BIOLOGICAL ASSAYS WITH SPECIAL REFERENCE TO BIOLOGICAL STANDARDS

By J. O. IRWIN

ADDRESS GIVEN AT THE SECOND INTERNATIONAL BIOMETRIC CONFERENCE, GENEVA, 30 AUGUST TO 2 SEPTEMBER 1949

(With 1 Figure in the Text)

1. HISTORY

A paper on biological assays and biological standards might be expected to start with a definition. What is a biological assay and what is a biological standard? However, I shall for the moment make the somewhat doubtful assumption that we all know the answer to those questions but shall return to them later, making a start meanwhile by discussing some aspects of the history of our subject.

(1) *History of standards*

An excellent account of the history of international biological standards was given by Sir Percival Hartley in this Dixon lecture of 1945. The story starts with diphtheria antitoxin which was discovered by Behring in 1890. I quote Hartley at this point:

> Roux and his colleagues, whose communication to the International Congress of Hygiene at Budapest in 1894 reporting the clinical efficacy of diphtheria antitoxin created such a deep impression, attempted with little success to determine potency by a complicated method in which the weight of the animal, the dose of antitoxin given and the time of survival following the injection of living culture all had to be taken into account, and while the early attempts of Behring and the German workers based on determining the amount of antitoxin which neutralized a certain number of so-called 'minimum lethal doses' of toxin was simpler, these like the French attempts failed because the highly complex nature of diphtheria toxin was not then understood. An appeal was made to Ehrlich, not specially interested at that time in problems of this kind and not yet the famous world figure which he subsequently became, who was thus led to a series of investigations which not only solved these immediate problems of antitoxin standardization but laid the foundations of what has properly come to be recognized as the modern science of biological standardization. Ehrlich revealed the complexity of diphtheria toxin and explored its various reactions with antitoxin, and his discovery that 'toxins' contained in variable amount a substance which, though without toxic action, nevertheless neutralized antitoxin provided an explanation of Behring's bewilderment and failures. As a result of these investigations Ehrlich showed that the only way out of the maze and the difficulties was to adopt a sample of diphtheria antitoxin as a 'standard' in comparison with which the potency of other samples of diphtheria antitoxin could be determined; and he also showed that the unit of diphtheria antitoxin was properly defined in terms of the standard, viz. as the specific biological activity (in this case the neutralizing or combining power of diphtheria toxin) contained in a given quantity of the standard—two important fundamental principles; and he passed to a third. He recognized, that, since the unit must be a fixed unvarying quantity, it was essential that the standard must also be fixed and stable. Accordingly, since liquid preparations of antitoxin lose their potency slowly with age, Ehrlich reduced the sample of diphtheria antitoxin selected for the standard to the absolutely dry condition and preserved it constantly at low temperatures *in vacuo*.
>
> The first important standard for diphtheria antitoxin was supplied from Ehrlich's Institute at Frankfort to laboratories all over the world....In 1905 another and separate standard for diphtheria antitoxin was established at the Hygienic Laboratory at Washington and the unit

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

> of diphtheria antitoxin which Ehrlich had established was defined as a weight of this standard also....Moreover, this work at Frankfort and Washington has proved so successful in regard to the assay of diphtheria antitoxin that in both these countries dry stable standards were established for tetanus antitoxin....Thus the American standard and unit for tetanus antitoxin were well established and in common use, particularly in the laboratories of the United States and of this country, for some years before the outbreak of war in 1914.

Hartley then goes on to describe the beginnings of international co-operation. The cessation during the first war of the Frankfort supply of diphtheria antitoxin led to the establishment of an international standard first for this substance and then for other substances. In 1921 an international conference held in London, at Thorwald Madsen's initiative, under the auspices of the Health Organization of the League of Nations, ascertained that the unit of diphtheria antitoxin based on the standard established at Washington did not differ from the Frankfort unit, and recommended the adoption of the original unit of Ehrlich as the international unit. A second international conference met in Paris in 1922 to explore the possibility of establishing standards and units for other antitoxins and antisera. Madsen suggested to Sir Henry Dale in 1922 that something on similar lines might be done for other substances. This led to the Edinburgh conference of 1923 and the Geneva conference in 1925—at both of which Dale was Chairman—and at which international standards for digitalis, pituitary (posterior lobe) extract, insulin and the arsphenamines were adopted. In 1924 the Health organization of the League of Nations formally instituted the Permanent Commission on Biological Standardization, a small body of experts representative of different countries. The personnel was enlarged somewhat in 1935.

> The international standards for the antitoxins, antisera and for tuberculin are maintained at the State Serum Institute at Copenhagen, while the international standards for all other substances, insulin, pituitary extracts, the arsphenamines, the vitamins, the sex hormones, the heart drugs, heparin and penicillin are maintained at the National Institute of Medical Research at Hampstead, London.

The main decisions on vitamin standards were reached at two international conferences held in London, the first in 1931 and the second in 1934. Sir Edward Mellanby was Chairman on both occasions. The first conference adopted provisional standards for vitamins A, B and D and defined units in terms of them. In particular, a solution of irradiated ergosterol was adopted as the standard for D. At the second conference β-carotene was adopted for the standard of vitamin A, the adsorbate on fuller's earth was continued for vitamin B_1, ascorbic acid was adopted for vitamin C, and it was suggested that calciferol (pure crystalline vitamin D_2) should eventually become the basic standard for vitamin D. The adsorbate was replaced by the pure synthetic vitamin B_1 in 1938.

A third international conference was planned for 1939 but had to be cancelled owing to the outbreak of war. The postponed conference has recently been held in London, this year, under the auspices of the Health Organization of W.H.O., which is continuing the functions of the old League of Nations Health Organization.

The conference recommended the replacement of the β-carotene standard by crystalline vitamin A acetate, and found unacceptable the suggestion made in 1934 that the standard of irradiated ergosterol should eventually be replaced by pure crystalline vitamin D_2 (calciferol). It was recognized even in 1934 that the vitamin D_2 standard was not a suitable standard for determining the vitamin D activity of

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

poultry. The conference accordingly recommended the replacement of the irradiated ergosterol by a preparation of crystalline vitamin D_3.

The present standards of vitamin B_1 and of vitamin C being pure substances, their biological assay is now seldom required. They are controlled by chemical and physical tests and the description of their assays has not been included in the *British Pharmacopoeia* for 1948. This is an example of the achievement or near achievement in practice of the consummation which Sir Henry Dale perhaps devoutly desired but at any rate formulated in the phrase: 'The ultimate aim of all progressive work on biological standardization, as in all progressive medicine, may be regarded as self-extinction.'

Hartley's Table of International Biological Standards is given in Table 1.

Table 1. *International biological standards*

Standard preparation	Adopted	International unit (mg.)	International centre
Diphtheria antitoxin	1922	0·0628	State Serum Institute, Copenhagen, Denmark
Tetanus antitoxin	1928	0·1547	
Anti-dysentery serum (Shiga)	1928	0·0500	
Staphylococcus α-antitoxin	1934	0·500	
Anti-pneumococcus serum (Type I)	1934	0·0886	
Anti-pneumococcus serum (Type II)	1934	0·0894	
Gas-gangrene antitoxin (Perfringens)	1931	0·2660	
Gas-gangrene antitoxin (Vibrion septique)	1934	0·2377	
Gas-gangrene antitoxin (Oedematiens)	1934	0·2681	
Gas-gangrene antitoxin (Histolyticus)	1935	0·3575	
Gas-gangrene antitoxin (Sordelli)	1938	0·1334	
Old-tuberculin	1931	—	
Diphtheria antitoxin for flocculation test	1935	—	
Vitamin A: Mixed carotenes	1931	0·001	National Institute for Medical Research, Hampstead, London
Pure β-carotene	1934	0·0006	
Vitamin B_1: adsorption product of vitamin B_1	1931	10·0	
Pure synthetic vitamin B_1	1938	0·003	
Vitamin C: l-ascorbic acid	1934	0·05	
Vitamin D: Irradiated ergosterol solution	1931	0·01	
Calciferol	1934	0·000025	
Vitamin E: α-tocopheryl acetate	1941	1·0	
Arsphenamine	1925	—	
Neoarsphenamine	1925	—	
Sulpharsphenamine	1925	—	
Insulin: Crude dry insulin hydrochloride	1925	0·125	
Pure crystalline insulin	1935	0·0455	
Pituitary (posterior lobe) powder	1925	0·5	
Digitalis	1925	80·0	
Ouabain	1928	—	
Oestrus-producing hormones:			
(1) Oestrone	1932	0·0001	
(2) Oestradiol monobenzoate	1935	0·0001	
Androsterone (for male hormone)	1935	0·1	
Corpus luteum hormone (progesterone)	1935	1·0	
Chorionic gonadotrophin	1938	0·1	
Serum gonadotrophin	1938	0·25	
Thyrotrophin	1938	—	
Prolactin (galactin or mammotrophin)	1938	0·1	
Heparin	1942	0·0077	
Penicillin	1944	0·0006	

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

(2) *History of statistical methods*

I shall not attempt to retrace the history of the development and application of statistical methods in biological assay technique. I think my 1937 (Irwin, 1937) paper gave a not unreasonable account of what had been done up to that time, but I should now only reckon it as a datum line from which to reckon advances by others. Furthermore, a very fine bibliography was published by Bliss & Cattell in 1943; Bliss also summarized the work done on 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, while in 1947 he gave us his fine text-book on *Probit Analysis*, which summarizes almost everything worth while done up to that time in the field of assays involving quantal responses. Still more recently C. W. Emmens has provided a text-book on biological assays (Emmens, 1948).

As Bliss & Cattell say, few references antedate the text-books by Burn and Coward; my impression is that very little was done in the twenties if we except Trevan's important paper in the *Proceedings of the Royal Society* for 1927. Trevan really inspired Gaddum (1933), who is the real inventor of the modern statistical technique of treating quantal responses.

If I were asked to say what are the most important advances in statistical methodology in the field since 1937 I should reply: 'In the first place, advances in design!' Here Bliss & Marks (1939) led the way with their now famous work on insulin, and I should find it difficult to enumerate all the fertile suggestions about design which have come from the former. Secondly, I should mention the advance in methods of stating errors, in other words, the use made of confidence or fiducial limits for ratios. Here I think the principal credit is Fieller's, though Bliss, Finney and the reader of this paper have all played their part. Thirdly, techniques for slope-ratio assays largely developed by Finney and Wood must not be forgotten. The remaining advances 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 or response scales (other than the probit transformation which came earlier) to effect linearity or equalize variance.

II. GENERAL IDEAS

Now let us return for the moment to the fundamental questions. What is a biological assay? What is a standard? What is a unit? In my 1937 paper, I did not attempt a definition but said: 'There are some therapeutic and other substances whose activity can only be tested by experiments on animals. The object of a biological assay is, in essence, to compare the potency of the particular preparation under test with that of a standard preparation of the same substance.' Reflexion since made me very conscious of the question-begging words 'activity, potency'. So I turned to other writers to see what they say. Bliss & Cattell say: 'Biological assays may be defined as determinations of potency or toxicity based upon the reaction of living matter, including biological reactions not involving intact cells, such as serological tests *in vitro*.' In the opening words of his text-book Finney says: 'The

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

term *biological assay*, in its widest sense, should be understood to mean the measurement of the potency of any stimulus, physical, chemical or biological, physiological or psychological, by means of the reactions which it produces in living matter. The biological method of measuring the stimulus is adopted either for lack of any alternative, or because an exact physical or chemical measurement of stimulus intensity may need translation into biological units before it can be put to practical use.' This seems to leave one quite uncertain as to whether one is measuring a stimulus or the potency of a stimulus, whatever that may be. For myself, I doubt whether a stimulus in the sense intended can be anything but physical, though its mode of production or its effects can be of any of the kinds mentioned. Emmens, perhaps wisely, does not attempt to define a biological assay but starts from the notion of a standard. This hardly seems to need definition. 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 definition 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 very often is what in fact happens.

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 (I use the term drug in a very general sense) is contained, per unit weight or volume, in a substance under test. I should like to emphasize this. I do not think the biological assay of a drug should 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.

I am speaking of the position prior to this year's W.H.O. conference on Biological Standardization. If we are in England we turn to the *British Pharmacopoeia*. 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

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

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 this year's W.H.O. 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 I think 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. I use the word *inevitably* 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. The sampling variation of the animals need not detain us, for this can be allowed for by modern designs. The rat-growth test gives a linear relation between response and log-dose, and we can assume, as is in fact the case, that the straight lines for the standard and the oil under test show no significant departure from parallelism. (If they did we should be on our guard at once and conclude, supposing a satisfactory design had been used, that at different dose levels, different proportions of the vitamin A in the oil were being used by the rats.) Nevertheless, 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. Parallelism suggests that the total amount of vitamin A utilized bears a constant proportion to the dose of oil given, but 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 utilize the D_3 and not the D_2. Consequently, if a mixture is assayed against D_2 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 unit of the new standard. Difficulties about irrelevant absorption are being rapidly surmounted, and when these are finally overcome it will be possible to state the vitamin A content of an oil in say μg./g. as soon as its spectrophoto-

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

metric value is known. When this stage is reached a standard will be unnecessary and Sir Henry Dale's consummation will be attained.

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 in 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 about the drug. A substance which initially has been regarded as though it were a pure chemical compound has later 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 synthesized we are approaching the stage when the standard will be unnecessary. To make it unnecessary should be 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.

III. STATISTICAL TECHNIQUE

In this section I propose to refer to some points in statistical technique which seem to me still to require further elucidation and on which perhaps I can shed a little light. They may seem to form rather a disconnected list, but I have met them all in my work and think they are worth putting on record.

(1) *Equivalence of formulae for fiducial limits*

The formula for the fiducial limits of the result of a single assay, when the response is linearly related to the logarithm of the dose, has been put into several forms whose equivalence is not immediately obvious.

In a usual notation let the logarithm of the result be given by

$$M = \bar{x}_1 - \bar{x}_2 + \frac{\bar{y}_2 - \bar{y}_1}{b},$$

where $\bar{x}_1, \bar{y}_1$ refer to the standard and $\bar{x}_2, \bar{y}_2$ to the preparation under test, and b is the slope. Let

$$M' = M - \bar{x}_1 + \bar{x}_2 = \frac{\bar{y}_2 - \bar{y}_1}{b}. \tag{1}$$

In a paper in the *Journal of Hygiene* (1943), I gave the following expression for the logarithm of fiducial limits:

$$\bar{x}_1 - \bar{x}_2 + \frac{b}{b^2 - t^2B}(\bar{y}_2 - \bar{y}_1) \pm \frac{t}{b^2 - t^2B}\sqrt{\{A(b^2 - t^2B) + B(\bar{y}_2 - \bar{y}_1)^2\}}, \tag{2}$$

which is equivalent to

$$M + \frac{t^2B(\bar{y}_2 - \bar{y}_1)}{b(b^2 - t^2B)} \pm \frac{t}{b^2 - t^2B}\sqrt{\{A(b^2 - t^2B) + B(\bar{y}_2 - \bar{y}_1)^2\}}. \tag{2 bis}$$

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

Here $$A = V(\bar{y}_2-\bar{y}_1) = s^2\left\{\frac{1}{S_1(W)}+\frac{1}{S_2(W)}\right\},$$

$$B = Vb = \frac{s^2}{S_1\{W(x-\bar{x})^2\}+S_2\{W(x-\bar{x})^2\}},$$

which with suitable choice of the weights W and of s^2 will cover either the quantal or the non-quantal case.

Now put $$\frac{(\bar{y}_2-\bar{y}_1)s}{\sqrt{A}} = D, \quad \frac{bs}{\sqrt{B}} = B'. \tag{3}$$

Making the substitutions $b^2 = BB'^2/s$, $(\bar{y}_2-\bar{y}_1)^2 = AD^2/s^2$ it is easy to see that (2) becomes
$$M+\frac{M't^2s^2}{B'^2-t^2s^2}\pm\frac{kts\sqrt{\{B'^2-t^2s^2+D^2\}}}{B'^2-t^2s^2}, \tag{4}$$

where $$k = \sqrt{\frac{A}{B}} = \frac{M'B'}{D} = \frac{\sqrt{[S_1\{W(x-\bar{x})^2\}+S_2\{W(x-\bar{x})^2\}]}}{1\Big/\sqrt{\left\{\frac{1}{S_1(W)}+\frac{1}{S_2(W)}\right\}}}, \tag{5}$$

a form which emphasizes that k^2 is the ratio of the variance of the average response difference to that of the slope. Fieller's (1940) form is

$$\bar{x}_1-\bar{x}_2+\frac{C(\bar{y}_2-\bar{y}_1)}{b}\pm tS',$$

where $$S'^2 = \frac{Cs^2}{b^2}\left[\frac{1}{N_1}+\frac{1}{N_2}+\frac{CB}{s^2}\left(\frac{\bar{y}_1-\bar{y}_2}{b}\right)^2\right]$$

and $$C = \frac{b^2}{b^2-t^2B} = \frac{B'^2}{B'^2-t^2s^2} = 1+\frac{t^2s^2}{B'^2-t^2s^2}. \tag{6}$$

Hence Fieller's form is

$$M+\frac{M't^2s^2}{B'^2-t^2s^2}\pm\frac{B't}{\sqrt{\{B'^2-t^2s^2\}}}\sqrt{\left\{\frac{As^2}{BB'^2}+\frac{M'^2s^2}{(B'^2-t^2s^2)}\right\}}. \tag{7}$$

Now $$M' = \frac{\bar{y}_2-\bar{y}_1}{b} = \left(\sqrt{\frac{A}{B}}\right)\frac{D}{B'}.$$

Hence (7) becomes

$$M+\frac{M't^2s^2}{B'^2-t^2s^2}\pm\frac{t}{\sqrt{(B'^2-t^2s^2)}}\sqrt{\left\{\frac{As^2}{B}\left(1+\frac{D^2}{B'^2-t^2s^2}\right)\right\}}$$

$$= M+\frac{M't^2s^2}{B'^2-t^2s^2}\pm\frac{kts\sqrt{\{B'^2-t^2s^2+D^2\}}}{B'^2-t^2s^2},$$

in agreement with (4).

Finney's (1946, 1947) form follows from my first form by putting

$$g = \frac{Bt^2}{b^2} = \frac{t^2s^2}{B'^2}.$$

If we make this substitution in (4) we obtain

$$M+\frac{M'g}{1-g}\pm\frac{kts\sqrt{\{B'^2(1-g)+D^2\}}}{B'^2(1-g)}.$$

As the estimate of variance increases in precision t^2s^2 becomes negligible compared with B'^2 and (4) becomes $$M\pm kts\sqrt{\{B'^2+D^2\}}/B'^2. \tag{8}$$

This agrees with the result given by Bliss (1940). His kI is equivalent to the k used here, since he has worked with log dose interval as a unit. His B is equivalent to our B'.

(2) *Combination of fiducial limits from individual assays to obtain fiducial limits for a pooled result*

For assays of the same type in which response is linearly related to the logarithm of the dose Fieller (1944) has shown how to form a pooled estimate of potency and to determine its fiducial limits provided:

(1) There are no significant differences in slope from assay to assay.

(2) There are no significant differences in error variance from assay to assay.

In this situation the estimates of error of the average response differences to test and standard and of the pooled slope are both based on the same estimate of variance and there is no difficulty. In other cases the logarithms of the estimated potency ratios are usually weighted inversely as their approximate sampling variances to form a pooled estimate whose error variance is taken as $1/S(W)$ if there is no heterogeneity between results and $\chi^2/\nu S(W)$ if there is heterogeneity; where

$$\chi^2 = SW(M-\bar{M})^2, \quad \nu+1 = \text{number of assays.}$$

But this leads to an underestimate of the fiducial limits.

We may sometimes need to pool the results of assays of different types; for example, the line test and the bone-ash method for vitamin D. No exact method of calculating fiducial limits in such cases has been given.

If the results of the individual assays differ significantly, I do not think there is an exact solution, and the same applies if the average response differences to test and standard or the slopes vary significantly from one test to another of the same type.

If we may assume that there is no heterogeneity of this kind, and if the error variances of average response differences and of slope in individual assays are based on a sufficient number of degrees of freedom (say $\geqslant 30$) to be regarded as known exactly, the following method would I think give a solution.

In the ith of h assays let the result be given by

$$M_i = (\bar{x}_1-\bar{x}_2)+K_i/L_i,$$

where K_i and L_i are respectively $(\bar{y}_2-\bar{y}_1)_i$ and b_i, or convenient multiples of them.

We shall assume that $\bar{x}_1-\bar{x}_2$ is constant throughout; this is true if in all assays corresponding doses of test and standard—and the numbers of animals on them—are in the same ratio on each dose in each assay.

Let the true result be $(\bar{x}_1-\bar{x}_2)+\mu$ and let $u_i = K_i-L_i\mu$. Then

$$\left(\sum_{i=1}^{h}\lambda_i K_i\Big/\sum_{i=1}^{h}\lambda_i L_i\right),$$

where the λ_i are suitably chosen, provides a pooled estimate of μ. Further $U = \sum_{i=1}^{h}\lambda_i u_i$ has expectation zero and variance

$$\sum_{i=1}^{h}\lambda_i^2(v_{K_i}+v_{L_i}\mu^2).$$

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

The λ_i are at our choice, but since K_i and L_i may be in units of entirely different kind from $K_{i'}$, and $L_{i'}$, it seems inevitable to take λ_i equal to

$$c_i(v_{K_i}+v_{L_i}\mu^2)^{-\frac{1}{2}},$$

where c_i is a pure number. The variance of U is then minimized when $c_i = c$ say, and can be made unity by taking $c^2 = 1/h$. Then our estimate of μ becomes

$$\bar{K}(\mu)/\bar{L}(\mu),$$

where

$$\bar{K} = \frac{1}{\sqrt{h}}\sum_{i=1}^{h} K_i(v_{K_i}+v_{L_i}\mu^2)^{-\frac{1}{2}},$$

$$\bar{L} = \frac{1}{\sqrt{h}}\sum_{i=1}^{h} L_i(v_{K_i}+v_{L_i}\mu^2)^{-\frac{1}{2}},$$

and must satisfy the equation

$$\frac{\bar{K}}{\bar{L}} = \mu. \tag{1}$$

If $\bar{\mu}$ be the solution of this equation, then $\bar{K}(\bar{\mu})-\bar{L}(\bar{\mu})\,\mu$ has expectation very close to zero and variance

$$\frac{1}{h}\sum_{i=1}^{h}\left\{\frac{v_{K_i}+v_{L_i}\mu^2}{v_{K_i}+v_{L_i}\bar{\mu}^2}\right\}.$$

Hence we find for the fiducial limits for μ, the quadratic equation

$$\{\bar{K}(\bar{\mu})-\bar{L}(\bar{\mu})\,\mu\}^2 = \frac{t^2}{h}\sum_{i=1}^{h}\left\{\frac{v_{K_i}+v_{L_i}\mu^2}{v_{K_i}+v_{L_i}\bar{\mu}^2}\right\}. \tag{2}$$

An approximate solution is

$$\frac{\bar{K}(\bar{\mu})}{\bar{L}(\bar{\mu})} \pm \frac{t}{\bar{L}(\bar{\mu})}. \tag{3}$$

Equation (1) can be solved by successive approximation. We start with an estimate of μ obtained, say, by weighting individual results with their approximate sampling variances. Substituting them in the left-hand side we get a first approximation to $\bar{\mu}$, and we then repeat the process as often as necessary.

The process is illustrated in Table 2 by combining the results of four different rat tests of two supposedly equivalent preparations of vitamin D. Tests (1) and (2) were line tests in the same laboratory, test (3) was a line test in another laboratory and test (4) was a test in a third laboratory using percentage ash content of bone as a response.

Different scales of healing for the line tests were used by the first two laboratories. The individual results were as follows:

	Slope/s.e. of slope	Potency ratio	Fiducial limits	
			$P=0{\cdot}95$	$P=0{\cdot}99$
(1) Line test P	3·9	0·963	0·695–1·298 (72·2–134·8 %)	0·571–1·520 (59·3–157·8 %)
(2) Line test W	5·2	0·814	0·531–1·165 (65·2–143·1 %)	0·432–1·331 (53·1–163·5 %)
(3) Line test Z	9·2	0·881	0·729–1·051 (82·7–119·3 %)	0·679–1·117 (77·1–126·8 %)
(4) % ash	11·0	0·901	0·764–1·054 (84·8–117·0 %)	0·720–1·113 (79·9–123·5 %)

Table 2 shows the combined results for tests (1) and (2), tests (3) and (4) and all four tests.

Table 2. *Pooled results and fiducial limits from four tests of vitamin D*

	Result		Fiducial limits					
			$P = 0{\cdot}95$			$P = 0{\cdot}99$		
Combination of tests	Weighted mean	'Exact' solution	From approx. variance of weighted mean	From 1st approx.	Exact solution from quadratic	From approx. variance of weighted mean	From 1st approx.	Exact solution from quadratic
(1) and (2)	0·907	0·895	0·735–1·120 (81·0–123·4 %)	0·723–1·109 (80·7–123·9 %)	0·704–1·108 (78·6–123·7 %)	0·688–1·197 (75·8–131·9 %)	0·676–1·186 (75·5–132·5 %)	0·639–1·193 (71·3–133·2 %)
(3) and (4)	0·893	0·892	0·798–0·999 (89·4–111·9 %)	0·793–1·002 (89·0–112·3 %)	0·791–1·000 (88·7–112·2 %)	0·770–1·034 (86·3–115·9 %)	0·765–1·039 (85·8–116·5 %)	0·760–1·038 (85·2–116·4 %)
All four	0·897	0·893	0·812–0·990 (90·6–110·4 %)	0·803–0·993 (89·9–111·3 %)	0·800–0·993 (89·6–111·1 %)	0·787–1·021 (87·8–113·9 %)	0·776–1·027 (86·9–115·0 %)	0·771–1·025 (86·3–114·8 %)

The weighted mean was obtained by weighting the logarithms of the individual results inversely as their approximate sampling variances; the 'exact' solution from equation (1) above. Estimates of the fiducial limits by three different methods are compared; the first method is the usual one of taking the sampling variance of the result as the reciprocal of the sum of the weights, the second uses equation (3) above, the third equation (2). It was not necessary to use more than two iterations to solve equation (1). For these data the usual approximation would have been adequate for practical purposes; it gives almost the same result as the exact method. However, these are better data than are often available, and one could not always assume that the approximate method would be adequate.

(3) *The χ^2 test in probit technique when the numbers are small*

In probit technique goodness of fit of the straight lines fitted by maximum likelihood is customarily tested by calculating $\Sigma nw(y-Y)^2$, where $w = Z^2/PQ$ and $y = Y + p/Z - P/Z$ in the customary notation. The sum is treated as a χ^2 value with $(k-2)$ degrees of freedom, where k is the number of dosage groups.

Thus it is assumed that when the hypothesis of a linear dosage-response relation is true, χ^2 follows the standard tabulated distribution for $(k-2)$ degrees of freedom. It has long been realized that the standard distribution is not followed for small numbers of responses, but the effect has not been examined very closely, though it is known that in the actual distribution there is an excess of very small and very large values of χ^2.

The following investigation may throw some further light on the subject. In fact $\chi^2 = \Sigma n(p-P)^2/PQ$, where P is the expected proportion of responses. If we suppose that the P's are known exactly, it is clearly possible for a specified number of groups and small n:

(i) to calculate all possible values of χ^2 and to examine their distribution,

(ii) to calculate theoretically the moments and cumulants of the actual distribution.

The fact that the P's have, in practice, to be estimated from the data is not likely seriously to disturb the general picture obtained; it will not be wildly wrong to assume that the distribution for k groups when the P's have to be estimated is—other things being equal—the same as for $(k-2)$ groups when they are known.

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

(3·1) With the aid of my colleague Miss Irene Allen, who carried out all the numerical calculations, I have carried out a sampling investigation for the extreme case when there is one animal on each dose. Though this case is not likely to be of *practical* importance in biology, it might well be so in other fields. For example, a single shell might be fired at each of a number of muzzle velocities and a record made of whether or not it penetrated armour plating of specified properties.

We chose eleven equally spaced dose levels between -2σ and $+2\sigma$ at intervals of $0{\cdot}4\sigma$, where σ is the S.D. of tolerances on the dosage scale. In other words the expected probit values were 3·0 (0·4) 7·0. The values of P were obtained from seven-figure probability integral tables. Now in this case, where there is only one animal on each dose, p is either 1 or 0, and hence the contribution to χ^2 from a particular dose is either P/Q or Q/P and takes these values with probabilities Q and P.

Hence χ^2 is the sum of P/Q for the negative responses and Q/P for the positive responses. With eleven doses there are only $2^{11} = 2048$ samples; hence it is an easy matter to work out χ^2 for each sample and by multiplication of the eleven appropriate probabilities to obtain the probability of its occurrence. We did this and obtained the distribution of χ^2 shown in Fig. 1. The sum of the 2048 associated probabilities came to 0·999995, which provides some check. A further check was obtained by calculating the first four cumulants theoretically as well as from the 2048 sample values. The theoretical method is described below.

The moments about zero of the sample values were obtained by separating them into two portions, the 144 values less than 25—which accounted for about 91 % of the frequency—and a tail. The moments of the first portion were calculated from the individual values, those of the tail by grouping with a class interval of 5 and then using Sheppard's corrections with abruptness coefficients. Finally, the two portions were added together and the cumulants calculated from the moments. There was agreement to three significant figures between these cumulants and those calculated by the theoretical method before the samples were tabulated. The values are as follows:

Cumulants of χ^2 for one animal per group. Eleven groups with
$Y = 3{\cdot}0\,(0{\cdot}4)\,7{\cdot}0$

	Calculated directly	Sample values
κ_1	11	11·02
κ_2	129·05	129·01
κ_3	3903·08	3899·56
κ_4	137279	138093 136713 with Sheppard's correction

It will be observed that the actual frequencies show oscillations (which are still more noticeable in the ungrouped distribution) rising to subsidiary maxima and dropping again. The 2048 values of χ^2 are not equally spaced but tend to cluster in certain positions, and they are, of course, not all different. Further, the probabilities of the distinct values do not always increase or decrease monotonically.

The histogram gives the actual distribution of χ^2 values grouped in groups of width 5, the curve with the higher mode is the 'normal theory' χ^2 distribution for

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

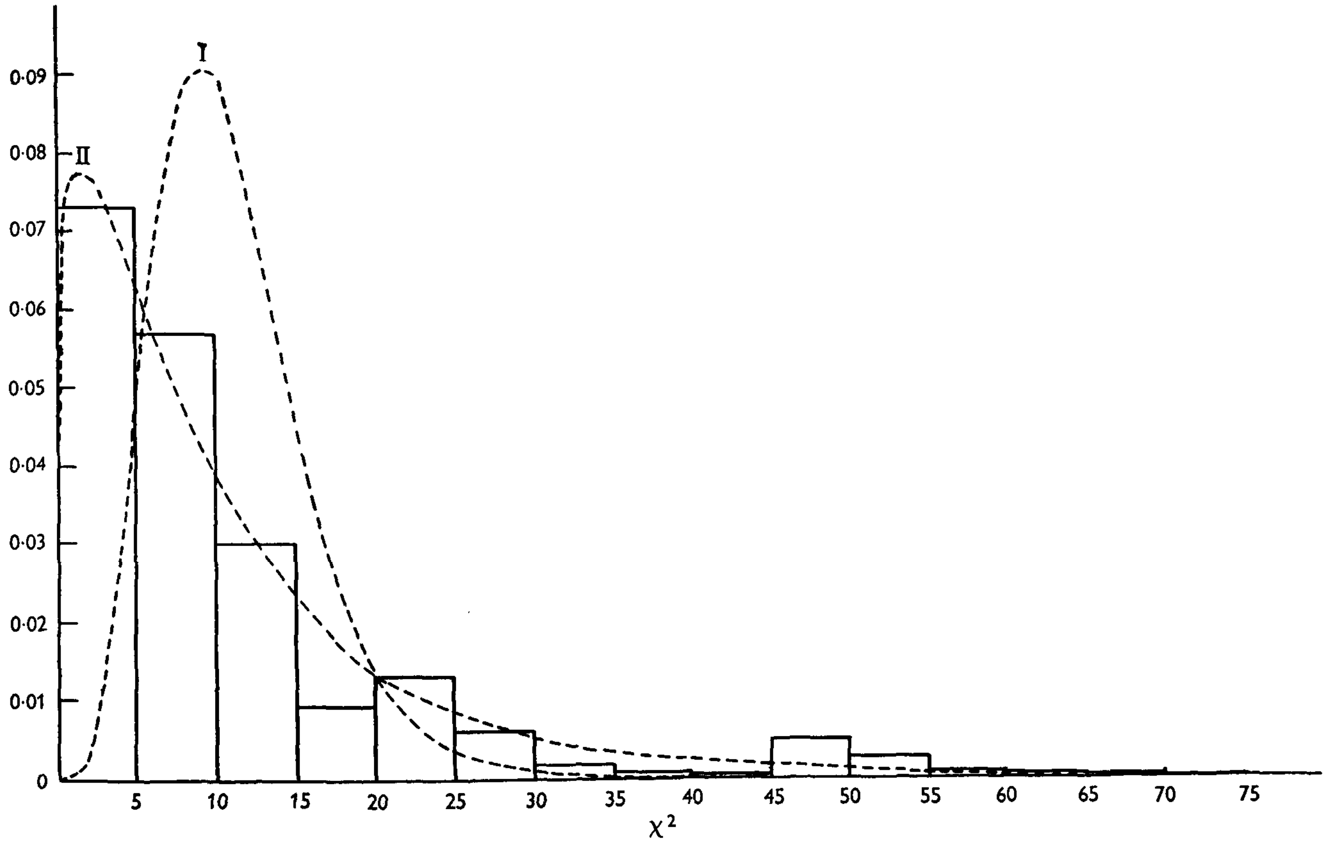

Fig. 1. Eleven doses with one animal on each dose. Distribution of χ^2. Histogram shows actual distribution. Curve I: 'Normal theory' with 11 degrees of freedom. Curve II:

$$y = 0{\cdot}03484\left(\frac{\chi^2}{11}\right)^{0{\cdot}2062}\left(\frac{78{\cdot}360+\chi^2}{90{\cdot}360}\right)^{-10{\cdot}9085}.$$

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

11 degrees of freedom, while the more skew curve is a Pearson Type VI curve starting at zero and fitted to the first three moments.

Its equation with origin at zero is

$$y = 0{\cdot}03484\left(\frac{\chi^2}{11}\right)^{0\cdot2062}\left(\frac{79{\cdot}360+\chi^2}{90{\cdot}360}\right)^{-10\cdot9085}. \qquad (1)$$

The values given by this curve for 95 and 99 % values of χ^2 are close to the true ones, as the following table shows:

		(2)	(3)
	(1)	Value of	True probability
	True value	χ^2 (P) from	corresponding
P	of χ^2 (P)	fitted curve	to (2)
0·95	36·3	32·9	0·946
0·975	49·5	41·6	0·952
0·99	56·0	54·0	0·987
0·995	64·1	65·1	0·996

It will be noted that between $\chi^2 = 35$ and 45 the true frequency distribution has a minimum; this accounts for the discrepancies at $P = 0{\cdot}975$.

The *ungrouped* distribution of the values of $\chi^2 < 25$ is shown in Table 3.

Table 3. *Ungrouped distribution of values of* χ^2, *less than* 25

(Eleven doses with one animal on each dose.)

Value of χ^2	No. of times occurring (f)	Total probability of occurrence
3·01165	2	0·178292
4·38802	4	0·187468
5·76438	2	0·049279
6·46305	4	0·095850
7·83941	8	0·100784
9·21578	4	0·026493
9·91445	2	0·012882
10·57201	4	0·046368
11·29081	4	0·013546
11·94837	8	0·048753
12·66718	2	0·003561
12·32473	4	0·012816
14·02340	8	0·024927
15·39977	16	0·026210
16·77613	8	0·006890
17·47480	4	0·003350
18·13236	2	0·003015
18·85117	8	0·003522
19·50872	4	0·003170
20·20209	4	0·020673
20·22753	4	0·000926
20·88508	2	0·000833
21·57845	8	0·021738
21·58376	4	0·001621
22·95481	4	0·005714
22·96012	8	0·001704
23·65349	8	0·011114
24·33648	4	0·000448
$\geqslant 25$	1904	0·088048
Total	2048	0·999995

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

(3·2) In the general case when there are varying numbers of animals in the dosage groups, the cumulants of χ^2 may be calculated as follows. Since

$$\kappa_r(\chi^2) = \Sigma\kappa_r(u),$$

where $u = n(p-P)^2/PQ$, and since

$$\mu_r'(u) = \mu_{2r}(np)/(nPQ)^2$$

and
$$\kappa_{2r}(np) = \sum_{i=1}^{2r}\{\Delta^i(0^{2r})(-1)^{i-1}P^i/i\}, \tag{1}$$

as many cumulants of u as are necessary can be calculated. These may then be summed for the different dosage groups to give the cumulants of χ^2. In this way we find

$$\left.\begin{aligned}
\mu_1'(u) &= 1,\\
\kappa_2(u) &= 2-\frac{1}{n}\left(6-\frac{1}{PQ}\right),\\
\kappa_3(u) &= 8-\frac{1}{n}\left(112-\frac{22}{PQ}\right)+\frac{1}{n^2}\left(120-\frac{30}{PQ}+\frac{1}{P^2Q^2}\right),\\
\kappa_4(u) &= 48-\frac{1}{n}\left(1824-\frac{384}{PQ}\right)+\frac{1}{n^2}\left(6720-\frac{2000}{PQ}+\frac{112}{P^2Q^2}\right)\\
&\qquad -\frac{1}{n^3}\left(5040-\frac{1680}{PQ}+\frac{126}{P^2Q^2}-\frac{1}{P^3Q^3}\right).
\end{aligned}\right\} \tag{2}$$

When $n = 1$ an alternative method may be used. Since $(u-1)$ takes values $-P\left(\frac{Q-P}{PQ}\right)$ and $Q\left(\frac{Q-P}{PQ}\right)$ with probabilities Q and P, it follows that the cumulant generating function of u is the same as that of the binomial with $\left(\frac{Q-P}{PQ}\right)t$ replacing t. Then

$$\kappa_r(u) = \left(\frac{Q-P}{PQ}\right)^r \sum_{i=1}^{r}\left\{\frac{\Delta^i(0^r)}{i}(-P)^i\right\},$$

giving
$$\left.\begin{aligned}
\kappa_2(u) &= \frac{(Q-P)^2}{PQ} = -4+\frac{1}{PQ},\\
\kappa_3(u) &= \frac{(Q-P)^3}{P^2Q^2} = 16-\frac{8}{PQ}+\frac{1}{P^2Q^2},\\
\kappa_4(u) &= \frac{(Q-P)^4}{P^3Q^3}(1-6PQ) = -96+\frac{64}{PQ}-\frac{14}{P^2Q^2}+\frac{1}{P^3Q^3}.
\end{aligned}\right\} \tag{3}$$

It may be observed that the expressions (2) reduce to (3) when $n = 1$. The values of $\kappa(\chi^2)$ on p. 226 were calculated by using equations (3).

(3·3) The sampling investigation described in §(3·1) suggested that in other, but less extreme, cases with small numbers a knowledge of the first four cumulants would lead to a satisfactory approximation to the form of the distribution.

We have worked out β_1 and β_2 for the distribution of χ^2 in eight cases:

(i) Two doses with $100P = 72·36$ and $27·64$ % for $n = 1\,(1)\,5\,(5)\,10$, 20, ∞. (Here $1/PQ = 5$.)

(ii) Case (i) repeated twice, the kind of situation that arises when two parallel

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

straight lines correctly represent the dosage-response relation in a simultaneous comparison of test and standard.

(iii) Three doses with $100P = 23{\cdot}89$, 50 and 76·11 % and n as above. (Here the mean value of $1/PQ = 5$.)

(iv) Case (iii) repeated twice.

(v) Three doses with $P = 30$, 50, 70 %.

(vi) Case (v) repeated twice.

Table 4 shows the results.

Table 4. *Values of β_2, β_1 and $\phi = (2\beta_2 - 3\beta_1 - 6)$ for the χ^2 distribution in certain cases*

	(i)				(ii)				(iii)			
n	β_2	β_1	ϕ	Type	β_2	β_1	ϕ	Type	β_2	β_1	ϕ	Type
1	2·50	0·5	−2·50	VI?	2·83	0·25	−1·08	VI?	2·75	0·75	−2·75	VI?
2	6·97	4·90	−6·75	VI?	4·99	2·45	−3·38	VI?	6·69	3·41	−2·85	VI?
3	10·05	4·96	−0·82	VI? or III	6·52	2·48	−0·41	VI? or III	8·19	3·36	0·29	VI or III
4	10·15	4·82	1·03	VI	6·87	2·41	+0·52	VI or III	8·43	3·24	1·13	VI
5	10·82	4·69	1·56	VI	6·91	2·35	+0·78	VI or III	8·40	3·15	1·36	VI
10	10·36	4·38	1·58	VI	6·68	2·19	0·79	VI or III	7·95	2·92	1·14	VI
20	9·79	4·20	1·00	VI or III	6·40	2·10	0·50	VI or III	7·54	2·80	0·68	VI or III
∞	9·00	4·00	0·00	III	6·00	2·00	0	III	7·00	2·67	0·00	III

	(iv)				(v)				(vi)			
n	β_2	β_1	ϕ	Type	β_2	β_1	ϕ	Type	β_2	β_1	ϕ	Type
1	2·88	0·38	−1·38	VI?	2·38	0·38	−2·38	VI?	2·69	0·19	−1·19	VI?
2	4·84	1·70	−1·42	VI?	4·36	1·41	−1·52	VI?	3·68	0·71	−0·76	VI? or III
3	5·59	1·68	0·14	VI or III	5·70	1·82	−0·07	VI? or III	4·35	0·91	−0·14	VI? or III
4	5·71	1·62	0·57	VI or III	6·17	2·03	0·26	VI or III	4·59	1·01	0·13	VI or III
5	5·70	1·57	0·68	VI or III	6·41	2·15	0·35	VI or III	4·70	1·08	0·17	VI or III
10	5·47	1·46	0·55	VI or III	6·76	2·41	0·29	VI or III	4·88	1·20	0·15	VI or III
20	5·27	1·40	0·34	VI or III	6·90	2·54	0·18	VI or III	4·95	1·27	0·09	VI or III
∞	5·00	1·33	0·00	III	7·00	2·67	0·00	III	5·00	1·33	0·00	III

If one fits a frequency curve to the moments, one would naturally choose a form starting at zero, so that only three moments are required. So, as far as Pearson's types are concerned, the choice is between Type III—that is, the usual χ^2 form with a modified exponent—and Type VI. Table 4 shows that in cases (v) and (vi) $|\, 2\beta_2 - 3\beta_1 - 6 \,|$ is considerably less than 0·5 for $n = 3$. In this case Type III would be adequate, while for $n = 1$ and $n = 2$, Type VI would almost certainly suffice to provide a reasonable approximation to the upper 5 and 1 % points, judging from the result of the sampling investigation in (3·1). Neither in the sampling investigation, nor in the two cases just mentioned, do β_1 and β_2 satisfy Pearson's criterion for Type VI; β_2 is very sensitive to departures from the Type VI form in the neighbourhood of its mode, and here, at any rate in the sampling investigation, the actual frequencies oscillate quite a lot, as may be seen from Table 3.

In cases (iii) and (iv) the values of $(2\beta_2 - 3\beta_1 - 6)$ are larger than in cases (v) and (vi). In case (iii) Type VI would, I think, be necessary up to $n = 10$, though Type III might suffice for $n \geqslant 3$ in case (iv). In cases (i) and (ii) the situation is somewhat similar. As soon as we begin to include groups with small values of P, the approach of the χ^2 distribution to its 'normal theory' form is slowed up.

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

When Type III suffices this means that a 'normal theory' χ^2 distribution

$$\{df = 1/\Gamma(\tfrac{1}{2}\nu)\, e^{-\frac{1}{2}\chi^2}(\tfrac{1}{2}\chi^2)^{\frac{1}{2}\nu-1}d(\tfrac{1}{2}\chi^2)\}$$

can be used with a modified value of ν estimated from $\nu = 8/\beta_1$.

It would be worth while comparing the 5 and 1 % points yielded by both distributions in all these cases, but this we have not so far found time to do. Neither have we found time to carry out a sampling investigation in which, for each value of χ^2, the expected frequencies are estimated from the data. This would involve fitting some hundreds of probit lines and would seem to require the resources of some institution such as the Mathematics Department of the National Physical Laboratory.

IV. THE PROBIT TECHNIQUE APPLIED TO LITTER-MATES

I have several times been asked how the probit technique should be applied to litter-mates, but have never happened to come across suitable data, though some are likely to be available to me in the near future.

It is quite clear that if the responses of litter-mates are correlated there must be a gain in precision by using them, placing one member of each litter on each dose. The point at issue is how the gain in precision should be measured. If the ordinary technique is used the weighted sum of squares of deviations from the fitted straight line will have its expectation reduced by a factor of order $(1-r)$, where r is the average correlation between litter-mates, and it is not difficult to see that the theoretical variance of an observation, which in the ordinary technique is given weight ($w = nZ^2/PQ$), is not $1/w$, but is reduced by the same factor.

If we are prepared to assume that the correlation between the tolerances of litter-mates (in litters of k) is the same for all pairs, we can in theory start from a normal multivariate distribution of tolerances with k variates, write down the maximum likelihood equations and endeavour to solve them. The equations are complicated. However, there is a simpler and more practical approach. Let us suppose we have n litters with k animals and we put one member of each litter on each dose. We can find the maximum likelihood solution for the probit-log dose line by the usual technique and, having done so, calculate for each animal the *corrected probit*, which will be either $Y-P/Z$ or $Y+Q/Z$, according as it gives a positive or negative response. With each of these will be associated the weight Z^2/PQ.

We notice that if we fit, by weighted least squares, a straight line to these individual values we are led to the same line as before. For at any one dose level there will be nq values $Y-P/Z$ and np values $Y+Q/Z$, and therefore the weighted mean response is

$$q\left(Y-\frac{P}{Z}\right)+p\left(Y+\frac{Q}{Z}\right) \quad \text{or} \quad Y+\frac{n}{Z}(p-P).$$

Therefore we need only carry out a weighted analysis of variance of the corrected probits into doses, litters and error. The error mean squares will take the place of unity in all estimates of the errors of the parameters, and in the test for goodness of fit, which becomes a variance ratio instead of a χ^2 test. Assuming a satisfactory fit, if s^2 is significantly less than unity, $r = 1-s^2$ is an estimate of the average correlation between litter-mates.

V. PROBITS, LOGITS AND THE ANGULAR TRANSFORMATION

There has been a good deal of discussion of the rival merits of the normal distribution function, of the logistic and the curve $P = \sin^2 Y$ in representing the dosage-response relation in the quantal case.

A number of other alternatives were listed in his 1947 paper by Finney, who showed how to apply the method of maximum likelihood to each.

There are two main points of view from which the discussion can be conducted; we may have theoretical reasons for preferring one model to another, or we may advocate a particular form of dosage-response relation on the ground that it is sufficiently close to the truth for practical purposes.

I do not think that there has been much support for the angular transformation on *theoretical* grounds; here it is the relative advantages of the normal distribution function and the logistic which have received most attention. The concept of a normal distribution of tolerances is easy to understand, while the arguments for the logistic have been based on physico-chemical conceptions which I personally do not feel competent to criticize. So I shall not go further now into this theoretical field, if field it be rather than forest, but content myself with expressing the hope that the discussion to follow will throw the light of lucidity on the whole area and not just illuminate a few particular trees with fitful beams. I will concentrate on the second question, adequacy for practical purposes.

In our research unit we have a programme for the reanalysis of a considerable body of data by all three methods, and my colleague, Mr P. Armitage, is undertaking this work. It has not got very far yet, but Table 5 presents the results Armitage has so far obtained for two series of data quoted by Berkson (1944).

The angular transformation has at first sight much to recommend it for ease of fitting and has been advocated for this reason (see, for instance, Knudsen and Curtis (1947)), because if

$$Y = \sin^{-1} P, \quad \frac{dY}{dP} = \tfrac{1}{2}/\sqrt{(PQ)}$$

and therefore

$$V(Y) = \tfrac{1}{4}n \text{ (radians}^2\text{)} \quad \text{or} \quad \frac{1}{4n} \times \left(\frac{180}{11}\right)^2 \text{ (degrees}^2\text{) approximately.}$$

This suggests fitting a straight line $Y = a + bx$ by least squares, weighting the transformed responses with the numbers of animals in the groups, so that an iterative process is unnecessary. However, we note that this is not the maximum likelihood solution—which is itself no easier to carry out for this form of dosage-response relation than any other—and that the method suggested may lead to estimates of the expected values of the transformed variable which are greater than 90° at high dosages. In one of the examples in Table 5 this happens. If it were taken at its face value, it would indicate a drop in the percentage of positive responses for sufficiently high doses, while the hypothesis under examination supposes an increase with dose up to a limiting value at and above which all responses are 100 %. This is an indication that the method of maximum likelihood should have been used in fitting.

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

Dr Joseph Berkson, the principal advocate of the logistic, prefers the logistic to the normal distribution function on theoretical grounds and prefers a minimum χ^2 to a maximum likelihood solution, I think, also on theoretical grounds, but in addition because he has discovered a good approximation to the minimum χ^2 solution which needs no iteration.* This is, I think, the strongest point in favour of the logistic. If for samples of the magnitude with which we have to work (i) we are unable to discriminate by a significance test between the two forms of relationship, (ii) if also we find that Berkson's approximation is sufficiently close to the true minimum χ^2 solution and, finally, (iii) if the maximum likelihood and minimum χ^2 solutions do not differ materially from one another (we know that they tend to equivalence in 'large' samples but we do not know how large '*large*' means), then I think there is a strong case for using the easiest method. By 'not materially' in condition (iii) I mean that the differences between the estimates by the two methods are small compared with errors of sampling. As far as the data of Table 5 go, these conditions seem to be satisfied, and I am inclined to think that this will be confirmed by examination of further data.

Table 5 shows that for these data the difference between maximum likelihood and minimum χ^2 is negligible, that the angular transformation is unsatisfactory and that both the normal and the logistic forms of dosage-response relation are consistent

Table 5. *Comparison of maximum likelihood and minimum χ^2 methods of fitting the normal and logistic forms of dosage-response relation*

(Data quoted by Berkson (1944).)

Data	Murray	Bliss Series I
No. of doses	11	6
Total no. of observations	5495	175
D.F. for χ^2	9	4

	Values of χ^2 $=\Sigma n(p-P)^2/PQ$		Values of LD 50	
	Murray	Bliss Series I	Murray	Bliss Series I*
Normal:				
Max. L	11·9	0·70	66·52	60·03
Min. χ^2	11·9	0·68	66·50	60·02
Logistic:				
Max. L	6·4	1·10	66·79	60·02
Min. χ^2	6·4	1·09	66·78	60·01
Berkson's min. χ^2 approximation	6·3	1·09	66·86	60·00
Angular transformation (no. iteration)	73·8	1·14†	65·01	59·90

* Lowest value of p corrected.

† χ^2 calculated from deviations from fitted line because one expected value of Y was greater than 90°.

* Clearly the maximum likelihood solution must give a larger χ^2 than the minimum χ^2 solution, but the distributions of χ^2 when the hypothesis tested is true and for samples of the size ordinarily used will differ for the two solutions and I should be surprised if the correct values of P (χ^2) differed at all greatly. The theoretical reasons for preferring maximum likelihood to minimum χ^2 in dealing with small samples depend on one's philosophy of statistical inference and raise questions of great difficulty in the theory of estimation.

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

with the data. The χ^2 values favour the logistic in one case and the normal distribution function in the other, but one should not interpret χ^2 values as a *measure* of goodness of fit, when the significance test does not contradict the hypothesis tested. In cases where the hypothesis *is* contradicted, χ^2/ν might serve as a measure of the magnitude of the departure from it. Berkson's minimum χ^2 approximation is quite satisfactory.

VI. MISCELLANEOUS RAPID METHODS IN THE QUANTAL CASE

Whenever a new substance is being assayed or when a new test has been devised for an existing substance, it is essential that the results should be analysed by a method which provides satisfactory errors of estimate, and this necessitates the employment of maximum likelihood methods of estimation or of methods such as Berkson's minimum χ^2 for the logistic, which yields virtually the same results.

However, a number of rapid approximate methods of estimating median effective doses have been suggested, and these may sometimes be used with advantage for routine tests which have already had a thorough statistical examination. My own experience is that these methods usually have errors of estimation which are small compared with errors of sampling for samples of the size usually available.

I do not propose to discuss the Wilson-Worcester method which provides a table based on the logistic for estimating median effective doses, nor that of Litchfield and Festig which provides empirical formulae for similar purposes.

Behrens's method and Kärber's method were discussed by Gaddum as early as 1933. The Reed-Muench method is essentially the same as that of Behrens. The Reed-Muench method and Kärber's seem to be the most frequently employed. These tests were originally used when the whole range of response between zero and 100 % was covered; when that is not the case certain special assumptions have to be made in order to get a result.

This has recently led Thomson (1947) to suggest a method of moving averages; in a test with equal dosage intervals he takes a moving average of say three consecutive percentage responses and interpolates between two consecutive values on either side of the 50 % point. This is quite a satisfactory procedure, and will avoid extrapolation whenever it is possible to do so; but I do not think he has established, as he claims, that the method has greater precision than Kärber's in his re-examination of Topley's data originally analysed by Irwin and Cheeseman.

My own preference is for Kärber's method, which has recently been put in a more favourable light by the work of Cornfield & Mantel (1948). They show that Kärber's method provides the maximum likelihood solution when the following conditions are satisfied:

(*a*) equal numbers of animals at each dose level,

(*b*) equally spaced intervals,

(*c*) an underlying log-logistic distribution of tolerances,

(*d*) the mean and standard deviation of the population are defined in terms of a frequency distribution with finite class intervals rather than an integral with infinitesimal class intervals,

(*e*) the whole range of response between zero and 100 % is covered.

VII. CONCLUSION: THE VITAMIN D INVESTIGATION

I will conclude by giving a short account of an investigation which, in spite of the fact that I played some part in it, seems to me to show how much can be done by combining in application the modern knowledge of biological standardization, of experimental planning, of statistical investigation and of international co-operation.

Up to the beginning of this year the International Standard of Reference for vitamin D has been a solution of irradiated ergosterol in olive oil. It contains 1000 international units of vitamin D per gram. In addition to calciferol (vitamin D_2) it also contains other products of irradiation of ergosterol. It would obviously be preferable to have a pure preparation of vitamin D as standard, and in 1934 the Second International Conference on Vitamin Standardization recommended that 'when the present international standard solution is exhausted, or if it should become unsatisfactory for any reason, it should be replaced by an equivalent solution of pure crystalline vitamin D in olive oil of such strength that 1 mg. contains $0{\cdot}025\,\mu$ of crystalline vitamin D'.

However, a further possibility recently presented itself, the adoption of a sample of vitamin D_3 as the international standard of reference. This can now be obtained in a pure state. The advantage of having vitamin D_3 as a standard of reference is that it could be used for determining the vitamin D content of oils intended either for human or for animal and poultry feeding. The antirachitic activity of cod-liver oil is largely due to the presence of vitamin D_3. Since vitamin D_3 and vitamin D_2 are, weight for weight, equally effective in the rat and probably also in the human being, vitamin D preparations intended for human beings can be standardized by tests on rats. But vitamin D_2 (calciferol) is relatively ineffective for the chick, and therefore the vitamin D_2 standard cannot be used in the determination of the vitamin D content of preparations intended for poultry feeding. The adoption of one international standard of reference for all vitamin D determinations would simplify matters for clinicians, pharmacists, veterinarians and farmers. The difficulty, in connexion with poultry, led the British Standards Institution to adopt a standard preparation of vitamin D_3 of their own, and this has been used for the assay of poultry preparations in Great Britain. Now a conference on biological standardization held in London by the World Health Organization at the beginning of this year has proposed the adoption internationally of D_3 as a standard; and there is little doubt that the W.H.O. will accept this recommendation.

Before this stage could be reached much preliminary work was necessary. The vitamin D Sub-Committee of the Accessory Food Factors Committee of the British Medical Research Council organized a collaborative experiment to see whether it would be possible to adopt a sample of vitamin D_3 as the international standard of reference for vitamin D. Five firms known to be making vitamin D_3 were invited to

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

contribute about 5 g. each. The vitamin D Sub-Committee thought it desirable to compare the following preparations for vitamin D activity by biological tests:

(1) The present international standard for vitamin D (irradiated ergosterol in olive oil).

(2) The new preparation of pooled samples of vitamin D_3.

(3) The British Standards Institution standard for vitamin D_3.

(4) A preparation of the purest sample of calciferol available.

The U.S. Pharmacopoeia Vitamin Advisory Board joined in the scheme. *Eighteen* laboratories in the U.S.A. and Canada, *nine* in Great Britain, *two* in New Zealand and *four* in the Scandinavian countries participated. Results of twenty-nine rat assays and twenty-five chick assays were received. The design of the chick assays was straightforward. There were, of course, no litter-mate complications, and from three to five doses of each of the solutions compared were used. The proposed New Standard and the B.S.I. Standard were compared in Europe; in the United States the U.S.P. Reference Oil was also included. Some of the laboratories used the T.M.T. test and others the percentage ash test. Not less than fifteen chicks on each dose were used.

The design of the rat assays needed some consideration. Eventually we decided to use at least ten litters at each of three dose levels (in the ratio 1, 2, 4), assigning one member of each litter to each of the four solutions. This meant that solution comparisons were isogenic but that the slope was determined from non-isogenic comparisons. We knew that the slope would be sufficiently accurately determined in this way, and deliberately decided to sacrifice a certain loss of accuracy to simplicity of design. This meant that the errors of the response differences and of the slope had to be determined from different terms in the analysis of variance, but a sufficient number of degrees of freedom for each were available to ensure that the numerator and denominator of the log-potency ratio could be regarded as normally distributed with a known variance; in other words, the 'normal' value could be assigned to 't' and fiducial limits calculated in the usual way. Most of the laboratories used the line test, two used the X-ray test and four percentage ash content of bone. It is not possible to present here the detailed results, which will be published elsewhere, but they were surprisingly uniform. A few points may, however, be mentioned.

In the rat tests with solutions 1, 2, 3 and 4, in only one case was there any evidence of significant differences between laboratories where a comparison of seven laboratories in Great Britain who used the line test gave a χ^2 of 13·8 against a 5 % point of 12·6. As this was the only case, we felt justified in regarding the results as homogeneous when they were pooled.

With the exception of one set of bone-ash results from a Scandinavian country and the two bone-ash results from New Zealand which gave significantly higher values for the potency ratios (New Standard/Old Standard) and (B.S.I./Old Standard), there were no significant differences between countries or between methods of testing. It may be noted that in all comparisons against the Old Standard these three laboratories gave results somewhat higher than the remainder. This suggests that the Old Standard sent to the Continent and New Zealand, though not apparently to the U.S., may have deteriorated slightly.

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

The results for the chick tests also showed a remarkable uniformity. However, in the tests of solutions 1, 2, 3 and 4 there were significant differences between five American laboratories which carried out ashing of the individual bones. This was allowed for by the appropriate reduction of (statistical) weight in calculating the general mean. It was not possible to test the significance of differences between ten assays in nine American laboratories which used 'pooled ashing'; here the mean and variance between laboratories were estimated by giving each log result equal weight. The reciprocal of the error variance of the mean obtained in this way was taken as the weight for the purpose of combining this result with other results.

Table 6. *Co-operative vitamin D investigation. Pooled results for four solutions*

	Rats		Chicks	
	Potency ratio	Fiducial limits ($P=0{\cdot}95$)	Potency ratio	Fiducial limits ($P=0{\cdot}95$)
New St./Old St.	0·966*	0·925–1·009* (95·8–104·4 %)	— —	— —
B.S.I./Old St.	0·942*	0·902–0·983* (95·8–104·4 %)	— —	— —
Calciferol/Old St.	0·933	0·896–0·972 (96·0–104·2 %)	— —	— —
New St./B.S.I.	1·043	1·001–1·084 (96·0–104·2 %)	1·090 —	1·046–1·137 (95·9–104·4 %)

* Excluding 'Other European' and New Zealand. The corresponding results including them are 0·989 and 0·960.

Table 6 shows the final results for rats and chicks separately. I think it will be agreed that fiducial limits at $P = 0{\cdot}95$ of less than 5 % leave little to be desired from the point of view of accuracy.

The U.S.P. reference oil caused a little more difficulty, *ten* laboratories using the line test obtained significantly different results, and, while the mean had fiducial limits ($P = 0{\cdot}95$) of the order of 95–105 % if heterogeneity was ignored, the true error was clearly considerably greater. There was a similar phenomenon in the percentage ash test for chicks. The pooled result for the potency ratio (U.S.P./B.S.I.) was 0·963 for rats with fiducial limits ($P = 0{\cdot}95$) 0·860–1·079 (89–112 %), and 0·828 for chicks with fiducial limits 0·766–0·895 (92–108 %); the difference is on the borderline of significance, so that as had been suggested earlier the U.S.P. Reference Oil may contain a little D_2.

If in places this paper has struck too critical a note, I must ask pardon, but I do think it is important that at a meeting such as this everything should be done to encourage clear and resolute thinking on the meaning of what we do. If there has been any failure on my part in this respect, I have no doubt that it will be pointed out in the discussion, which will, I hope, be lively.

But, so that there may be no mistake, may I conclude by emphasizing how much I admire the magnificent achievement of the pioneers who succeeded in getting standards established, people like Dale, Gautier, Gaddum, Hartley and Trevan,

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press

thereby enabling many of the newer discoveries of medicine to be utilized on a comparable basis throughout the world to the immense advantage of many thousands of sufferers; also the splendid contribution made to this end by the biometricians and statisticians, too many of whom are here to make it anything but invidious to mention them by name.

REFERENCES

Berkson, J. (1944). Applications of the logistic function to bio-assay. *J. Amer. Statist. Ass.* **39**, 357–65.

Berkson, J. (1949). Minimum χ^2 and maximum likelihood solution in terms of a linear transform, with particular reference to bio-assay. *J. Amer. Statist. Ass.* **44**, 273–8.

Bliss, C. I. (1940). Factorial design and covariance in the biological assay of vitamin D. *J. Amer. Statist. Ass.* **35**, 498–506.

Bliss, C. I. (1945). Confidence limits for biological assays. *Biometrics*, **1**, 58–65.

Bliss, C. I. & Cattell, McK. (1943). Biological assay. *Ann. Rev. Physiol.* **5**, 479–539.

Bliss, C. I. & Marks, H. P. (1939). The biological assay of insulin. *Quart. J. Pharm.* **12**, 82–110.

Cornfield, J. & Mantel, N. (1948). Simplified calculation of the dosage response curve. Div. Publ. Health Methods Pub. Health Service, Bethesda, Maryland. Unpublished Communication.

Emmens, C. W. (1948). *Biological Assay.* London: Chapman and Hall.

Fieller, E. C. (1941). The biological standardization of insulin. *J. Roy. Statist. Soc. Suppl.* **7**, 1–64.

Fieller, E. C. (1944). A fundamental formula in the statistics of biological assay and some applications. *Quart. J. Pharm.* **17**, 117–23.

Finney, D. J. (1946). Principles of biological assay. *J. Roy. Statist. Soc. Suppl.* **9**, 46–91.

Finney, D. J. (1947). *Probit Analysis.* Cambridge University Press.

Gaddum, J. H. (1933). Reports on biological standards. III. Methods of biological assay depending on a quantal response. *Spec. Rep. Ser. Med. Res. Coun., Lond.*, no. 183.

Irwin, J. O. (1937). Statistical method applied to biological assays. *J. Roy. Statist. Soc. Suppl.* **4**, 1–60.

Irwin, J. O. (1943). On the calculation of the error of biological assay. *J. Hyg., Camb.*, **43**, 121–8.

Knudsen, L. F. & Curtis, J. M. (1947). The use of the angular transformation in biological assay. *J. Amer. Statist. Ass.* **42**, 282–96.

Thomson, W. R. (1947). Use of moving averages and interpolation to estimate median effective dose. *Bact. Rev.* **11**, 115–45.

(*MS. received for publication* 15. ii. 50.)

https://doi.org/10.1017/S0022172400015023 Published online by Cambridge University Press