PART IV: SAMPLING

A sample is a selection from a given set or "population" of items. The purpose of sampling is generally to save time, effort, and money by dealing with a sub-group instead of the whole population.

To be useful, a sample must be more or less representative of the population from which it was selected. There is always *some* loss in accuracy by dealing with a sample, but this may be acceptable if the amount of error is known.

A major contribution of modern statistics has been the theory and practice of random sampling. Procedures of taking a random sample are outlined in Chapter 16. The major theoretical concept is the *sampling distribution.* This describes how different samples vary, as discussed in Chapter 17. Making a statistical inference from a specific sample to the population is discussed in Chapter 18. In general, the larger the sample the smaller the possible sampling errors. Problems of statistical inference therefore matter most when dealing with small samples.

CHAPTER 16

Taking a Sample

In this chapter we discuss the physical process of selecting a sample from a population. The detailed work is not necessarily difficult, but tends to be carried out either by specialists or by people with previous experience. However, the general reader needs some appreciation of sampling operations in order to understand the nature of the results.

16.1 The Purpose of Sampling

In Part III we discussed the incidence of boys and girls in different families. The results were based on a total of 5,017,632 children: all the legitimate births recorded in Saxony in the years 1876–85 (Geissler, 1889). It seems obvious that virtually the same conclusions would have been obtained from a tenth of the children, still half a million readings, or even from less, if the right kind of sample had been taken.

A good sample of a few thousand is accurate enough for almost any purpose, and in many cases a few hundred or less will do. Some studies involve much larger numbers, and complete "censuses" of every item are still not uncommon. This is often because adequate information is required about many separate sub-groups of the total population, and so the total number of readings builds up.

Little is lost by taking a sample and cutting down the time, effort, and money required to collect and analyse extensive data. Another advantage is that more care can usually be devoted to the individual measurements in a small sample study. Sometimes sampling is more than a convenience, as when the act of measurement seriously interferes with the object being measured, e.g. in testing matches to see that they will burn. Sampling then becomes essential.

But just any sub-set of the population will not do. A sample must be more or less representative of the population being studied. But even with the best forms of sampling some degree of accuracy is usually lost. Thus sampling is largely a matter of economics, balancing the size and cost-implications of

the possible sampling errors against the costs of full data collection and analysis.

The statistically safe way of selecting a sample is to pick items "at random". But in many situations no one would dream of "putting all the items in a hat, shuffling them, and pulling out the required number". Most sampling situations are not statistical in this sense at all.

For example, we usually just take a sip from a cup of tea to see if it is hot. A doctor takes a blood sample by deliberately selecting a vein near the surface. In seeing whether a goose is cooked, we usually check with a small cut in one of its legs or its breast. In starting up a new barrel of beer, the barman runs off a couple of pints and then checks the *third* pint to see that the beer is not cloudy.

In none of these cases is the sample "statistically" representative. Yet the information remains valid. This is because we have a great deal of prior knowledge and understanding of the situation. We either know that the material is more or less homogeneous, so that it does not matter greatly which sample item one selects, or we know how different parts of the system are related.

Statistical sampling is required when there is *substantial and unpredictable variability*. An example is measuring the proportion of households which own freezers. We know that some households own one and some do not: there cannot, in a sense, be greater heterogeneity than that. We also know that the better-off are more likely to own one, but otherwise we are rather ignorant about who owns what. This is the kind of situation where statistical sampling is usually required to obtain a representative result.

16.2 **Types of Statistical Sampling**

A sample is a selection of objects or measurements taken from a specific *population* of such items. The aim is to make the results from the sample tell us more or less what we would have found by measuring the whole population. There are various ways of selecting a sample, but only with random or probability sampling is it possible to know how representative the sample results are likely to be.

Two popular forms of selecting a sample are "convenience sampling", e.g. selecting the first 100 on a list because that is easiest, and "plausibility sampling", e.g. selecting the middle child in each family because it should be "average". Both types of procedure can give consistently wrong results. Furthermore, neither the nature nor the size of the sampling errors are usually known. These methods are not acceptable as possible forms of *statistical* sampling, i.e. selection procedures which on their own provide representative results.

An alternative is to select a sample *systematically, e.g.* by going down a list and taking every other or every tenth item or whatever, depending on how large a sample is needed. This might appear to provide a good cross-section of the population, but it can still lead to biased results.

For example, if we aim to sample half the houses in a street, we could systematically select every other house. This might result in all the houses with even numbers. These could all be on one side of the street and would generally have their front doors facing in the same direction; they could also be the older and larger houses, say. The results could well be very different from the other houses in the street, and hence be unrepresentative. Furthermore, we would not generally know how big a bias and even what *kind* of bias is involved. We would therefore not know how well, or how badly, the sample represents the particular population, here the whole street.

However, there are many situations where systematic sampling is used successfully. This is when enough is already known about the population to be sampled to suppose that it has no obvious or dramatic regularities. Certain forms of "purposive" sampling can also give representative results. An example is "quota sampling" in social and market surveys involving human populations. Here each interviewer is set certain quotas of different types of people to interview (e.g. five white-collar workers aged 45 to 65, three non-working housewives aged under 35 with husbands in a manual occupation, etc.). Each interviewer personally selects the individuals within each "quota". This is where bias can enter despite additional rules and regulations; e.g. not more than two persons to be selected in one street, and the interviews restricted to specified districts (usually selected on a random or probability basis).

It has been shown empirically, by checks against known data, that well-controlled quota sampling gives representative results for quite a wide range of topics (e.g. certain forms of mass-behaviour which are not closely related to those demographic factors where quota samples might still be biased, such as for people in particular occupations). Quota sampling is generally cheaper than the main alternative, random sampling. But the crucial point is that quota sampling can only carry much conviction in those limited situations where previous experience has shown it to work.

Random Sampling

Sample bias is generally avoided by using a random or probabilistic procedure for selecting the sample. This is a method that works whatever the nature of the population sampled. Leaving things strictly to chance involves very precise procedures, it is not to be confused with being haphazard. It certainly does not eliminate sampling errors, but their likelihood can be calculated theoretically. An error which is *known* is no longer simply

an error, it can be allowed for. In particular, there are many situations in which random sampling eliminates the risk of any consistent bias in the results.

Suppose again that we have to sample half the houses in a certain street and that 60% of the houses have a garage. We want to avoid samples which are heavily biased in this or any other respect. This can be done by tossing an unbiased coin for each house in the population and selecting it for the sample if the coin comes down heads. There is still a chance that ***all*** the houses in the sample will be ones with a garage, but the probability will be small.

The chance that the first house picked will have a garage is .6 or 60%. The chance that the next house selected for the sample will have a garage is also .6. But this selection is independent of the first (independent tosses of a coin), so that the chance of *both* houses having a garage is only .6 x .6 = .36. The chance of three houses all having a garage is about .2. The probability that a sample of n houses would be made up entirely of those with a garage would be $.6^n$ (the first term in a Binomial Distribution). Less than one in every 100 possible random samples of 10 houses would consist only of houses with garages. The probability would decrease to less than 1 sample in a million (.0001%) for samples of 30 houses. Similar results apply to the chances of a sample being greatly atypical in any other respect,

We have here three basic properties of random sampling:

(i) the chance of an unrepresentative sample is relatively small;
(ii) this chance decreases as the size of the sample increases;
(iii) the chance can be calculated.

16.3 Random or Probability **Sampling**

Selection of a random sample proceeds with something like the following sequence of steps:

(i) all members of the specific population to be sampled are individually identified (e.g. on a list);
(ii) each item is numbered or otherwise coded;
(iii) the numbers are put on separate slips of paper;
(iv) these slips are put in a hat and thoroughly shuffled;
(v) the required number of slips is then selected blindly.

This identifies the members of the population who are to make up the sample.

The crucial step in this sequence is shuffling all the slips in the hat. This aims to produce a completely irregular or effectively random mixture. All the slips should have the same probability of being selected, and it should not matter how the slips are picked.

But, in practice, when selecting slips from a hat, most people do not simply select one after the other in a systematic manner. We usually follow some

irregular practice, to be fair, like one from near the top, then one from near the bottom, and so on. We do this because we do not completely trust the shuffling of the slips. We realise that complete irregularity or randomness is difficult to achieve and add an additional stage of irregular selection.

The concept of randomness is in fact a *theoretical* one. There is no physical process which is exactly random, or if it were, we would have no means of telling. But as with other forms of applied mathematics, certain observable phenomena can be modelled by the theoretical concept *to a close degree of approximation.*

Selections from numbered slips shuffled in a hat can be tested for systematic patterns If no noticeable ones emerge, the process can be regarded *as if* random. But in practice, such shuffling is seldom very good (slips close together when put into the hat tend somewhat to stay together). In any case, this particular procedure is clumsy when large samples are required, as is tossing a coin.

Therefore most random samples are drawn using specially prepared tables of so-called random numbers or computerised procedures for producing random number sequences. Tables of "random numbers" are prepared from processes where experience has shown that the results appear predictably irregular (e.g. certain kinds of electronic phenomena, or the results of certain mathematical calculations, like the last digits in successive square roots of a given number). But none of these processes are "really" random, and all have to be tested empirically to establish that they adequately approximate *theoretical* randomness.

The reason for all this paraphernalia is that if the physical selection process can be successfully approximated by the theoretical concepts of randomness, then the theory of probability can be used to make useful inferences about the results (as will be discussed in Chapters 17 and 18).

14.4 **Technicalities of Random Sampling**

In practice, random samples are usually selected using a published table of random numbers (e.g. Fisher and Yates, 1957; Lindley and Miller,1966). Table 16.1 illustrates an extract from such a table.

TABLE 16.1 A Small Extract of "Random Numbers"

41	03	59	24	78	54	14	48	27	05
53	26	08	33	10	98	62	46	16	94
96	17	25	92	41	17	55	13	73	59
43	61	20	39	65	62	18	15	70	66
65	04	96	78	37	13	98	90	62	28

Starting at any haphazardly chosen point, one reads off successive numbers in some direction, horizontally or vertically, until n numbers are accumulated for a sample of size n. For example, suppose we want to select a sample of 5 from a list of 400 items coded in 3-digit figures running from 20 (i.e. 020 to 419). We might start with the 08 in the third column and second row, and read off groups of three digits, going from left to right. This gives

083 310 (986) 246 169 (496) 172.

The numbers in brackets are simply ignored because there is no corresponding item in the population.

Multi-stage Sampling. Various refinements to this kind of simple random sampling are used in many situations. For example, in sampling human populations it is common first to select a sample of towns (or districts in towns), and then to select individuals in the chosen towns. This is *multi-stage* sampling: first towns, then individuals. It avoids having to use a list of all individuals in the population. Instead, only a list of all towns is needed, and then lists of all individuals in the selected towns. Random sampling would be used at each stage, i.e. in selecting towns as well as individuals.

Cluster Sampling. If more than one individual is sampled per town, we have a form of *cluster* sampling, each town providing a "cluster" of individuals. Cluster sampling is often less expensive than simple random sampling. For example, in surveys involving personal interviews it is often cheaper to interview several people in the same district. But in general, cluster sampling is somewhat less efficient *statistically.* The chances of sampling errors are not reduced as much as the sheer size of the sample might imply, because people in the same district may tend to resemble each other, The cost advantages therefore have to be balanced against the estimated loss of statistical accuracy.

Stratified Sampling. The population may in the first instance be divided into sub-groups or "strata" and an appropriate number sampled at random from each stratum. For example, instead of selecting a simple random sample of 10 children from a population where it is known that boys and girls occur in a 50 : 50 mixture, two samples of exactly 5 boys and 5 girls may be randomly selected. With simple random sampling, only 25 % of all possible samples of 10 would be split exactly into 5 boys and 5 girls. Stratified sampling ensures that every sample is strictly in the right proportion of boys and girls. The representation of any factor related to sex would also be improved,

But to select stratified samples one needs prior information on the appropriate split of the population. It must also be physically possible to assign each member of the population to his stratum before the sample is selected. In stratified sampling our prior knowledge about the population can be used

without risk of bias, unlike the earlier case of judgment sampling. Returning to the example of sampling houses in a street, we can group all the houses by sides of the street and by whether they are corner houses, and then sample each stratum separately.

Weighting. If an unstratified sample of 10 has given 6 boys and 4 girls and we *know* that boys and girls are 50: 50 in the population, the results in each stratum can be "weighted" to bring the sample into line with the population proportion (e.g. by multiplying all the girls' readings by 1.5). This kind of "posterior" stratification is usually less effective than prior stratification. The weighted portion of the sample has an undue effect on sampling errors (e.g. an untypical girl would count for 50% more than an untypical boy).

Weighting of sample results is also widely used when the sample is out of proportion for other reasons than the errors inherent in random sampling. A sample of adults may have too low a proportion of very young and very old people, who for different reasons are difficult to interview in a survey. In such cases the beneficial effect of corrective weighting on sampling errors is often less striking than seems to be thought. Furthermore, while the weighted sample will have the right proportion of old and young people it will not necessarily be representative of the right *kind* of old or young people.

Probability Sampling. The situation where each item in the population is given the same probability of independent selection is called simple random sampling. The possible advantages of random sampling (unbiased results with calculable chances of error) also arise in more general *probability sampling,* where items have unequal probabilities of selection. As long as these different probabilities are *known,* then the sample results can be appropriately weighted.

For example, if individuals are selected by randomly choosing one person per household from a sample of households, people from large households will be under-represented. But this occurs to a known degree if the size of each household is recorded. Multiplying the results for each sampled individual by the number of people in his household will then redress the imbalance. Alternatively, households can first be selected with probabilities proportional to their *size.* This is often the more efficient procedure in multistage sampling (for example in selecting towns and then clusters of people within them).

Variable Sampling Fractions. Instead of sampling the same proportion of readings from each stratum or cluster of the population, the proportion can be made to vary. Selecting one individual per household, whether large or small, was a case in point. In other cases a higher proportion is taken in strata which are of particular interest or where the variability of the data is known to be exceptionally great, in order to achieve greater statistical

accuracy there. This occurs particularly if some important strata are numerically small in the population. For example, there may be far more patients suffering from a relatively mild attack of a certain illness than from a severe one. In studying the illness one might then sample a much higher proportion of those severely affected in order to achieve an adequate sample size of this group.

Generally the *overall* statistical accuracy of the *total* sample will then be less than that of a straight sample of the same size. Before the numbers from the different strata can be added and analysed they have to be brought back into line with the proportions in the population by weighting. The advantages of increased accuracy for certain sub-groups therefore have to be balanced against the likely reduction in accuracy in the overall results.

Sampling with Replacement. In the normal process of simple random sampling from a population of N members, exemplified by putting N numbered slips into a hat and successively selecting a sequence of n slips, the slips do not all have the same probability of being selected.

With the first selection each slip has a chance of l/N, but with the second selection each of the remaining $(N-1)$ slips now has a chance of $1/(N-1)$, and so on. This kind of "sampling without replacement" is therefore a form of "probability sampling", i.e. sampling items with different but known probabilities. This can be allowed for in the analysis, but it makes parts of it considerably more complex.

In contrast, each sampled slip could be replaced in the population before the next slip is selected. Then at each stage the probability of any slip being selected is l/N. This is called "sampling with replacement". (An item in the population may then be picked more than once, but it would not actually have to be *measured* more than once, as long as the relevant numerical result is counted the appropriate number of times.) Because all the probabilities of selection are equal, the theoretical mathematics of sampling *with* replacement is much easier than that of sampling without replacement.

In practice, however, most sampling is carried out without replacement because it is usually physically more convenient. This can generally be ignored in the analysis and the simpler forms of analysis appropriate to sampling *with* replacement used instead. This "fudging" is possible because the numerical results of the two types of sampling are almost the same unless n, the size of the sample, is large compared with N, the size of the population. If not, fairly simple correction formulae can often be used.

Randomised Experiments. Random sampling also plays a major role in statistically controlled experiments. To assess the effectiveness of some drug for an illness, or the effect of some fertiliser on a crop-yield, experimental control can be introduced by taking a group of patients or some plots of

ground and dividing them into two sub-groups *at random.* One sub-group is treated and the other is left alone as a control.

This is an important procedure when the material under examination is very uneven. Some patients are inherently more likely to improve than others, and some plots of ground are more fertile than others. This could greatly affect the apparent results of the study. But in a randomised experiment the chance of the control group having substantially fewer spontaneous recoveries or fewer fertile plots than the treated group is relatively small and can be calculated. The theory and practical applications of the design of statistical experiments have been highly developed and are further discussed in Chapter 19.

Further Reading. The techniques of probability sampling are described more fully in various specialised texts (e.g. Kish, 1965; Moser and Kalton, 1971; Yates, 1960; Cochran, 1943: Deming, 1960; Hansen, Hurwitz, and Madow, 1953).

16.5 The Results of Taking a Sample

We now consider the results of taking a single random sample. The population sampled are the 491 households whose half-yearly purchases of Corn Flakes were analysed in Chapters 12 and 13. Suppose we have taken a simple random sample of 10 of these 491 households and measured their half-yearly Corn Flakes purchases, as set out in Table 16.2. This is now all the direct information we have about the population.

TABLE 16.2 Results for a Random Sample of 10 Households from the Population of 491

(Purchases set out in order of their size)

Half-yearly purchases of Corn Flakes	Av.
0, 0, 0, 0, 1, 1, 2, 3, 5, 7	1.9

The sample made an average of 1.9 purchases. Although we hardly expect this result to represent the population exactly, it does tell us *something.* For example, the average number of purchases in the population is unlikely to be 19 or 190, even a sample of 10 seems to tell us that. It is the role of statistical sampling theory to set more precise limits on the information that a sample gives us.

In interpreting such data we might have some background information. For example, we might know that in the previous year annual sales of Corn

Flakes in Great Britain amounted to almost 150 million packets. With about 20 million households in the country, this gave an average rate of 3.7 packets per household per half-year.

Last year's national rate of sale was therefore twice as high as the average rate of 1.9 packets in our sample data now. But we also know that there are no really dramatic seasonal trends in the consumption of breakfast cereals, and that the dominant brand's sales do not usually drop by 50% from one year to the next. That kind of thing does not happen. It therefore looks as if the specific population of 491 households did not behave like the country at large.

If we had measured all 491 households we would know. But having measured only a sample of 10 of these households, another possibility is sampling error. Perhaps the population of 491 *did* make an average purchase of about 3 or 4 packs, like the country in general, and it was only our small sample which was different. This is a typical question for sampling theory to deal with: how likely is a result of 1.9 packs for a random sample of 10 if in fact the rate was 3.7 for the population of 491?

Sampling theory can never tell us more than if we had measured the whole population, in this case all 491 households. Its role is to bridge the gap between a sample and what we want to know about the population. This we discuss in the next two chapters.

16.6 Summary

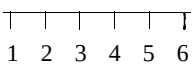
A sample can never give us more information than would the population from which it was selected. Samples are usually taken to save time, effort, and cost in data collection. With a random sample one can calculate the chance of an unrepresentative result and reduce it by choosing an adequately large sample-size.

Various elaborations, such as multi-stage stratified sampling, can be used to reduce costs for any given level of accuracy. Random sampling also plays a fundamental role in controlled statistical experiments.

CHAPTER 16 EXERCISES

Exercise 16A. Judgment Sampling

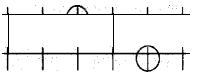
Relying on "judgment" to avoid bias in selecting a sample can be difficult. Here are 6 equally-spaced points (e.g. houses in a street:



Consider how people would set about selecting a "representative sample" of two of these points.

Discussion.

Most people would first select one of the two "middle" points, say number 3. Next they would generally choose a more extreme point, over to the right like number 5, to "balance things out":



We now have a sample of two readings with a mean of 4, compared with the "population" mean of 3.5. This is close. But the *scatter* of the sample readings is smaller than that in the population. In particular, the end values 1 and 6 have been omitted completely. Most people would not pick these, they are "too extreme" or "not typical", although they account for a third of the population (and corner houses tend to be more expensive).

One could try to allow for this particular kind of bias, but generally it is found that systematic biases in judgment sampling are difficult to eliminate completely and with certainty.

Exercise 16B. "Drawnat Random"

"A set of IQ cards numbered 1 to 10 is shuffled, and three cards are drawn at random." (From "Experiment I", Brookes and Dick, 1969, p. 8.) Comment.

Discussion.

The three cards cannot have been drawn *at random.* Instead, they must have been selected in some more or less deliberate or haphazard way. The approximate randomness of the process arose from the *shuffling.* (For the three cards to have been selected literally at random, ten slips of paper numbered 1 to 10 would have had to be put into a hat, shuffled, and three selected haphazardly, giving three numbers. These would then identify a random sample of three cards. These three cards could then be selected by looking at the numbers on all *ten* cards and picking out the relevant three !) Statisticians who confuse explicit randomness with haphazard selection make a difficult concept almost impossible.

As another example, a recent examination question had a housewife go to a deep-freeze cabinet in a supermarket and "select three packets at random". But a housewife certainly does not pull out all the packets, number them, put correspondingly numbered slips of paper in a hat, shuffle, etc. Apart from the insult to housewives, the phrasing of the question ignores the whole nature of the problem, that the selection of packets is generally *biased.* Packets at the bottom of the cabinet are picked less frequently.

Exercise 16C. A RandomSample

"Table X gives the weight of hearts and the weight of kidneys in a random sample of twelve adult males between the ages of 2.5 and 55." (From "Example 16.1", Moore, 1969, p. 253.) Comment.

Discussion.

It seems unlikely. One can take a random sample of 12 men from the population, but how does one weight their hearts and kidneys unless they are *dead*? The sample must therefore have been from a population of dead men (or from a set of appropriate records), so it will not be representative of men generally. One also wonders if the population was aged from 2.5 to 55, or if that was merely the age-range for the 12 men.

A further question is why a sample of only 12 was taken instead of some more "adequate" sample like 50 or 100. It looks as though this was not a proper sample at all, random or otherwise, but just all the readings available. It is a common practice to upgrade some haphazard collection of readings by calling it "a sample".

(In point of fact, the 12 sets of readings here were a random sample from the records for male patients who had died at a certain hospital. But some scepticism seemed called for.)

Exercise 16D. Defining the Population

What is a *population* in statistical sampling?

Discussion.

A population has to be defined in terms of the operational conditions of observing it, and not merely by specifying a particular group of living individuals or material items.

For example, a nation's population would have to be defined *operationally* not merely as all persons "living" there, but as all those who can be observed to be alive at a given point in time, using a specified method of observation. This might then exclude nationals who were abroad, and include foreigners, whether resident or visitors. Is that what was intended?

For the practical purpose of measuring every individual in such a population or of taking a sample from it, the population must defined in terms of a listing of all individuals (or some equivalent operating procedure). When examined closely, such a list might not include the army, prisoners, and people in hospitals, whilst other people might be listed more than once (e.g. students at home *and* at their university). A list will inevitably be out **of** date (some members of the list will have died, newborns will not be included, and some people will have changed addresses).

Some members of the population, although explicitly included in the list, will be difficult or impossible to observe or measure. In measuring human populations (e.g. in a sociological survey), people abroad or ill, travelling salesmen, the old and the rich, are often difficult to contact. Others may *refuse* to be measured or interviewed.

None of these problems has anything to do with sampling as such. But it emphasises that the population to be sampled is not necessarily the one we would like to sample. Instead, care has to be taken in (i) explicitly defining the population actually sampled, (ii) how this might differ from the population one had in mind, and (iii) how one can reduce or allow for any biases involved. These are matters for the technical expert experienced in dealing with the particular type of population in question.

There will always be some fuzziness in defining a population. For example, we may aim to define it as all people alive and in the country at

12 noon on a certain day. But for some people we will not know whether they were alive, or in the country, *exactly* at 12 noon. A test of whether we have a reasonably well-defined population is whether we could physically list them all and draw a *proper random sample.* If not, we need to reconsider what we are talking about.

Exercise 16E. Sampling in Time

How can we take a representative sample of events occurring over time'?

Discussion.

Time is rarely sampled statistically. We would have to be able to take a proper random sample of all the relevant moments in time, the "population". This is only possible with deliberate planning. For example, suppose we wanted to make traffic counts in a certain street for a random selection of 100 one-minute intervals tomorrow. We would have to select 100 random numbers between 1 and 1,440 and then make the corresponding observations.

Again, the number of deaths per year due to measles from 1964 to 1973 is not a *sample* of such readings. No population of years has been explicitly defined from which these 10 years were selected, randomly or otherwise. Instead, the 10 years amount to a "mini-population" of their own.

Exercise 16F. Sampling the Same Population Again

Having measured the heights of a sample of boys from a certain school, can we measure another sample the next day?

Discussion.

No. Suppose that the first sample were taken by appropriately selecting a number of boys from the school list, calling them to a certain room, and measuring their heights.

Following the same procedure the next day will not necessarily give the same results. This is not because of the sampling (the samples could be large) but because the *populations* on the different days are not the same. This may seem odd because by and large we would expect the measurements of heights to be very similar on different days. But insisting that the populations differ is not splitting hairs.

We only know the measurements of boys' heights will be much the same from one day to the next because we know that people's heights do not vary in the short run. This knowledge has nothing to do either with sampling or with the definition of the populations, but with the particular measurement : heights.

If our observations concerned what the boys wore, or what they did, results could differ markedly from day-to-day, and especially between a Friday and a Saturday, say. In general, the lessons the boys took, their ages, or anything of any kind would be different. To suppose that different populations are the same is to beg the entire question. Two populations can have the same properties in some respects (e.g. the same

average height, etc.), but this does not mean that they are operationally the same population.

Exercise 16G. TakingMorethanOneSample

Is it in fact possible to take more than one sample from the same population?

Discussion.

This can only be done in rather artificial and limited ways. For example, if from some population we take a random sample of 500 boys, numbered consecutively from 1 to 500, then the even-numbered and the odd-numbered boys selected each constitute a (smaller) random sample from the same population.

If the even-numbered boys were measured in the morning and the odd ones in the afternoon, we would be dealing with two samples from *two distinct populations*, the boys in the morning and the boys in the afternoon. (Their heights might be the same, but their activities, the state of their metabolism, etc., would not be.)

This implies that if the 500 boys were measured at one-minute intervals at 9.00, 9.01, 9.02 and so on, one would be dealing with samples of 1 from 500 populations. But this is only the case if the times when each boy is measured are recorded, and even then one may choose deliberately to oversimplify by ignoring this information and regarding the 500 boys as a random sample from "boys that day", which they are.

Exercise 16H. AnIndefinite Population

"If an observation, such as a simple measurement, is repeated indefinitely, the aggregate of the results is a population of measurements." (Fisher, 1950, p. 2). Comment.

Discussion.

The concept of an indefinite population is ill-defined. How do we know that the measurements will not change in the course of time? Is the experimenter allowed to rest every now and then in carrying out the indefinite sequence of measurements? Does he have to work at night and at weekends? What happens when he dies? Who is he anyway, or is more than one experimenter involved?

It is not clear what the population is, nor how any proper (random) sample selection can take place. The items in an indefinitely large population cannot be listed and sampled. This view of a population appears to be meaningless for any practical purpose.

The alternative is to define explicitly some *limited* populations of observations (and possible samples therefrom). One can see whether such different populations have the same observed properties.

Exercise 161. Games of Chance

Are 10 successive throws of a coin a random sample?

Discussion.

No, not according to the definitions used so far. Firstly, there is no formal random sampling procedure involved in selecting the ten throws from some larger number of throws. Secondly, there is no firmly defined population of throws.

In games of chance the outcome of each event (e.g. throwing a coin, throwing dice, etc.) can be at best a "quasi-random" event, under suitable empirical conditions (see Section 13.2). A sequence of throws can therefore behave *like* a random sample (and indeed be used in selecting "random" samples from a real population, e.g. half the houses in a street). But this is because of something in the inherent nature of each physical observation, not because any explicit random sampling procedure has been used in selecting a sequence of throws.

CHAPTER 17

Sampling Distributions

A sampling distribution summarises the variation of all possible random samples of a given size from a population. In practice, one rarely takes more than one random sample from a population, so the sampling distribution mostly remains a theoretical concept. It is used in estimating the accuracy of a single random sample from the population, as will be discussed in Chapter 18.

There are a variety of sampling distributions for different summary measures like the mean, the variance, the correlation or regression coefficients, and so on. In this chapter we concentrate on sampling distributions of the mean because they are the most important and the simplest.

17.1 Some Empirical Sampling Distributions

We now consider what happens when we measure the purchases of Corn Flakes made by different random samples taken from the 491 households in Chapter 12 (Table 12.9). These bought on average 3.4 packs per household in the half year. But different samples will have different mean purchase rates. A run of samples of a given size then yields a frequency distribution of sample means. This is the sampling distribution of the mean.

Tables 17.1 and 17. la illustrate this, for random samples of size 1, 2, 10 and 40. (Exercise 17A sets out the details for samples of 10 households.)

TABLE 17.1 Empirical Distributions of Sample Means for Random Samples of Different Sizes

	Value of Sample Mean											Ave-rage*	Stand. Dev.*
	0-	1-	2-	3-	4-	5-	6-	7-	8-	9-	10-	11+		
Samples of 1	% 40	17	11	4	4	5	3	1	4	0	2	9	3.4	4.7
Samples of 2	% 29	21	12	8	5	4	5	4	1	1	2	8	3.4	3.9
Samples of 10	% 5	5	30	15	20	15	10	-	-	-	-	-	3.4	1.5

*Calculated from the ungrouped values

TABLE 17.1a The Distribution of the Means of Samples of 40 Households

	Value of Sample Mean								Average*	Stand. Dev.*
	<1.8	1.8-	2.2-	2.6-	3.0-	3.4-	3.8-	4.2- 4.6-		
Samples of 40 %	-	10	10	15	25	10	20	10 -	3.4	.8

* Calculated from the ungrouped values

Few of the individual smaller samples have means anywhere near the population mean of 3.4. For example, 40% of the samples of $n = 1$ have mean values of zero. A single sample of one reading cannot provide very accurate information about the population. But for larger samples, such as with $n = 40$, most of the sample means lie between about 2.6 and 4.5, only a unit or so from the population mean. Therefore in most cases any single sample of 40 will indicate *fairly* accurately the actual population mean of 3.4.

To describe the empirical variability of the various sample results, the sampling distributions must be specified, i.e. their shapes and summary measures like their own means and standard deviations. The illustrations in the tables show three important features.

(i) The *average* value of each sampling distribution is 3.4, which is equal to the population mean.

(ii) The standard deviations of the sampling distributions decrease markedly as the sample sizes increase, but less than proportionately; e.g. doubling the sample size does not halve the scatter.

(iii) The shape of the distributions changes as the sample sizes increase. For $n = 1$ or 2, the distribution is highly skew, but for $n = 40$ the distribution is fairly close to Normal; 70% of the readings lie less than $\pm$ 1 standard deviation of the average mean value of 3.4.

These illustrative empirical results generalise and can be developed theoretically.

17.2 The Sampling Distribution of the Mean

Suppose we have a population with a mean μ ("mu", the Greek m) and a standard deviation σ ("sigma':, the Greek s), to use the conventional statistical notation of Greek letters for population values.

Then it can be shown that the distribution of the means of all possible random samples of size *n* from that population has the following properties :

(i) its mean is equal to the population mean, μ;

(ii) its standard deviation is $\sigma/\sqrt{n}$;

(iii) its shape is approximately Normal, except for small samples if the population is not Normal.

These results are fundamental in statistical sampling and have a firm theoretical basis. They follow mathematically from the fact that each reading in a random sample is selected according to the rules of probability.

But in practice we do not know the values of the population parameters μ and σ. It is therefore of little immediate use to know that the sampling distribution of the mean is approximately Normal with (unknown) mean μ and standard deviation $\sigma/\sqrt{n}$.

However, we have our sample of n readings with mean m and standard deviation s. We can use these sample values to estimate the mean and standard deviation of the sampling distribution. But unfortunately it is then not true that this distribution is Normal with standard deviation of $s/\sqrt{n}$ (its standard deviation is the unknown $\sigma/\sqrt{n}$). How then can we describe the sampling distribution of the mean?

Publishing under the pseudonym "'Student", W. S. Gossett (a Brewer of Arthur Guinness Ltd.) presented a solution to this problem in 1908. This marked the first breakthrough in the exact statistical treatment of small samples. He took the ratio of the sample mean m to the observed value $s/\sqrt{n}$, a quantity known as Student's "t"

$$t = \frac{m}{s/\sqrt{n}}$$

and established the mathematical nature of its sampling distribution for samples from a Normal population.

The t-distribution has a larger scatter than a Normal Distribution. For example, with samples of size $n = 5$, about 95% of the readings of a t-distribution lie within ± 2.6 standard deviations of the average. When $n = 10$ and $n = 20$, 95% of the readings lie within $\pm$ 2.2 and 2.1 standard deviations of the average. In a *Normal* Distribution, 95% of the readings lie within only ± 2 standard deviations of the average. The t-distribution therefore dependson the sample size n (or on the *degrees of freedom* of the data, which here are $(n - 1)$, as described for the variance in Exercise 11H).

The important point illustrated by the above results is that the t-distributions do not differ radically from the Normal Distribution, except for very small samples (n less than 10). Thus whether 95% of the readings lie within $\pm$ 2.2 or $\pm$ 2.0 standard deviations is only a small difference. Even for rather small samples of 10, 20 or so, having to estimate the population σ from the sample values therefore hardly affects the sampling distribution of the mean. (For large samples, say $n = 1000$ or more, this would not be surprising since s must then be a close estimate of σ anyway.)

The practical implication of Student's achievement is therefore not so much that he provided the exact answer to the sampling distribution of the mean for very small samples (since these seldom occur), but that it follows

from his findings that the problem he tackled does not greatly matter. The Normal Distribution will serve in most cases as a close approximation even when using the *estimated* standard deviation $s/\sqrt{n}$.

The Standard Error of the Mean

The quantity $s/\sqrt{n}$ is commonly referred to as the *standard* **error of the** *mean* and tends to be regarded as a property of the one sample actually observed, implying a standard or "average" level of sampling error. Nonetheless, it has to be remembered that it is by definition an estimate of the true standard deviation $\sigma/\sqrt{n}$ of the distribution of the means of all possible random samples of size n from the same population.

It is a remarkable achievement to be able to estimate from a single sample how different samples might vary. We can do it because of the nature of random sampling, where each item in a sample is selected independently of all previous items. Thus, a single random sample of size n is effectively composed of n separate random samples of size 1, or $n/2$ random samples of size 2, etc. Therefore, one single sample *does* tell us about the variability of different samples, but smaller ones. The formula $s/\sqrt{n}$ then provides the link for samples of different sizes n.

The standard error formula $s/\sqrt{n}$ fits in with common sense. As the sample size n increases, the scatter of the sample means decreases. More of the sample means lie close to the population mean, μ. But the scatter does not decrease in direct proportion to the increase in sample size, since additional readings provide only marginal extra information. As we saw from the illustration in Section 17.1, the sampling distribution for $n=1$ had a standard deviation of 4.7, but for $n=40$ the standard deviation was .8. To estimate this figure from the first, we have to divide 4.7 by $\sqrt{40}$, which gives as an estimate $s=.75$ or .8. Thus the standard error formula uses as the divisor the *square-root* of the sample size, $\sqrt{n}$. To *halve tie* standard error the sample size has to be *quadrupled.*

One can use the standard error formula to determine the sample size required for a particular degree of accuracy. For example, if we want a sample mean with a standard error of about .5 and know from previous results that the standard deviation of the data is about 4, then-we would have

$$\frac{4}{\sqrt{n}} = .5.$$

A sample size of about 64 would be required. However, since the formula involves the square-root of n, the precise sample size is not critical. For $n=100$ the standard error of the mean would be .4, and for $n=50$ it would be .6, neither of which is very different from .5.

17.3 The Difference Between Two Sample Means

The sampling distribution of the difference between the mean values of two independently drawn samples is of great practical importance. For example, one could have an experimental group and a control group, each consisting of a random sample from a larger population. The sampling distribution of the difference of two means can then be used to assess the accuracy of the observed difference between the two samples.

If the sizes of the two independent samples are n_x and n_y, then the sampling distribution of the differences in their means $(m_x - m_y)$ will have a standard error of

$$\sqrt{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)},$$

where σ_x^2 and σ_y^2 are the variances of the x and y variables. This formula contains the sum and not the difference of the two separate variances because the distribution of $(m_x - m_y)$ is subject to the sampling errors of both x and y. Therefore there will generally be *more* error than for either m_x or m_y alone.

The numerical value of the standard error has again to be estimated, by using the observed standard deviations s_x and s_y to replace σ_x and σ_y. The sampling distribution follows a t-distribution with $(n_x + n_y - 2)$ degrees of freedom, on the null hypothesis that the population means and standard deviations are equal. Unless the degrees of freedom fall well below 10 or the population is highly non-Normal, the t-distributions are almost identical to a Normal Distribution.

17.4 Other Sampling Distributions

Any summary measure of sample data varies from sample to sample and this variation is described by a sampling distribution. Examples are the variance, standard deviation or range of a sample, the proportion of zeros in the data, or the correlation and regression coefficients for two variables. When n is very large these sampling distributions also tend to be Normal, but otherwise they are generally more complicated than that of the mean. The subject can get very technical, but for the practical user the broad principles already discussed remain the same. For specific applications one can usually refer to suitable numerical tables showing the proportion of samples with values exceeding some particular level. We now discuss a particular example.

The χ^2-Distribution

The χ^2-distribution (called chi-squared) is an example of a more complex sampling distribution. One practical use is in assessing how well theoretical

models fit sample data. This is discussed in Chapter 18 ; here we consider the nature of the relevant sampling distribution.

Suppose we have a sample of 200 readings with a mean of 9 and a standard deviation of 3, as shown grouped in the top line of Table 17.2. Suppose further that we *know* that the sample comes from a Normal population (or it could be from some other type of distribution, e.g. an NBD or a Poisson, etc.). The theoretical values for a Normal Distribution are shown in the bottom line (i.e. 68 % of the 200 readings should lie between 6 and 12). The observed values clearly differ somewhat from the population ones. The question here is how much will the results of *different* random samples vary from these theoretical norms?

TABLE 17.2 Observed and Theoretical Numbers for a Sample of 200 Readings from a Normal Distribution (Mean 9, Standard Deviation 3)

	Values						Total No. of Readings
	<3	3-	4-	9-	12-	15-	
No. of readings							
Observed	7	30	61	73	25	4	200
Theoretical	5	27	68	68	27	5	200

To answer this question the usual approach is to calculate an overall measure of agreement of fit between the observed and theoretical values and then to consider how the value of this measure might vary from sample to sample. One possible measure to use is the mean deviation, between the observed and theoretical values, which here is $20/6 = 3.3$. But the sampling distribution of this quantity is technically difficult to establish.

A more tractable measure is Sum {(Observed $-$ Expected)^2/Expected}, i.e. the sum of the squared differences between the observed and theoretical frequency in each group, divided by the theoretical value. In our example, this would give us

$$\frac{(7-5)^2}{5} + \frac{(30-27)^2}{27} + \frac{(61-68)^2}{68} + \frac{(73-68)^2}{68} + \frac{(25-27)^2}{27} + \frac{(4-5)^2}{5} = 2.6$$

(The smaller the value, the closer the theoretical model fits the data.)

This measure is useful because its distribution for different samples can be calculated, a result due to Karl Pearson in 1900. It approximates to a χ^2-distribution, whose mathematical properties are already well-known. (The χ^2-distribution is a special case of the Gamma-distribution mentioned earlier, which occurs in several distinct ways in statistical theory.)

For our 200 readings the appropriate χ^2-distribution is one with three "degrees of freedom". This is the difference between the six categories into which the data have been grouped and the three parameters used in fitting the theoretical model (the standard size n, the mean, and the standard deviation). The theoretical χ^2-distribution is shown in Table 17.3 (from Elderton, 1902). Data with different degrees of freedom have different χ^2-distributions.

TABLE *17.3* *The* f-distribution with 3 Degrees of Freedom

3 degrees of freedom	Values of χ^2												Total
	<1	1-	2-	3-	4-	5-	6-	7-	8-	9-	10-	11-	
No. of readings %	20	23	18	13	9	6	4	2	2	1	1	1	100%

This distribution is very skew, with a long positive tail. But it shows that 18 % of samples of 200 from a Normal Distribution would give a χ^2-value of between 2 and 3. Thus the data in Table 17.2 seem quite typical : many other random samples would give similar χ^2-values.

The shape of the χ^2-distribution depends only on the number of degrees of freedom, not on the sample size. The number of different χ^2-distributions that need to be tabulated for reference is therefore limited to the possible levels of degrees of freedom. A further simplifying feature is that for relatively large degrees of freedom the distribution becomes increasingly humpbacked and symmetrical and tends to an approximately Normal form. This tendency can be accelerated by mathematically transforming χ^2 to $\sqrt{(2\chi^2)}$. Then the skewness of the distribution is reduced and the distribution is almost Normal even for relatively low degrees of freedom (i.e. 20 or 30). The distribution of $\sqrt{(2\chi^2)}$ has a mean of $\sqrt{(2\nu - 1)}$ and a unit standard deviation, where v (Greek n, or "nu") is the conventional, symbol for the number of degrees of freedom. The tendency of most sampling distributions to approximate a Normal shape for reasonably large sample sizes (or numbers of degrees of freedom as here) is a major simplifying feature in the general statistical theory of sampling.

17.5 Sampling **Theory for Other than Simple Random Sampling**

Until now our discussion has been concerned with sample data obtained by simple random sampling, but in practice most sampling is not done this way. As discussed in the last chapter, sampling is almost always done *without* replacement, and in sample surveys some form of stratified multi-stage cluster sampling is usually employed. This affects the sampling errors that occur.

The formula $\sigma/\sqrt{n}$ applies to sampling *with* replacement ; sampling without replacement decreases the standard error. Dividing the population into strata before sampling should also reduce the size of sampling errors (but the effect is small if the stratifications are not very discriminating). In contrast, multi-stage cluster sampling (e.g. selecting a number of towns and then a "cluster" of individuals in each town) generally *increases* the size of sampling error, compared with a simple random sample of the same size.

Unfortunately, statistical theory has not yet told us very much about the numerical effects of most of these types of sampling. Instead, the formulae for simple random sampling tend to be used, in the hope that they will adequately approximate the true answers. Sometimes estimates are made of the "design factor", the extent to which the sample design has affected the size of the sampling errors. However, delving more deeply into these matters is largely a subject for the professional statistician and the specialist textbook.

One case that is clear is sampling without replacement. The standard error formula $\sigma/\sqrt{n}$ implies that the size of the sampling error depends on the sample size n and not on that of the population. But many people instinctively feel that a large population can only be adequately represented by a large sample. This is in fact correct for sampling without replacement : the larger the population, the larger the sampling error, the sample need therefore to be larger to give the same degree of accuracy. However, the *numerical* effect of the population size on the sampling error is trivial if the sample is a small portion of the population, as is almost always the case.

The statistician's usual assertion is that sampling error only depends on sample size and not on population size. For sampling with replacement (which nobody practises) this is true; for sampling without replacement it is in principle false, but in practice a close and simple approximation to the truth.

17.6 **Summary**

Statistical sampling theory deals with the way a particular summary measure of a sample, such as its mean, varies from sample to sample.

For simple random samples of size n from the same population, the sampling distribution of the mean has three simple properties. Its own mean equals the population mean μ; its form tends to be Normal except for very small samples; and its standard deviation is $\sigma/\sqrt{n}$, where σ is the standard deviation of the individual readings in the population.

The value of σ can be estimated by the standard deviation s of the observed sample. The sampling distribution of the mean then strictly requires use of Student's i-distribution, but except for very small sample sizes it differs little from a Normal Distribution with standard deviation $s/\sqrt{n}$.

The quantity $s/\sqrt{n}$ measures the scatter of the means of different samples of size n and is usually called the standard error of the mean. It is the basic formula in statistical sampling theory.

The difference between the means of two independent samples of size n_x and n_y, drawn from the same population, will have an estimated standard error

$$\sqrt{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)},$$

where s_x and s_y are the standard deviations of the x and y readings.

Sampling distributions for other summary measures are usually more complicated than those for the mean. However, for large sample sizes they generally tend to be Normal. This fact is the major simplifying feature in the statistical theory of sampling.

CHAPTER 17 EXERCISES

Exercise 17A. Empirical Samples

To illustrate the nature of random samples, draw a series of such samples from a specified population and consider the results.

Discussion.

As an example, consider taking samples of 10 from the population of 491 households whose half-yearly purchases of Corn Flakes were described in Table 12.9. The households can be arbitrarily labelled from 1 to 491, then sets of 10 numbers between 1 and 491 inclusive can be read off from a table of random numbers or the like. Doing this for 20 samples gave the results in Table 17.4.

Each sample broadly resembles the population, where most households bought either no Corn Flakes or only one or two packs, and few households bought many packs. But the resemblance is not very precise, since the results for the different samples are so variable. For instance, the numbers of non-buyers vary from 2 to 6 out of 10, and the means vary from .6 to 5.8.

This kind of sampling generates a sampling distribution for any aspect of the data one may wish to consider, e.g.

the sample means, 2.2, 3.1, 2.3, 1.7, etc.;
the percentage of zeros or non-buyers, 30 %, 20 %, 50 %, 40 %, 50 %, etc.;
the standard deviations of the samples, 2.2, 4.8, 3.5, 2.1, 4.9, etc.;
the k-parameter of the Negative Binomial Distribution, 1.8, .5, .5, 1.1, .7, etc. (using the formula $k = m^2/(s^2 - m)$ in terms of the sample mean m and variance s^2) if an NBD is fitted to the data.

The sampling distributions of these other statistics are in many cases highly skew. For example, the k-values for the 20 samples in Table 17.6 are, in order of size,

.2, .3, .3, .4, .4, .4, .4, .5, .5, .7,
.7, .8, .9, .9, 1.0, 1.1, 1.4, 1.5, 1.8, 9.0.

TABLE 17.4 Twenty Random Samples of 10 from the Given Population of 491 Households in Table 12.9

(Half-yearly Purchases of Corn Flakes)

Samples of 10	Number of purchases made 0	1	2	3	4	5	6	7	8	9	10	11+*	Average
1st sample	3	2	1	2	-	-	2	-	-	-	-	-	2.2
2nd "	2	4	1	-	2	-	-	-	-	-	-	(17)	3.1
3rd "	5	-	2	1		1	-	-	-	-	-	(12)	2.3
4th "	4	3	-	1	-	1	1	-	-	-	-	-	1.7
5th "	5	-	1	1	-	-	-	-	1	-	1	(14)	3.7
6th "	6	1	1	1	-	-	-	-	-	-	-	(17)	2.3
7th "	3	3	2		-	1	-	-	1	-	-		2.0
8th "	6	1	1	-	1	-	-	-	-	-	-	(13)	2.0
9th "	5	1	1	-	1	-	-	-	-	-	-	(14, 22)	4.3
10th "	2	3	-	1	-	1	1	-	1	-	-	(16)	4.1
11th "	3	1	1	-	-	-	-	-	-	2	-	(12, 13, 14)	5.8
12th "	6	2	2	-	-	-	-	-	-	-	-		.6
13th "	4	1	1	-	-	2	-	-	-	1	-	(29)	5.0
14th "	2	3	2		-	1	1	-	-	-	-	(36)	5.2
15th "	3	1	1	-	1	2	-	-	-	1	1		3.6
16th "	3	2	1	-	1	1	-	-	1	-	-	29	5.0
17th "	6	1	-	-	-	-	-	1	-	-	-	(20, 24)	5.7
18th "	6	1	-	-	-	-	-	1	-	-	-	(13, 13)	4.1
19th "	2	3	3		-	-	-	-	-	-	1	(11)	3.0
20th "	5	1	1	-	-	2	-	1	-	-	-	-	2.0

* The actual values of the 11+ readings

The range of these sample values is large, from .2 to 9.0, and the average is 1.2. This compares badly with the k-value of SO for the whole population (see Section 12.3). Thus using the sample value of k to estimate the population value gives biased results, i.e. the wrong answer on average.

Often the size of this type of bias can be established by theoretical analysis, so that a correction factor can be devised. This, and other methods of deriving better estimators, is part of the more advanced theory of statistics and typifies some of the more complex problems in sampling. The sample mean is a somewhat exceptional measure because it gives an unbiased estimate of the population mean and hence its sampling theory is particularly simple.

Exercise 17B. The Expected Value of the Sample Mean

Outline a proof that the average value of the sampling distribution of the mean, m, of a sample is equal to the population mean, μ. Is the equivalent result true for the range?

Discussion.

Consider random samples of two readings, selected with replacement. For a particular sample, the readings are x_1 and x_2. The sample mean is $(x_1+x_2)/2$. The sampling distribution of the means of such samples therefore has a mean equal to the average, or statistically "expected", value of $(x_1+x_2)/2$ over all possible samples of two readings. (The expression "expected value" is a useful way of describing the average value of a reading across all possible samples. It is what one "expects" to obtain on average.)

Because the two readings x_1 and x_2 were sampled independently of each other, the expected value of $(x_1+x_2)/2$ equals the expected value of $x_1/2$ across all possible samples, plus the expected value of $x_2/2$ across all possible samples. Now the expected value of a single reading x_1 across all possible samples is μ, by definition the average value of all the readings in the population. Similarly, the expected value of x_2 is μ. Hence the expected value of the sample mean $(x_1+x_2)/2$ is $2\mu/2=\mu$, the population mean, which is the required result. The argument generalises readily to samples greater than 2.

This argument may seem almost simple-minded. But it is only possible because in random sampling with replacement the two items, x_1 and x_2, are selected independently of each other. Without this independence we could not consider the expected value of $x_1/2$ separately from that of $x_2/2$.

For instance, this kind of proof is not possible for sampling without replacement (although for the sample mean the same *result* still holds). More generally, the equivalent result will not hold for a statistical measure that is not a "linear function" of the n sample readings, i.e. a function of the form $a_1x_1+a_2x_2+a_3x_3+\cdots+a_nx_n$, where the coefficients $a_1, a_2, a_3, \ldots$ are fixed *a priori*, (For the sample mean, all these coefficients are equal to $1/n$.) The variance, for example, is not a linear function of the readings in this sense. Nor is the range, because the coefficients of all but the highest and lowest readings are zero, but are not determined *a priori*.

The expected value of the range for a sample size n is always biased. For any particular sample it is always either the same or smaller than the range of all the readings in the population, it cannot be *larger*. Hence the average value of the range of all possible samples of size n cannot be equal to the population value of the range.

Exercise 17C. The Standard Error of the Mean

Outline a proof that the standard error of the mean of a single random sample of n readings is $\sigma/\sqrt{n}$, where the population standard deviation is σ.

Discussion.

The theoretical formula for the standard error of the mean holds because the items in a simple random sample are selected independently. Consider again a sample of two readings x_1 and x_2, with a mean of $(x_1+x_2)/2$. The variance of the sampling distribution of such means is then by definition the expected value of

$$\left(\frac{x_1+x_2}{2}-\mu\right)^2.$$

This expression can be written as

$$\left(\frac{x_1 - \mu}{2} + \frac{x_2 - \mu}{2}\right)\left(\frac{x_1 - \mu}{2} + \frac{x_2 - \mu}{2}\right)$$

$$= \frac{(x_1 - \mu)^2}{4} + \frac{2(x_1 - \mu)}{2}\frac{(x_2 - \mu)}{2} + \frac{(x_2 - \mu)^2}{4}.$$

The average or "expected" value of $(x_1 - \mu)^2$ over all possible samples is simply the population variance σ^2: that is how the variance is defined. The same holds for $(x_2 - \mu)^2$. Thus the expected value of $(x - \mu)^2/4$ is $\sigma^2/4$.

Since the values of x_1 and x_2 are selected independently, we can start considering the middle term with one particular value for x_1, say $\hat{x}_1$. Then the expected value of $(\hat{x}_1 - \mu)(x_2 - \mu)$ across all possible values of x_2 must be zero, since the average of x_2 is μ. Similarly for any other value of xi. Hence the expected value of the middle term is zero.

It follows that the expected value of the above expression is $\sigma^2/4 + 0 + \sigma^2/4 = 2\sigma^2/4 = \sigma^2/2$. The same kind of argument can be used to show that the variance of the sampling distribution of the mean for samples of size n is σ^2/n. -The standard error of the mean, i.e. the standard deviation of its sampling distribution, is therefore $\sigma/\sqrt{n}$.

Exercise 17D. The Distribution of the Sample Mean

Explain why the sampling distribution of the mean tends to be approximately Normal for large enough samples.

Discussion.

The mean of a random sample of n readings is a variable made up of the average of n independent random variables. This will tend to follow a Normal Distribution for large n, because of the Central Limit Theorem (see Section 13.2 and Exercise 13K). The detailed mathematics required to prove the Central Limit Theorem are complex, but at least it is easy to illustrate that in taking reasonably large samples from a highly skew distribution, the sampling distribution of the mean must tend to be humpbacked.

Consider the population of 491 households and their purchases of Corn Flakes, as set out in Table 12.9a. This distribution was very skew: 39% of the households made 0 purchase, 14% bought 1 pack, and only 11% bought 10 packs or more.

A sample with a large mean, say 10 or 20, would have to include large numbers of heavy purchasers. Because individual households are sampled at random and independently, the chance of this happening is very small. For example, since only 11% (0.11) of all households buy 10 or more packs, the chance of getting 10 such heavy buyers in a random sample of 10 households is $(0.11)^{10}$ or roughly 1 in 4 thousand million. By a similar argument, very small sample means will be rare. The sampling distribution of the mean will thus tend to be humpbacked.

Exercise 17E. The Standard Error of the Difference of Two Means

Outline a proof of the formula

$$\sqrt{\left\{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right\}}$$

for the standard error of the difference of the means, $(m_x - m_y)$, for independent samples of size n_x and n_y from populations with means μ_x and μ_y and standard deviations σ_x and σ_y.

Discussion.

The variance of the sampling distribution of the quantity $(m_x - m_y)$ is the expected value of the squared deviations

$$\{(m_x - m_y) - (\mu_x - \mu_y)\}^2$$

across all possible samples of n_x and n_y readings. We can write this expression as

$$\{(m_x - \mu_x) - (m_y - \mu_y)\}\{(m_x - \mu_x) - (m_y - \mu_y)\}$$
$$= (m_x - \mu_x)^2 - 2(m_x - \mu_x)(m_y - \mu_y) + (m_y - \mu_y)^2.$$

The expected value of the middle term is zero because the x and y samples are selected independently. (For any given sample value of m_x, the expected value of $(m_y - \mu_y)$ is $(\mu_y - \mu_y) = 0$, etc.). Therefore the expected value of the above expression reduces to the expected value of $(m_x - \mu_x)^2 + (m_y - \mu_y)^2$, which is $\sigma_x^2/n_x + \sigma_y^2/n_y$ from Exercise 17C.

Exercise 17F. Sampling Without Replacement

Discuss the effect of sampling without replacement on the sampling error of the mean. As an example consider sampling from a very small population.

Discussion.

We are sampling n items from a population of N items. If each selected item is replaced before the next item is picked, each item in the population has an equal chance of l/N of being selected. But if a selected item is *not* replaced, the chances for successive items are l/N, $1/(N-1)$, $1/(N-2)$, etc. The sample mean will still be equal to the population mean when averaged across all possible samples of n, but the standard error of the mean will be smaller than in sampling with replacement.

To illustrate, consider a very simple population which consists of only three items with values 1, 3, and 5. We aim to sample two of these. Then in sampling without replacement, the first item selected might be "1", and the next item would have to be either "3" or "5". There are in fact six possible samples of two: 1 & 3, and 1 & 5, 3 & 1 and 3 & 5, 5 & 1 and 5 & 3. The sample means are 2, 3, 2, 4, 3, and 4. These six values differ on average by .67 from the population mean of 3.

In sampling *with* replacement, however, after selecting the "1" for the first item, there are still three equally likely possibilities for the second item, namely 1 again, or 3, or 5. There are therefore nine equally likely samples:

the six earlier ones plus 1 & 1, 3 & 3, and 5 & 5. The sample means are 1, 2, 2, 3, 3, 3, 4, 4 and 5. These nine values differ on average by .89 units from the population mean, which is greater than the value .67 when sampling with replacement. Thus the "unlucky" chance of hitting the same item twice increases the average size of the sampling error. However, this effect is only sizeable if the sample is a large proportion of the total population.

It can be shown by simple but relatively lengthy mathematics that the standard error of the mean for sampling without replacement is

$$\frac{\sigma}{\sqrt{n}}\sqrt{\left\{\frac{(N-n)}{(N-1)}\right\}}.$$

When the population size N is very large compared with the sample size n, the quantities $(N-n)$ and $(N-1)$ are both virtually equal to N, and the value of their ratio is very close to 1. The standard error formula is then almost equal to the value $\sigma/\sqrt{n}$ for sampling *with* replacement.

For example, taking a sample of 100 from a population of 1,000 leads to a factor of .95. Thus the *correct* standard error formula for sampling without replacement is only 5% smaller than the value given by the simpler $\sigma/\sqrt{n}$ formula. The difference is not large. If one is sampling 1,000 items from a population of one million, the correction factor would be about .9995, i.e. a difference of less than 1%, which is clearly negligible.

These results account for the fact that while it is common practice to sample without replacement (usually the easier operation physically), this is ignored in the analysis (which is then easier).

Exercise 17G. Binomial Sampling

Discuss the sampling distribution that arises in industrial quality control when items are randomly selected from a large batch of manufactured items of which a proportion p are defective.

Discussion.

Random sampling from a population in which each observation can take one of two values (e.g. Defective or Non-defective, or boy or girl) is a common way for the Binomial Distribution of Section 12.4 to arise.

Consider the first item randomly selected. It will have a chance p of being defective and $(1-p) = q$ of not being defective. The next item selected will have an equal chance p of being defective, if sampling is with replacement or if the population is large. If the two items are selected independently, the probability of both being defective is $p \times p = p^2$, that of one being defective is $p \times q + q \times p = 2pq$, and that of neither being defective is q^2. This is the (positive) Binomial Distribution with $n = 2$.

Thus it can be shown that the sampling distribution in such a case takes the form of the Binomial Distribution for any sample size n.

CHAPTER 18

Statistical Inference

When dealing with sample data one needs to do three things: (1) estimate the characteristics of the population sampled; (2) assess the accuracy of these estimates; and (3) check one's prior hypotheses about the data to determine whether any deviation in the sample result is due only to the likely errors of sampling.

The concept of the sampling distribution is used to infer from a single random sample the characteristics of the population from which it has come. In the discussion in this chapter we shall assume that the data have been selected from a specified population by simple random sampling. Calculations for other forms of probability sampling are more complicated, but the general principles remain the same.

18.1 Estimation

To estimate a population characteristic from a sample of readings commonsense suggests that we might simply calculate the corresponding value for the sample. For instance, the mean of the sample gives a good estimate of the population mean.

But for other kinds of summary measures it is not always as easy. For example, if the variance of a sample of n readings is defined as $\Sigma(x-\bar{x})^2/n$, the average or "expected" value of this quantity across all possible samples will be fractionally smaller than the population value of the variance. This systematic error or "bias" in the sample estimate can be eliminated by defining the variance as $\Sigma(x-\bar{x})^2/(n-1)$, using the divisor $(n-1)$ as described in Chapter 11. Such problems typically occur when trying to argue from a sample to the population from which it came.

Absence of statistical bias, i.e. giving the right answer ***on average*** across all possible samples, is one possible criterion in judging a good estimator. An alternative is aiming at the most ***accurate*** estimate from the single sample we have. Another widely used estimating principle, especially in complex situations, is "maximum likelihood". This means picking an estimate of

the population value which, if it were true, would give the highest probability or "maximum likelihood" that one would have observed the particular sample data actually observed. These three criteria do not always lead to the same answer. But despite these theoretical difficulties, in many common situations there are no important problems of how to estimate the population value. The corresponding sample values often serves as a fairly adequate estimator.

18.2 Confidence Limits

Having chosen a sample estimate of the population value we now have to consider how accurate it is. Because we are dealing with random samples, the answer will be in the form of probabilities, i.e. how *likely* we are to be wrong.

For example, suppose we want to estimate the population means from a random sample of 100 readings with a mean of 15, a standard deviation of 4, and an estimated standard error of $4/\sqrt{100} = .4$. We saw in Chapter 17 that we can make rather precise statements about the variability of the results obtained from different samples of this kind. For instance, because the sampling distribution of the mean would be approximately Normal, about 95 % of the sample means would lie within twice the standard error ($2 \times .4 = .8$) of the population mean μ. But since we do not know the value of μ, this does not tell us how close μ is to our observed sample mean of 15.

To get a better answer, suppose that our sample were actually one of the 95 % of all possible samples whose means lie within the two standard error limits $(\mu - .8)$ and $(\mu + .8)$. The difference between the unknown μ and our observed value 15 must then be less than .8. We can therefore turn this statement around and say that μ must in this case be less than .8 away from the observed value 15, i.e. that μ must lie between 14.2 and 15.8. Since this case occurs in 95 % of all samples, we can say that μ must lie within the 2 standard error limits (here $\pm .8$) of the observed sample mean for 95 % of all samples.

If we make this statement, we will be correct 95% of the time, i.e. the probability of being right is .95. This is commonly referred to as one's "confidence" in being right in saying that the population mean lies between 14.2 and 15.8, and these two-standard error limits are referred to as the "95 % confidence limits".

This kind of result is more complex than it appears on the surface. We are *not* saying that the unknown value μ has a probability of .95 of lying between the two-standard error limits. The value of μ cannot have a probability distribution, it has one particular value (with probability 1, if one likes). Instead, we have to make the more convoluted statement that μ will lie between the two-standard error limits for 95 % of all possible samples. Even

this is quite an achievement. From a single sample we can tell how accurate it probably is!

The Level of Confidence

By the same form of argument, we can determine other confidence limits. Since the sampling distribution of the mean is approximately Normal for sample sizes of 30 or more, we can use results such as those in Table 18.1. For example, we can say that the population mean will lie between $\pm$ 3.3 times the standard error and expect to be right in 99.9% of all possible samples and wrong in only 1 in a 1,000 cases.

TABLE 18.1 Descriptive Characteristics of the Normal Distribution

Distance from the mean in terms of the standard deviation	The proportion of readings lying within the stated limits
$\pm$ 2.0 s.d.	95%
$\pm$ 2.6 s.d.	99%
$\pm$ 3.0 s.d.	99.7%
$\pm$ 3.3 s.d.	99.9%

In our numerical example with a standard error of 0.4, we can therefore say that we expect the population mean to lie

between 14.2 and 15.8 with 95% confidence,
between 14.0 and 16.0 with 99% confidence,
between 13.7 and 16.3 with 99.9% confidence.

The notable feature of these results is that the risk of being wrong decreases sharply, but the size of the confidence limits increases relatively little. Thus although the population mean will lie outside the limits 14.2 and 15.8 in 5% of all samples, in all but 1 in 1,000 cases it will lie only just outside these limits, by up to 0.5 units. Even in that one case μ will mostly be *just* beyond the 13.7 and 16.3 limits. This means that even if our original sample had been that 1 in 1,000 case, our sample estimate of μ would not be much more inaccurate. We can therefore be pretty sure that the population mean lies roughly between 14 and 16, or fractionally outside these limits.

Prior Knowledge

In our numerical example we have supposed so far that we merely have a sample of 100 readings with a mean of 15 and standard deviation of 4, without saying what the data refer to. In practice, we would generally have some degree of prior knowledge of the situation,

For example, suppose that the population consists of the 50,000 employees of a large firm and we want to determine the average number of days in 1973 that each employee was absent from work. Without analysing any data we already know certain facts.

Firstly, assuming a 5-day week, we know the answer cannot lie outside the limits 0 and about 250. Secondly, we know the average rate almost certainly lies well below 100 unless there was a special circumstance, such as a 5-month strike, which someone would already have noticed and told us about. Thirdly, there will be results on absenteeism in previous years, for other firms in 1973, and so on. For example, if the firm's figures in the three previous years were 19, 17 and 18, the 1973 result should be something like 18, otherwise something exceptional must have occurred.

To get much closer to the truth, we have to measure what actually happened in 1973. We could reach the answer through the attendance records of all 50,000 employees (subject to any problems of measurement inherent in such records), but even a relatively small random sample of 100 would tell us a great deal.

If the sample gives an average of 15 days absent and the variation from employee to employee has a standard deviation of 4, the average for the total work-force is almost certainly not as high as 50 or as low as 5. This seems obvious. But how much narrower can we make the limits while still feeling "almost certain"?

The more specific we make the estimate, i.e. the narrower the limits, the more risk we run of their being wrong. For example, if we set the limits at 14.5 to 15.5, we cannot be "almost certain" since the population value may well lie outside. The important contribution of sampling theory is that it can give us a more precise measure of our risk of being wrong; e.g. 1 in 20 or 5 % for the ± 2 standard error limits of 14.2 to 15.8 noted above, and 1 in 1,000 for the ± 3.3 limits 13.7 to 16.3.

In recent years, certain theoretical procedures have been developed to try to improve such inferences from sample data still further by explicitly taking into account one's prior knowledge of the situation, such as the implication of last year's results, etc. The basic step in the so-called *Bayesian* approach is to try to translate this prior information into "prior probabilities" about the likely value of the unknown 1973 rate of absenteeism. Suppose one can somehow determine a zero probability that it is 13 or less, a .01 probability that it is 14, .1 that it is 15, and so on (with a peak probability of say .6 that it is 18 as in the preceding years). The "posterior probabilities" of the likely value are then-obtained by combining the prior probabilities with the information contained in the sample, using a well-known result in probability theory due to Thomas Bayes in the middle of the eighteenth century. This result essentially says that given a sample value of 15, the probability of the population mean taking some particular value, say 16,

is proportional to the *prior* probability that μ would be 16, multiplied by the probability that a sample result of 15 would occur if the population mean were *in fact* 16.

This method of "adjusting" prior probabilities in the light of sample evidence seems at first to make sense, but it is not widely used in ordinary statistical inference. One difficulty is fixing on the prior probabilities themselves. There is usually no objective way of determining them and they are generally referred to as "subjective probabilities". In fact ordinary people (scientists, administrators, etc.) do not usually think explicitly in *precise* probabilistic terms. Therefore adjusting prior probabilities which no one has been explicitly thinking about is after all perhaps not a very obvious method to use.

The usual place for prior information is not in sample estimation and the fixing of probabilities, but in determining the kinds of *hypotheses* one wishes to test.

18.3 **Testing a Statistical Hypothesis**

In analysing data one usually has some presupposition or hypothesis to test or explore. Typically, we might expect average absenteeism in 1973 to be 18 days, as in previous years. If the actual result for 1973 is 15 days, then our expectation was wrong and the hypothesis that it would be 18 in 1973 is rejected. This is straightforward.

However, problems arise if the 1973 result of 15 is based only on a *sample.* Perhaps absenteeism in 1973 really was 18 but we were "unlucky" in that our particular random sample of 100 employees happened to give a way-out result. How likely is it that the difference between the observed sample result of 15 and our initial hypothesis of 18 is only due to random sampling error? It is this narrow type of uncertainty problem that is tackled in testing statistical hypotheses.

The specific statistical hypothesis that is tested is usually called the "null hypothesis". In our example suppose the null hypothesis is that the population mean $\mu = 18$ (with a standard deviation of 4). Then the means of samples of 100 would have a Normal sampling distribution with mean 18 and standard error $4/\sqrt{100} = .4$. It follows that our observed value of 15 differs from the mean of 18 by $7\frac{1}{2}$ times the standard error. This is well beyond the 1 in a 1,000 probability level noted in Table 18.1. In fact this value would be observed less than once in a million random samples of 100, if the samples really came from a population with mean 18.

Given that the observed sample is so very unlikely if the null hypothesis were true, we "reject" this hypothesis. After all, it was only a hypothesis to be pitted against the facts. Sampling error *might* have accounted for the

discrepancy between 15 and 18, but we have now calculated that this is extremely unlikely.

To illustrate the opposite kind of result, suppose we started with a null hypothesis that the population mean $\mu = 15.5$. Then the observed sample value of 15.0 lies just over one standard error away. For a Normal Distribution, values more than one standard deviation away from the mean occur in almost 30% of all samples (about two-thirds of the readings lie *within* one standard deviation). Therefore the observed sample result is quite likely if the hypothesis that $\mu = 15.5$ is true, and we have no reason to reject this hypothesis.

The Level of Significance

In the two cases just discussed, the null hypotheses were either highly unlikely or highly likely in the light of the sample data.

We now consider a less clear-cut type of result, a sample mean which is two standard errors away from the hypothesised population value. With a Normal Distribution sample results further from the mean occur in only 5% of all possible samples. This is *fairly* unlikely: in any 20 empirical studies there would be only one such result. Therefore at the 5% probability level it is conventional to reject the null hypothesis and to call the observed result "significantly" different.

But such a cut-off point is arbitrary. One should never think of a two-standard error result as being strongly "significant", while with a 1.9 sample result the null hypothesis is "acceptable". Instead, anything like a 1 in 20, a 1 in 15, or a 1 in 25 chance should be considered rather unlikely.

The difficulty is that a 1 in 15 or a 1 in 20 result is not "impossible", therefore one might wrongly reject a null hypothesis when it is in fact true. One could cut the chances of rejecting a true null hypothesis by adopting a more stringent level of significance, say 1 in 100 or 1 in 1,000. But this would *increase* the chance of accepting a wrong null hypothesis.

These so-called "errors of the first and second kind" (either rejecting a true null hypothesis or accepting a false one), together with the notion of the "power" of a test of significance in discriminating between alternative hypotheses, are part of the theory of statistical inference which was highly developed by Jerzy Neyman and Egon Pearson around 1930, following Fisher's lead in the 1920's.

The technical arguments simplify considerably because a small change in the difference between an observed sample value and the null hypothesis has a marked effect on the probability of whether the difference occurred merely by chance. Problems of interpretation therefore only arise when sample observations differ from the null hypothesis by about $1\frac{1}{2}$ to $2\frac{1}{2}$ times the standard error. This is quite a narrow "twilight" range. More discrepant

sample values almost unambiguously lead to the *rejection* of the null hypothesis, with a probability of less than 1 in 100 of its being true, while sample values differing by well under two standard errors must equally unambiguously be regarded as being *consistent* with the null hypothesis.

When sample values fall into the twilight range one usually either rejects the null hypothesis "tentatively", or "has doubts about it". There is always the possibility of taking another set of readings to reduce these doubts.

18.4 **The Choice of Hypothesis**

The choice of the null hypothesis is the crucial feature in tests of significance. A "significant" result means that the null hypothesis has to be rejected, it was the wrong hypothesis. The analyst was therefore wrong in choosing it, often presumably either through ignorance of his subject matter or incompetence. The results are not as he had thought. With sample data this might be caused by a rather "unlucky" random sample, but the purpose of establishing a result as "significant" is to show that this particular possibility is highly unlikely. Thus the null hypothesis was almost certainly really wrong.

Occasionally an unexpected observation can be important; that is how some discoveries are made (Fleming's discovery of Penicillin is a popularly quoted example). But it is not advisable to make a habit of collecting data that differ from one's expectations. It only shows that one is consistently incompetent in selecting the appropriate hypotheses to investigate.

Despite this, some tradition has grown up in the last few decades that "significant" results are "good" results. Findings are reported as being 5%, 1% or even 0.1% significant, and the symbols *, **, *** tend to be attached to the results (as in hotel guides for the ignorant tourist, as Sprent (1970) has put it).

But it is easy to choose a hypothesis that will amost certainly differ from an observed result. The more absurd the hypothesis the more "significant" the observed sample result will be. Instead, one's choice of hypothesis should depend on one's prior knowledge or expectations. Then there is little problem. For example, if the rate of absenteeism in previous years has been about 18 days, *that* is the relevant hypothesis, unless additional prior information leads one to expect something different now. Again, in analysing the height and weight data of some group of children, the generally appropriate hypothesis is the result $\log w = .02h + .76$, unless one has relevant additional information (such as that the children are older girls, or babies, or undernourished). However, if one's prior knowledge is not clear-cut, then *that* is part of the situation and one needs to say so. The empirical study will then be more of a fishing expedition to throw some light on the situation, rather than a rigorous test of some crucial hypothesis. All one needs to do is to attach confidence limits to the estimated values.

The problem of choosing an appropriate null hypothesis is highlighted by a particular form of null hypothesis: that the population value should be zero (probably the reason for the name "null hypothesis"). A typical example is that there should be no difference between two mean values, e.g. between the responses of a treated and a control group in a clinical trial.

But in most cases there is nothing objective about the choice of such a null hypothesis of zero. The analyst usually chooses it to provide "a fig-leaf of scientific respectability", not because of anything he *knows.* He does not generally expect the null hypothesis of zero difference to be true. Few clinical trials are carried out because the drug is expected to have no effect. Testing such a no-effect null hypothesis is mostly a game: the analyst wants to prove himself "scientific".

In a clinical trial of a drug, the scientist's expectation or hypothesis might be that the drug will *decrease* blood pressure (not increase it), and that it will do so by about 10 units. That is what other more or less similar studies, or theory, or his gut-feeling, lead him to expect. In dealing with a sample of treated patients, a decrease of 10 units is therefore the appropriate statistical null hypothesis to test. Only if he is checking to confirm the previous *failure* of a drug, or the absence of unwanted side-effects, would the "no-difference" type of null hypothesis be relevant.

The fear remains that the analyst might mislead the reader (or himself?) by ignoring the possibility that an apparently positive sample result. might still only represent a zero situation in the population as a whole. But establishing that an observed result agrees with one's prior hypothesis within the limits of sampling error does not absolve one from taking note of this sampling error. If the sampling error is so large that zero is included in the two standard error confidence limits, then the sample evidence on its own cannot exclude the possibility that there really is no difference. The implication is that the investigation was badly designed-the samples should have been large enough to lead to a clear distinction between the expected "positive" null hypothesis and zero.

"No-difference" tests of significance are widely carried out with correlation and regression coefficients. The usual null hypothesis tested is that of a zero value (no correlation) in the population. For example, given that the observed correlation between x and y in the sample is 0.3, could it be that there is really *no* correlation in the population? But the analyst who reports fifty "significant" correlation coefficients has merely picked the wrong hypothesis fifty times. As already said, he should not make a habit of it.

One reason for the popularity of the zero null hypothesis here is that its meaning is clear: x and y are *unrelated.* If the results are established as "significant" (i.e. *not* zero), too often the meaning of the result is left unexplored. What does a correlation of 0.6 *mean,* or a regression coefficient of 2.5? How do these values compare with previous experience? Is a generalisable

pattern of results building up? Unfortunately such questions tend to remain unanswered if the emphasis is merely on establishing that the coefficients are effectively non-zero.

There is nothing remarkable in finding that x is related to y in some particular set of data. But somehow reporting a "significant" correlation of 0.3 based on a sample of 100 seems to be treated as more important than finding a correlation of 0.3 in the whole population, or in a large sample of 10,000 (where anything is "significant").

The question raised by such a "significant" result is "so what?". Finding a coefficient which is significantly different from zero is at best a *starting-point* in the analysis. As Gatty (1966) has put it

> "Statistical significance of a correlation or regression coefficient merely means that there is a pretty good chance that it is in fact a number different from zero. One should not exaggerate the worthwhileness of a coefficient simply because it probably differs from zero."

18.5 Empirical Variation

Except in the earliest stages of studying a topic, there is generally a great deal of previous empirical information about the variability of the material in question. Thus the reliability or statistical significance of a sample result can be judged in other ways than just from that result itself.

For example, consider data on the heights and weights of a random sample of 100 girls from some specified larger population

Average height : 49 inches,

Average weight : 56 lbs.

Without any other information about the girls the only null hypothesis that can be used is the general relationship

$$\log w = .02h + .76 \pm .01,$$

discussed in Fart II of this book. We are effectively predicting that this equation should hold again.

The logarithm of 56 is 1.75, so the new result deviates from the hypothesis by

$$(1.75 - .02 \times 49 - .76) = .01 \text{ log lb units.}$$

This deviation is very much in line with all the earlier results, where the means of various groups of boys or girls gave a fit to within average limits of about .Ol log lbs. In that sense, the deviation for the new data is not different from the null hypothesis.

But if we look at the new data purely from the point of view of random sampling, we have to note that the deviations for *individual* children generally have a standard deviation of about .04 (see Section 6.4). Hence the standard error of the mean for the sample of 100 girls is $.04/\sqrt{100} = .004$. The observed discrepancy of .Ol is therefore statistically significant at almost the 1% level of probability. This means that if we measured a larger sample or the whole population of these girls, we would have to expect the result to deviate from the line $\log w = .02h + .76$. Thus the deviation was not only due to sampling error.

We can resolve the apparent contradiction between these two conclusions by noting that most of the height and weight means discussed in Part II also differ "significantly" from the equation $\log w = .02h + .76$. The deviations there were also not due to sampling alone; there were other factors involved. General experience shows that observed data do not fit any model exactly. At best they fit only within some close, more or less irregular, limits. Sampling errors usually account for only a small part of these deviations.

One usually judges discrepancies in new data against the general run of discrepancies found previously, If these earlier discrepancies have not yet been explained, i.e. their causes *in addition* to sampling error, one can hardly say more about the *new* result than whether it fits in with previous experience.

But if the new result is beyond the usual limits, e.g. a height/weight discrepancy of .06 log lbs, one would have to ask if it were merely due to sampling errors (which it could be in a very small sample or for an *individual* reading) or whether some additional "real" factor were also involved. In most reasonably well-developed areas of study, tests of significance therefore perform mostly a negative function, a form of hygiene, to establish whether some unusual result is merely an unlucky sampling error.

18.6 Specific Tests of Significance

There are various commonly used procedures to test statistical significance, i.e. to test whether the difference between a hypothesis and a sample result is due to a *real* difference in the population or merely to the errors of random sampling. We shall start here with tests involving sample means and then briefly cover tests involving variances, correlation and regression coefficients, goodness-of-fit procedures, and contingency tables.

A technically useful device in many tests of significance is the number of "degrees of freedom" in the data. This often identifies the particular sampling distribution to be used. For the χ^2-distribution discussed in Chapter 17 the degrees of freedom were the number of groupings into which the data were classified, minus the number of "constraints" on the data caused by fitting a theoretical model or comparable calculations. For detailed *quantitative* data (in contrast to such grouped or *qualitative* data), the degrees of freedom are generally defined as the sample size minus the number of constraints. This accounts for the divisor $(n-1)$ commonly used to calculate the variance of the

deviations from the mean; one degree of freedom has been used to estimate the sample mean.

The Mean of a Sample

To test the statistical significance of hypotheses about the mean of a sample of n readings, we have to use the l-ratio of the sample mean m to its estimated standard error, $s/\sqrt{n}$:

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

This ratio follows Student's t-distribution with $(n-1)$ degrees of freedom if the population sampled is Normal with zero mean. Table 18.2 summarises the three most commonly used significance levels for t-distributions with various degrees of freedom. Thus 95 % of the means of samples of $n=6$ readings (i.e. 5 degrees of freedom) lie less than 2.6 times the standard error from the population mean. The t-distribution significance levels are very similar to those of the Normal Distribution with unit standard deviation except for very small sample sizes ($n=10$ or less). Thus, in practice, one can generally use the Normal Distribution values, as shown in the last column of Table 18.2.

TABLE 18.2 Common Significance Levels for the t-distribution
(Multiples of the sample standard deviation within which 95%, 99% or 99.9% of the values lie)

	Degrees of freedom							
	1	2	5	10	15	20	30	Large*
95%	13	4	2.6	2.2	2.1	2.1	2.0	2.0
99%	64	10	4.0	3.2	2.9	2.8	2.7	2.6
99.9%	640	32	6.9	4.6	4.1	3.8	3.6	3.3

* As for the Normal Distribution

If the population data are non-Normal, the distribution of sample means for small samples will not follow a t-distribution. But for large enough samples the distribution will still be approximately Normal. What is "large enough" depends on the nature of the population distribution. Even in fairly extreme cases, e.g. sampling from a skew Poisson or Negative Binomial type of distribution, the distribution of sample means for $n=50$ tends to be close to Normal.

To test whether the mean m of an observed sample differs significantly from a hypothesised value μ, we look in Table 18.2 to find the probability with which the value

$$t = \frac{m-\mu}{s/\sqrt{n}}$$

will be exceeded. For example, if a sample of 6 workers has an average absentee level of 14 days with a standard deviation of 4, the l-value against the hypothesised rate of 18 days is $(14-18)/(4/\sqrt{6})$, or approximately 2.5. This virtually reaches the 5 % significance level of 2.6 for $(n-1)=5$ degrees of freedom. Thus if the population mean were 18, values at least as different as 14 would occur in only 1 out of 20 samples of $n=6$. The probability that this would occur by chance errors in the sampling is therefore sufficiently unlikely that one would generally reject the null hypothesis that $\mu=18$.

An alternative hypothesis with μ somewhat less than 18 would make the observed result appear more probable. It is not necessarily clear *what* alternative hypothesis one should consider. But in the absence of other information, the most likely value would be 14, i.e. simply the observed sample mean.

The Difference Between Two Means

To determine the significance of the difference between two means, m_x and m_y, of two independent samples of n_x and n_y readings with variances s_x^2 and s_y^2, we calculate the t-statistic

$$t = \frac{(m_x - m_y) - (\mu_x - \mu_y)}{\sqrt{(s_x^2/n_x + s_y^2/n_y)}},$$

Here $(\mu_x - \mu_y)$ is the hypothesised difference in the two population means. We then assess the numerical value of the t-statistic against a t-distribution with $(n_x + n_y - 2)$ degrees of freedom, since *two* means have been fitted. (This is again virtually identical with a Normal Distribution with unit standard deviation when the degrees of freedom are greater than 10 or 20 or so.)

If the null hypothesis is that there is no difference in the population means, then $\mu_x - \mu_y = 0$ and the numerator of the t-statistic simplifies to $(m_x - m_y)$. In such cases the hypothesis being tested usually says that the populations sampled also have the same variances and the same shape, i.e. that their properties are altogether the same. Therefore one could calculate a single "pooled" estimate of the variance for the two samples combined (on the basis that the hypothesis is true). This consists of the average squared deviations of the $n_x + n_y$ readings from their *overall* mean $(n_x m_x + n_y m_y)/(n_x + n_y)$. This leads to a t-distribution with $(n_x + n_y - 1)$ degree of freedom because only *one* mean is fitted. The t-test is then slightly more sensitive, but the gain is almost negligible, especially with relatively large samples.

More than Two Means-Simple Analysis of Variance

To test whether there are significant differences among the means of three or more samples one can use the Variance-Ratio or F-statistic (named after Sir Ronald Fisher):

$$F = \frac{\text{Variance estimate based on means}}{\text{Variance estimate based on individual readings'}}$$

Table 18.3 sets out data for random samples of four readings from three larger populations, A, B, and C.

TABLE 18.3 Three Samples of 4 Readings Each

	From Population		
	A	B	C
	2	3	4
	4	4	6
	4	6	6
	6	7	7
Mean	4	5	6

The variances of the three samples are 8/3, 10/3, and 8/3, which look very similar. They can therefore be averaged into an overall estimate of the variance σ^2 of the individual readings within each population, giving (26/3)/3 or 2.9, on the null hypothesis that the three populations have the same distributions (i.e. the same means, the same variances, and the same shapes). This is the denominator of the F-statistic above.

We can also estimate the variance σ^2 of the individual readings from the mean values of the three samples. The observed variance of these three means is

$$\frac{(4-5)^2+(5-5)^2+(6-5)^2}{2}=1$$

The null hypothesis says that the three populations are the same. So, if the null hypothesis is true, in effect we have three sample means from the same population, For samples of $n=4$ the variance of the sampling distribution of these means would be $\sigma^2/4$. Therefore we can estimate σ^2 by multiplying the observed variance of the sample means by 4, giving a value of 4 x 1 = 4.0, the numerator in the F-statistic above.

We now have two possible estimates of the population variance, 4.0 and 2.9. The question is whether the difference is merely due to sampling error (i.e. the null hypothesis) or whether the larger value of 4.0 reflects real differences among the population means. We can test the null hypothesis by forming the variance-ratio

$$F=\frac{4.0}{2.9}=1.4,$$

(where the larger value of the variance value is always put on top). This is a useful procedure because the sampling-distribution of the F-statistic is known.

If the null hypothesis is true, then these two variances should be equal in the population and the ratio will be 1. But in random samples the estimates will vary and their ratio will follow an F-distribution with, in our case, 2 and 9 degrees of freedom. (The variance estimate based on the means has *two* degrees of freedom since the overall mean 5 has been estimated. The variance estimate based on the individual readings has three degrees of freedom for each sample and hence a total of *nine*.)

From tables of the F-distribution (as given in books of statistical tables, e.g. Fisher and Yates, 1957; Lindley and Miller, 1966) we see that an F-value of 1.4 with 2 and 9 degrees of freedom occurs quite often. So although the three sample means in Table 18.3 differ, we could expect these differences to occur with random sample data, given the quite large variability of the *individual* readings. There is therefore no reason to reject the null hypothesis that the three populations have the same means.

If the null hypothesis were *not* true and the population means *were* different, the variance calculated from the sample means would be greater. Then the F-ratio would be greater than expected from random sampling alone. Thus in our example we would expect an F-ratio above the 5% probability value of 4.3.

The test procedure outlined here is a simple example of the Analysis of Variance. It is called this because it analyses the total variance of the data into separate variance components (here "between samples" and "within samples"). These procedures were first developed by Fisher in the 1920's. Since then they have been greatly elaborated in connection with statistically designed experiments and described in various specialist texts (e.g. Fisher, 1935; Cochran and Cox, 1957; Cox, 1958).

***The Problem* of Selection.** If the F-test is significant, one or more of the population means must be different from the others. Thus if the means in Table 18.3 had been 4, 5, and 10 (with corresponding individual readings), the F-test would have been significant. This

is presumably because population C with its sample mean of 10 differs from populations A and B, as common sense suggests.

But such a conclusion is technically difficult to test for its precise degree of significance. Ordinary r-tests or standard error calculations for the difference of two means cannot be applied. These test procedures are not designed for situations where a large difference has been specially selected for testing after inspection of the data, or where *many* differences are tested (e.g. A against B, A against C, etc.).

For example, suppose we have 7 sample means to analyse. We have to compare $(7 \times 6)/2 = 21$ pairs of means. If all the population means were equal, 1 in 20 of the sample results would still be beyond the appropriate 5% limits. If we select the biggest observed difference out of the 21 pairs of means, it may be no more than the *normal* 1 in 20 case. There are additional complications because the 21 comparisons are not all independent. *One* exceptional mean value out of seven would lead to six exceptional *differences.*

Although certain procedures have been put forward for dealing with such problems, none seems to be commonly accepted in practice. The problem is less serious than it might seem because such comparisons are at most needed at early stages of a study. Once the subject matter becomes structured in the light of previous findings, the *uninformed* search for significant differences becomes unimportant.

Correlated Readings

Suppose x and y are paired readings, e.g. heights of brothers and sisters, or readings on the same patients before and after a clinical drug trial, as in Table 18.4. The differences in the sample means m_x and m_y could be due either to the treatment or to the effects of random sampling.

TABLE 18.4 Correlated Readings: Five Patients Before and After Treatment

	Patients					
	A	B	C	D	E	Mean
Before	28	19	19	17	17	20
After	20	15	14	17	14	16
Difference	8	4	5	0	3	4

To test the significance of the differences in sample means for *n pairs* of readings we can use the t-statistic

$$t = \frac{m_x - m_y}{s_{x-y}/\sqrt{n}},$$

with $(n-1)$ degrees of freedom. Here s_{x-y} stands for the standard deviation of the paired difference $(x-y)$, i.e.

$$s_{x-y} = \sqrt{\left[\frac{\text{Sum } \{(x-y)-(m_x-m_y)\}^2}{n-1}\right]}.$$

In our numerical example, s_{x-y} is $\sqrt{(34/4)} = \sqrt{8.5} = 2.9$, so that

$$t = \frac{4}{2.9/\sqrt{5}} = 3.1.$$

With 4 degrees of freedom this falls almost exactly at the 5% level of significance (interpolating in Table 18.2). Since the difference in means is larger than would usually occur from sampling error alone, the drug therefore appears to have been effective.

There is positive correlation between the before and after scores (e.g. A is high on both, and E low), and so the assessment of the mean difference, $20 - 16 = 4$, is less subject to sampling error variation than if *different* patients had been tested before and after the treatment (where on the same readings the standard error would have been 5.2 instead of 2.9). The gain in sensitivity is typical of the more efficient statistical "design" of the study, using the same patients before and after. The design makes use of the fact that in this case individual differences in response levels before the treatment tend to recur *after* the treatment.

Because the analysis here essentially involved only the readings in the "difference" row, the appropriate degrees of freedom are 4, i.e. five differences minus 1 for the mean of the differences. (Sometimes the same test procedure is described in terms of the individual before and after readings plus the correlation coefficient between them.)

A More Complex Analysis of Variance

A further question for the data in Table 18.4 is whether the patients really differ *significantly* from each other. Their *apparent* tendency to differ (e.g. A high both before and after, and E low) might only be a fluke due to random sampling variation for the particular 5 patients sampled and not typical of the population sampled.

This can be tested by calculating an F-ratio with a numerator based on the variance of the mean values offigures for each patient (i.e. 24, 17, 16.5, 17 and 15.5) multiplied by 2, since each value is a mean of 2 readings (the σ^2/n effect). The denominator is again s^2_{x-y}. The appropriate degrees of freedom of the F-ratio are 4 and 4. (We could also have tested the before and after treatment effect, $m_y - m_x$, by the F-ratio. In the case of two means, the F-ratio is the square of the t-statistic used in the preceding section.)

These are simple examples of more advanced types of Analysis of Variance. The basic concepts will be outlined further in Chapter 19 in connection with the design of experiments.

Lawlike Relationships

Tests of significance for lawlike relationships are mainly tests of the mean values $\bar{x}$ and $\bar{y}$ of the different sets of data analysed.

An early problem is that we may need to establish whether there is any relationship at all between variables x and y. To do this, one can first test the difference of the means $\bar{x}_1 - \bar{x}_2$ for two (or more) sets of data against the null hypothesis of no difference, along the lines already discussed here. Secondly, one similarly tests the differences of the means $\bar{y}_1 - \bar{y}_2$. If both tests are significant, i.e. significant variation in x and significant variation in y, then the apparent correlation between $\bar{x}$ and $\bar{y}$ in the various samples must be significant.

To establish whether a particular pair of mean values $(\bar{x}, \bar{y})$ based on n pairs of readings differs significantly from a previously established relationship $y = ax + b$, one tests the l-statistic

$$t = \frac{\bar{y} - a\bar{x} - b}{\text{stand. dev. } (y - ax - b)/\sqrt{n}}$$

with $(n - 1)$ degrees of freedom. The analysis is similar to that for correlated readings because it depends on the differences between x and y. However, here one needs to

adjust the scales of measurement of x and y, i.e. working with the deviations $(y - ax - b)$ instead of $(y - x)$.

The same test procedure can be used to establish whether an *individual* pair of readings (x, y) is significantly different from the relationship. Then the values of x and y are inserted in the numerator instead of the means and $n = 1$. The test determines whether the observation differs from the line by more than 95% or 99% of the readings generally do, or whatever cut-off criterion one wishes to use.

Correlation and Regression Coefficients

Tests of significance of correlation and regression coefficients are mostly of the null hypothesis of zero correlation or regression.

The standard error of the product-moment correlation coefficient r for a sample of n from a bivariate Normal Distribution with zero correlation in the population can be estimated by the formula $(1-r^2)/\sqrt{(n-1)}$, and the sampling distribution of r is then approximately Normal for samples of $n = 100$ or more. For an observed correlation coefficient of $r = .2$ based on a sample of $n = 100$, the estimated standard error is therefore $(1-.04)/\sqrt{99} = 0.1$; the observed sample value of .2 is therefore twice the standard error from 0 and hence significant at the 5% probability level. For smaller samples, one can use the fact that the quantity $r\sqrt{(n-2)}/\sqrt{(1-r^2)}$ is distributed as Student's t-distribution with $(n-2)$ degrees of freedom, if the population value ρ of the correlation is zero.

To test hypotheses of *non-zero* values of the correlation ρ in a bivariate Normal population, Fisher showed in 1915 that for different sample values r, the quantity $\frac{1}{2}\log_e\{(1+r)/(1-r)\}$ follows a Normal Distribution to a close degree of approximation, with a mean of $\frac{1}{2}\log_e\{(1+\rho)/(1-\rho)\}$ and a variance $1/(n-3)$. This sampling distribution can therefore be used for tests of non-zero null hypotheses.

The theory of testing regression *coefficients* has various complexities, especially if one is dealing with time-series as in econometrics or other forms of data where successive readings may be serially correlated. But for Normally distributed independent residuals, the basic standard error formulae for the slope-coefficient $a = \text{cov}(xy)/\text{var}(x)$ in the linear regression $y = ax + b$ is

$$\frac{\text{stand. dev. } (y - ax - b)}{\text{stand. dev. } (x)\sqrt{(n-2)}},$$

and the standard error of the intercept-coefficient $b = \bar{y} - a\bar{x}$ is

$$\frac{\text{stand. dev. } (y - ax - b)}{\sqrt{n}}.$$

These expressions could therefore be used in tests against a specified null hypothesis for large samples, where the sampling distributions should be approximately Normal.

The Variance

So far we have concentrated on tests related to mean values. These include correlational analyses of how one variable tends *on average* to vary with another. But in studying the scatter of readings about a mean, we may need to test hypotheses about variances. For example, does a particular sample variance s^2, based on a sample of n readings, differ significantly from a hypothesised population value σ^2, or is the observed difference likely to be due only to random sampling errors?

If the readings come from a more or less Normal Distribution with variance σ^2, we test the ratio of the observed to the hypothesised value of the variance:

$$\frac{s^2}{\sigma^2}.$$

If the null hypothesis is true, the sampling distribution of this ratio follows a χ^2-distribution with $(n-1)$ degrees of freedom. No general test procedure has been developed for variances from markedly non-Normal data.

Two or More Variances

When comparing two or more samples of readings, we want to know not only whether their *means* differ significantly, as drawn out before, but also whether their *variances* are the same. A simple test of whether the variances, s_1^2 and s_2^2, of two samples of n_1 and n_2 readings from two different (Normal) distributions differ significantly uses the variance ratio

$$\mathrm{F} = \frac{s_1^2}{s_2^2}.$$

This ratio follows an F-distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom (using the larger variance as the numerator).

For more than two variances, s_i^2, all having the same number of degrees of freedom, r, one can use Bartlett's index M for "variance homogeneity":

$$\mathrm{M} = r\{\log \bar{s}^2 - \mathrm{Sum}\,(\log s_i^2)\}$$

where $\bar{s}^2$ is the average of the $i = 1$ to q different variances. This index can be tested against a χ^2-distribution with $\{1+(q+1)/3qr\}$ degrees of freedom, where q is the number of variances.

If the result is not significantly large, the different samples should come from populations with more or less the same variance (e.g. Bartlett, 1937). However, if the samples come from non-Normal distributions, this test can give apparently significant results even though the different population variances might be equal. Nowadays the test does not seem to be in common use, but there appears to be no adequate substitute.

The general problem of establishing whether different sets of sample data have the same scatter (and whether the observed differences in scatter are significant) can be important. In practice variance heterogeneity is often linked to non-linear relationships, so that the problem is also solved by a non-linear transformation of the variable (e.g. taking logs), which can largely eliminate the heterogeneous scatter.

Goodness of Fit

The scatter of readings about some fitted theoretical model is 'often arranged in n groupings of readings, as we did in Section 17.4. With sample data we can test the "goodness of fit" of the model by assessing the statistic

$$\mathrm{Sum}\frac{(\mathrm{Observed} - \mathrm{Theoretical\ Frequency})^2}{\mathrm{Theoretical\ Frequency}}$$

against a χ^2-distribution with $(n-k-1)$ degrees of freedom, where k is the number of parameters that had to be fitted in the theoretical model. A non-significant χ^2-value means that the observed deviations from the model are likely to be due to chance alone.

Thus they would not be expected to generalise to other samples or to the population as a whole.

The value of χ^2 must be calculated from the actual frequencies and not from percentages or proportions. The data can be grouped arbitrarily but the grouping must be essentially decided *before* the data are collected (orsummarised) to avoid any danger of deliberately biasing the χ^2-value. Another requirement is that the theoretical frequency in any one grouping interval should not be less than 5, otherwise the χ^2-approximation fails to apply. If necessary, adjacent groups can be combined to produce the minimum theoretical frequency and the degrees of freedom reduced accordingly.

The arithmetical calculations for this test are the same as for contingency tables, as is explained below.

Contingency Tables

A special case of goodness of fit problems arise with a "contingency table". This is a two-way cross-classification of *qualitative* data. For example, one classification in a "2 x 2" table might be between male and female, the other between brown-eyed and not-brown-eyed. In general, one may cross-classify k classes one way by m classes the other.

The only null hypothesis that is commonly testedstatistically is that the classifications are independent of each other, i.e. that in the population, the proportion of men who are brown-eyed equals the proportion of women who are brown-eyed. With sample data the two proportions will usually not be precisely equal, and the question is whether this difference is only due to the sampling errors.

Table18.5 gives an example of a 2 x 2 table. Two machines produce the same product and each machine tends to yield a proportion of defective items. From a week's production of each machine, 20 and 30 items are sampled at random. When tested, 10 items from the first sample and 5 items from the second are found to be defective. The samples suggest that the first machine, with 10/20 or 50% defectives, is worse than the second, with only 5/30 or about 17% defectives. The question is whether this difference is significant, i.e. was the first machine worse than the other *throughout the week?*

TABLE 18.5 The Incidence of Defectives in Random Samples of Items Produced by Two **Machines**

	1st machine	2nd machine	Total
Defective	10	5	15
Non-defective	10	25	35
Total samples	20	30	50

On the null hypothesis that there was no difference in the machines' proportions of defective items during the week, the best estimate of this general proportion of defectives is from the total column, that 15/50 or 30% were defective. This leads to the theoretical "expectation" that 30% of the items for each machine should have been defective, 6 out of 20 for the first and 9 out of 30 for the second, as shown in Table 18.5a. (These values tend to be called "expected" rather than "theoretical" because they merely refer to what is expected on the null hypothesis and do not refer to any deep theory or model.)

TABLE 18.5a The Observed and "Expected" Frequencies

	1st machine		2nd machine		Total
	Obs.	Exp.	Obs.	Exp.	
Defective	10	6	5	9	15
Non-defective	10	14	25	21	35
Total	20	20	30	30	50

To test the hypothesis we calculate a χ^2-statistic as in a goodness-of-fit test. This is the sum of the quantities

$$\frac{(\text{Observed} - \text{Expected Frequency})^2}{\text{Expected Frequency}}$$

for each "cell" of the 2 x 2 table. Adding these numbers gives

$$\chi^2 = \frac{(10-6)^2}{6} + \frac{(10-14)^2}{14} + \frac{(5-9)^2}{9} + \frac{(25-21)^2}{21} = 6.4$$

This figure can be tested for significance against a χ^2-distribution with 1 degree of freedom. [Given the marginal totals (15, 35, 20, and 30) in the 2 x 2 table, only 1 figure in the body of the table can vary freely; the others are determined. For a k x m table generally, the appropriate degrees of freedom are $(k-1)(m-1)$.]

The 1% value of χ^2 is 6.6, so a value as high as 6.4 would occur in only just over 1% of samples. Hence we reject the null hypothesis as being unlikely and accept that there was a difference between the two machines in the week's production sampled.

18.7 Summary

Statistical inference means using the data in a single sample to estimate the values of the population sampled.

In simple cases like the mean, the sample value provides a good estimate of the corresponding population value. But this simple approach may not work as well with other parameters, so more complex estimation procedures have to be devised.

In general, an estimate from a sample will not equal the population value. If the sampling distribution is more or less Normal, the standard error of the estimate provides the basis for its "confidence limits"; this is so for the mean and most other descriptive measures, if the sample size is large enough. For example one would be right 95% of the time in asserting that the population value lies within ± 2 standard errors from the observed sample value. Even if the population value were outside these limits (which occurs for 1 sample in 20), it would not lie *far* outside.

In tests of significance, the question posed is whether the difference between the observed sample value and some hypothesized population value (the "null hypothesis") is due only to random sampling errors, or whether the difference would generalise to other samples and the population as a whole.

The hypotheses tested should generally derive from previous empirical data and should reflect what such prior knowledge has led one to expect. A significant difference therefore means that one's expectation was wrong.

If previous data exist, then extensive *empirical* evidence about the general variability of the data should have been built up. One therefore need not have to rely on *theoretical* sampling theory to provide an inference about the new sample's likely variability.

CHAPTER 18 EXERCISES

Exercise 18A. Significant with a Larger Sample?

An observed sample mean of 5 has a standard error of 3. If the null hypothesis is zero, the sample mean is within the $\pm$ 2 standard error limits and hence not significantly different.

With double the sample size, the standard error would be reduced by a factor $\sqrt{2} = 1.4$, i.e. from 3 to $3/1.4 \div 2$, and a value of 5 would then be more than 2 standard errors away from zero.

Is it right to suppose that the observed sample mean of 5 would have been significant with a larger sample?

Discussion.

No. Another sample would generally have a different mean. And a *larger* sample would (on the null hypothesis) generally have a mean closer to zero. The standard error would be smaller, but so would the observed mean value, so it would generally still not be "significant".

Exercise 18B. The Nature of Statistical Hypotheses

Are statistical hypotheses different from the normal hypotheses of science?

Discussion.

Yes. Examples ofscientific hypotheses are Einstein's theoretical deduction that light waves are bent by gravity (a popular example), that a certain drug will reduce people's blood pressure, or that the height and weight of some children will follow the relationship $\log w = .02h + .76$. In contrast, a statistical hypothesis is concerned with whether in random sample data an observed deviation from a scientific hypothesis is real or only due to an error in the sampling. (The question of statistical inference really arises only with small samples. With large samples the standard error of the sample estimate is generally small, so that anything except the smallest overt differences from the hypothesised value will generally be real.)

Exercise 18C. The Full Null Hypothesis

In the example in Section 18.3, the null hypothesis was an average rate of absenteeism of 18 days. Is it possible to deduce a sampling distribution from this information about the population?

Discussion.

In principle, the answer is no; in practice, it is often yes.

To deduce the sampling distribution of the average rate of absenteeism (along the lines discussed in Chapter 17), the null hypothesis must state *fully* what the population is expected to be like. It must give the form of the frequency distribution, e.g. Normal, Poisson, or whatever, and the parameters of the distribution, e.g. its standard deviation as well as its mean μ.

However, certain simplifications arise. Unless the sample size n is very small, the sampling distribution will be approximately Normal whatever the shape of the population distribution. Next, the sampling distribution will have a standard deviation effectively equal to $s/\sqrt{n}$, where s is the observed standard deviation of the *sample.* Hence, for purposes of testing statistical hypotheses about the mean, it is usually unnecessary to specify either the form of the population distribution or its standard deviation (unless the sample is very small). This explains why the null hypothesis is usually described as merely saying something about the population mean.

Nonetheless, the analyst will be concerned with the "shape" of the population distribution and the amount of scatter, in order to describe and understand his data. If he has prior expectations or hypotheses about these, he may need to "test" them if the observed sample data look very different.

Exercise 18D. One 95% Confidence Interval or Many?

The 95 % confidence limits in the numerical example in Section 18.2 were 14.2 and 15.8. Does this mean that μ lies between **14.2** and 15.8 for 95 % of all samples?

Discussion.

No. The calculation of a particular set of confidence limits depends on the particular sample results used. The sample mean of 15 and standard error of .4 gave the 95 % limits of 14.2 to 15.8. Another sample from the same population might have a mean of 15.3 and a standard error of .3. This would lead to 95% confidence limits of 14.7 and 15.9.

Probability statements about confidence limits therefore do not refer to a single set of numbers. The situation is more complex. If for any particular sample one can say that the population mean lies between the 95 % confidence limits, then for 95 % of all samples the corresponding statement for each sample will be correct (i.e. 14.2 to 15.8 for the first sample, 14.7 to 15.9 for the second one, and so on).

This may seem almost intolerably complex. But in practice most confidence limits are numerically similar because most sample means are fairly similar. After all, 95 % of them lie within $\pm 2\sigma/\sqrt{n}$ of the population mean μ! It follows that the rough-and-ready interpretation of confidence limits, that the population mean lies in the range 14.2 to 15.8 with a probability of .95, will be close to the truth.

The choice is between making a statement which is true but so complex that it is almost unactionable, and one which is much simpler but not quite

correct. Fortunately the *content* of the two kinds of statement is very similar.

Exercise 18E. A 4 × 6 Contingency Table

Table 18.6 shows the incidence of light, medium, or heavy attacks of influenza in random samples from six occupational groups. Does the incidence of influenza in the six populations really differ?

TABLE 18.6 Severity of Influenza in Random Samples of Adults from 6 Occupational Groups

	Occupation						
	White collar	Skilled working	Unskilled working	House-wife	Un-employed	Retired	Total
Severity							
Light	8	15	5	11	6	5	50
Medium	11	30	14	30	5	10	100
Heavy	9	22	16	40	5	8	100
None	32	53	35	99	14	17	250
Total	60	120	70	180	30	40	500

Discussion.

The null hypothesis that the differences in the table are only due to random sampling can be tested by the χ^2-procedure described for a 2 x 2 table at the end of Section 18.6. If the incidence of the various degrees of influenza does not differ among the six populations sampled, the best estimate of their incidence is given by the totals in the right-hand column. Thus 50/500 or 10 % of the population would have had a light attack, 20 % a medium one, 20 % a heavy one, and 50 % no attacks at all.

It follows that in the absence of sampling errors the "expected" incidence of degree of illness in each occupational group would be given by applying these percentages to each sample. Thus of the 60 white collar workers, 6 would be expected to have had a light attack, 12 a medium one, and so on.

We now calculate the χ^2-measure of differences between each observed and expected value. Adding for all the items in the 4 x 6 table we have

$$\text{Sum}\frac{(\text{Observed} - \text{Expected Frequency})^2}{\text{Expected Frequency}}$$

$$= \text{Sum}\frac{(8-6)^2}{6} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \text{etc.} + \frac{(17-20)^2}{20}$$

$$= 15.41.$$

Two of the theoretical values are less than 5 (for light attacks among the retired and unemployed). But with a large table as here, this can only marginally affect the χ^2-approximation. [One could combine the retired and unemployed categories, giving a χ^2-value of 13.6 with 3 x 4 = 12

degrees of freedom. The approximation to a χ^2-distribution can also be improved by *Yates Correction*, which consists of reducing the difference between the observed and expected frequencies by half a unit, i.e. $(7.5-6)^2/6+(11.5-12)^2/12+\cdots$). This helps because the observed values are discrete, i.e. whole numbers, and the expected ones are continuous.]

The above quantity will be distributed a χ^2 with $(4-1)(6-1)=15$ degrees of freedom for different samples. From tables of th χ^2-distribution it will be seen that a value of 15 is not significant, more than % of possible sample values are greater than 15. The conclusion is that the incidence of influenza does not really vary greatly (if at all) between the different occupational groups in the population sampled.

It should be noted that percentaging each column of figures in Table 18.6 would have shown at a glance that there are no vast differences in the incidence of influenza even in the *sample* data. From this it would seem unlikely that there could be large differences between the populations. Tests of significance are more valuable in establishing whether an apparently dramatic deviation in a sample really represents something real, rather than whether trivial sample differences are real but negligible.

Exercise 18F. χ^2 as a Measure of Correlation

The larger the value of χ^2 (or of any other test statistic), usually the more "significant" the observed result is. Does the value of χ^2 then provide a measure of the "importance" of this result, i.e. of the degree to which one variable varies with the other (e.g. the incidence of influenza with the occupational classification)?

Discussion.

The value of χ^2 is calculated on the basis of the null hypothesis being true (i.e. *no* association). Once this hypothesis has been rejected because of a large χ^2-value, the basis on which this numerical value has been calculated is no longer relevant. In any case, one is usually not concerned with merely measuring the "strength" of the relationship, but with describing its nature.

Exercise 18G. One-tailed Tests of Significance

In a standard test of significance based on the Normal Distribution, a sample mean m is regarded as significantly different from the hypothesised population mean μ if it is more than 2 standard errors *below* or *above* μ. This is a "two-tailed" situation.

Discuss the practical use of "one-tailed" tests of significance, where only sample values lying sufficiently far *above* the mean, say, are regarded as significant.

Discussion.

If the analyst expects that his sample observations will either agree with the null hypothesis or differ in one direction only, a one-tailed test of significance is often advocated. For example, in testing the effect of a drug or a fertiliser one might expect either little or no effect, or an *increase* in yield, but not a decrease. Sample values showing a decrease could there-

fore only occur because of sampling errors and should never lead to the rejection of the null hypothesis.

A one-tailed test is usually more "sensitive" in the sense that a smaller difference in the expected direction will be regarded as "significant". Thus 5% of sample values lie more than 1.6 times the standard error *above* μ, and any such value would then be significant at the 5% level. (With a two-tailed test, $2\frac{1}{2}\%$ of sample values lie more than 2 times the standard error above μ, and $2\frac{1}{2}\%$ lie more than 2 standard errors below μ, making 5% in all. The differences from μ generally have to be larger, but in *either* direction, to be significant at the 5% probability level.)

Three warnings about one-tailed tests need to be made. Firstly, the precise level of significance is taken too seriously. Consider a particular sample observation which is 1.6 times the standard error above the mean. This has a 1 in 10 chance in a two-tailed test but only a 1 in 20 chance in a one-tailed test. It is relatively unlikely (but not "impossible") to have occurred by chance whichever way we look at it. (A more extreme observation might have a 1 in a 1,000 chance under a one-tailed test, and a 1 in 500 chance in a two-tailed test; does this difference influence any conclusion?).

Secondly, would the analyst *really* accept all sample results in the unexpected direction (i.e. *below* the mean, say), however large the deviation? Can they *really* only be due to random sampling? What about experimental or computational errors, or some quite irrelevant but real factor (a power-cut, a change in the government)?

Thirdly, if one already knows enough to be sure that a truly negative result cannot happen, is it necessary to test the null hypothesis that *nothing* happened? If a positive result is firmly expected, one should be testing *that* as the null hypothesis.

Exercise 18H. More Complex Estimators

Discuss using estimators from sample data to determine the parameter k of the Negative Binomial Distribution (see Section 12.3).

Discussion.

This illustrates the more complex problems of statistical estimation that occur when one uses measures other than the sample mean.

One approach is to note that the variance σ^2 of a Negative Binomial Distribution with mean μ is given by $\sigma^2 = \mu(1 + \mu/k)$, where k is the second parameter of the distribution. This equation can be rewritten as $k = \mu^2/(\sigma^2 - \mu)$. One can thus estimate k by substituting the observed sample mean m and variance s^2; i.e. by writing $k = m^2/(s^2 - m)$. This is another example of estimating by the "method of moments" (see Chapter 12).

A second approach is to note that the proportion of zeros in an NBD is $(1 + \mu/k)^{-k}$. One can therefore use the sample mean m and the sample proportion p_0 of zeros in the equation $p_0 = (1 + m/k)^{-k}$ and solve for k along the lines referred to in Exercise 12I.

A third approach is to calculate the "maximum likelihood" estimate of k. The maximum likelihood principle is widely regarded as providing the "best" estimates of population parameters from sample data, but the mathematics are very cumbersome for the NBD.

These various estimates, and there are others, will generally give numerically different values for any one sample. For large samples the differences should be small (if the estimators are what is technically called "consistent") as long as the population follows an NBD exactly. But additional problems arise if there are systematic discrepancies, even if small, from the theoretical model.

Blind reliance on any one method should be avoided. Thus the "method of moments" is popular in statistical practice, but it is not very accurate for an NBD with a large proportion of zeros; the distribution is skew, so that the occasional large value markedly affects the variance. In contrast, the estimate using p_0 and m is then a good one because p_0 is observed rather accurately.

Even dogmatic reliance on the maximum likelihood principle is not safe, although is often regarded as the theoretically "best" method. It is widely used in complex situations where the results are not easy to-judge, but the principle can give nonsense results under certain circumstances where other procedures work adequately (e.g. Neyman and Scott, 1948; Ehrenberg, 1950, 1951). Statisticians in fact virtually never use the maximum likelihood results in relatively simple situations where "common-sense" judgment *can* be applied. For instance, the maximum likelihood estimate of the variance for a normal distribution is Sum $(x-\bar{x})^2/n$ and not the universally used Sum $(x-\bar{x})^2/(n-1)$. (An apparent exception is the sample mean, which is the maximum likelihood estimate of the population mean. But the sample mean is the "best" estimate of the population mean from almost every possible point of view.)