REVIEW ARTICLE

STATISTICAL METHOD IN THE FIELD OF BIOLOGICAL ASSAY

By J. O. Irwin, Sc.D.

Medical Research Council's Statistical Research Unit

I. General Ideas

Two recent review articles in this journal have told the story of the International Standards for Vitamins D and A (Coward[1], Morton[2]). No careful reader of those articles can avoid seeing the importance of the part which the statistical treatment of biological assays has played in the development of satisfactory ways of measuring vitamin content. The same is true of all those substances for which international standards have been established, and in an even wider field. Here it is proposed to treat the subject in its own right.

A biological assay, strictly speaking, is an attempt to find out from a trial with living creatures of some particular species how much of a given substance is present, per cent., weight or volume, in a preparation containing the substance. For example, from an assay with rats we may make an estimate of the vitamin A content of a cod-liver oil. Biological assays are usually used, originally, when the exact chemical constitution of the substance is unknown, but it is often convenient to go on using them after the constiution has become known, because of the difficulty of actually isolating and measuring the extremely minute quantities which are present.

In a slightly more general sense if two substances each produce a characteristic effect on members of a particular species of living creature, a comparison of the amounts necessary to produce the same effect—amounts which may or may not be in the same ratio at all levels of response—though they usually are in actual applications—may be called a biological assay.

The field of biological assay is thus wider than the field in which standards are available, but standards are so important in pharmacological work that it is desirable to give special emphasis to this part of the subject.

Any intelligent person can understand what is meant by a standard yard or a standard pound, and has no real difficulty in grasping the implied extension of the notion when we speak of a standard for vitamin D or a standard for insulin, namely, a preparation of the substance in question such that the properties and effects of a given amount of it do not change in time and with which the properties of given amounts of more or less similar substances can be compared. This points the way to a definition of potency. The potency of any preparation is the inverse ratio of the amount of it which produces a given effect to the amount of the standard required to produce the same effect. As far as this defi-

nition goes potency might vary with the type of effect under consideration and with its intensity or—which is the same thing—with the amount of standard which produces an effect of that intensity. This is not what we want to happen, but it quite often is what in fact happens.

Once we have a standard a unit may be defined as a given amount of the standard, though even here there may be complications.

Let us illustrate the difficulties by a particular example. We have, say a standard preparation of vitamin A. We are presented with a cod-liver oil, and we want to know how much vitamin A it contains. No question at first sight could seem clearer! We will suppose that for one reason or another a chemical or physical determination is impracticable, so we have to use a biological method. That is the real object of a biological assay, to find out how much of a given drug (the term " drug " is used here in a very general sense) is contained, per unit weight or volume, in a substance under test. This should be emphasised. The biological assay of a drug, we think, should not as such be concerned with the therapeutic effects of the drug in man. That is a different question, and confusion arises unless the two questions are separated conceptually. Compromises in practice, for reasons which will shortly become apparent, will sometimes be necessary.

Suppose that we are in the same position as before the 1949 World Health Organisation Conference on Biological Standardisation. If we are in England we turn to the British Pharmacopœia. We find the following statement: " The standard preparation of vitamin A is a quantity of pure β-carotene. The unit is the same as the international unit. It is the specific activity contained in 0·6 μg. of the standard preparation in use." We do not need to be Socrates to ask " Specific for what? " No clear guidance is given, but as the method of assay suggested is based on the increase of growth in rats, we have to assume that the ratio of the amounts of the cod-liver oil in question and of vitamin A which produce the same effect on the growth of rats (a ratio assumed to be the same at all levels of dosage) remains the same if for the rat test we substitute any other *bona-fide* biological test that might be suggested.

Now there is a rather special difficulty here because β-carotene is not vitamin A, and this has led last year's WHO Conference to recommend the replacement of the β-carotene standard by a preparation of vitamin A acetate. This difficulty has occurred on several occasions, when it has been found that a substance originally assumed, it may be tacitly, to be a pure chemical compound of a particular type was not so in fact. The assay of digitalis is in this position, because digitalis is a mixture of several compounds in unknown proportions; at present therefore the assay of digitalis has to be an assay of " activity " if it is to fulfil as well as possible the practical end of enabling safe and efficacious doses to be prescribed. Here, the ultimate aim should be the ability to state exactly what compounds—and in what proportions—any given preparation contains. Until this is achieved statements about the " activity " of any preparation of digitalis are inevitably to some extent tendentious. The word inevitably is used on purpose, for this is not

meant as a criticism of the efforts of those who carry out assays as well as they can, it is merely a plea for the effort to think out clearly what is being done.

But let us return to vitamin A and suppose we are referring to the new standard—which is what it is intended to be—and see what difficulties remain. Even if the data of the test satisfy the usual statistical criteria (we shall see later what they are), caution is still required. "The vitamin A in the oil" is an ambiguous phrase. It may not and usually will not all be in the form of preformed vitamin A, it may be in the form of β-carotene and be converted into vitamin A in the animal body. If the statistical criteria are satisfied, we know that the total amount of vitamin A utilised bears a constant proportion to the dose of oil given, but this provides in itself no proof that all the β-carotene is converted into vitamin A and that all the vitamin A is used. If this is not the case, a test with a different species of animal might give different results for the vitamin A content of the oil.

This actually happens with vitamin D, which may be a mixture of vitamin D_2 and vitamin D_3. Amounts of vitamin D_2 and vitamin D_3 which are equivalent for rats are far from equivalent for chicks, which can utilise the D_3 and not the D_2. Consequently if a mixture is assayed against D_2 (or against D_3) one obtains different results for rats and chicks.

In the case of vitamin A, fortunately, a check on the biological assay exists. Vitamin A can be assayed spectrophotometrically, and in ordinary practice now is always so assayed; while the value of the conversion factor is implicit in the definition of the new standard. Professor Morton[2] has shown how difficulties about irrelevant absorption can be surmounted, so that it will become possible to state the vitamin A content of an oil in say μg./g. as soon as its spectrophotometric value is known. When this stage is reached a standard will be unnecessary. Sir Henry Dale one said: "The ultimate aim of all progressive work in biological standardisation, as in all progressive medicine, is self-extinction." Vitamin B_1 and vitamin C, being pure substances whose constitution is now known, have already reached this stage. They are controlled by chemical and physical tests and the description of their biological assays has not been included in the *British Pharmacopœia*, 1948. But biological assay will nevertheless remain an indispensable method in pharmacological work for many years to come.

To sum up: If we are given a standard there is no difficulty in defining a unit. The unit is defined as the specific biological activity of a given amount of the standard. It cannot be defined as the given amount itself, because we may want to assay against the standard substances which exhibit the "specific activity" but are not necessarily of the same chemical form. "Specific activity" although somewhat tendentious is an unavoidable phrase. It has as its background a working hypothesis which often has to be abandoned as more is learnt of the drug. A substance which initially has been regarded as though it were a pure chemical compound has often been found to be a mixture of several. The ideal thing is then to enable each of these to be assayed separately, either by

biological or preferably by physical or chemical means. When the constitution of each is known and they can be synthesised we are approaching the stage when the standard will be unnecessary. To make it unnecessary is the ultimate aim of research.

There is no difficulty in defining potency provided we are prepared to admit that it may vary at different levels of dosage or in tests with different species of animals. When this happens the definition is deprived of much of its practical utility, but the results are an indication that more fundamental research is required until the situation is cleared up.

II. Statistical Technique and Design

(*i*) *Technique.* What makes statistical technique necessary in dealing with biological assays is animal variability. No two animals of the same species are exactly alike in their response to any stimulus, and even among litter-mates the variability is usually considerable though less than among unrelated animals. For example the coefficient of variation of increase in weight of 100 female rats of a stock colony between the ages of 5 weeks and 10 weeks was 23 per cent. In a line test of vitamin D with female rats the coefficient of variation of "area of healing" was 23 per cent. for non-litter mates and 14 per cent. for litter-mates.

When it is proposed to introduce a new test for biological assay purposes, after deciding on the response to be observed, the first thing to do is to examine the relation between the average response of a group of animals to the dose of standard given and the dose itself. This is called the dosage-response relation. The dosage is the mathematical function of the dose actually used in calculation. It may be the dose itself, it is commonly the logarithm of the dose but may be some other function. Responses are of two kinds "measurable" and "all or none" or "quantal" as the latter are termed technically. The statistical treatment of assays based on the two kinds of responses are in many respects different. In the former case the dosage-response relation is between the mean response to the dose given and the dose itself, in the latter between the percentage of animals responding positively and the dose. Such a percentage can of course always be regarded as the mean value of a variable which is 100 if the animal responds positively and 0 if it responds negatively.

Dosage-response curves in any satisfactory test are of the same general form for both standard and unknown and retain this form when the test is repeated subsequently. They will of course differ in position according to the relative strength of standard and unknown. If there were no animal variability these curves would be smooth and invariable for one and the same preparation. But this is far from being the case. Statistical technique is enormously simplified if the dosage-response curves are straight lines. In most biological assays for which standards exist and for which the response is measurable, it is found to be linearly related to the logarithm of the dose over a sufficient range for working purposes. This means that in a satisfactory test, the dosage-response relations for the preparation under test and the standard will be repre-

sented by parallel straight lines. The horizontal distance between the two straight lines therefore gives the difference between the logarithm of doses of test and standard which produce the same effects—and this immediately provides the potency ratio of the two preparations. The function of statistical method is to estimate from the data of the test, the position and slope of the dosage-response straight lines—and from this the estimate of potency follows at once—then to estimate the accuracy of the result obtained.

Because the animals are variable the mean values of the responses at the different dose levels will not lie smoothly on straight lines, but will have "sampling errors." Not only is there variation in response to a given dose from animal to animal, but a whole colony of animals may undergo fluctuations in sensitivity over a period of time. In order therefore that the estimates of potency and of accuracy (or "error") may be valid, two conditions must be satisfied. The animals selected for each of the dosage groups of both preparations must be selected at random from the stock, and in every assay there must be a simultaneous comparison of the unknown and standard preparations. At any rate until enough is known of the test for it to become a routine procedure there should be at least three dose levels of each preparation (more are sometimes desirable). Even when the test has become a routine test, there should be two doses of each preparation. These conditions ensure that the slopes and positions of the lines are not biased and enable a check to be kept on variations in slope, and any tendency to depart from straightness.

Many accounts have been given of the statistical procedure necessary in fitting straight lines to the data, and in determining the potency and its error, for instance in papers by Irwin[3], Fieller[4], Finney[5,6] and in the text-books of Coward[7] and Emmens[8]. It is only necessary to say here that accuracy or "error" is measured by calculating fiducial limits of error. These are limits calculated by a rule such that the true value would lie between them in a specified percentage usually (95 per cent.) or (99 per cent.) of repetitions of the assay under essentially the same conditions.

When responses are quantal the relation between response and log-dose can often be transformed into a straight line. How this is done requires a little explanation. Any individual animal has a tolerance (or minimum individual effective dose) which may be defined as the least dose to which he will respond positively. Sometimes the tolerance may be measured directly. For instance in the cat assay of digitalis, the preparation is injected continuously until the cat dies so that the least amount required to kill is obtained separately for each cat. In this case no elaborate statistical treatment is necessary. If the required number of cats are selected and half, chosen randomly, are put on the standard and half on the test preparation the ratio of the mean tolerances—the tolerance is here the individual lethal dose, or the difference between mean log-tolerances will provide the potency ratio. This only assumes that the tolerances of the same animal to the two preparations are always in a constant ratio. The error is then obtained by the usual elementary statistical methods.

As a rule however the individual tolerances cannot be measured directly, but some function of them (often the logarithm) will have a normal frequency distribution in animals of the type used. In this case the proportion of animals P who respond positively to any dose are those whose tolerances are less than the dose in question. If m and σ are the mean and standard deviation of the distribution of tolerances on the dosage scale used it is known that

$$P = \int_{-\infty}^{Y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

where $Y = (x - m)/\sigma$ and x is the dosage. Hence if a value of P is given and the corresponding Y obtained from it (many tables exist for the purpose) the relation between Y and x is linear. The curve of Y against x is a straight line whose slope is the reciprocal of σ. Y is called a *normal equivalent deviation* and $Y + 5$ a *probit*.

Hence if groups of animals are put on to a series of doses of the test and standard preparations, if the percentage of positive responses are noted and the corresponding normal equivalent deviations or probits are obtained from them, straight lines may be fitted to the data for the standard and unknown preparations.

Departures of the observed values from the lines fitted should not be more than can be accounted for by the sampling variation of the animals, nor should any departure from parallelism of the two lines. From the actual data of the assay it is possible to test whether this is true, with a sufficient probability. The calculation of the potency ratio and of fiducial limits can then be performed in much the same way in *principle* as for the case when the response is not quantal but measurable. There are complications in detail as regards the weighting of the observations and methods involving successive approximation have to be used. Finney's textbook on probit analysis[6] gives an admirable account of the technique.

In a rather exceptional class of cases in which some microbiological assays are included response is linearly related to the dose itself. In that case the potency ratios will be given by the ratio of the slopes of the two lines. The lines may be estimated by the usual statistical technique of regression, and fiducial limits for the potency ratio may be calculated. It is interesting to note that whether the response is linearly related to the logarithm of the dose or to the dose itself, the problem of estimating error reduces to the statistical problem of calculating fiducial limits for a ratio.

(*ii*) *Design.* The need for consultation with a statistician over the design of an assay is now generally recognised. The amount of information that a biological experiment or test of any kind will provide, for a given number of animals used, depends largely on the design. If the latter does not satisfy certain criteria, it will be impossible to obtain a valid estimate of the accuracy of the result.

The necessity of randomisation has already been mentioned. A definite randomisation procedure is necessary so that each animal has an equal chance of being allotted to every dosage group. This is greatly facilitated by the existence of tables of random numbers such as that given in Fisher and Yates Tables[9]. An excellent passage from Emmens' textbook may be quoted here. "Many workers have been under the impression that such a procedure as taking the first 20 animals that come to hand from a cage and allotting them to the first dosage group, taking the next 20 and allotting them to the second dosage group and so on constitutes random selection. This is most definitely not the case, the first 20 animals that come to hand will often be the tamest animals. They may be the biggest animals and they will quite rarely be representative of the group as a whole. A striking instance of this occurred when an assistant was requested to select groups of mice at random, and it was afterwards possible to demonstrate a highly significant correlation between the order in which the animals were taken from the cage, and the weight of the animals. Similarly it is not random selection to allot the top-rack in an animal room to one dose, the second rack to another and so on, because the position of the animals in the room will sometimes affect their response in the tests. The top of the room may be lighter than the bottom; one wall may be warmer than another; and animals in the one position may receive more food than those in another if assistants feed them in a set routine, and these are factors likely to affect the results of a large number of tests. The order in which doses are administered may also affect results; this is particularly likely to happen when the response is measured within a short time after administration or if the drug is given at a certain period after preparation of the test object. Thus whenever such factors are even remotely likely to affect results, the order of administration of doses should also be determined by a process of randomisation. It should be noted also that attempts to adjust groups of animals so that their mean weights shall be approximately the same are open to criticism. Methods of making such adjustments and of allowing for differences which may be found to exist, which are more statistically acceptable, will be described later on." The last two sentences are a reference to the statistical technique of analysis of covariance, which is the best way of allowing for uncontrolled concomitant variation.

The advantage of using litter-mates has also been stressed. If for example we have two dosage groups of the standard preparation and two of the unknown, and litter-mates of four (preferably of the same sex) are available, one member of each litter may be placed on each dose.

Comparisons between the two preparations are then unaffected by litter differences, and the error of the assay is reduced. The correct error may be estimated by use of what are now well-known analysis of variance procedures. Here we have an example of randomisation subject to a simple restriction. More complicated restrictions are often useful.

For example, an assay of insulin using rabbits with percentage blood sugar reduction or final blood sugar as a response might have a Latin

Square design. Six rabbits might each be tested on 6 days and 3 dosage levels used for both the standard and unknown preparations. The arrangement might be as follows:—

Days	1	2	3	4	5	6
Rabbits :—						
I	S_1	S_2	U_2	U_1	U_3	S_3
II	U_3	U_2	S_1	S_3	U_1	S_2
III	S_2	S_1	U_1	U_3	S_3	U_2
IV	S_3	U_3	S_2	S_1	U_2	U_1
V	U_1	S_3	U_3	U_2	S_2	S_1
VI	U_2	U_1	S_3	S_2	S_1	U_3

The Latin Square is one chosen at random from the possible 6×6 Latin Squares. Each rabbit has every dose once, and each dose is given once on each day. Thus day to day and rabbit to rabbit variations in sensitivity are eliminated from the comparisons, and may, by statistical analysis, be eliminated from the estimate of error. The analysis of variance procedure necessary is now a standard technique and needs no elaboration here. If say 24 rabbits were available, 4 groups of 6 rabbits with 4 separately randomised Latin Squares could be used.

Many variations in design are possible to meet differing experimental circumstances; but all designs must satisfy the requirement of adequate randomisation and replication. Some useful examples are given in the textbooks of Finney and Emmens. R. A. Fisher's "Design of Experiments"[10] lays down the principles necessary in the wider field of biological experimentation which includes that discussed here; the recent textbook of Cochran and Cox, "Experimental Designs,"[11] describes and gives examples of all the types of design hitherto suggested.

III. History

An excellent account of the history of biological standards was given by Sir Percival Hartley[12] in his Dixon lecture of 1945. As regards the development and application of statistical methods, the reviewer's 1937 paper[3] gave a not unreasonable account of what had been done up to that time; he would now only regard it as a datum line from which to reckon advances made by others. A very fine bibliography was published by Bliss and Cattell[13] in 1943; Bliss[14] also summarised the work done on fiducial (or confidence) limits, in the first volume of Biometrics in 1945. In 1946 Finney gave the Research Section of the Royal Statistical Society an account of progress since 1937, particularly mentioning Fieller's work published in 1941. His textbook and that of Emmens have already been mentioned.

As Bliss and Cattell say, few references antedate the textbooks by Burn and Coward; very little was done in the twenties if we except Trevan's important paper[15] in the Proceedings of the Royal Society for 1927. Trevan really inspired Gaddum who is the real inventor of the modern statistical technique of treating quantal responses in biological assay.

It is interesting to note, however, that the statistical ideas behind the

quantal response technique goes back to the work of the psychophysicists Fechner[16] and Müller[17] in the last century, and that of Urban[18] and Thomson[19] about 1910. Fechner seems to have been the real discoverer of the method, which he used in discussing the distribution of just perceptible differences in weight.

The most important advances in statistical methodology since 1937 have been advances in design. Bliss and Marks[20] led the way with their now famous work on insulin and in a very long series of papers, many fertile suggestions have come from the former. Next advances in methods of stating errors must be mentioned, the use of confidence or fiducial limits for ratios. Fieller, Bliss, Finney and the reviewer have all played their part in this work. The statistical techniques necessary for dealing with slope-ratio assays where the response is linearly related to the actual dose given have been developed by Finney and Wood[21]. Recently Irwin has reconsidered the problem of the combination of results from different assays which Fieller was the first to deal with in any exact way; Armitage[22] and Irwin[23] have compared the results obtained from the alternative assumptions of logistic and normal tolerance distributions in the quantal case, and Irwin has examined the adequancy of the usual χ^2 test for the satisfactory fit of linear probit-dosage response curves. The remaining advances since 1937 have been in the nature of particular applications of general advances in statistical technique, such as the use of covariance to allow for concomitant variation and the transformation of dosage-response scales (other than the probit transformation which came earlier) to effect linearity or equalise variance.

The development of methods suitable for biological assay has been an outstanding example of the value of scientific co-operation. Previous review articles in this Journal by Coward and by Morton have elaborated particular instances of how this co-operation developed and shown to what useful results it led. The whole subject was fully discussed at the First International Conference of the Biometric Society at Geneva in 1949. The writer of this article is a statistician and the names of the leading statistical contributors to the subject have been mentioned in its course. He would like to conclude by emphasising his own personal admiration for the magnificent achievement of the pioneers who succeeded in getting standards established, people like Dale, Gautier, Gaddum, Hartley and Trevan, thereby enabling many of the newer discoveries of medicine to be utilised on a comparable basis throughout the world to the immense advantage of thousands of sufferers.

References

1. Coward, *J. Pharm. Pharmacol.*, 1949, **1**, 737.
2. Morton, *J. Pharm. Pharmacol.*, 1950, **2**, 129.
3. Irwin, *J. roy. statist. Soc., Suppl.*, 1937, **4**, 1.
4. Fieller, *J. roy. statist. Soc., Suppl.*, 1941, **7**, 1.
5. Finney, *J. roy. statist. Soc., Suppl.*, 1947, **9**, 46.
6. Finney, *Probit Analysis*, Cambridge University Press, 1947.
7. Coward, *Biological Standardisation of the Vitamins*, Baillière, Tindall and Cox, London, 1947, 2nd ed.

8. Emmens, *Principles of Biolcgical Assay,* Chapman and Hall, London, **1948.**
9. Fisher and Yate, *Statistical Tables for Biological, Agricultural and Medical Research,* Oliver and Boyd, London, 1948, 3rd ed.
10. Fisher, *The Design of Experiments,* Oliver and Boyd, London, 1947, 4th ed.
11. Cochran and Cox, *Experimental Designs,* John Wiley, New York; Chapman and Hall, London, 1950.
12. Hartley, *Proc. R. Soc. Med.,* 1945, **39,** 45.
13. Bliss and Cattell, *Ann. Rev. Physiol.,* 1943, **5,** 479 .
14. Bliss, *Biometrics,* 1945, **1,** 57.
15. Trevan, *Proc. roy. Soc.,* 1927, **101B,** 483.
16. Fechner, *Elemente der Psychophysik,* Breitkopf and Härtel, Leipzig, **1860.**
17. Müller, *Pflüg. Arch. ges. Physiol.,* 1879, **19,** 191.
18. Urban, *Arch. ges. Psychol.,* 1909 **15,** 261.
19. Thomson, *Biometrika,* 1919, **12,** 216.
20. Bliss and Marks, *Quart. J. Pharm. Pharmacol.,* 1939 **12,** 82.
21. Finney, *Quart. J. Pharm. Pharmacol.,* 1945, **18,** 77.
22. Armitage and Allen, *J. Hyg., Camb.,* 1950, in the Press.
23. Irwin, *J. Hyg., Camb.,* 1950, **48,** 215.