Rothamsted Repository Download

A - Papers appearing in refereed journals

Kleczkowski, A. 1955. The statistical analysis of plant virus assays: a transformation to include lesion numbers with small means. *Journal of General Microbiology.* 13 (1), pp. 91-98.

The publisher's version can be accessed at:

- https://dx.doi.org/10.1099/00221287-13-1-91

The output can be accessed at: https://repository.rothamsted.ac.uk/item/9706w/the-statistical-analysis-of-plant-virus-assays-a-transformation-to-include-lesion-numbers-with-small-means.

The Statistical Analysis of Plant Virus Assays: a Transformation to Include Lesion Numbers with Small Means

By A. KLECZKOWSKI

Rothamsted Experimental Station, Harpenden, Hertfordshire

SUMMARY: The numbers of local lesions (x) produced by different leaves when inoculated with the same virus preparation deviate greatly from a normal distribution, and the standard errors of x's depend on their magnitude. Before customary statistical analyses are applicable to results of infectivity assays, x's must be suitably transformed. When mean values of x are greater than about 10, the transformation $z = \log_{10}(x+c)$ (where c is a constant), is satisfactory but inapplicable with smaller numbers. In some work the use of poorly infective inocula is unavoidable, and to allow statistical analysis of results in such work a new transformation $z = \log_{10}\frac{1}{2}[x+c+\sqrt{(x^2+2cx)}]$ is derived. This operates satisfactorily when mean numbers of x are greater than about 1·5. The results of the two transformations converge as x increases. Values of z are tabulated to make the use of the new transformations as quick as that of the other transformation.

Numbers of local lesions produced by inoculating plant viruses to leaves of host plants cannot be used directly in routine statistical analyses, such as analysis of variance, because their standard errors depend on their magnitude and their frequency distribution deviates very considerably from normal. Therefore, the transformation

$$z = \log_{10}(x+c) \tag{1}$$

was proposed (Kleczkowski, 1949), in which x's are numbers of lesions on individual leaves or half-leaves (whichever is used as a unit) and c is a constant whose value usually lies between 5 and 15. The values of z are approximately normally distributed with standard errors independent of their magnitude, when values of x are not too small. An exact limiting value cannot be given, for it varies with experimental conditions and with different virus-host systems. However, the transformation (1) usually makes the standard error of z constant when the mean number of lesions is greater than about 10, but the standard error begins to decrease as the mean decreases between 10 and 0. Thus transformation (1) can only be used safely when the mean numbers of lesions given by individual inocula are greater than 10 and never when they are smaller than 5.

When infectivities of several preparations of a virus are to be compared, a test should be so designed that as many sources of variation as possible are eliminated. This is done by suitably allotting leaves or half-leaves for different inocula. When at least one of the tested inocula gives a mean number of lesions too small for the transformation (1) to be applied, the usual analysis of variance of the whole experiment using the transformation (1) may give false results. As in many kinds of problems the use of inocula that form only

a few lesions per leaf cannot be avoided, a transformation that would still be applicable is needed if the results are to be analysed statistically; this paper describes the derivation and the use of such a transformation.

Derivation

The relationship between the means ($\bar{s}$) of estimated standard errors (s) of numbers of lesions (x) and of the means of numbers of lesions (m) is given by the regression equation

$$\bar{s} = b(m+c),\tag{2}$$

where b is the regression coefficient and c is a constant identical with that of the transformation (1) (Kleczkowski, 1949). The relationship between the standard error on the mean number of lesions is thus given by the equation

$$s = \beta(m+c),\tag{3}$$

the value of β being slightly greater than that of b of equation (2) because mean values of standard errors are smaller than square roots of mean values of variances; the ratio of β to b depends on the size of samples from which variances and standard errors are computed, and approaches 1 as the size of the sample increases. Thus, the transformation, that could be expected to make the standard error of transformed values constant and approximately equal to 1, is

$$y = \int \frac{dx}{\beta(x+c)} = \frac{1}{\beta} \log_e (x+c).\tag{4}$$

The use of the transformation (1) instead of (4) would make the standard error of z approximately equal to $0\cdot4343\beta$.

The values of z given by the transformation were, indeed, found to have a standard error independent of their magnitude and were approximately normally distributed (Kleczkowski, 1949). However, the standard error can be considered to be the linear function of the mean given by (3) only when the mean is not smaller than about 10, and, as the mean becomes smaller, the value of the standard error deviates appreciably from the straight line (3) and approaches 0 as the mean approaches 0. Moreover, discontinuity of the transformed variable (x's being integers) becomes more pronounced as the variate decreases, and this may eventually upset conditions that assume an approximately normal, and thus continuous, distribution of the transformed values. However, disregarding for the moment any possible effects of discontinuity, the deviation of the standard error from the line (3) alone makes the transformation (1) or (4) unsuitable when the deviation becomes appreciable. A transformation was, therefore, derived which is based on the assumption that the standard error of x follows a curve that starts at 0 when $m=0$ and then approaches asymptotically the straight line given by (3) as m increases. A curve that fulfils these conditions is the hyperbola

$$s = \beta\sqrt{[(m+c)^2 - c^2]}.\tag{5}$$

Assuming (5) to be true, the transformation that may be expected to equalize the standard error would be

$$y = \int \frac{dx}{\beta\sqrt{[(x+c)^2 - c^2]}} = \frac{1}{\beta}\log_e[x+c+\sqrt{(x^2+2cx)}].$$ (6)

The variance of the values of y would be expected to be approximately equal to 1. In practice the transformation

$$z = \log_{10}\tfrac{1}{2}[x+c+\sqrt{(x^2+2cx)}]$$ (7)

is more convenient. As x increases, the value of z given by (7) approaches that given by (1), so that for sufficiently high values of x the values of z can be computed more simply using (1) than using (7). The standard error of z would be expected to be approximately 0.4343β.

There is no evidence that (5) does truly express the dependence of the standard error of numbers of lesions on their means, but the standard errors probably follow a curve not very different from (5); in the absence of anything better, it was assumed to be true and the transformation (7) was applied to experimental data to see whether the results would be satisfactory.
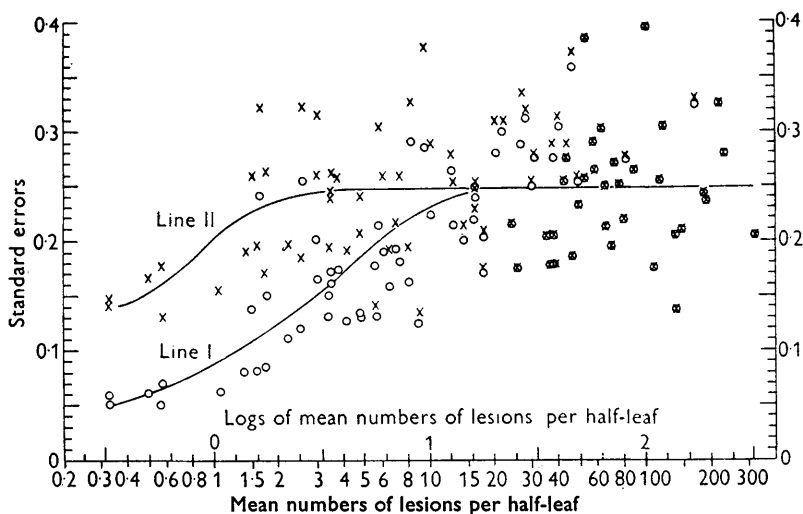

Fig. 1. Relationship between standard errors of transformed lesion numbers and the means of the numbers. Transformation $z = \log_{10}(x+c)$: points ○ and line I. Transformation $z = \log_{10}\tfrac{1}{2}[x+c+\sqrt{(x^2+2cx)}]$: points × and line II.

Application to experimental data

Fig. 1 compares results of transformations (1) and (7) of numbers of lesions produced by tobacco mosaic virus on leaves of *Nicotiana glutinosa* in several separate experiments. In each experiment six different inocula containing virus at different dilutions, or at different levels of activity remaining after exposure to different amounts of ultraviolet radiation, were inoculated on to half-leaves of six plants. Each plant had six inoculable leaves. The treatments were so distributed that each inoculum occurred twice on each plant and on

each leaf position. There were, therefore, 12 replications of each treatment (inoculum).

The constant c was taken arbitrarily as 5 for all the transformations, for previous experience showed that this value lies within the range of values that usually give satisfactory results. The standard errors of the transforms were computed separately for each treatment without making adjustments to eliminate variation between plants and leaf positions. It is important that the transformation should give satisfactory results even when variation between plants and leaf positions cannot be eliminated, because some experiments can only be designed so that variations between 'blocks' comprising a few plants or between leaf 'levels' comprising a few leaf positions, can be eliminated, and in some experiments even this cannot be done.

Fig. 1 shows standard errors of the transforms (1) and (7) plotted against the logarithms of mean numbers of lesions per half-leaf. To make inspection of the distribution of the standard errors easier, lines I and II were passed approximately through mean values of standard errors of the transforms (1) and (7), respectively, in different regions of the logarithms of mean numbers of lesions. The course of the lines shows that standard errors of the transforms (1) and (7) can be considered independent of their magnitude only when the mean numbers of lesions per half-leaf are not smaller than about 10 and 1·5, respectively.

In the region where the standard errors of the transforms can be considered independent of the mean numbers of lesions, the mean value of the standard errors was about 0·25. If the distribution of numbers of lesions does approximately correspond with the assumptions on which the transformations are based, the mean value of the standard errors of the transforms should be near the value of $0·4343b$, where b is the coefficient of regression of standard errors of untransformed numbers of lesions, whose means are not smaller than 10, on the means. The value of b in most experiments was found to be about 0·5, so that the value of $0·4343b$ is 0·22 which is, indeed, near to the mean value of the standard errors of the transforms.

The conclusion is that the transformation (1) can be used only when mean numbers of lesions given by different treatments are not smaller than about 10, whereas the corresponding limit for the transformation (7) is about 1·5.

The transformation (7) is more laborious than (1), which is simply read from a logarithmic table or from a slide rule. To make the transformation (7) as quick to use as (1), Table 1 is appended. As only two first decimal figures need be used, the table gives the values of the transform (7) for several values of the constant c only as far as two first decimal figures of the transform (7) differ from those of the transform (1). For all higher values of x the transformation (1) would give the same results as (7) and so can be used instead of (7).

An important condition before either transformation (7) or (1) can be applied is that the test plants should not contain any exceptionally resistant to infection. Including a plant that produces very few lesions increases the ranges of variability of numbers of lesions produced by all treatments. However, the increase will be disproportionately greater with treatments that

produce many lesions on the other plants than with treatments that produce few lesions. Consequently, the transformation may fail to equalize the standard errors.

Table 1. *The values of* $z = \log_{10} \frac{1}{2}[x + c + \sqrt{(x^2 + 2cx)}]$

For values of x higher than those given in this table $z = \log_{10}(x + c)$.

x	z	x	z	x	z	x	z	x	z	x	z	x	z	x	z
c=3		c=5		c=6		c=10		c=10		c=15		c=20		c=20	
0	0·18	6	1·02	21	1·43	0	0·70	43	1·72	33	1·67	7	1·35	50	1·84
1	0·52	7	1·06	22	1·44	1	0·89	44	1·73	34	1·68	8	1·38	51	1·84
2	0·65	8	1·10	23	1·46	2	0·97	45	1·74	35	1·69	9	1·40	52	1·85
3	0·75	9	1·13	24	1·47	3	1·03	46	1·75	36	1·70	10	1·42	53	1·86
4	0·82	10	1·16	25	1·49	4	1·08	47	1·75	37	1·71	11	1·44	54	1·86
5	0·89	11	1·19	26	1·50	5	1·12	48	1·76	38	1·72	12	1·46	55	1·87
6	0·94	12	1·22	27	1·52	6	1·15	49	1·77	39	1·72	13	1·47	56	1·87
7	0·99	13	1·25	28	1·53	7	1·19			40	1·73	14	1·49	57	1·88
8	1·03	14	1·27			8	1·22	c=15		41	1·74	15	1·50	58	1·89
9	1·07	15	1·29	c=7		9	1·25	0	0·88	42	1·75	16	1·52	59	1·89
10	1·11	16	1·32			10	1·27	1	1·03	43	1·76	17	1·53	60	1·90
11	1·14	17	1·34	0	0·54	11	1·30	2	1·10	44	1·76	18	1·55	61	1·90
12	1·17	18	1·36	1	0·77	12	1·32	3	1·15	45	1·77	19	1·56	62	1·91
13	1·20	19	1·38	2	0·87	13	1·34	4	1·19	46	1·78	20	1·57	63	1·91
		20	1·39	3	0·93	14	1·36	5	1·22	47	1·79	21	1·59	64	1·92
c=4		21	1·41	4	0·99	15	1·38	6	1·25	48	1·79	22	1·60	65	1·92
0	0·30	22	1·43	5	1·04	16	1·40	7	1·28	49	1·80	23	1·61	66	1·93
1	0·60	23	1·44	6	1·08	17	1·42	8	1·31	50	1·81	24	1·62	67	1·93
2	0·72	24	1·46	7	1·12	18	1·43	9	1·33	51	1·81	25	1·63	68	1·94
3	0·80			8	1·15	19	1·45	10	1·35	52	1·82	26	1·64	69	1·94
4	0·87	c=6		9	1·18	20	1·47	11	1·37	53	1·83	27	1·65	70	1·95
5	0·93	0	0·48	10	1·21	21	1·48	12	1·39	54	1·83	28	1·66	71	1·95
6	0·98	1	0·72	11	1·24	22	1·49	13	1·41	55	1·84	29	1·67	72	1·96
7	1·03	2	0·82	12	1·26	23	1·51	14	1·43	56	1·85	30	1·68	73	1·96
8	1·07	3	0·89	13	1·29	24	1·52	15	1·45	57	1·85	31	1·69	74	1·97
9	1·10	4	0·95	14	1·31	25	1·54	16	1·46	58	1·86	32	1·70	75	1·97
10	1·14	5	1·00	15	1·33	26	1·55	17	1·48	59	1·87	33	1·71	76	1·98
11	1·17	6	1·05	16	1·35	27	1·56	18	1·49	60	1·87	34	1·72	77	1·98
12	1·20	7	1·09	17	1·37	28	1·57	19	1·51	61	1·88	35	1·73	78	1·99
13	1·22	8	1·13	18	1·39	29	1·58	20	1·52	62	1·88	36	1·73	79	1·99
14	1·25	9	1·16	19	1·41	30	1·60	21	1·54	63	1·89	37	1·74	80	2·00
15	1·27	10	1·19	20	1·42	31	1·61	22	1·55	64	1·89	38	1·75	81	2·00
16	1·30	11	1·22	21	1·44	32	1·62	23	1·56	65	1·90	39	1·76	82	2·01
		12	1·24	22	1·46	33	1·63	24	1·57			40	1·77	83	2·01
		13	1·27	23	1·47	34	1·64	25	1·59	c=20		41	1·77	84	2·01
c=5		14	1·29	24	1·49	35	1·65	26	1·60	0	1·00	42	1·78	85	2·02
		15	1·31	25	1·50	36	1·66	27	1·61	1	1·14	43	1·79	86	2·02
0	0·40	16	1·33	26	1·51	37	1·67	28	1·62	2	1·19	44	1·80	87	2·03
1	0·67	17	1·35	27	1·53	38	1·68	29	1·63	3	1·24	45	1·80	88	2·03
2	0·78	18	1·37	28	1·54	39	1·69	30	1·64	4	1·27	46	1·81	89	2·03
3	0·85	19	1·39	29	1·55	40	1·70	31	1·65	5	1·30	47	1·82	90	2·04
4	0·92	20	1·41	30	1·57	41	1·70	32	1·66	6	1·33	48	1·82	91	2·04
5	0·97			31	1·58	42	1·71					49	1·83	92	2·05

To choose the value of the constant c, standard errors of untransformed numbers of lesions obtained with different treatments, whose means are not smaller than about 10, are plotted against the means, and a regression line is roughly passed through the plotted points. The value of c is the distance

between the origin and the point of intersection of the regression line with the axis of abscissae, which should be on the negative part of the axis. The choice of c need not be based on experimental results that are actually analysed. It can be based on results of other experiments made with the same virus-host system in similar conditions, or the value of c may be found from a special experiment. No great accuracy is usually required in choosing the value of c, for any value from a rather wide range gives satisfactory results. For example, any value between 5 and 10 is usually satisfactory with numbers of lesions produced by tobacco mosaic virus or tomato bushy stunt virus on leaves of *Nicotiana glutinosa*. As the choice of c is to some extent arbitrary and need not be based on any information obtained from the experiment whose results are actually analysed, obviously it does not cause a loss of a degree of freedom in the final analysis of the results.

If an experiment is done only to see whether different inocula have different infectivities, or whether plants differ in their susceptibility to a given virus, the usual analysis of variance followed by a customary test of significance is simply carried out using the transforms (7). However, if an experiment is done to test a hypothesis that predicts that a treatment reduces activity of a virus preparation to a certain proportion of the original activity, special methods of analysis are necessary. A method that can be used for such tests was described for the transformation (1) (Kleczkowski, 1953). The transformation (7) requires a modification of the method.

Testing results obtained for compatibility with hypotheses

Let us assume, as previously (Kleczkowski, 1953), that a hypothesis is tested which predicts that a treatment reduces activity of a virus preparation to a proportion p, which is a function f of a variable ψ and an adjustable constant k, so that

$$p = f(\psi, k). \tag{8}$$

To test this, inocula, some containing variously diluted untreated virus and others containing virus that has been subjected to the treatment under test, are inoculated to leaves (or half-leaves) of test plants, distributing the different inocula so that analysis of variance can subsequently be applied.

It is usually possible, by suitably adjusting the values of three constants, N, ξ and λ, to express the relationship between the numbers of lesions (Y) and the concentration of a virus in the inoculum (h) by the equation

$$Y = \frac{N}{\lambda\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left\{-\tfrac{1}{2}\left(\frac{r-\xi}{\lambda}\right)^2\right\} dr, \tag{9}$$

where $u = \log_{10} h$ (Kleczkowski, 1950, 1953). If the virus preparation has been subjected to the treatment under test and its activity is expected to be altered by the factor p of equation (8), $u = \log_{10} ph$.

The value of Y of equation (9) is considered to be the postulated mean number of lesions expected to be formed by a given inoculum. Now, in the transformation (7) logarithms are taken of experimental values of

$$v = \tfrac{1}{2}[x + c + \sqrt{(x^2 + 2cx)}].$$

The mean value of v (say $\bar{v}$) is half the sum of the mean value of $(x+c)$, which is $(m+c)$, and of the mean value of $t=\sqrt{(x^2+2cx)}$. To find the mean value of t (say $\bar{t}$) use is made of the fact that var $(t)=E(t^2)-\bar{t}^2$. Therefore

$$\text{var } (t) = E(x^2+2cx)-\bar{t}^2 = E(x^2)+2cm-\bar{t}^2 = \text{var } (x)+m^2+2cm-\bar{t}^2.$$

Thus $\bar{t}=\sqrt{[m^2+2cm-\text{var } (t)+\text{var } (x)]}$.

The variance of x is given by (5) and equals approximately $\beta^2(m^2+2cm)$. To obtain the variance of t we notice that $dt/dx=(x+c)/\sqrt{(x^2+2cx)}$, which means that t is obtained from x by the transformation

$$t=\int \frac{(x+c)dx}{\sqrt{(x^2+2cx)}}\,. \tag{10}$$

As the transformation (6) gives the values of y with a variance approximately equal to 1, the transformation (10) will make the variance of t approximately equal to $\beta^2(m+c)^2$. It can be assumed, therefore, that

$$\text{var } (t) - \text{var } (x) = \beta^2 c^2, \tag{11}$$

and, therefore, $\bar{t}=\sqrt{(m^2+2cm-\beta^2c^2)}$, so that $\bar{v}=\frac{1}{2}[m+c+\sqrt{(m^2+2cm-\beta^2c^2)}]$. Thus, if Y of equation (9) is the postulated mean number of lesions, the postulated value of $\bar{v}$ will be

$$\overline{V}=\tfrac{1}{2}[Y+c+\sqrt{(Y^2+2cY-\beta^2c^2)}]. \tag{12}$$

The values of z, i.e. the means of experimental values of z of the transformation (7), are decimal logarithms of geometric means of v. Thus, to be compared with the values of $\bar{z}$, the postulated arithmetic means of the values of v must be converted into decimal logarithms of geometric means. Let $\overline{Z}$ be the logarithm of the postulated geometric mean of v. Assuming the values of z to be normally distributed with the mean $\overline{Z}$ and variance σ^2, the postulated arithmetic mean of $v=10^z$ is

$$\overline{V}=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} 10^z \exp\left\{-\tfrac{1}{2}\left(\frac{z-\overline{Z}}{\sigma}\right)^2\right\}dz = 10^{\overline{Z}}\,e^{\frac{1}{2}(2\cdot3\sigma)^2},$$

so that

$$\overline{Z}=\log_{10}\overline{V}-1\cdot15\sigma^2 = \log_{10}\tfrac{1}{2}[Y+c+\sqrt{(Y^2+2cY-\beta^2c^2)}]-1\cdot15\sigma^2. \tag{13}$$

The estimate of variance of a single value of z, say s^2, obtained from analysis of variance of the results, can be used as σ^2. The value of β^2 can be obtained from s^2 if it is remembered that s^2 is approximately equal to $(0\cdot4343\beta)^2$, so that β^2 can be taken as equal to $5\cdot29s^2$. Substituting these values into (13) we obtain

$$\overline{Z}=\log_{10}\tfrac{1}{2}[Y+c+\sqrt{(Y^2+2cY-5\cdot29s^2c^2)}]-1\cdot15s^2. \tag{14}$$

The value of Y is given by (9) and c is the constant of the transformation (7).

To avoid obtaining false values of $\overline{Z}$ by equation (14) the value of c should be more carefully chosen than when experimental results are merely tested for significance of differences between treatments. Although the choice of c need not be based on the data of the experiment whose results are actually analysed, a check should be made to see whether the chosen value of c does indeed correspond with the conditions of the experiment. A convenient method of doing this is based on computing variances of untransformed numbers of lesions (x) and also of the values $t=\sqrt{(x^2+2cx)}$ separately for each treatment

7

(inoculum), without making any adjustments to eliminate any known sources of variation such as between plants or leaf positions. The differences between variances of t and x should not tend to increase or decrease with the increasing means of x and should on average be equal to $\beta_1 c^2$ (see (11)). (β_1 may be greater than β of equations (12) and (13), because the latter may be based on results of analysis of variance, which would be expected to decrease the value of β.) The value of β_1 can approximately be obtained by plotting standard errors of x's for different treatments, whose means are not smaller than about 10, against the means and passing roughly a regression line through the plotted points. The value of β_1 can be assumed approximately equal to the slope of the line. In fact it should be slightly greater than the slope, for the same reason that β of (3) is greater than b of (2), but the difference decreases with increasing size of samples and can be neglected when the number of replications is 10 or more. With tobacco mosaic virus inoculated to leaves of *Nicotiana glutinosa* in our conditions the value of β_1 was usually about 0·5 and the value of c about 6.

Further procedure is similar to that previously described for the transformation (1) (Kleczkowski, 1953). The values of the constants k, N, ξ and λ of equations (8) and (9) are adjusted by the method of least squares, i.e. by minimizing the value of $\Sigma(\bar{z}_i - \bar{Z}_i)^2$, where $\bar{z}_i$'s are the mean values of z obtained with different inocula and $\bar{Z}_i$'s are the corresponding values of $\bar{Z}$ of equation (14). If each $\bar{z}_i$ is a mean of n values of z, $n\Sigma(\bar{z}_i - \bar{Z})^2/W$, where W is the number of degrees of freedom equal to the number of inocula minus the number of adjustable constants of equations (8) and (9), should be an estimate of the variance of a single value of z independently of that given by s^2, obtained by the customary procedure of the analysis of variance. If the hypothesis on which equation (8) is based is true, the two estimates should be compatible, and this can be tested by the variance ratio test.

The values of the constants of equations (8) and (9) can be adjusted as described previously (Kleczkowski, 1953). If it becomes necessary to use 'normal' equations, the partial derivatives of $\bar{Z}$'s can be obtained from those previously listed for the procedure based on the transformation (1) by substituting $[Y + c + \sqrt{(Y^2 + 2cY - 5\cdot29s^2c^2)}]$ for $(Y + c)$ and multiplying by $(1 + (Y + c)/\sqrt{(Y^2 + 2cY - 5\cdot29s^2c^2)})$.

REFERENCES

KLECZKOWSKI, A. (1949). The transformation of local lesion counts for statistical analysis. *Ann. appl. Biol.* **36**, 139.

KLECZKOWSKI, A. (1950). Interpreting relationships between the concentrations of plant viruses and numbers of local lesions. *J. gen. Microbiol.* **4**, 53.

KLECZKOWSKI, A. (1953). A method of testing results of infectivity tests with plant viruses for compatibility with hypotheses. *J. gen. Microbiol.* **8**, 295.