Studies in Biology no. 43

Introductory Statistics for Biology

Second Edition

R. E. Parker

The Institute of Biology's
Studies in Biology no. 43

Introductory Statistics for Biology

Second Edition

R. E. Parker
B.Sc.
Senior Lecturer in Botany,
The Queen's University of Belfast

Edward Arnold

First published 1973
by Edward Arnold (Publishers) Ltd,
41 Bedford Square, London WCIB 3DP

Reprinted 1975, 1976
Second edition 1979

Paper edition ISBN: 0 7131 2729 5

British Library Cataloguing in Publication Data

Parker, Reginald Ernest
Introductory statistics for biology. – 2nd ed.
– (Institute of Biology. Studies in biology;
no. 43 ISSN 0537–9024).
1. Biometry
I. Title II. Series
519.5′02′4574 QH323.5

ISBN 0–7131–2729–5

Photo Typeset by The Macmillan Co. of India Ltd
Bangalore
and
printed and bound in Great Britain at
The Camelot Press Ltd, Southampton

General Preface to the Series

Because it is no longer possible for one textbook to cover the whole field of biology while remaining sufficiently up to date the Institute of Biology has sponsored this series so that teachers and students can learn about significant developments. The enthusiastic acceptance of 'Studies in Biology' shows that the books are providing authoritative views of biological topics.

The features of the series include the attention given to methods, the selected list of books for further reading and, wherever possible, suggestions for practical work.

Readers' comments will be welcomed by the Education Officer of the Institute.

1979

Institute of Biology
41 Queen's Gate
London SW7 5HU

How to Use this Book

Statistics is not presented here as a branch of mathematics but as a logical commonsense development from very simple beginnings. The difficulties, such as they are, do not lie in mathematical manipulation but in grasping a few simple but unfamiliar concepts. Learning to apply statistical methods is rather like learning to swim or drive a car. One does not become fully proficient immediately and certainly not by just reading and thinking about it. Practical experience is all important. Through practical experience one acquires practical skills and real understanding.

Do not wait until you have some numerical data to analyse before turning to this book. Start now, at the beginning, and try to understand why biologists need to think in terms of statistics and to employ statistical methods. Then, whether you are convinced or not, work steadily on through the book. Master each section as you go, attempt *all* the problems at the ends of the chapters, and check your numerical solutions and logical conclusions with the solutions provided. You will not regret making the effort: the understanding and skill that you acquire will be of lasting value to you, both in your biological work and outside.

Belfast 1978 R. E. P.

1 Probability and Statistics

1.1 Why statistics?

The first problem that most students of biology have with statistics is in understanding why they need to study the subject at all. It is possible, during the first few years of biology at school, to learn a great deal about animals, plants and biological systems in terms of discoveries made by biologists *in the past*. It takes more than this, however, to make a real biologist. A biologist must understand how biological knowledge has been obtained and is being constantly modified and extended by research. It is here that an appreciation of the role of statistics becomes meaningful. For a biologist aiming to make his or her own contribution to biological knowledge some understanding of statistics is essential.

An appreciation of the role of statistics in biology comes most easily through personal involvement in biological investigation, hence the importance of project work, provided that its objective is to discover something new. In part, the role of statistics is direct, enabling us to make statements and draw conclusions of scientific significance from the limited evidence we have obtained by the examination of one or more relatively small samples. For example, suppose as part of a study of the effect of geographical isolation we wish to make statements about the body measurements of the population of field mice on a certain island. We could never hope to catch all of the mice, there might be many thousands, so we trap a sample of them (perhaps 50) and measure this representative group. But when reporting our finding of, for example, 'mean tail length', we would want this to relate to the population as a whole and not just to our small sample. We could do this by using *confidence limits* (see Chapter 2).

Again, suppose that we wished to compare two 'wetting agents' as additions to foliar fertilizers. In our experiment we would only be able to treat a limited number of, for example, lettuce plants, but our conclusions would need to apply to all lettuces of the particular variety used in the experiment. Before concluding that there was no real difference between the wetting agents or that one was more effective than the other we would be likely to need a *test of significance* (see § 1.4 and Chapter 3).

Another role of statistics lies in the simplification of data with the detection and definition of trends or relationships. In biology observed relationships are rarely clear-cut. Even when an underlying relationship is simple our picture of it is often confused by uncontrolled variation. This role can be seen in a simple form in Chapter 8, where the linear component of the relationship between temperature (°C) and water-loss (mg) of a group of mice is extracted and defined by means of *linear regression*. With advances in techniques of environmental sensing, and in the recording and processing of data, the role of statistics in simplification and extraction of trends is increasing. Even when a computer is used for the data processing it must be programmed with the appropriate statistical instructions. The use of statistical methods as tools in biology has had several important repercussions. The introduction of new physical and chemical methods, such as the electron microscope, radioisotopes, and chromatography, led to the opening up of new fields of biological enquiry. The same is true of statistics: there are branches of biology which only became possible with the development of the necessary statistical tools, for example, quantitative population genetics. More generally, advances in the design of biological surveys and experiments have come about as a direct result of development in statistical techniques. The close relationship between experimental design and data analysis is discussed in Chapters 9 and 10.

1.2 Probability

Whenever we draw conclusions relating to whole populations from the evidence of samples, for example, when we fit confidence limits or make tests of significance, these conclusions are always couched in terms of probability. In statistics *probability* takes on a full quantitative meaning, having values ranging from zero which is equivalent to impossibility, to unity which is equivalent to complete certainty. There are two ways of estimating the probability of a particular kind of outcome: one, *a priori*, is from some knowledge of the underlying process, or at least some hypothesis concerning it, e.g. for a cross between a heterozygote (Gg) and a homozygote (gg) we can estimate the probability of a single random offspring being (Gg) as 0.5, the same as its being (gg). The other way, *empirical*, is by observation of the outcome of a number of actual trials, thus:

$$\text{estimated probability, } p = \frac{\text{number of successes}}{\text{number of trials}}$$

where a 'success' is the kind of outcome in which we are interested. For example, the probability of a seed, taken at random from a population, being capable of germination may be estimated by carrying out a germination test on a sample of seeds from that population. The results of such a test are conventionally expressed in the form of a percentage (i.e. $100 \times p$). In general, the larger the sample examined the more closely will the estimated probability approximate to the true one.

In order to find out why observations tend to fall into classes with the frequencies they do, we compare the *observed* frequencies with the frequencies which would be *expected* on the basis of a particular hypothesis. To obtain an expected frequency we simply multiply the expected probability by the total number of trials. For example, if we wished to discover whether two dominant genes were linked we could backcross double heterozygotes (AaBb) with double recessives (aabb) and count the number of organisms in each of the four phenotype classes. If the genes segregate independently (i.e. if there is no linkage) the expected probability for each class is the same and equal to 0.25. If we have a total of 400 offspring the expected frequency of each class would be 400 × 0.25 = 100. Note that every individual *must* belong to one of the four classes and that with equal probability of it falling in each one we partition the total probability of 1.0 into four equal parts of 0.25. (If we had chosen to make the $F_1 \times F_1$ cross the partition would have been into: 9/16, 3/16, 3/16, and 1/16.)

1.3 Probability distribution

In partitioning the total probability of 1.0 into several components, each corresponding to a different kind of outcome, we are exposing a simple probability distribution. We need to look more thoroughly into this because the distribution of probability is the key to a great deal of statistics. Let us begin with the simplest kind of situation, one in which at a single trial there are only two kinds of outcome.

Suppose that in an investigation of the relationship between a certain beetle species and its environment we wish to test the hypothesis that its known sensitivity to humidity is located in its antennae. We could take a beetle, remove its antennae, and place it in a choice-chamber where two different levels of humidity were maintained. If our hypothesis is true the probability of beetle choosing the high humidity is the same as the probability of its choosing the low, and is equal to 0.5. Provided that the choice-chamber was of suitable design two beetles could be introduced at the same time. What would the probability distribution be now? There are three different possible types of outcome: both move to high humidity, both move to low humidity, or one moves to high and the other to low. The corresponding probabilities are given in the table on p. 4.

Note that the probability of both beetles behaving in the same way is the *product* of the individual probabilities, as two conditions must be fulfilled before the outcome is attained. Also note that the probability of a mixed outcome is the *sum* of the probabilities of the different ways in which the same outcome can be attained. The table also shows the possible types of outcome and their probabilities for groups of three and four beetles. In this simple example the expected probabilities can be

Number of beetles	*Possible outcomes*					*Number of possibilities*
1		H($\frac{1}{2}$)			L($\frac{1}{2}$)	2
2		HH(1/4)	HL LH ($\frac{1}{2}$)		LL(1/4)	3
3	HHH(1/8)	HHL HLH LHH (3/8)		HLL LHL LLH (3/8)	LLL(1/8)	4
4	HHHH (1/16)	HHHL HHLH HLHH LHHH (1/4)	HHLL HLHL HLLH LHHL LHLH LLHH (3/8)	HLLL LHLL LLHL LLLH (1/4)	LLLL(1/16)	5

easily worked out from first principles. More generally we compute the probabilities by expanding the expression $(p+q)^n$ where p is the probability of an individual, taken at random, falling into one of the two mutually exclusive catagories, $q = 1-p$ (i.e. the probability of its falling into the other), and n is the number of individuals in the group. For $n = 4$ this gives:

$$(p+q)^4 = p^4+4p^3q+6p^2q^2+4pq^3+q^4$$

You may recognize this as the binomial expansion. The distribution of probability is discontinuous because however large the group (n) there will always be a finite number ($n+1$) of categories. In the present example the distribution is symmetrical. This is because $p = 0.5 = q$. More generally $p \neq 0.5$ and the distribution is asymmetrical.

1.4 Tests of significance

Suppose that we have conducted an experiment in which we placed eight beetles from which the antennae had been removed in a choice-chamber with high and low humidity compartments, and that all eight had moved to high humidity. What would we conclude? With such an extreme result we would probably be left in no doubt that the beetles were still sensitive to humidity differences and that our hypothesis should be rejected. But first examine Fig. 1–1 which represents in the form of a histogram* the probability distribution corresponding to $(p+q)^8$ where $p = 0.5 = q$. You will see that the outcome with the

* Conventionally such discontinuous distributions are represented by bar charts. Histograms are used here to emphasize the similarity between discontinuous distributions and continuous distributions to be met with later.

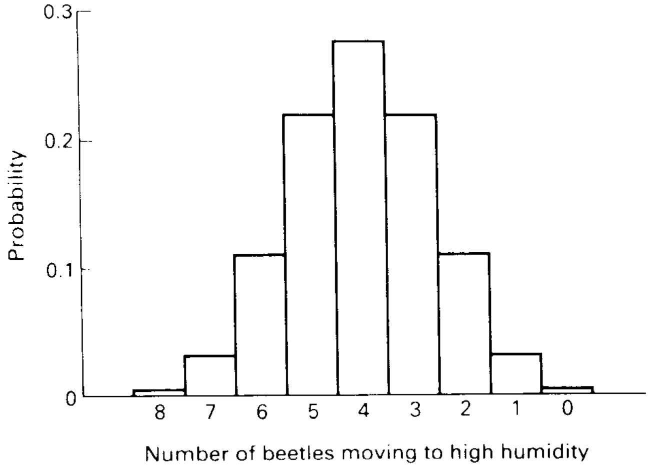

Fig. 1·1 Probabilities of different numbers of beetles, from 8 to 0, moving to high humidity: $p = 0.5$.

highest probability, the *expected* outcome, is the one with equal numbers of beetles moving to high and low humidity. The observed result, if the beetles are insensitive to humidity (i.e. if $p = q$), has a probability of $p^8 = 1/256$ (c. 0.004). We have therefore *either* to retain our hypothesis and accept that a very unlikely event has occurred *or* to reject our hypothesis and replace it with one in which $p > 0.5$. With such a low probability we would normally modify our hypothesis and conclude that the beetles were still sensitive to humidity differences, even with antennae removed.

The test consists of the computation of a probability corresponding to the result observed on the assumption of a particular hypothesis. If the probability is low (conventionally $\leqslant 0.05$) we conclude that the hypothesis is incorrect. If the probability is high (conventionally > 0.05) we conclude that the departure from expectation is not great enough for the hypothesis to be rejected. In general, it is necessary to take into account other possible outcomes which depart an equal or greater amount from expectation. In the above example there is only one other outcome with as great a departure, i.e. eight beetles moving to low humidity. We do not need to take account of this because intact beetles are known to move to *high* humidities and so the possibility of the opposite kind of behaviour does not arise. We are making what is known as a 'one-tailed' test (see also § 4.4).

It might be argued that in the example above we had no need of statistics because the conclusion was self-evident. Results are not always as extreme. Suppose that only six beetles had moved to high humidity the remaining two moving to low. It could be argued that as the beetles have shown a marked preference for high humidity they must still be sensitive to humidity differences and our hypothesis should be rejected. It might also be argued that as some beetles moved in each direction our hypothesis should be retained. This is just the sort of situation in which statistics can help. If we again turn to Fig. 1·1 we see that the probability corresponding to the observed result is now much greater. It is in fact $28p^6q^2 = 28/256 = 0.1094$. Before making the test, however, we must take into account the two more extreme results, i.e. of seven and eight beetles moving to high humidity. The total probability of the three types of outcome is $28/256 + 8/256 + 1/256 = 0.1445$ (c. 1/7).* This is well in excess of 0.05 (1/20) and so we would regard the observed departure from expectation as *not significant* and retain the hypothesis that the beetles have been rendered insensitive to humidity differences. In fact, tests of this kind, if repeated, would be expected, on average, to give as large or larger departure one in seven times, even though the beetles had been rendered insensitive.

In this first chapter we have become acquainted with the ways in which probability can be treated quantitatively, in particular how it can be partitioned to yield discontinuous probability distributions. We have also seen how we can use such a distribution in making tests of significance. Discontinuous distributions are related to observations of discrete events. Such observations are of great importance in biology and we will be examining the methods for their analysis in some detail in Chapters 4 to 7. In Chapters 2 and 3 we will direct our attention to observations which take the form of measurements on a continuous scale. The probability distributions of such measurements are continuous and this presents us with rather different problems and opportunities for analysis.

Problems

1·1 In a choice-chamber experiment like the one examined above what is the smallest number of beetles that could be used in an experiment and still give a significant result in a 'one-tailed' test?

1·2 A reasonable *a priori* probability of a child, taken at random, being a boy is 0.5, this being consistent with equal numbers of male and female sex chromosomes on gametogenesis in the male parent. How large a family, consisting of children of the same sex, would a biologist need to have before rejecting the general hypothesis of $p_{\text{boy}} = 0.5 = p_{\text{girl}}$ as applicable to his or her own relationship? (Assume the conventional significance level of 0.05.)

* This is, of course, again a 'one-tailed' test.

2 Continuous Distributions: Confidence Limits

2.1 A population represented by a sample

The populations about which we wish to make statements and draw conclusions are represented in biological surveys and experiments by samples (see § 1.1). These samples consist of individuals. In some investigations these individuals are whole plants or animals but more generally they range from individual cells, or even organelles, to plots of forest trees. They can take the form of organs, tissue preparations, extracts, and even environmental locations. Despite this diversity they have in common the fact that they contribute an item of information relating to one or more of their attributes. For the moment we are concerned only with situations in which there is information relating to one kind of attribute and in which the information consists of measurements on a continuous scale, such as weight, volume, area, length, concentration, rate, pH, etc.

Suppose that as part of an autecological study of the bracken fern (*Pteridium aquilinum*) we wished to assess the performance of the fern within a certain area, Area 1. Frond (leaf) length could be included among the performance parameters. There are thousands of fronds in the population of the area and they vary conspicuously in length. We might decide to measure a sample of 100 fronds. Clearly the sample should be representative and must therefore be selected without bias. This is more difficult to achieve than it might appear, simply measuring one frond here and another there in an irregular manner will not do. Ideally we should take a *random* sample but we will leave this problem for now and assume that the lengths of 100 fronds have been measured. The 100 measurements are added together and divided by the number of fronds (100) to obtain an average or mean length. This is a *sample mean* and it is our best estimate of the *population mean*. A simple statement of our sample mean, with the unstated implication that the population mean is likely to be rather similar, is not likely to be very satisfactory when we come to make comparisons between different areas and try to draw meaningful conclusions. We need to assess and state in some way the reliability of our sample mean as an estimate of the population mean. We can do this by attaching confidence limits.

2.2 The normal distribution

Clearly the reliability of a sample mean is bound up with the variability of the individual measurements and with the number of them that we have to average. We need then some measure of variability. The measure that we use is related in an important way to the kind of probability distribution shown by the individual measurements. Fortunately there is a strong tendency for the measurements of individuals in different populations to show the same kind of distribution, the *normal distribution.*

If we had measurements for a large number (for example 1000) of individuals in a population, e.g. frond length as in our example, we could group the measurements into size classes, count the number falling into each size class, and plot a frequency histogram to show how the individual measurements were distributed. (NB Measurements fall automatically into size classes if they are made to a certain degree of precision only.) As probability is the ratio of frequency to the total number of measurements (see § 1.2) the frequency histogram could be rescaled to illustrate the corresponding probability distribution. It would now indicate the distribution of probability of an individual measurement, taken at random, falling into each size class. The histogram produced is likely to be more or less symmetrical and to have a bell-like outline. In some respects it would resemble the histogram for a symmetrical binomial distribution, (Fig. 1–1), but would differ from this in two important respects. The range of a binomial distribution is fixed by the number of events recorded in each trial but the range of the population of measurements is not limited in this way. The number of classes in the binomial distribution is also fixed, $(n+1)$, and the distribution of probability essentially discontinuous. We could not read from Fig. 1–1 the probability of $5\frac{1}{2}$ beetles moving to high and $2\frac{1}{2}$ beetles moving to low humidity. On the other hand, by making measurements to a high degree of precision and making enough of them a histogram could be prepared with a very large number of very narrow size classes and an outline that would approach closely a smooth curve. The theoretical curve towards which many natural probability distributions tend is called the *normal curve.*

2.3 Mean and standard deviation

Clearly the normal distribution cannot be defined in the same terms as a binomial distribution. The normal curve is defined by the expression:

$$Y = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-[(X-\mu)^2/2\sigma^2]}$$

Fortunately we can use the properties of the normal distribution without using this expression but there are several important points to note about it. The variables X and $Y$are related through two parameters μ and σ. μ is the *mean*, the point about which the distribution is symmetrical, and σ is the *standard deviation*, a measure of the variability or spread of the measurements about the mean. The important point is that a normal distribution is completely defined by these two parameters; the other two quantities π and e are, of course, constants. Thus, if we know the standard deviation of a population we have the key to the distribution of probability about its mean. In practice we rarely know the population standard deviation, σ, but have to estimate it as s from the sample measurements. We can estimate σ as $s = \sqrt{[\Sigma(X - \bar{X})^2/(N - 1)]}$, where the numerator is the sum of the squares of the deviations of X from its mean, and the denominator is one less than the number of measurements. You may wonder why the denominator is not N. If we knew the population mean (μ) the correct denominator would be N, but in practice we have to estimate μ as $\bar{X}$. If we have N values of X and compute $\bar{X} = \Sigma X/N$, then we have only $(N - 1)$ values of X which are independent of $\bar{X}$, in other words we have only $(N - 1)$ *degrees of freedom*. Having determined $\bar{X}$ the Nth value of X is also fixed. The above formula is rarely used to evaluate s because to do so is unnecessarily laborious and usually introduces rounding errors. Instead we use the algebraically equivalent expression:

$$s = \sqrt{\left(\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}\right)}$$

NB ΣX^2 denotes the sum of the squares of all values of X taken singly, and $(\Sigma X)^2$ denotes the square of the sum of all values of X. The quantity $\Sigma(X - \bar{X})^2 \equiv \Sigma X^2 - ((\Sigma X)^2/N)$ is so often met with that it is referred to as the *sum-of-squares of* X and is denoted by Σx^2.

We have already seen in section 1.4 that given a hypothetical probability (p) and a group size (n) we can compute the terms of the corresponding binomial distribution and draw a histogram illustrating the probability distribution. From this histogram we can read off the probabilities of different kinds of outcome in terms of the heights or areas of the corresponding rectangle or rectangles. Figure 2–1 shows the probabilities of the seven different kinds of family composition for a family of six children, assuming $p_{\text{boy}} = 0.5 = p_{\text{girl}}$. The probability of a family including one or two boys is indicated by the total area of the two shaded rectangles. Now, in much the same way, the probability of an individual measurement, taken at random, falling between two stated values of X is given by the area of the figure bounded by the appropriate normal curve and the X axis, and lying between the verticals corresponding to the two X values. Figure 2–2 shows the probability distribution for heights of adult human males in a population. The probability of a man,

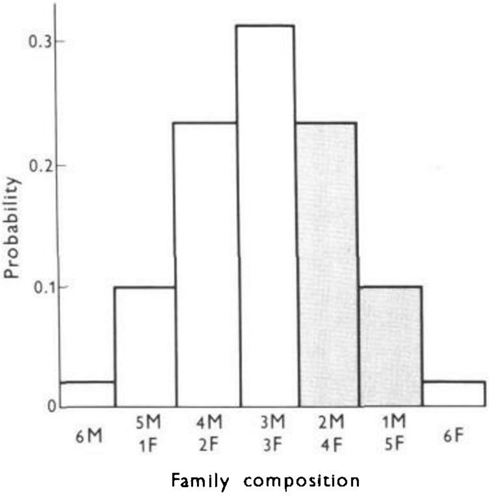

Fig. 2 1 Probabilities of the 7 different kinds of family composition for a family of 6 children: $p_{\text{male}} = p_{\text{female}} = 0.5$. The probability of a family including 1 or 2 boys is indicated by the area of the shaded rectangles.

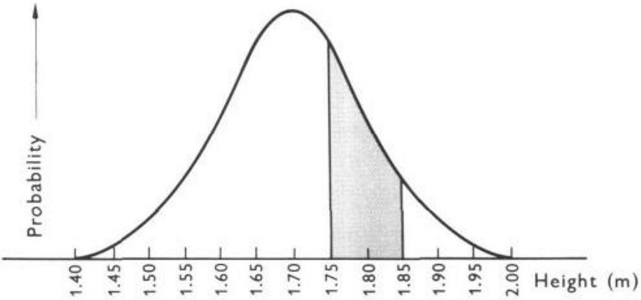

Fig. 2–2 Probability distribution for the heights of human adult males in a population. The probability of a single individual taken at random having a height falling between 1.75 m and 1.85 m is indicated by the area of the shaded part of the figure.

taken at random, having a height between 1.75 and 1.85 m is indicated by the area of the shaded part of the figure.

The shape of a particular normal curve, and therefore the corresponding distribution of probability, depends entirely on the standard deviation of the normal distribution which it represents. It is possible therefore to draw up tables for the areas beneath the curve (in terms of probability) between certain limiting values of X, provided that such values are expressed in terms of standard deviation units. Such tables are then true for all normal distributions. For example, we can read from such tables that the area between the line of symmetry (at $X = \mu$) and the line at one standard deviation above ($\mu + \sigma$) *or* below ($\mu - \sigma$) corresponds to a probability of 0.3413. Thus we can say that there is a probability of 0.6826 of a single value, taken at random, falling within one standard deviation of the mean. Similarly, a single value has a probability of 0.95 of falling within 1.96σ of the mean. In terms of frequency this means that 68.26% of all values lie between the limits $\mu \pm \sigma$ and 95% of all values lie between the limits $\mu \pm 1.96\sigma$ (Fig. 2–3). The probability that an individual value, taken at random, would fall within this range is 0.95: conversely given an individual value we can say that the probability of the true mean falling within 1.96σ of *it* must also be 0.95. It might appear at first sight that we have here a way of describing the reliability of a single random observation in estimating the population mean. Our estimate of the mean would be the single measurement and there would be a probability of 0.95 of the population mean lying within $\pm$ 1, 96σ of it. The snag, of

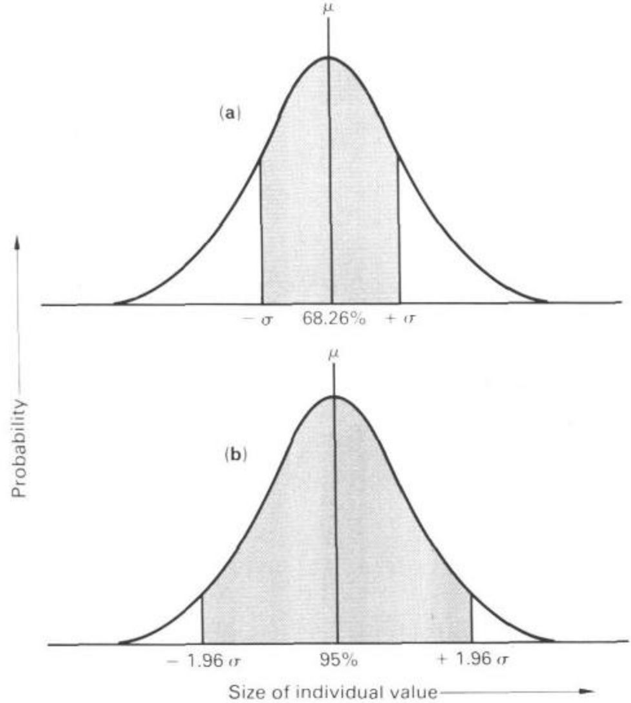

Fig. 2–3 Probability limits for the normal distribution defined in terms of standard deviation (σ): (**a**) 68.26% and (**b**) 95%.

course, is that a whole series of measurements would be needed for the estimation of the standard deviation.

2.4 Standard error of the mean and confidence limits

If a series of measurements were available we would use their mean to indicate the population mean rather than rely on a single measurement to do so. If a sample mean is to be used in this way we would need to know something about the distribution of sample means. We have seen that if we take a very large sample of measurements from a population, divide it into a series of small size classes and plot a frequency histogram, the outline of the histogram approaches a normal curve. If we were now to take these measurements in random groups (of N), calculate the means of these groups and prepare a frequency histogram from these, its outline would again approach a normal curve but one with less spread, i.e. with a smaller standard deviation (Fig. 2–4). It may be shown that the standard deviation of the means of N measurements from a population with a standard deviation of σ is $\sigma/\sqrt{N}$. Conventionally the standard deviation of a mean is known as its *standard error*. In the same way that we could use the standard deviation to describe the reliability of a single random measurement in indicating the population mean so we can use the standard error to indicate the reliability of a sample mean in doing so. Thus, having computed $\bar{X}$ (as $\Sigma X/N$) we can state that this is our best estimate of the true mean (μ) and attach *confidence limits* at a chosen level

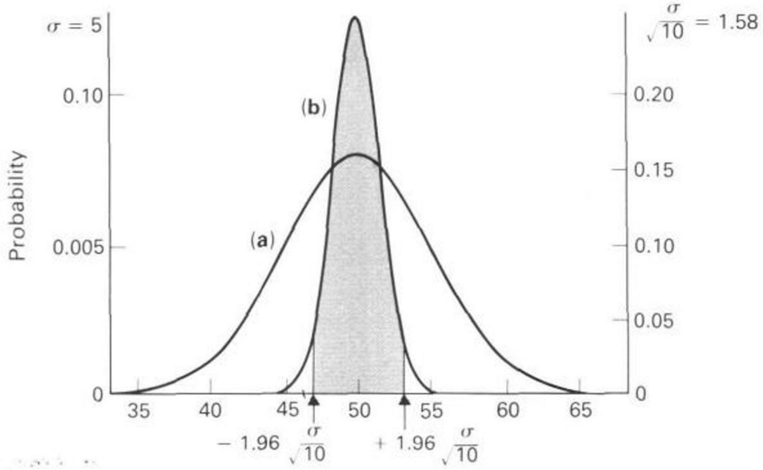

Fig. 2–4 Normal curves (**a**) for individual values of X with $\mu = 50$ and $\sigma = 5$, and (**b**) for means of ten values. 95% probability limits for means of 10 values.