informant.

Yet no matter how well designed and conducted, such surveys would always leave some residual doubt about what was cause and what effect. The cause-and-effect could be the opposite of what is hypothesised; a propensity to contract lung cancer may actually cause a craving for cigarettes. Or the two characteristics may be independently associated with a third, possibly unknown, characteristic. For instance, a certain psychological trait may (i) tend to make its possessors want to smoke *and* (ii) tend to induce cancer. In both these cases one could remove the smoking

(by persuasion or even compulsion) and not affect the propensity to contract cancer. The association found in the survey would be true, but *irrelevant.*

Other Methods.

Some of the minor shortcomings in such a survey method can be removed by refinements in the analysis. For instance, the sub-samples of smokers and non-smokers can be checked on other characteristics which *might* be influencing the apparent association : e.g. more city dwellers than rural dwellers might be sufferers and it might be hypothesised that the general atmospheric pollution in the cities was causing the lung cancer rather than the smoking, or there might be a difference in inherited propensity to lung cancer. (Sometimes such analyses are made by *post-weighting* the samples of smokers and non-smokers to "match" on rural/urban and parental background factors, etc. But post-weighting of the aggregate results is pointless unless the analyst has established by analysing the separate sub-groups that the factors matched *do* have any effect, and if so, what it is.)

This procedure can be followed for any characteristics where smokers and non-smokers are shown to differ. (It may be necessary to collect further descriptive data about the informants in order to explore all the suggested hypotheses about possible differences.) These other characteristics can then slowly be eliminated from the association, one by one.

Many doubts about cause-and-effect would be eliminated by organising a controlled *experiment* in which two random sub-samples of the human population (preferably after stratification by other relevant characteristics) were artificially induced to smoke and not to smoke regardless of their "natural" behaviour.

Such a procedure would virtually eliminate all the disadvantages of a typical survey as described since (a) the two sub-samples *must* be matched within the terms in which they were stratified, (b) it is highly probable (and can be tested by replication) that they are matched in other unstratified (and possibly unknown) characteristics, and (c) smoking is the one characteristic for which the two sub-samples were definitely and completely unmatched. Therefore it must be this difference in smoking which *either directly or indirectly* causes any difference in lung cancer incidence between the two sub-samples shown in the experiment (and replicated to establish a degree of generalisability).

One difficulty with such an experimental approach is inducing a human population to adopt a behavioural pattern they do not voluntarily wish to adopt. If the inducement is too artificial, it may affect the conclusions to be drawn from the experiment: "forced" smokers may not be equally prone to contract lung cancer as "natural" smokers. Apart from that, it would in this case be considered improper to force, or even to encourage, a sample of the human population to adopt a form of behaviour that might be injurious to their health.

Therefore such experiments have been conducted on animals, e.g. rats. It is possible, and appears ethically acceptable, to induce animals to "smoke" under reasonably realistic conditions (e.g. by bringing them up in a "smoking-machine"). But such experiments are flawed by the fact that rats are not human beings, physiologically or psychologically, and therefore

do not necessarily react in the same way as human beings to any particular treatment. Controlled experiments on rats demonstrating that *their* smoking causes a propensity to contract lung cancer are strong corroborative evidence of the human survey evidence, but still do not prove the cause-and-effect relationship in humans. The main function of such animal experiments is in fact not to "prove" that smoking causes lung cancer, but rather to establish an increasing depth of understanding of the mechanisms involved.

Such understanding is desired partly for "applied:' reasons; it could lead to ways of reducing or eliminating any bad effects of smoking without necessarily stopping people from smoking altogether. A more important "basic" reason is that even fully controlled randomized experiments on *humans* would not prove that it is the sheer mechanical act of smoking that causes lung cancer. The cause could be "third factors" which happen to be associated with the act of smoking, e.g. frequent movements with one's hand, the effect of touching cigarette packets, frequent visits to tobacconists, or the psychosomatic fear that smoking will cause lung cancer. Thus depth of understanding, rather than an apparently simple one-to-one relationship, is required.

CHAPTER 20

Description and Explanation

This book has concentrated on the description of data. But the purpose of description is to gain understanding, to find out how and why things happen and hence also to gain some *control* over the phenomena in question.

We now discuss the link between description and explanation. We start by considering the nature of descriptive relationships. For example, what sort of statement is it to say that y varies approximately as $(ax+b)$?

20.1 Descriptive Relationships

We have already noted that scientific laws are not universal. They merely describe how particular observed phenomena behave, with exceptions. For example, Boyle's Law, $PV=C$, describes how pressure and volume vary together under certain conditions. As already discussed, the law does not hold when the temperature changes, when there is condensation or a chemical reaction, etc.

Nor is a scientific law directly *causal.* The relationship $PV=$ C does not claim that pressure causes volume, or even that changes in one variable directly cause changes in the other. The law merely describes the values that Pressure P and Volume V take when some third factor, like the piston in Figure 20.1, is moved from position X to position Y.

If we infer more explanation from Boyle's Law it is because we are aware of a wider range of later knowledge about the movements of molecules inside gases, about Avogadro's Hypothesis, and so on. But a descriptive relationship between two or more variables is never a direct statement of causal effects.

One reason why scientific laws in themselves cannot be directly causal is that they are not exact. They are all deliberate oversimplifications and hence are not strictly true. Science aims to find one general&able equation that covers a wide range of different phenomena with some known degree of error rather than a great number of different specific equations that give a

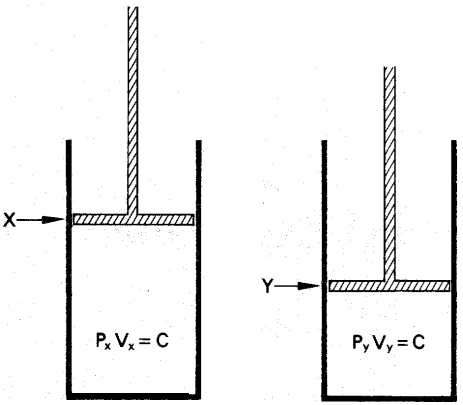


Figure 20.1 $PV = C$ for Positions X and Y of the Apparatus

closer descriptive fit to isolated sets of readings. The result is more useful, but it is also necessarily oversimplified.

We have followed this precept throughout this book. We have forced straight lines onto data which we knew were slightly non-linear, and have fitted Normal Distributions as simplifying approximations to data which could not conceivably be ***exactly*** Normal, and so on. It is the common process of science.

Generalisations are usually reached by excluding local complicating factors. This is not merely a case of smoothing away irregular errors, but may even be one of by-passing consistent biases, of very ***deliberate*** over-simplification and reformulation.

For example, the practical usefulness and intellectual glory of Newtonian mechanics is that the same laws cover vastly different phenomena: falling bodies, the swing of a pendulum, the motion of the stars, balls rolling down inclined planes, the movement of the tides, the trajectory of cannon balls, and so on. But it can do this only because the calculations generally ignore complicating factors like friction, air resistance, and the gravitational pull of other bodies. Balls always roll down. inclined planes more slowly than is stated in simple mechanics where friction is ignored (and where they would in any case not roll at all, but *slide*).

The size and direction of the errors has still to be established under all the relevant conditions, just like the basic relationships themselves. We have to learn when the theory approximates the data closely enough to ignore the errors, and what correction factors to use when it does not. Once an error is known it is no longer merely an error.

The simplifying approach also shows up in the fact that most well-developed laws in science have no numerical coefficients (other than l's and

certain "absolute constants" like π, e, the absolute zero of temperature $-273°C$, and the velocity of light c in physics). For example, Boyle's Law $PV = C$ also reads $P_X V_X = P_Y V_Y$, without any numerical coefficient. Again, the buyer behaviour relationship in Chapter 10, $w(1-b) = \text{constant}$, is free of coefficients in the form $w_X(1-b_X) = w_Y(1-b_Y)$.

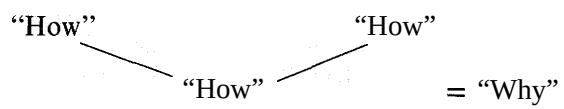
Numerical coefficients are avoided not merely because of the mathematician's delight in the elegance of form ("no coefficients !"), but because a law is only ready for use when it contains no coefficients whose numerical values still have to be established. This applies both to practical and theoretical uses.

20.2 Explanation

If scientific laws merely describe how certain observed phenomena behave, most of us still want to know *why* this behaviour occurs, We are more comfortable with a finding which we can understand than with one based on empirical fact alone: "It is all very well in practice, but how does it work in theory?"

Explanations often take time to develop. Newton was criticised for not having explained why bodies move. He responded to the effect that he had shown *how* they moved, and that should be enough for the moment. In fact he provided much understanding, e.g. by interrelating the movement of the tides and the swing of a pendulum. The same law holds for both. But even today, three hundred years later, there is still no accepted explanation of the *nature* of gravity.

In fact, explanations are often more superficial than is thought. A close look at explanations shows that they are really no more than descriptive linkings of how one thing varies with how another thing varies:



For example, Boyle's Law works because, roughly speaking, the smaller the volume of a given amount of gas, the more often the molecules of the gas hit the side of the containing vessel and create more pressure. The explanation is descriptive.

A typical explanation of a very simple kind is that "Ice floats because it is lighter than water". At first sight this seems little better than the virtual tautology that "Johnny is good at playing the piano because he has a talent for it". But the explanation actually accomplishes two things.

Firstly, it relates a very specific phenomenon-ice floating-to a larger generalisation which we already know about, the notion of specific gravity, that things generally float if they are lighter than water. Hence the phenomenon has been "explained", because it has been linked to something which

we know and accept. (Similarly, in Chapter 8 the cube-root formulation of the relationship between children's heights and weights was more comfortable than the logarithmic one. It linked up with, or was "explained" by, our more general knowledge that weight varies as volume and that volume is three-dimensional.)

Secondly, the explanation rules out all other mechanisms which might have accounted for ice floating. For example, it excludes the Aristotelian doctrine that objects float or sink essentially according to their *shape.*

Many explanations are deeper than this because they touch on the mechanics involved. An example is that "hydrogen burns in air because it combines with oxygen". But saying that hydrogen combines with oxygen is not a blow-by-blow causal explanation. None of us knows what actually happens when two atoms of hydrogen meet one atom of oxygen to make water. Similarly, we accept the molecular explanation of Boyle's Law even though we do not fully understand it.

We seek explanation and understanding. But this is never simple because there are always "third factors", and it is never complete because there are always "black boxes" that elude our grasp. And when we use laws and relationships for theoretical work and practical applications we do so only because they actually describe what happens. We know that if we press a light switch, the light will generally come on. We do not need to know more until the light does not work and we have to search for a cause: a blown fuse, a burnt-out bulb, a power-cut, or an unpaid bill? Even then we usually do not know all that much about the precise mechanism, there are still "third factors" and "black boxes". We may not know how fuses work or what actually happens when we pay our overdue bill. These things can function without our understanding them.

Explanations often do not even take the form one might have expected. There would have been nothing obvious to Sir Robert Boyle about the molecular explanation of his law. The mere idea that gases consist of molecules moving around at high speed was effectively unknown in his time. More parochially, consider the equation $w(1-b) =$ constant, that large brands are bought more frequently by their buyers than small brands are bought by theirs. The explanation might have been sought in terms of people% incomes, the sizes of their households, and their exposure to advertising, but the result actually follows from (or is "explained by") quite different considerations, as was outlined in Chapter 10.

20.3 From Facts to Theory and Back Again

The development of scientific knowledge depends on the interaction between facts and theory, Basically it has to start with facts. Fact-collecting

must of course be preceded by some hypothesis, as Popper (1959) has stressed. But initially this only needs to be some very loose, imprecise notion of which facts it might be interesting or relevant to collect and which to ignore. At the beginning these ideas are largely based on ignorance. When they turn out to be wrong it matters little except to the investigator.

Having found some low-level patterns or generalisations among the early facts, one tries to generalise them further and also to account for the exceptions that will arise. Interrelating the different findings leads to the beginning of *theory based on facts*. This in turn often suggests hypotheses, i.e. ideas about new ways of looking at the available facts or about new kinds of facts to collect. Hence we have the continuing interplay between fact and theory and back again.

Theory here is empirically based. But nowadays it has become rather fashionable to formulate behavioural or mathematical models *before* looking at the facts, These explanatory models are based on abstract ideas of how nature might work, using various intuitively reasonable assumptions. ("Reasonable" is a telling word here. It is generally used only when no reasons can in fact be given.)

There is nothing like a few facts to eliminate any number of speculative assumptions. For example, the old view was that the planets move in circles. This was logical enough since the circle is the most perfect geometrical shape, God is good, and therefore planets have to move in circles. Similarly, it is not difficult to think of reasons why black children have a different shape from white children, or why the leading brand of breakfast cereals appeals particularly to heavy users of the product. But just a few facts show these statements are simply not true.

Given the rehabilitation of speculative theorising in recent decades, it is not surprising that theory has earned itself a bad name, as something which does not work. Despite the extraordinary success of science in the last few hundred years, words like "theoretical", "academic", and "intellectual" are nowadays tinged with pity. But there is good theory and bad theory, good scientists and lesser ones.

Theory and applied mathematics are not supposed to be substitutes or precursors of empirical knowledge. Their function is to help model and explain the known facts, but the facts have to come first. Einstein is often quoted as the supposed counter-example to this. (People usually pick on dramatic extremes which they do not understand.) But Einstein's work concerned minor discrepancies in results that were based on centuries of detailed empirical observation and analysis (e.g. that light rays generally travel in straight lines, except, Einstein predicted, very fractionally near the sun). As he himself put it (Einstein, 1949): "... before a theory explaining a process can be tested that process must be known".

There is a straightforward test for the modern mind-over-matter tendency to build mathematical models and theories before any facts are known:

> "Take away the mathematical language and what generalised factual knowledge of the process in question remains? If the answer is none, the mathematical symbol for *that* is very simple."

The pay-off of Sonking, the Scientification of Non-Knowledge, is only in the model-builder's neo-Cartesian self-regard: "I sonk, therefore I am". The question is not what the theoretician thinks, but what he *knows*.

Science means knowledge. But particular subject matters, like physics or psychology or economics, may be in very different states of development. Three main stages of knowledge can usefully be identified, depending on the existence of a generally accepted conceptual frame-work or "paradigm" (Kuhn, 1970).

First there are the initial fishing-expeditions: trying to find some facts, some patterns, and some generalisable notions. Subject areas at this early pre-paradigmic stage of development usually lack stable foundations of knowledge. They are often distinguished by intense concern with techniques and apparently endless arguments over methodology. But this is unnecessary: this is where the low-level procedures discussed in this book mostly apply, the notion of empirical generalisation, of seeing whether the results in one study agree with all the previous ones, an approach which is low-key but nonetheless exacting.

The second stage of development is "normal science", firmly based on acknowledged past achievements where its practitioners share some general theoretical viewpoint. The greater part of scientific effort takes place at this level: filling in gaps, linking up pieces, puzzle-solving. This work can be very exciting, like atomic physics in the 1920's and 30's or the discovery of the structure of DNA.

The third stage is the occasional revolution, where some accepted scientific view-of-the-world or paradigm is turned upside down. For aeons people thought the sun moved round the earth. That was the way it looked, a view which could not be changed by any directly discernible evidence (Wittgenstein's "What would it have looked like if it had looked as if the earth rotated?"). That revolution had to wait for Copernicus' creative insight and nerve.

These latter qualities are also needed at the level of this book. It takes more insight and courage to report an average or an abstract relationship than merely to present all the facts.

20.4 **Summary**

A scientific law is a theoretical abstraction, but inherently it still remains only a descriptive generalisation of the facts. Understanding and explanation are continuous processes that grow as different laws are interrelated.

CHAPTER 20 EXERCISES

Exercise20A. The Main Conclusion

What is the main point this book makes about the analysis of data?

Discussion.

That data need to be *summarised.*

The criteria of a *good* summary are that it be

(i) succinct,

(ii) complete (i.e. that the original data could be reconstructed, within the stated or implied limits of approximation),

(iii) usable (e.g. that the results can readily be used when analysing further data),

Exercise 20B. StatisticiansasTheir OwnCustomers

To what extent do statisticians in fact use previous results when analysing new data?

Discussion.

The statistical literature contains little explicit discussion of the use of previous results in analysing new data. (An apparent exception is the Bayesian approach to statistics mentioned in Chapter 13, but this is not widely practised.)

A small-scale check of statisticians' use of previous results was therefore carried out a few years ago (Ehrenberg, 1969). A number of eminent past presidents of the Royal Statistical Society and editors of its Journal were asked whether they or some nominee could prepare a methodological paper or case-history about using the results of previous statistical analyses for a statistical conference on "Consumer Satisfaction". The question was whether statisticans were satisfied customers of their own results.

The approach was unsuccessful, as is illustrated by the following replies (quoted with permission):

(a) "Customer Satisfaction in any sense of the words is not really me, I think. Although the idea you suggest is an interesting one, I am not really competent to speak on it."

(b) "The proposal to have a session of the kind you describe seems to me an excellent one. However, I feel it essential that speakers should at some time have been in a position where quantitative prior data were available for use in the analysis of new material. As I have never really had such experience, though often felt the lack of it, I am afraid that I could not make any useful contribution to the discussion,"

(c) "No, not I; nor yet any other. The job is not one for a statistician *per se,* but for a science-historian."

(d) "I have put out a few feelers in the department [in agricultural statistics] but nobody seems to have a case-history which they think would make a sufficiently interesting paper. I wonder whether the accumulation of prior information would be more regular in an *industrial* environment ['The grass on the other side of the fence. . .']. I agree that it would be a

fascinating topic with the right case-history to initiate the discussion."

(e) "Incorporating previous data? This just hasn't come my way."

(f) "I think the proposal you make is a very interesting one, but I shall not be able to take it on as this is clearly one which would require a good deal of thought and preparation."

(g) "Participating in the conference at Sheffield will not, I am afraid, be possible. I have nothing useful to conribute to this subject."

(h) "We use other people's *conclusions*, but these are qualitative results. We don't use other people's *figures* very much, if at all, in the sense of other people's estimates of parameters. We don't combine our statistical test results with theirs either, in a numerical way. The reason is I think that we are never sure that parameters have not varied or, more generally, that our statistical set-up is strictly comparable with that which went before. There is also the question of *authority*. I don't think I would trust more than 20% of statisticians that I know not to miss the obvious. Too many are so blinded by their theory that they don't look at the data. And as for the statisticians I don't know... "

And so on and so forth. Apparently one must never use previous results because things *might* have been different (so one will never find out whether they were or not). Statisticians with 20 or 50 years' experience claim never to have used previous results! This is wrong. Why *produce* statistical results, if even the statistician himself never expects to use them?

Exercise 20C. Description Without Explanation

Can relationships which are merely descriptive and have not been explained be valid and useful?

Discussion.

The history of science shows that explanations often come long after a result has been descriptively well-established, e.g. the gap of more than one hundred years between Boyle's Law and Avogadro's molecular explanation.

Farmers have used fertilisers since time immemorial without knowing how they work. Are they absorbed into the plant (and then what happens?) or do they stimulate soil bacteria that break down inorganic nitrogen compounds; or do they work in some other way completely? Agricultural scientists and manufacturers now know more about fertilisers, but they still do not understand many of the detailed mechanisms.

Exercise 20D. Assumptions Without Justification

Is it proper to make an assumption without direct justification or explanation?

Discussion.

Except in speculative theorising, assumptions must be founded in fact. However, the empirical justification does not have to be direct, and it may come much later. Nor does the assumption have to be fully understood.

As long as the assumption links up *other* empirical results, it performs a useful descriptive function. (A typical example of this process was provided by the justification of the Poisson-Gamma theory which underlies the NBD model of repeat-buying, as outlined in Exercise 13P.)

Exercise 20E. Testing an Explanatory Theory

To what extent do theories have to be tested against the facts?

Discussion.

Many theories develop organically by piecing together empirically well-established but low-level relationships. The basic results therefore do not require testing, as they are already empirically established. But the process of linking different results together and the development of explanations involves working assumptions and hypotheses that do require testing.

In contrast with such theories which start with a basis of factual generalisation there are the speculative -theories that are developed as possible explanations or models of certain phenomena before it is empirically known just how these phenomena actually behave and what there is to explain. Most writers suggest that such models need to be tested against the facts, but they frequently miss the right emphasis, as in the *Principle of Empirical Viability* (Montgomery and Urban, 1969, p. 89).

> "It seems reasonable to require that a model be demonstrated to be empirically viable for at least one set of data. That is, the model should be consistent with (fit) at least one set of data."

Saying something twice does not make it any more true, but requiring a model to hold for *two* sets of different empirical data would at least serve to establish whether it could in fact generalise. (Many theories can hold for one selected set of data, but are they known to apply to a sufficiently wide range of circumstances to be of any practical or academic interest?) The real problem is that the intending model builder here did not know anything by way of empirical generalisations before he started building the model. All he intends to do is to "test it" against some new and undigested data.

Exercise 20F. The Timing of an Explanation

Could explanations be arrived at earlier than they are?

Discussion.

In principle the explanation of some new result could often be reached earlier, but in practice it takes time for the new result to become familiar enough for its explanation to be "seen".

Often an explanation requires additional kinds of findings, and then a delay is inevitable. (This happened with both the illustrations referred to in Exercises 20C and D.) Sometimes one can speculate theoretically on the new kinds of findings that might be needed to explain the given result, and this may speed the process. However, creatively guessing at new explanatory factors is often difficult. Part of the problem is that empirical models are not *exact*. Possible connections cannot therefore be deduced by logical or mathematical arguments alone. Seeing the nature of the necessary approximation in the argument usually requires a major imaginative jump.

Consider a simple result, e.g. that a particular plant in the garden has died. One looks for an explanation such as lack of water, too much heat, certain pests or disease, mere age, or whether someone trod on it. But if none of these appear to apply, some new kind of explanatory factor has to be found. To do so requires creative insight as well as technical knowledge.

Exercise 20G. Knowing the Causal Direction

The precise form of certain descriptive analyses is sometimes said to depend on the causal direction. What is the justification for this?

Discussion.

An example occurs in regression analysis (Chapter 14). To apply this technique the analyst has to choose between the two possible lines, y on x or x on y. To do this he often makes a prior assumption about the direction of the "causation", e.g. that x influences y and not the other way round.

But he is then claiming that he knows the causal direction even though he does not know the coefficients in the equation, whether the equation is really linear, or even whether it exists. (He usually puts much effort into testing the null hypothesis of zero correlation!) This seems untenable, an analyst's half-baked ideas about causal connections determining the results of the analysis. (Not surprisingly, therefore, this approach does not seem to have led to any lasting results.)

Exercise 20H. Correlation Is Not Causation

It is commonly said that just because two variables are correlated, this does not necessarily mean that there is causation.

Can you think of an example to the contrary, where a correlation between two variables *does* mean that one variable directly causes the other?

Discussion.

No. None of the laws of science seem in themselves to tell us directly about what causes what. At best there always seem to be "black boxes" or "third factors" involved.

The medical drug Thalidomide is known to be related to the incidence of malformed babies. An obvious explanation is that the drug causes the deformities.

But it is also known (i) that a proportion of human foetuses are naturally malformed; (ii) that many of these are rejected by the mother's body during the early stages of pregnancy; and (iii) that Thalidomide helps to suppress the body's rejection of strange or foreign tissues. We could therefore form the alternative hypothesis that the correlation is due to a *negative* effect, that Thalidomide prevents the natural abortion of naturally deformed babies.

These different hypotheses about the correlation cannot be elucidated in any simple manner, but only by interrelating the descriptive results of many different types of studies (e.g. the *kinds* of deformities observed under different conditions, the length of time for which the drug is taken, etc.).

Exercise 201. TheCausalDirection

Is it possible to establish the causal direction between two variables?

Discussion.

It is generally very difficult to make a precise, watertight statement about the causal connection between any variables occurring in a quantitative scientific law. An electric current running along a wire does not cause resistance, nor vice versa. The current does not even "cause" a difference in voltage; if anything, we might think of a difference in voltage causing the current to flow, but it is actually due to a third factor or black box mechanism (e.g. a battery or dynamo), and this does not appear in the equation at all.

Among logicians and philosophers, the notion of causation is at best controversial. Among scientists it is hardly used, except as a loose and superficial shorthand, such as that a charged battery, or a dynamo when driven and suitably connected, will cause a voltage gradient to exist and a current to flow. But this is not a causal interpretation of the explicit relationship between the observed variables R, V, and I in Ohm's Law $v = RI$.

For another example, it is often thought that people's expenditure depends on their income. You have to *have* money in order to spend it. But in fact people's earnings are often influenced by what they need to spend. And some people spend more than they earn, thus ruling out any *simple* cause-and-effect relationship.

As a further example, it is widely thought that price determines the volume of sales. But often it is also the other way round, prices for high-volume goods can sometimes be reduced because of economies of scale.

Again, it might be thought that a product's sales level is influenced by the amount spent on advertising. But in practice advertising budgets are frequently set as a proportion of sales income, and advertising is generally the first item of expenditure cut if sales drop. The causal direction is not simple and unambiguous_

In driving a motor-car, it might seem that turning the steering-wheel causes the road-wheels to turn. But if the car is running in a rut, the opposite occurs. If there is a bend in the road, then the road-wheels need to follow it and this usually causes the steering-wheel to be turned. (In any case,

few drivers understand the steering mechanism in their cars, rack and pinion, etc., or what happens to the causal chain when the steering fails.)

Nevertheless, it can be helpful in everyday terms to say that a current flows if there is a difference in voltage or that, by and large, a car is directed by turning the steering-wheel. At least this describes what generally happens!

Exercise 20J. Spurious Relationships

How can one distinguish spurious relationships from the proper laws of science?

Discussion.

It is common to deride so-called "spurious relationships", like the correlation between the number of pigs and the production of pig-iron over the years.

But the association between pigs and pig-iron is no less real than that between the pressure and volume of a gas. It has happened over and over again, and typifies a growth-pattern that is desired by almost every developing nation ("more of everything"). The relationship seems humorous partly because we know of circumstances where it does not work: e.g. that nothing much happens to the number of pigs if the government interferes with the production of pig-iron, or if there is a strike of foundry workers. But the same is true of Boyle's Law; there are circumstances where pressure and volume are not related.

Nor is the association between pigs and pig-iron merely a blind correlation. We actually know a good deal about the underlying mechanism, that the growth in the number of pigs and in pig-iron production is related to the growth in the population and the growth in productivity. The association is therefore due to "third factors". So is that in Boyle's Law, the piston in Figure 20.1 and what the experimenter does to it.

It is easy to think of examples of relationships which are spurious, and difficult or impossible to think of one which is not. In this sense then, all lawlike statements are "spurious". They are in themselves merely descriptive generalisations.

Exercise 20L. Empirical Generalisations

Few statistical texts refer to the idea of empirical generalisation, let alone emphasise it. Why all the fuss about it in this book?

Discussion.

A result which only holds for one set of data remains an isolated historical event.