Reprinted from the University of Leeds Medical Journal, Vol. IX, *No.* 2, 1960, (*pages* 50 *to* 60).

LOGIC AND THE INTERPRETATION OF OBSERVATIONS

DAVID COLQUHOUN, B.SC.
Department of Pharmacology, University of Leeds

"There is no more common error than to assume that, because prolonged and accurate mathematical calculations have been made, the application of the result to some fact of nature is absolutely certain" A. N. WHITEHEAD.

Figures are not enough. They have to be interpreted; and the result may be equivocal, especially when living material is involved, because of its variability. Experiments should be suitable for testing the hypotheses in which we are interested and as many as possible of the assumptions we have made, with maximum precision and with the greatest possible certainty that the conclusions are valid. To achieve this it is best to make quantitative measurements whenever possible. As these will be subject to variation, statistical analysis will often be necessary before a proper judgement of their implications can be made. I shall attempt to discuss some of the assumptions invoked by the use of these methods and, in particular, the conclusions that can logically be drawn from the results of their application.

The statistical methods encountered in the biological sciences are very largely those of Sir Ronald Fisher and his school, based on the earlier work of Karl Pearson: methods expounded in many books, e.g. Fisher (1951 and 1954) and Snedecor (1956). There are, however, at least two other important schools of thought on statistical method and scientific inference — those of H. Jeffreys and of J. Neyman and E. S. Pearson. Although not entirely distinct, they — especially Jeffreys — differ on certain points. These systems of inference have been discussed by, e.g. Anscombe (1948), Neyman (1941) and Hogben (1957).

PROBABILITY

It is perhaps surprising that there is still no satisfactory definition of probability. A subjective definition as the "degree of rational belief" or "degree of rationality of belief" or "degree of legitimate conviction" is upheld by some, e.g. Keynes. The classical (Laplacean) definition identifies probability with the "ratio of the number of favourable cases to the total number of equiprobable cases". This is open to criticism on logical grounds in that the concept to be defined is introduced as part of its own definition by the word "equiprobable". Nevertheless when the reference set of "the total number of equiprobable" cases is finite, this definition is always used and accepted in practice. For instance, if one-quarter of the students of the University of Leeds are female, we would assert that the probability of a student chosen from this finite population being female is 0.25, provided that it was equally probable that any one of the students would be picked, i.e. that the choice was made at random. When the reference population is infinite, however, this definition is obviously unsatisfactory.

The frequency theory of probability defines the probability P of an event as the limiting relative frequency of its occurrence in a random sequence of trials; e.g., as the number of trials is extended indefinitely, as when a penny is tossed, if the frequency of heads is 50 per cent. we would say that the probability that a penny will come down heads is 0.5. This type of definition seems reasonable, and is often invoked in practice, but again it is by no means satisfactory as a complete, objective definition (see Kneale 1949; Popper 1959). R. von Mises, the principal exponent of a frequency theory, has found it necessary to restrict the definition to sequences which fulfil the following two conditions: (i) The sequence is random, i.e. it is equally probable that any admissible subsequent event will follow any specified antecedent and contains no relevant subset that would lead to a different value of P; (ii) the relative frequency of occurrence of the event converges to a fixed limit, P, as the sequence is increased indefinitely. Such an infinite sequence he calls a "collective". However the first condition invokes the concept of "probability"—again a *circulus in definiendo*. A second difficulty is raised by the "axiom of convergence". Mathematically, the concept of convergence is applicable only to infinite sequences constructed according to a rule, e.g. $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$, such that all the terms of the sequence after a certain term keep within any given interval, however small, of the limit. However our sequence, by definition, follows no rule, and consequently it has been argued that the conditions of convergence

and irregularity required by von Mises are incompatible. This objection has been met by Popper (1959) in his translation of his *Logik der Forschung* (1934) by modifying these requirements. But Kneale (1949) agrees with von Mises that they are necessary for a frequency theory of probability.

In spite of such objections it is quite usual for those who use probability as an aid to judgement in research to conceive its meaning in terms of frequency in repeated trials, and indeed if it is to be any use at all some such concept is necessary unless we adopt a subjective definition. Attempts to obtain an objective definition free from objections usually seem even further from the world of reality than the "collective" of von Mises (see, e.g., Kneale, 1949).

Deduction, Induction and Knowledge

Deduction may be defined as a process of reasoning from known causes, or set of axioms, to their results or, more precisely, as an argument in which it would be inconsistent to affirm the premises and deny the conclusion. This would not necessarily be so in an inductive argument which is concerned with the discovery of causal relations or with the drawing of conclusions about a population from observations on samples from it.

Philosophers have devoted much attention to the problem of when, if ever, we can say that we "know" something. Ayer (1956) has said that, except where the validity of a statement is a condition of its being made at all (such as Descartes' proposition "*Cogito ergo sum*": "I think, therefore I am"), then it is always possible that we take something to be true when it is false. The only way in which new knowledge can be arrived at in practical scientific work is by an inductive process. Such a process can never be certain. The possibility of error always remains and, if we are to be strictly logical we must say that no scientific hypothesis has been established as true, but that it has merely not yet been falsified. Obviously this does not mean that we cannot feel that, for all practical purposes, a proposition has been established as true and act as though it were true, though reserving the possibility that we may be subsequently proved wrong. It is interesting that it is equally impossible to prove the validity of deductive reasoning, as we cannot show that the axioms on which it is based are correct, all we can hope is that the set of axioms is consistent, i.e. they cannot lead to contradictory deductions, but even if this is so it does not prove their correctness.

Popper (1959) is of the opinion that scientific argument does not involve inductive reasoning except in so far as it is used to arrive at hypotheses, as the process of testing hypotheses is deductive, i.e., we deduce from the hypothesis the results we would expect if it was true, and we do an experiment to see whether these predictions are correct. If they are we retain the hypothesis; if not we reject it. However, as Ayer (1956) points out, it would not be rational to reject a hypothesis because it had been falsified by our experiment unless we assume that the result of repeating the experiment will be the same in the future. This assumption is only justifiable by inductive reasoning.

It is obvious that the evidence in support of some hypotheses is greater than for others. The question then arises as to whether it is possible to attach a probability to the assertion that a hypothesis is true. Such statement would be called one of *inverse probability*.

The theorem of the Rev. Thomas Bayes, published posthumously in 1763, has been the basis of many attempts to justify inductive reasoning. Its proper use is in the situation when we have a valid knowledge of the *a priori* probabilities of certain contingencies. The theorem can then be used to convert the *a priori* probability into an *a posteriori* probability in the light of experimental evidence. For example (Fisher 1959*a*), if two known heterozygous (Bb) black mice are mated, then according to Mendelian theory, a black mouse picked at random from the progeny has an *a priori* probability of being homozygous (BB) of $\frac{1}{3}$ and of being heterozygous, $\frac{2}{3}$. If on mating this mouse with a brown mouse (genotype bb) it yields seven offspring, all black, we can validly use Bayes' theorem to calculate the *a posteriori* probability in the light of the evidence, that the black mouse chosen was homozygous or heterozygous. These turn out to be 64/65 and 1/65 respectively. Thus in these circumstances the probability that we are correct in asserting that the mouse was homozygous is 64/65, and we have validly attached probability statements to the hypotheses (the possible genotypes) in the light of their results (the experimental evidence). However Bayes' method is not applicable in the majority of experimental work simply because

the *a priori* probabilities are not known. The practical worker would find it difficult to understand how this lack of knowledge can be avoided by saying, as did Laplace, or as would Jeffreys, that, when there are two mutually exclusive possibilities (e.g. homozygous or heterozygous), and with no prior information to favour one or the other, it is axiomatic that their *a priori* probabilities are equal, and that Bayes' theorem should be used on this basis.

Natural Variation and Error

The binomial probability distribution may be stated formally thus: the probability that r "successes" will occur in n independent trials of an event which may have either of two mutually exclusive results ("success" or "failure") at each trial is $\frac{n!}{r!\,(n-r)!}p^r q^{n-r}$ where p is the probability of a success, and $q = 1-p$ of a failure at a single trial. This follows from the classical theory of probability. The values of the above expression for the different possible values of r are the terms of the expansion of the binomial $(p + q)^n$. The sum of all these terms from $r=0$ to $r=n$ must therefore be 1 (as $p+q=1$) which is equivalent to saying that it is certain that in n trials r will have some value between 0 and n. For example if we toss two coins ($n=2$) the possible results may be represented HH, HT, TH and TT. If the probability of "heads" at a single throw is $p = \frac{1}{2}$ and of tails is $q = 1 - p = \frac{1}{2}$ then the probability of each of the above results is $(\frac{1}{2})^2 = \frac{1}{4}$. However there are two possible ways in which one head can occur in two trials (this is given by the coefficient of the expression above: $\frac{2!}{1!1!} = 2$), hence the probability of two heads out of two throws is $\frac{1}{4}$, of one out of two is $2 \times \frac{1}{4} = \frac{1}{2}$, and of none is $\frac{1}{4}$.

It can be seen that this distribution deals only with discontinuous or discrete values, the proportions of successes. The values calculated from it can be plotted as a block diagram, or histogram, with calculated probabilities or observed frequencies as ordinate, and proportion of successes $\left(\frac{r}{n}\right)$ as abscissa. If the total area of the blocks is defined as 1 then the probability of a given value of $\left(\frac{r}{n}\right)$ is equal to the area of the block corresponding to this value on the histogram.

If the number of trials, n, is increased the width of each individual block, corresponding to a value of $\left(\frac{r}{n}\right)$ must become narrower, until eventually it is so narrow that the outline of the histogram forms a smooth bell-shaped curve. This curve is the "normal curve of error". It is represented by the formula

$$y = \frac{1}{\sigma\sqrt{2\pi}}\, exp.\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad (1)$$

where y is the ordinate, μ is the mean of the distribution (which is the same as its mode, as it is symmetrical) and σ is its standard deviation (which is a measure of the variability of the observations). The square of the standard deviation is called the variance. The area under such a curve is unity. As the abscissa is now a continuous variable x, the blocks of the histogram have become indefinitely narrow and the element of probability represented by the area of such a block must now be written $dP = ydx$. The probability that x will lie between these limits on the abscissa, is obtained by the integration of the expression $dP = ydx$ between the appropriate limits.

The normal curve was first used, e.g. by Laplace, simply as an approximation to the binomial distribution to avoid the necessity of calculating its individual terms when n is large. It was applied by Gauss to the instrumental errors in making physical observations. He found that when a physical constant was being measured the errors of repeated observations were often well represented by a normal curve, the mean of the distribution being the required true value of the constant.

The situation in the biological sciences is rather different however. We are not concerned so much with *errors* of measurement due to the fallibility of the observer and his instruments, but with natural variation which is *real* and inevitable and in no way due to the experimenter's fallibility. Under these conditions the mean value of a distribution ceases to have any significance and in no sense is it a *true* value of the variable being measured—for instance the heights of men—and Hogben (1957) argues that if we have no *true* value for a variable the concept of error loses its meaning, and in the

application of the theory of errors to natural variation we have moved beyond its terms of reference. However most people who apply statistical methods to natural variation, and indeed most statisticians, would feel justified in applying procedures depending on the "normal law" if it had been actually observed that the variable in question was distributed according to the mathematical formula above. The fact that the mean of the distribution often has no real significance does not matter. It may simply be used as a device for defining the distribution, which, if normal, can be completely defined by its mean and standard deviation. It can be shown mathematically that if errors to which the variable is subject are of a certain type, the variable will be normally distributed. However the justification for the use of the normal law is observation that it holds good. The mathematics provide a possible explanation for this observation. Gaddum (1945) quotes the remark "Everybody firmly believes in it" (the normal law, *Ed.*) "because the mathematicians imagine it is a fact of observation, and observers that it is a theory of mathematics". In fact observations as such are often not normally distributed and it will be necessary to discover by trial and error some function of them which is, if statistical methods depending on normality (as the majority do) are to be validly used. Frequently we find the logarithm of the variable is distributed normally (Gaddum, 1945). For instance it is well known that the individual effective doses of drugs are often "lognormally" distributed in this way, so the logarithms of doses must be used for calculations. This has been shown to be so for the dose of sodium salicylate required to produce toxic symptoms in man (Hanzlik, 1913; Gaddum, 1954). Also Wootton and King (1951 and 1953) have shown that the distributions of the levels of various blood constituents are sometimes normal, and sometimes lognormal. In assuming either without a prior investigation we might very well be mistaken.

It is fortunate then that most procedures are not very sensitive to small deviations from normality; and that means of observations, even if the observations themselves are from a non-normal population, tend to be normally distributed as the sample size increases. It may be noted that a logarithmic transformation removes the lower limit of zero for the dose and so may be expected to remove the positive skewness of the original observations.

Another difficulty in the application of the theory of errors to natural variation is that a biological population is never static. If an experiment is done on a sample of animals or men we consider that these have been picked at random from a population of possible animals or men. However the assumption that the results obtained can have any application in the future, outside the realm of this particular experiment, involves assuming that in future experiments on different, or even on the same sample, we can look upon it as a sample from the same population as in the first case. This is difficult to justify simply because a biological population is not static and the conditions and circumstances of the first experiment can never be exactly repeated. Consequently we must regard the sample as being from a *hypothetical* infinite population. This population is defined by certain (hypothetical) *parameters* or constants, for instance, its mean and standard deviation if it is normally distributed. Estimates of these parameters, viz. *statistics*, are calculated from the observations on the sample. As we have said that we cannot make probability statements about values of parameters, how are we to decide which value best represents the data? The method most widely used at present is the "method of maximum likelihood".

Method of Maximum Likelihood

This was developed mainly by R. A. Fisher, although it was known in the last century. It states that the relative plausibility of hypotheses is best measured by their "likelihood". The "likelihood" that a parameter has any particular value is proportional to the probability that the observed data would have occurred if in fact the parameter had this value. We then choose the particular value of the parameter for which the likelihood is a maximum. For instance if we have n observations, x_i, which are normally distributed according to equation (1) what is the best estimate of μ, the mean of the distribution?

The likelihood of any value of μ is proportional to the probability that all the observed values of x_i would have occurred if this was the true value, i.e. it is the product of all the

terms resulting from substitution of each of the values of x_i in equation (1) i.e.

$$\text{Likelihood} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp.\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \text{constant} \times \exp.\left[\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]$$

where $\prod$ and Σ are the product and summation signs respectively. To find the value of μ for which this is a maximum we differentiate the expression (or more conveniently its logarithm L) with respect to μ and equate to zero (as the slope of the curve obtained by plotting L against μ is zero when L is at a maximum), and solve for μ.

$$\log_e(\text{likelihood}) = L = \text{Constant} - \frac{\sum_{1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L}{\partial \mu} = \frac{\sum_{1}^{n}(x_i-\mu)}{\sigma^2} = \frac{\sum_{1}^{n}x_i - n\mu}{\sigma^2}$$

This is zero when $\sum_{1}^{n}x_i = n\mu$, i.e. $\mu = \Sigma x_i/n$; i.e. the arithmetic mean of the observations is the maximum likelihood estimate of μ.

This method is more complicated when applied to problems, e.g., of biological assays with quantal responses (see Finney, 1952), but the results of its application can be used quite easily.

The Method of Least Squares

This is a much older method. It states that the best estimate of a parameter is such that the sum of the squares of the deviations of the observations from the estimate is a minimum. In the above example these two methods give the same result, as $\exp.\left[\frac{-\Sigma(x_i-\mu)^2}{2\sigma^2}\right]$ reaches a maximum when the sum of squared deviations $\Sigma(x_i-\mu)^2$ is a minimum, and in fact when the method of least squares is applicable it is equivalent to the method of maximum likelihood if the variation is normal.

Uncertainty of Estimates

Having made an estimate of a quantity, e.g. the potency ratio of two drugs, by one of the above methods we would then like to be able to state the probability of our being correct. Unfortunately such a statement of inverse probability is not usually possible. It is, however, possible to calculate certain limits for the value which we have estimated, within which we can be fairly certain that the 'true' value lies. Such limits are called "Confidence Limits" by J. Neyman and E. S. Pearson and "Fiducial limits" by R. A. Fisher. These schools differ in the precise interpretation which they put on the words "fairly certain" which I have used above.

Of the confidence limits of Neyman and Pearson, when calculated at the 95 per cent. level of significance using the appropriate tables, we may say that, if we consistently assert that the value of the unknown parameters lies within the calculated limits, then, in the long run, we would be correct 95 times out of a 100 (see, e.g. Neyman, 1941). A statement of this sort is in fact generally accepted in spite of the difficulty in conceiving what we mean by 'in the long run' when we have to invoke an infinite hypothetical population from which we sample for a single experiment which is not repeatable under exactly the same conditions. This statement can however also be interpreted as meaning that if we make such an assertion each time we do an experiment throughout our lives, then we will be mistaken only about 1 in 20 times. The difficulty arises when we try to make a statement about a single experiment. It seems an easy step from the above to say that, so long as we can feel justified in assuming that we can consider the particular experiment in question as a *random* sample from all the experiments which might hypothetically have been done, i.e. that 'nature has done the shuffling for us', the probability that we are correct in asserting that the unknown parameter lies within these particular calculated limits is 95% or that, unless 'a 1 in 20 chance come off', we shall be correct on this particular occasion. Such a statement would, I think, be acceptable to most people, though some prefer to restrict themselves to a statement of frequency in repeated trials.

Fisher however goes even further than this and states that the fiducial argument, where applicable, justifies the explicit statement that 'the probability that the unknown parameter lies within the limits calculated in any particular experiment is 95 per cent. This sounds superficially like a statement of inverse probability.

Jeffreys (1940) says that it is, and asserts the equivalence of his approach, through inverse probability, and Fisher's. Fisher states that it is not, and that he believes that "the theory of inverse probability is founded upon an error". It seems to be a common experience, however, after studying Fisher's justifications of his approach, to agree with Hogben (1957 p. 504) that "again and again we seem to sidestep the notion of inverse probability". In fact Yates (1939) has said of the use of the fiducial argument "we must frankly recognize that we have here introduced a new concept into our methods of inductive inference which cannot be deduced by the rules of logic from already accepted methods".

It may also be objected that as the parameter has a fixed constant value then it must be either inside or outside the limits and therefore the only correct statement of this form must involve a probability of 0 or 1 (as opposed to statements of the form; "the probability that we are *correct in asserting* that . . . . " which I do not think are open to this objection).

The Interpretation of Tests of Significance

The tests of significance in most common use in the medical and biological sciences are Pearson's χ^2 test, "Student's" t test and Fisher's z test (and the closely related variance ratio, or F test). Let us consider the t test, used for example, to test for the significance of the difference between two means, or the difference between an observed mean and some hypothetical value (e.g. zero). The value of t is calculated, in the former case, as the quotient of the difference between the two means and the standard deviation of this difference and this value referred to a table of the distribution of t (given e.g. by Fisher and Yates 1957), taking account of the size of the sample, and from this is discovered a value of P. What assumptions have we made and how are we to interpret the result?

Firstly we have assumed that our observations, (or some function of them, e.g. their logarithms), are normally distributed. Furthermore we have calculated the variance (and hence the standard deviation) of the difference between the means of the two samples by adding their separate variances. This involves the assumption that they are *independent*, i.e. that the values of the variable obtained in one sample are not correlated with those obtained in the other. We then set up a *null hypothesis*, that is we say 'assuming that both the samples had come from the *same* population, what is the probability that, owing to random sampling error, we should observe a deviation from this hypothesis as large as, or larger than that which we have observed?' This probability is the P value we have calculated. If it is very small we then conclude that either

(*a*) an improbable event has occurred or

(*b*) our samples were not chosen randomly or

(*c*) that our hypothesis was not true and that in fact the samples came from *different* populations.

If we decide that the last conclusion is right our justification is that if our hypothesis had been true then it is improbable that we should observe such a large deviation from it. What degree of improbability we are prepared to accept as evidence for a 'significant difference' is entirely a matter of personal judgement. Commonly a value of $P = 0.05$ is interpreted as 'probably significant', $P = 0.01$ as 'significant' and $P = 0.001$ as 'highly significant', but this is quite arbitrary and the results of significance tests should always be stated as a probability (loosely speaking, the probability that the observed difference was due to chance) so that the reader can decide for himself whether he chooses to regard the difference as significant or not. This choice will depend, for instance, on how important it is that he should not be mistaken: on how much he stands to lose if he makes a wrong decision, and how much will be gained if he is right.

Testing two samples with different variances

It is often said that it must be assumed, in doing the t test, that the variances of the two samples are the same and that the hypothesis tested is that "the 'true' *means* of the two samples are equal", so that we calculate the probability that a difference in sample means as large as, or larger than, that observed would have arisen if this were right; and in fact if the variances of the two samples are similar, this is virtually so. However Fisher (1954) has said that the equality of the variances is tested as part of the hypothesis (in the form in which it was first stated above) and that in theory at

least, a deviation from this hypothesis could be due to a difference in the population variances, rather than their means. He recommends that if a test for difference in *means* is required we should use the Behrens-Fisher ('d') test if there is evidence of difference between the sample variances, as this test takes account of the difference. However this test has given rise to controversy among statisticians as, when we use the 5 per cent. significance level (i.e. $P = 0.05$) and apply the test to repeated samples from two fixed populations, it will *not*, in the long run, lead to rejection of the null hypothesis that the population means were equal when it is true, in 5 per cent. of cases. Fisher says that the fact that we cannot use a frequency interpretation of the probability in no way affects the validity of the test, but other statisticians (as well as experimenters) feel the view adopted in this particular problem to be inconsistent with the usual statistical practice followed in other situations (see, e.g. Welch, 1937; Fisher, 1959*a*).

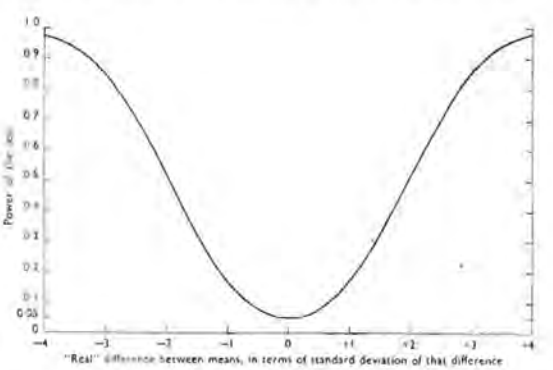
A test has however been devised by Dr. B. L. Welch, on the basis of which we *can* say, at least in the moderate sized samples usually encountered in practice, that the frequency of rejection of the null hypothesis if it is true will be 5 per cent. (or any other selected value) in the long run; i.e. we can use the commonly accepted frequency interpretation of the probabilities calculated in testing for significance or estimating confidence limits (Welch, 1947, and 1956). Tables of the v function (analogous with the t function, but taking into account the difference in sample variances) have been computed for the $P = 0.10$ and $P = 0.02$ significance levels (Aspin, 1949, reproduced as Table 11 of Pearson and Hartley, 1954), and for $P = 0.05$ and $P = 0.01$ (Trickett, Welch and James, 1956).

For most other tests of significance we can make a frequency statement of the above sort as well; and consequently if, every time we find $P < 0.05$, we say we have found a significant difference then we can expect, in the long run, to be mistaken anything up to 1 in 20 times.

If, after all this, we conclude that we think that our difference is real, and not due to chance (and furthermore that is sufficiently large to be of practical as well as statistical significance), then this conclusion can be properly applied only to the particular sample tested. If we wish to induce from this a prediction of what will happen in future tests on a different, or even on the same sample (and this would presumably be the purpose of the experiment) we must endow nature with a predictability which it has often belied. Such a prediction could be made with any security only after having looked closely to ensure as far as possible that the circumstances are similar in all relevant respects on each occasion.

If, however, our result is non-significant, it would be quite irrational to assume that because we have not rejected our null hypothesis we may accept it. For the test discussed above we must say that we *cannot demonstrate* evidence for there being a real difference between the samples, not that there *is* no difference. Likewise when we do a biological assay involving the estimation of dose response curves for a standard and unknown drug we test to see whether the lines deviate significantly from parallelism, or from linearity, and whether their slope is significant. If the assay passes these tests we say that it is '*not demonstrably invalid*', but we can never assert that it is valid.

This is illustrated by Neyman's concept of the 'power' of a significance test. This is defined as the probability that we shall reject the null hypothesis, and can be plotted against different values of the 'true' difference in means (see, e.g., Quenouille, 1958, p. 142). This is illustrated in the diagram. The abscissæ represent various hypothetical values of the deviation from the null hypothesis, e.g. of the difference between the "true" means of two populations, expressed in units of its standard deviation (i.e. the difference divided by its standard deviation). It can be seen that when the null hypothesis (that there is really no difference between the means)



"Power" of a statistical test in relation to deviations from expectation derived from the null hypothesis.

is true, then we would reject it (wrongly) in only about 5 per cent. of cases in the long run, i.e. the power of the test in this situation is 0.05. If however the null hypothesis was *not* true, and the samples came from different populations whose means differed by an amount equal, say, to one standard deviation (of the difference), then we would expect to reject (rightly this time) the null hypothesis that the difference was zero, in about 17 per cent. of cases in the long run, i.e. in about 83 per cent. of tests we would not detect a "true" difference of this size. Clearly if we accepted the null hypothesis as true each time we got a "non-significant" result we should be mistaken in more than 80 per cent. of our decisions! The power of a test, i.e. its ability to detect small deviations from the null hypothesis, can be increased by increasing the size of the sample, or by adopting a more elaborate sort of experimental design by means of which the effects of certain types of variability can be eliminated (see, e.g. Fisher 1951; Finney 1955).

It should be noted that we make no statements of inverse probability about hypotheses on the basis of significance tests. As Fisher (1959*a*) says, if we reject a hypothesis at the 1 per cent. probability level it is not because the probability of the hypothesis is 1 per cent., but because its correctness would entail an event of this low probability.

Correlation and Regression

Fitting Curves to Points

The fitting of lines to two related variables is best done statistically, usually by the method of least squares, in order to obtain the line best representing the data. Curved or straight lines can be fitted in this way but straight lines are preferable because the arithmetic is easier, and usually it is possible to transform one or both of our variables so that a straight line can be fitted over at least part of the extent of the data. For example we frequently find that a graded response is linearly related to the logarithm of the dose of the drug over at least the central part of the dose-response curve. Sometimes we have to use other transformations, e.g. Lauson *et al.* (1939) found that the *square root* of the uterine weight was linearly related to the logarithm of the dose of œstradiol in rats.

We normally fit the simplest curve compatible with the data, because this is the most easily refuted if it is wrong, as well as because it is most convenient. For instance if we have three points we could find a quadratic equation to fit them exactly (in general if we have n points they can be fitted exactly by a polynomial of order n-1); we could equally easily draw a sigmoid curve through them, but nevertheless we will always fit a straight line, if we have no *a priori* evidence to the contrary, unless our test of significance tells us that the points deviate more from the straight line than could be expected by chance.

It is also necessary to test for significance of the slope of the line; it is quite possible that an apparent slope might have varied from the horizontal (or from another slope which we may be interested in) by chance alone.

We must distinguish between the independent (x) and dependent (y) variables when fitting such a line. The former, e.g. dose of drug, is *assumed* to be measured with negligible error, while the latter, e.g. response is variable. Usually there are several measurements of the latter for each value of the former, and we must assume that variance of the dependent variable, e.g. response, does not depend on the independent variable, e.g. dose—i.e. that it is similar in each dose group. This can be tested by Bartlett's test (Bartlett, 1937). It may be necessary to use a transformation to ensure that this is so; e.g. Perry (1950) found it necessary to plot log dose against *log response* in order to obtain a straight dose response line complying with this condition of homogeneity of variances when using survival time as a response for toxicity tests, and Finney (1952) illustrates the use of the angle transformation ($\varphi = \sin^{-1} p^{\frac{1}{2}}$, where p is the response expressed as a proportion) for a similar purpose.

Having fulfilled these conditions we now fit a line. There are in fact two lines that we can fit depending on whether we minimize the sum of squares of the deviations of the values of x, or of y from the line. These lines cross at $(\bar{y}, \bar{x})$ but may have different slopes. We must always fit the line so that the sum of squares of the deviations of the *dependent* variable from the fitted line is at a minimum (see, e.g. Eisenhart, 1939). This is called the regression line of y on x. Sometimes however we cannot distinguish a dependent and an independent variable, for instance if we plot weight against height. In this case we take the geometric mean of the slopes of the two possible regression lines,

($\sqrt{b_1 b_2}$). The slope of this line is called the *correlation coefficient* (r) of the two variables. If y increases as x increases r is positive; if y decreases, r is negative.

Correlation

If two variables A and B are found to be correlated to a greater degree than would be expected by chance we may tentatively conclude that either

(i) A causes B

or (ii) B causes A

or (iii) some other common factor causes both A and B

or (iv) the apparent correlation is a coincidence. Presumably this accounts for the large and positive correlation between the annual divorce rate and the import of apples! (Fisher, 1959*b*).

For instance, there is no doubt that there is a correlation between smoking and the incidence of lung cancer and it would be very interesting to know whether this meant that smoking could cause lung cancer. But we don't. Certainly if a patient asks his doctor whether he should smoke or not it would be quite reasonable that the doctor should recommend that he should not smoke as smoking *might* increase his risk of contracting the disease, but does the evidence justify the conclusion (B.M.J. 1957) that the "dangers of cigarette smoking . . . . must be brought home to the public by all the modern devices of publicity"? Fisher (1959*b*) thinks that it does not and presents several good reasons for this. As he says, the increased incidence of the disease which seems to have taken place over the last 50 years (even allowing as far as possible for improved diagnosis, etc.) is frightening; but it shouldn't be used to frighten people.

The situation would be very different if we could take two large groups of teenagers *at random* and instruct one of them to smoke and the other not to smoke. It is true that the incidence of lung cancer is greater among heavy than among light smokers and that people who give up smoking tend to have the lower incidence of non-smokers, but these facts certainly do not exclude the other possibilities which we must consider—viz. that cancer, or a pre-cancerous state encourages smoking, or that it is some other factor, e.g. the individual genotype, which influences both. Some of the factors which must be explained are that:

(*a*) Although smoking has increased much more in women than in men in the last 50 years, lung cancer has increased more rapidly in men than in women. There seems little reason, therefore, to associate the secular increase in lung cancer with the increase in smoking.

(*b*) Smokers who inhale have a "significantly" ($P \simeq 0.01$) *lower* incidence of lung cancer than those in the same tobacco consumption bracket who don't inhale, by about 10 per cent. (or about 13 per cent. if pipe smokers are excluded).

(*c*) There is evidence (Fisher, 1959*b*) that monozygotic twins are much more alike in their smoking habits than dizygotic twins and furthermore the proportion of 'alike' and 'unlike' among monozygotic twins is much the same whether they have been brought up separately or together, so the effect cannot be ascribed to the greater mutual influence of monozygotic twins. This indicates that the genotype may influence smoking habits.

These interesting facts serve well to underline the already well known, but nevertheless considerable, danger of confusing correlation with cause.

Designing of Experiments

This is a large subject and increasingly recognised to be of great importance if valid conclusions are to be drawn from experiments. I have only enough space here to mention a few points of interest. The most important concepts are control, replication and randomisation.

Control

The need for control groups is now widely recognised. The number and variety of responses to pharmacologically inert dummy tablets is amazing even when the patient and doctor are unaware of the nature of the medicament. For instance Beecher (1955) records an average response rate to dummy treatment in over 1,000 patients in 15 different drug trials, mostly by the "double blind" method, of *c.* 35 per cent. This was a fairly constant figure suggesting a common basic mechanism. Lasagna *et al.* (1958) show that the curve relating analgesic efficacy in post-partum pain to

time after medication is very similar in form for both aspirin and dummy tablets. Both have a time of maximal effect after which the effect declines.

It might seem that dummies would make good therapeutic agents (as in fact they sometimes may); but they can also produce toxic effects, occasionally quite severe, e.g. overwhelming weakness, dermatitis medicamentosa, urticaria, angioneurotic œdema and diarrhœa (Wolf and Pinsky, 1954). Similarly Kerr and Davidson (1958) found that the incidence of gastro-intestinal side-effects was much the same in subjects taking oral iron preparations as in those taking dummy tablets which they thought to be another form of iron, whereas when taking tablets which were labelled "Control Pills", hardly any complained of side effects. Girdwood (1952) showed that white compound ferrous sulphate tablets produced few side effects in patients who "were quite certain that they could not tolerate Fersolate" (green tablets).

Obviously then any group receiving a new treatment must be compared with a control group of subjects receiving a dummy treatment, or alternatively the existing standard treatment for the condition (except perhaps in some diseases, such as acute leukæmia when experience may be sufficient control).

It is also necessary that neither the doctor assessing the effects nor the patient should be aware of the treatment given (double blind method) as is illustrated by Wolf (1959) who quotes a case of a patient receiving dummy tablets for asthma. When the doctor was under the impression that the tablets were not dummies he recorded an improvement, when he knew they were dummies the condition got worse.

Randomisation and Replication

The necessity for replication is common sense, but the importance of proper randomisation is perhaps less appreciated. It is necessary to prevent bias of the results by sources of variability which cannot be controlled. For instance in the example above we were interested in the effect of smoking on lung cancer, not in the possible effects of pre-cancerous states or genotype. If then we could pick out two groups of people at random we might assume that the proportion possessing a given genotype or harbouring a pre-cancerous condition was almost the same in each group, i.e. the *only* difference between the groups was that one smoked and the other did not. Only in this way could a causal relation be established in the face of the apparent contradictions mentioned.

A reasonable method of forming two groups, for instance to be treated by different methods, is to allocate subjects alternately to the two groups. However this may lead to bias if, for instance, it means that the person who decides whether a particular subject is to be included in the trial or not knows which treatment the subject will have before deciding (Doll, 1959). The best method of randomisation is to use a table of random numbers (Fisher and Yates, 1957) or the toss of an unbiased coin.

For the most precise results each subject should act as his own control. For instance in skin disorders it may be possible to try different treatments simultaneously on the same subject. With conditions that recur in a reproducible way, e.g. migraine or epilepsy, treatments can be given to the same patient at different times. The order in which each patient receives each treatment must then be randomised, to prevent any possible effect of the order of treatment from biassing the results. The method of statistical analysis depends on how this randomisation has been done. If each patient receives each treatment the same number of times it is possible to eliminate the variability due to differences between patients (e.g. by subtracting this component from the total variability in an analysis of variance); and if the randomisation is of the right type we can also eliminate variability due to differences in the response depending on order of treatment.

Often however it will be necessary to give the active and control treatments to different patients. It may then be useful to arrange patients in pairs as similar as possible before the test and use the differences between responses of the pair as a measure of the effect of the treatment. This is equivalent to eliminating the variability in response difference between pairs. Patients can in fact be grouped, *before* the experiment, according to any criterion, e.g. prognosis (e.g. good, fair or poor) and if there are significant differences between the groups the variability due to these can be eliminated by appropriate analysis of the results so increasing the precision of the comparison between the treatments.

Conclusion

The use of statistical method and logical experimental design is a valuable tool by means of which the experimenter can make the most of his observations and avoid making unjustifiable claims on the basis of them, as long as the results of the calculations are interpreted with all due respect for their logical basis.

I should like to thank Dr J. G. Dare, Mr. J. R. Lucas (Department of Philosophy), Dr. G. A. Mogey, Mr. W. H. Trickett and Dr. B. L. Welch (Department of Mathematics) for their invaluable advice and help, and criticism of the manuscript.

References

Anscombe, F. J. (1948). "The Validity of Comparative Experiments". *J. Roy. Statist. Soc.* A, **111**, 181.

Aspin, A. A. (1949). "Tables for use in Comparisons whose Accuracy involves two variances, separately estimated" (with an appendix by B. L. Welch). *Biometrika*, **36**, 290.

Ayer, A. J. (1956). "The Problem of Knowledge". Pelican.

Bartlett, M. S. (1937). "Properties of Sufficiency and Statistical Tests". *Proc. Roy. Soc.* A, **160**, 268.

Beecher, H. K. (1955). "The Powerful Placebo". *J. Amer. med. Assoc.*, **159**, 1602.

Brit. Med. J. (1957). "Dangers of Cigarette Smoking". **i**, 1518.

Doll, R. (1959), *in* "Quantitative Methods in Human Pharmacology and Therapeutics". p. 213. Pergamon Press.

Eisenhart, C. (1939). "The Interpretation of Certain Regression Methods and their Use in Biological and Industrial Research". *Ann. Math. Statist.*, **10**, 162.

Finney, D. J. (1952). "Statistical Method in Biological Assay". Charles Griffin.

Finney, D. J. (1955). "Experimental Design and its Statistical Basis". Cambridge U.P.

Fisher, R. A. (1951). "The Design of Experiments". Edinburgh, Oliver and Boyd.

Fisher, R. A. (1954). "Statistical Methods for Research Workers". Edinburgh, Oliver and Boyd.

Fisher, R. A. (1959*a*). "Statistical Methods and Scientific Inference". Edinburgh, Oliver and Boyd.

Fisher, R. A. (1959*b*). "Smoking. The Cancer Controversy". Edinburgh, Oliver and Boyd.

Fisher, R. A. and Yates, F. (1957). "Statistical Tables for Biological, Agricultural and Medical Research". Edinburgh, Oliver and Boyd.

Gaddum, J. H. (1945). "Lognormal Distributions". *Nature*, **156**, 463.

Gaddum, J. H. (1954). "Clinical Pharmacology". *Proc. Roy. Soc. Med.*, **47**, 195.

Girdwood, R. H. (1952). Letter to the Editor. *Brit. med. J.*, **i**, 599.

Hanzlik, P. J. (1913). "A Study of the Toxicity of the Salicylates based on Clinical Statistics". *J. Amer. med. Assoc.*, **60**, 957.

Hogben, L. (1957). "Statistical Theory". London. Allen and Unwin.

Jeffreys, H. (1940). "Note on the Behrens-Fisher Formula". *Ann. Eugen., Lond.*, **10**, 48.

Kerr, D. N. S., and Davidson, S. (1958). "Gastrointestinal Intolerance to Oral Iron Preparations". *Lancet*, **ii**, 489.

Kneale, W. (1949). "Probability and Induction". Oxford. Clarendon Press.

Lauson, H. D., Heller, C. G., Golden, J. B., and Sevringhaus, E. L. (1939). "The Immature rat uterus in the Assay of Estrogenic Substances, and a Comparison of Estradiol, Estrone and Estriol". *Endocrinology*, **24**, 35.

Lasagna, L., Laties, V. G., and Dohan, J. L. (1958). "Further Studies on the Pharmacology of Placebo administration". *J. clin. Invest.*, **37**, 533.

Neyman, J. (1941). "Fiducial Argument and the theory of Confidence Intervals". *Biometrika*, **32**, 128.

Pearson, E. S., and Hartley, H. O. (1954). "Biometrika Tables for Statisticians". **i**. *Cambridge*: The University Press for the Biometrika Trustees.

Perry, W. L. M. (1950). "The Design of Toxicity Tests". Medical Research Council. Special Report Series No. 270. (Reports on Biological Standardisation VI).

Popper, K. R. (1959). "The Logic of Scientific Discovery". Hutchinson.

Quenouille, M. H. (1958). "Fundamentals of Statistical Reasoning". Charles Griffin.

Snedecor, G. W. (1956). "Statistical Methods". Iowa State College Press.

Trickett, W. H., Welch, B. L., and James, G. S. (1956). "Further critical Values for the Two Means Problem". *Biometrika*, **43**, 203.

Welch, B. L. (1937). "The Significance of the Difference between Two Means when the Population Variances are Unequal". *Biometrika*, **29**, 350.

Welch, B. L. (1947). "The Generalisation of Student's Problem when Several Different Population Variances are Involved". *Biometrika*, **34**, 28.

Welch, B. L. (1956). "Note on Some Criticisms made by Sir Ronald Fisher". *J. Roy. Statist. Soc.* B, **18**, 297.

Wolf, S. (1959). "The Pharmacology of Placebos". *Pharmacol. Rev.*, **11**, 689.

Wolf, S., and Pinsky, R. H. (1954). "Effects of Placebo Administration and Occurrence of Toxic Reactions". *J. Amer. med. Assoc.*, **155**, 339.

Wootton, I. D. P., and King, E. J. (1953). "Normal Values for Blood Constituents. Interhospital Differences". *Lancet*, **i**, 470.

Wootton, I. D. P., King, E. J., and Maclean Smith, J. (1951). "The Quantitative approach to Hospital Biochemistry". *Brit. Med. Bull.*, **7**, 307.

Yates, F. (1939). "An Apparent Inconsistency arising from Tests of Significance based on Fiducial Distribution of unknown Parameters". *Proc. Camb. Phil. Soc.*, **35**, 579.

PRINTED BY
JOWETT & SOWRY LTD.
LEEDS 1