

Biometrika Trust

On the Constants of Index-Distributions as Deduced from the Like Constants for the Components of the Ratio, with Special Reference to the Opsonic Index

Author(s): Karl Pearson

Source: *Biometrika*, Vol. 7, No. 4 (Nov., 1910), pp. 531-541

Published by: [Biometrika Trust](#)

Stable URL: <http://www.jstor.org/stable/2345380>

Accessed: 23/06/2014 06:31

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



Biometrika Trust is collaborating with JSTOR to digitize, preserve and extend access to *Biometrika*.

<http://www.jstor.org>

ON THE CONSTANTS OF INDEX-DISTRIBUTIONS AS DEDUCED FROM THE LIKE CONSTANTS FOR THE COMPONENTS OF THE RATIO, WITH SPECIAL REFERENCE TO THE OPSONIC INDEX.

By KARL PEARSON, F.R.S.

(1) GIVEN x and y two variables, the frequency constants of both of which are known, we often require in statistics the frequency constants of their ratio or index: $i = x/y$.

If the coefficients of variation are small, we have with the usual notation for means, standard deviations, moment coefficients, etc.*:

$$\bar{i} = \left(\frac{\bar{x}}{\bar{y}}\right) \left\{ 1 + v_x^2 - r_{xy} v_x v_y - \frac{\mu_3'}{\bar{y}^3} + \frac{p_{12}}{\bar{x}\bar{y}} + \frac{\mu_4'}{\bar{y}^4} - \frac{p_{13}}{\bar{x}\bar{y}^3} + \text{etc.} \right\} \dots\dots\dots(\text{i}),$$

$$M_2 = \sigma_i^2 = \left(\frac{\bar{x}}{\bar{y}}\right)^2 \left\{ v_x^2 + v_y^2 - 2v_x v_y r_{xy} + \frac{4p_{12}}{\bar{x}\bar{y}^2} - \frac{2p_{21}}{\bar{x}^2\bar{y}} - \frac{2\mu_3'}{\bar{y}^3} + \frac{3p_{22} - p_{11}^2}{\bar{x}^2\bar{y}^2} + \frac{3\mu_4' - 2\mu_2'^2}{\bar{y}^4} + \frac{2\mu_2' p_{11}}{\bar{x}\bar{y}^3} + \text{etc.} \right\} \dots\dots\dots(\text{ii}),$$

$$M_3 = \left(\frac{\bar{x}}{\bar{y}}\right)^3 \left\{ \frac{\mu_3}{\bar{x}^3} - \frac{\mu_3'}{\bar{y}^3} - \frac{3p_{21}}{\bar{x}^2\bar{y}} + \frac{3p_{12}}{\bar{x}\bar{y}^2} + \frac{9p_{22} - 6p_{11}^2 - 3\mu_2\mu_2'}{\bar{x}^2\bar{y}^2} - \frac{3(p_{31} - \mu_2 p_{11})}{\bar{x}^3\bar{y}} - \frac{9(p_{13} - \mu_2' p_{11})}{\bar{x}\bar{y}^3} + \frac{3(\mu_4' - \mu_2'^2)}{\bar{y}^4} + \text{etc.} \right\} \dots\dots\dots(\text{iii}),$$

$$M_4 = \left(\frac{\bar{x}}{\bar{y}}\right)^4 \left\{ \frac{\mu_4}{\bar{x}^4} + \frac{\mu_4'}{\bar{y}^4} + \frac{6p_{22}}{\bar{x}^2\bar{y}^2} - \frac{4p_{31}}{\bar{x}^3\bar{y}} - \frac{4p_{13}}{\bar{x}\bar{y}^3} + \text{etc.} \right\} \dots\dots\dots(\text{iv}).$$

These formulae go to the order of the 4th power in the coefficients of variation, but of course this is not to the same order of approximation in M_2 , M_3 and M_4 .

(2) It will be seen at once that these approximate formulae would be practically unworkable if x and y were correlated, as we should have to find 3rd and 4th order product moments.

* A rule denotes a mean value, σ a standard deviation, $v = \sigma/\text{mean}$, is a coefficient of variation; μ_2, μ_3, μ_4 are the moment coefficients for x , μ_2', μ_3', μ_4' for y , M_2, M_3, M_4 for the index i , and $p_{uv} = S(x - \bar{x})^u (y - \bar{y})^v / N$, where N is the total number of pairs. Thus $p_{30} = \mu_3$, $p_{02} = \mu_2'$, etc.

If they be uncorrelated we have :

$$\begin{aligned}\bar{v} &= \left(\frac{\bar{x}}{\bar{y}}\right) \left\{ 1 + v_x^2 - \frac{\mu_3'}{\bar{y}^3} + \frac{\mu_4'}{\bar{y}^4} - \text{etc.} \right\} \dots\dots\dots(\text{i})^{\text{bis}}, \\ M_2 = \sigma_x^2 &= \left(\frac{\bar{x}}{\bar{y}}\right)^2 \left\{ v_x^2 + v_y^2 - \frac{2\mu_3'}{\bar{y}^3} + 3v_x^2 v_y^2 + \frac{3\mu_4' - 2\mu_2'^2}{\bar{y}^4} + \text{etc.} \right\} \dots\dots\dots(\text{ii})^{\text{bis}}, \\ M_3 &= \left(\frac{\bar{x}}{\bar{y}}\right)^3 \left\{ \frac{\mu_3}{\bar{x}^3} - \frac{\mu_3'}{\bar{y}^3} + \frac{6\mu_2\mu_2'}{\bar{x}^2\bar{y}^2} + \frac{3(\mu_4' - \mu_2'^2)}{\bar{y}^4} + \text{etc.} \right\} \dots\dots\dots(\text{iii})^{\text{bis}}, \\ M_4 &= \left(\frac{\bar{x}}{\bar{y}}\right)^4 \left\{ \frac{\mu_4}{\bar{x}^4} + \frac{\mu_4'}{\bar{y}^4} + \frac{6\mu_2\mu_2'}{\bar{x}^2\bar{y}^2} + \text{etc.} \right\} \dots\dots\dots(\text{iv})^{\text{bis}}.\end{aligned}$$

Formulae (i) to (iv) are extensions of formulae given by me in a paper on *spurious correlation**. Formulae (i)^{bis} to (iv)^{bis} are due to Dr M. Greenwood, Jun., who obtained them in dealing with the problem of the distribution of the opsonic index. They show at once two noteworthy but not yet noted points, namely, (a) if the distribution of both x and y be symmetrical, i.e. μ_3 and $\mu_3' = 0$, M_3 will not be zero or the distribution of indices must be skew; (b) the mean of the ratio of two numbers picked out of the same series is certainly greater than unity if the series be symmetrical, and will probably be always greater than unity even if it be not, i.e. $\bar{v} > \bar{x}/\bar{y}$, which is unity for the same x and y series†. Dr Greenwood found, however, that these formulae did not give with sufficient accuracy the constants of the index distribution. This was probably due to two causes: (a) clearly we ought only to keep to the square order in M_2 and the cubic order in M_3 if we retain only to the 4th order in M_4 ; or if we keep to the higher terms in M_2 and M_3 , we must go further with M_4 ; and (b) the values of v_x or v_y are not so small, that the convergency is sufficient when we take these lowest terms of the expansions. It seemed accordingly desirable to find some other way of attacking the problem, and Dr Greenwood asked me for suggestions. The problem he had in view was the distribution of the opsonic index when the blood of the same individual taken in the same manner at the same time was treated as test and as normal. If a wide range of values of the opsonic index could thus be obtained, it would cast some light on what deviations from unity must be looked upon as significant, when test and normal were different individuals.

(3) The idea that suggested itself to me was a fairly simple one, namely to tabulate the y -frequencies to a variable $z = 1/y$. The units of the z -frequency groups will not be equal, but they will all be sufficiently small for us to concentrate their frequencies at their mid-points. We can then calculate their moments easily. Let $\nu_1, \nu_2, \nu_3, \nu_4$ be the moments of x about the zero value of x , and $\nu_1', \nu_2', \nu_3', \nu_4'$ be the moments of z about the zero value of z . Then

$$i = z \times x,$$

* *R. S. Proc.* Vol. LX. p. 492.

† Given two dice, it would be advantageous to bet that the ratio of the number of pips on the two at a cast will exceed unity.

and if m_1, m_2, m_3, m_4 be the moments of i about its zero value, we have :

$$m_1 = P_{11}, \quad m_2 = P_{22}, \quad m_3 = P_{33}, \quad m_4 = P_{44},$$

where P_{uv} is the uv th product moment of z and x about axes through their zero values. In the special case in which z and x are uncorrelated, as in the opsonic index,

$$m_1 = \nu_1 \nu_1', \quad m_2 = \nu_2 \nu_2', \quad m_3 = \nu_3 \nu_3', \quad m_4 = \nu_4 \nu_4',$$

or we can obtain any moment about the zero of i by multiplying the corresponding moments of x and z .

These moments are then transferred by the usual formulae

$$M_2 = m_2 - m_1^2, \quad M_3 = m_3 - 3m_2 m_1 + 2m_1^3, \quad M_4 = m_4 - 4m_3 m_1 + 6m_2 m_1^2 - 3m_1^4$$

to the mean as origin and the type of frequency calculated in the usual way from the corresponding β_1 and β_2 .

In Greenwood and White's data we have, for the three series discussed below, elementary subranges rising by .32, .24 and .16 of a bacillus per leucocyte for the distribution of the means of counts of 25, 50 and 100. I find that for such distributions, the value of the variate z will only be affected by about a unit in the third place of decimals in the worst cases, i.e. the lowest values of y in samples of 25, whether we use for z (i) the mean of the inverses of the start and finish of the subrange, (ii) the mean of the inverses of all the 32, 24, or 16 hundredths in the subrange, or (iii) the inverse of the mid-point of the subrange. I have accordingly adopted the last as the simplest value of z for practical purposes.

(4) *Illustration of the method.* I. *Greenwood and White's 200 samples of 100 counts.*

The data are given in Table I. The moments of the frequency distributions for y and z as variates about the zero of those variates were then found by tables of powers of numbers and a calculating machine*.

I. *Distribution of 40,000 indices for 200 samples of 100 counts.*

For x : $\nu_1 = 3.67620$, $\nu_2 = 13.67643$, $\nu_3 = 51.50333$, $\nu_4 = 196.40357$.

For z : $\nu_1' = 0.275165$, $\nu_2' = 0.076603$, $\nu_3' = 0.021577$, $\nu_4' = 0.0061494$.

These give:

$$m_1 = 1.01156, \quad m_2 = 1.04766, \quad m_3 = 1.11126, \quad m_4 = 1.20776,$$

which transferred to the mean give for frequency constants of the 40,000 possible indices:

$$\begin{array}{lll} \mu_2 = .02440, & \sigma = .1562, & \kappa = .2651, \\ \mu_3 = .00213, & \beta_1 = .3123, & \text{Mean} = 1.01156, \\ \mu_4 = .00235, & \beta_2 = 3.9472, & \text{Mode} = .9774. \end{array}$$

The distribution is accordingly of Type IV.

* I have cordially to acknowledge help from Alice Lee, D.Sc., Julia Bell, M.A., and Amy Barrington, who have each worked out nearly the whole of one distribution for me, and from H. Gertrude Jones, who has prepared the diagrams.

TABLE I. Greenwood and White's distribution of 200 samples of 100 counts. Opsonic Index. Tuberculosis.

Variate x ...	2·745	2·905	3·065	3·225	3·385	3·545	3·705	3·865	4·025	4·185	4·345	4·505	4·665	4·825	4·985	5·145	Total
Frequency ...	2	9	4	13	29	36	34	34	15	11	5	3	1	1	2	1	200
Variate z ...	·364	·344	·326	·310	·295	·282	·270	·259	·248	·239	·230	·222	·214	·207	·201	·194	—

TABLE II. Greenwood and White's Distribution of 400 samples of 50 counts. Opsonic Index. Tuberculosis

Variate x ...	2·35	2·59	2·83	3·07	3·31	3·55	3·79	4·03	4·27	4·51	4·75	4·99	5·23	5·47	5·71	5·95	Total
Frequency ...	2	9	21	30	58	83	71	64	31	15	10	3	1	—	1	1	400
Variate z ...	·426	·386	·353	·326	·302	·282	·264	·248	·234	·222	·211	·200	·191	·183	·175	·168	—

TABLE III. Greenwood and White's Distribution of 800 samples of 25 counts. Opsonic Index. Tuberculosis.

Variate x ...	1·90	2·22	2·54	2·86	3·18	3·50	3·82	4·14	4·46	4·78	5·10	5·42	5·74	6·06	6·38	6·70	7·02	Total
Frequency ...	1	14	38	74	128	145	147	113	74	37	19	7	2	—	—	—	1	800
Variate z ...	·526	·450	·394	·350	·314	·286	·262	·242	·224	·209	·196	·185	·174	·165	·157	·149	·142	—

We find: $r = 16.5111$, $\nu = -9.9138$ (since μ_3 is positive),
 $a = .52743$, $y_0 = 8638.6$,
 and the distribution of the 40,000 indices possible is given by:

$$Y = 8638.6 \left(1 + \frac{X^2}{.27818} \right)^{-9.2556} e^{9.9138 \tan^{-1} \frac{X}{.52743}} \dots\dots\dots(i),$$

where Y is the frequency at distance X from the origin of the curve which is at the opsonic index .6949, the unit of X being absolute opsonic index measurement.

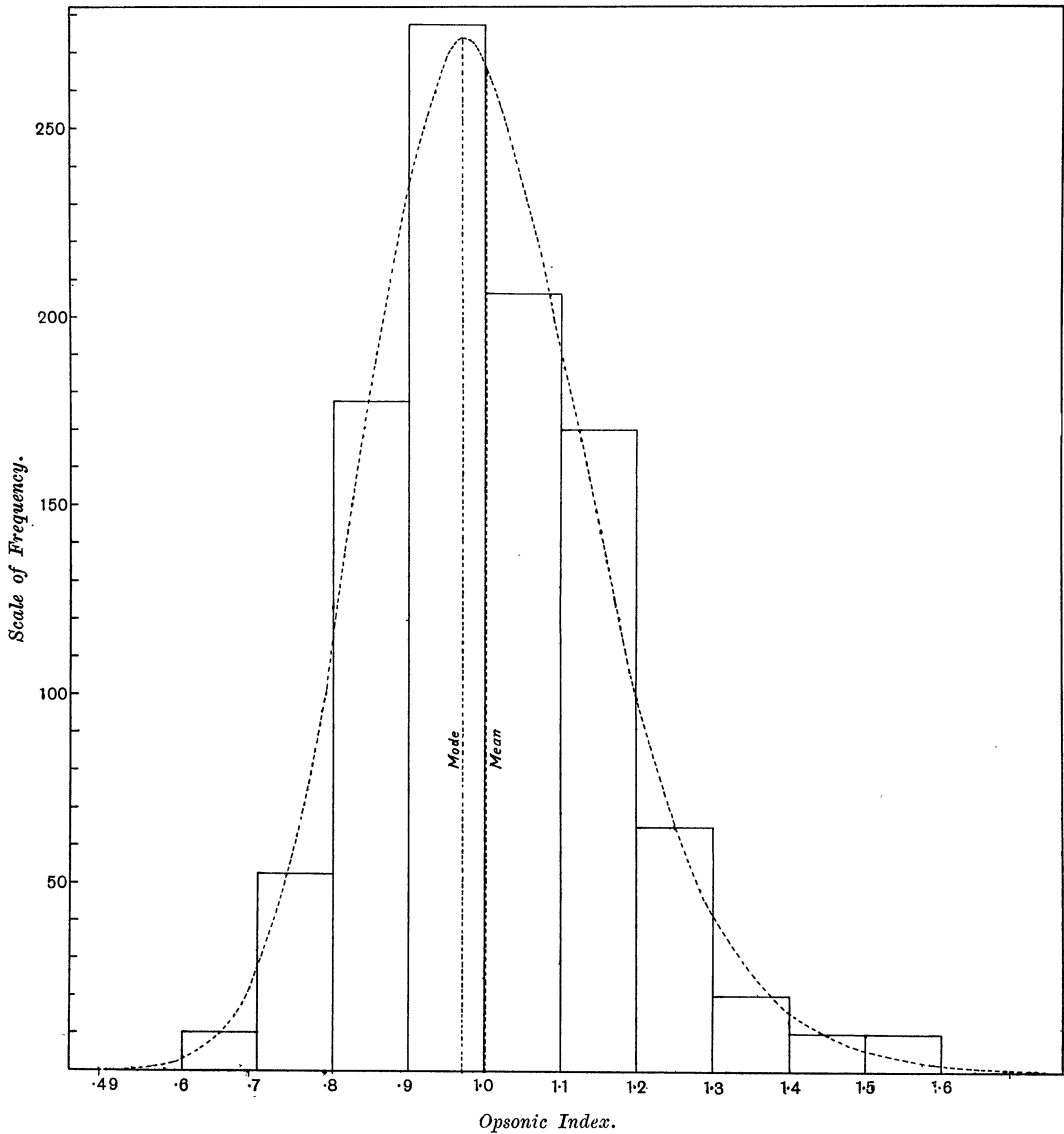


DIAGRAM I. Samples of 100 counted.

It will be seen, in accordance with the results drawn from the approximate formulae, that the mean value of the opsonic index is greater than unity and that there is marked skewness, i.e. no approach even with samples of 100 to a Gaussian distribution. Clearly the most probable value of the index is less than unity, and is rather farther from unity than the mean.

Greenwood and White have made an experimental determination by 200 random drawings. In Diagram I the frequencies of the above curve are reduced to 1000 total and plotted against their experimental data increased to 1000. This of course exaggerates the apparent deviations, but enables our three diagrams to be compared among themselves.

(5) *Further Illustrations.* The following frequencies are also deduced from Greenwood and White's results for 400 samples of 50 and 800 samples of 25.

With the same notation as before the constants of the distribution of the opsonic indices are as follows:

II. *Distribution of 160,000 indices for 400 samples of 50 counts.*

For x : $\nu_1 = 3.68140$, $\nu_2 = 13.82272$, $\nu_3 = 52.92733$, $\nu_4 = 206.69803$.

For z : $\nu'_1 = .277213$, $\nu'_2 = .078453$, $\nu'_3 = .022683$, $\nu'_4 = .006705$,

$m_1 = 1.02053$, $m_2 = 1.08443$, $m_3 = 1.20055$, $m_4 = 1.38591$,

$\mu_2 = .042948$, $\sigma = .2072$, $\kappa = .5415$,

$\mu_3 = .006191$, $\beta_1 = .4837$, $\text{Mean} = 1.0205$,

$\mu_4 = .007567$, $\beta_2 = 4.1022$, $\text{Mode} = .9611$.

The distribution is of Type IV:

$$r = 20.86018, \quad \nu = -22.67044,$$

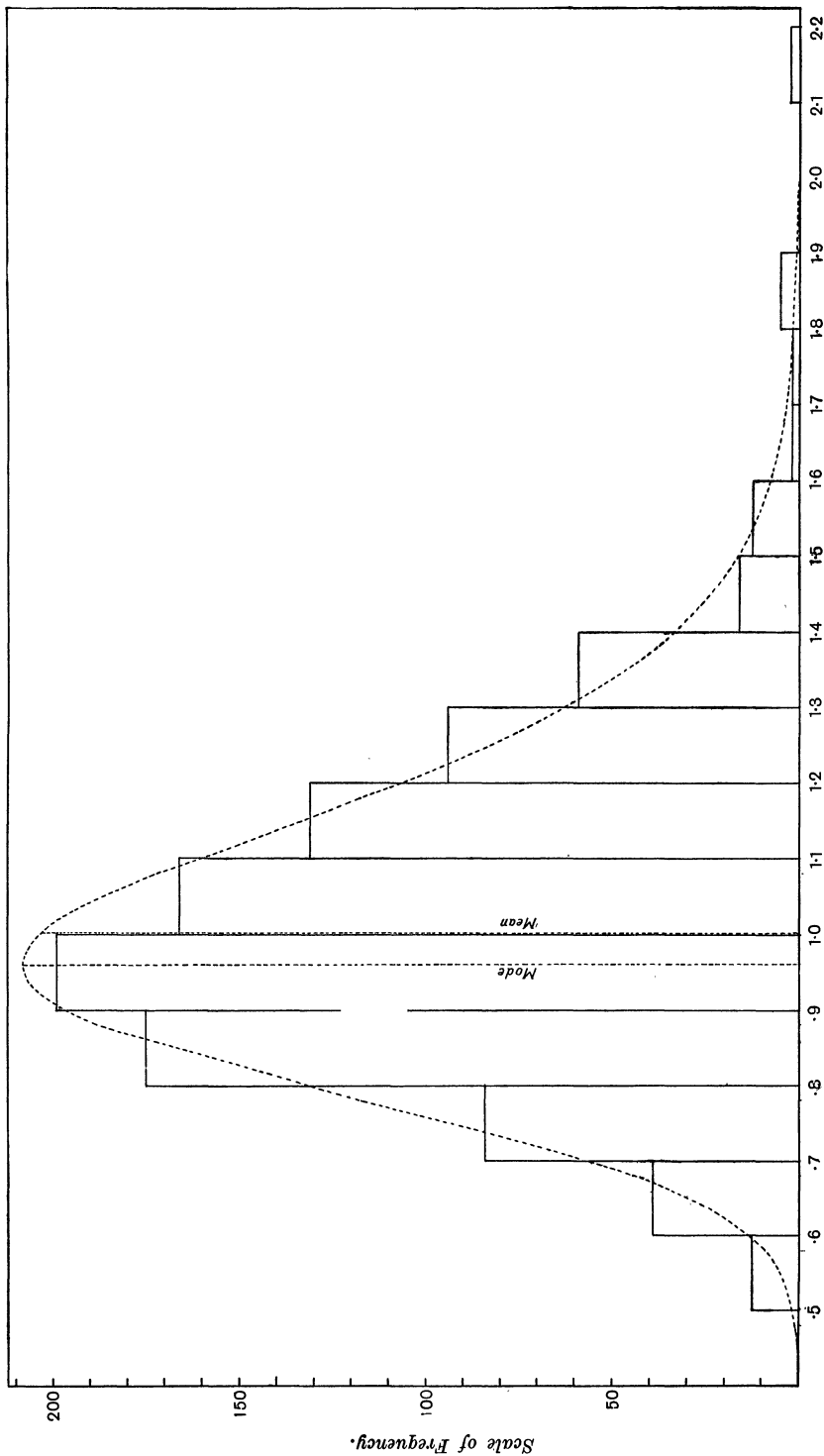
$$a = .62536, \quad y_0 = 16.9740,$$

and the equation to distribution of the 160,000 indices possible is given by:

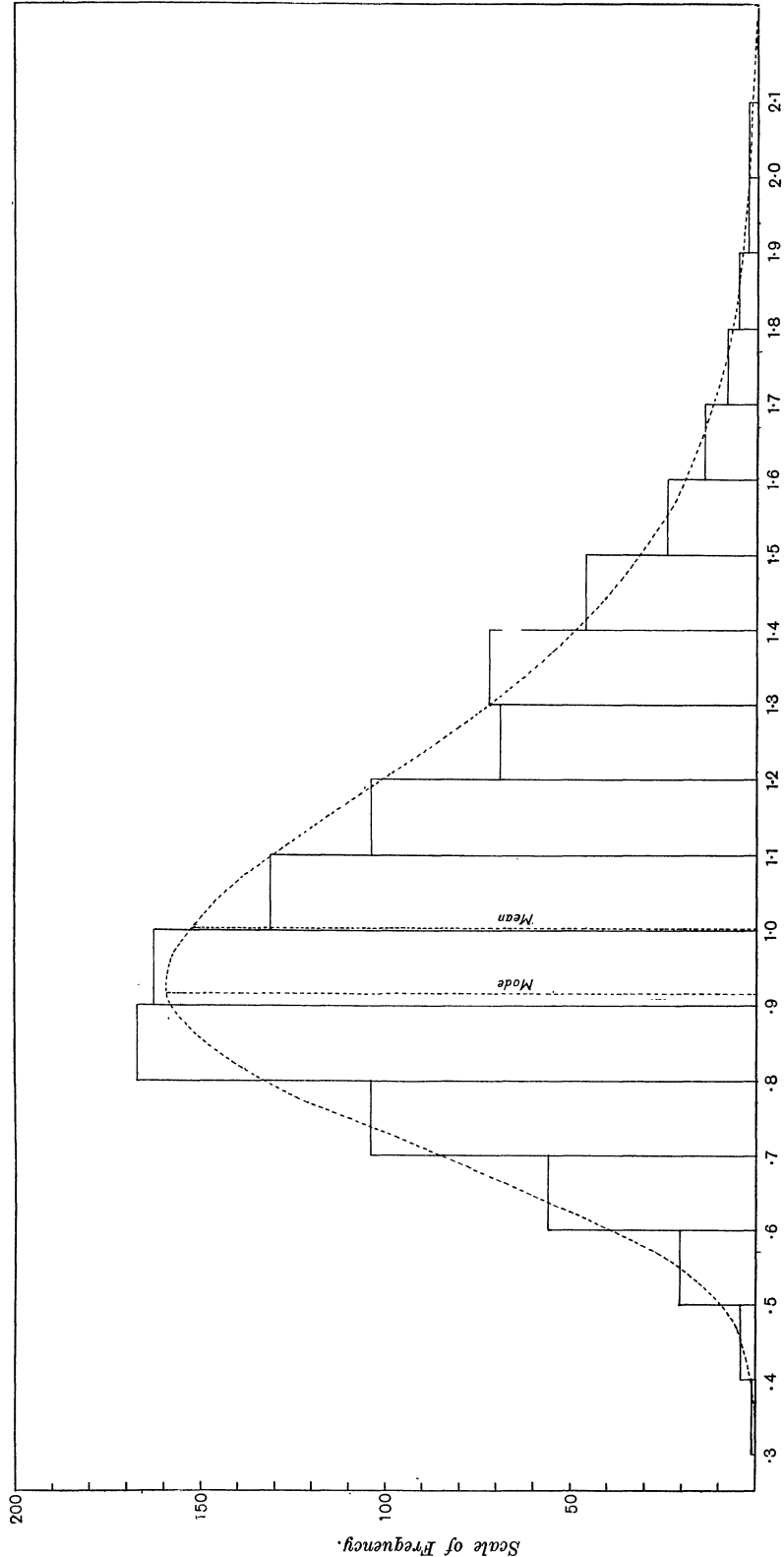
$$Y = 16.9740 \left(1 + \frac{X^2}{.39107} \right)^{-11.43009} \frac{e^{22.67044 \tan^{-1} \frac{X}{.62536}}}{e} \dots\dots\dots(ii).$$

The origin of the curve is at the opsonic index .3409.

Diagram II gives this theoretical distribution reduced to 1000 cases and compared with Greenwood and White's experimental data for 400 increased to 1000 also.



Opsonic Index.
DIAGRAM II. Samples of 50 counted.



Opsonic Index.
DIAGRAM III. Samples of 25 counted.

III. *Distribution of 640,000 indices for 800 samples of 25 counts.*

For x : $\nu_1 = 3.68160$, $\nu_2 = 14.00990$, $\nu_3 = 55.03056$, $\nu_4 = 222.90230$,
 For z : $\nu'_1 = .281298$, $\nu'_2 = .082055$, $\nu'_3 = .024861$, $\nu'_4 = .007834$,
 $m_1 = 1.03563$, $m_2 = 1.14959$, $m_3 = 1.36811$, $m_4 = 1.74622$,
 $\mu_2 = .077063$, $\sigma = .2776$, $\kappa = 1.2480$,
 $\mu_3 = .017943$, $\beta_1 = .7060$, Mean = 1.0356,
 $\mu_4 = .025583$, $\beta_2 = 4.3088$, Mode = .9330.

The distribution is of Type VI*:

$$r = -31.25271, \quad q_1 = 51.67822, \quad q_2 = 18.42551, \\ a = 1.52075, \quad \log y_0 = 26.6813462,$$

and the equation to the distribution of the 640,000 indices possible is given by:

$$Y = \text{antilog } 26.6813462 \times (X - 1.52075)^{18.42551} X^{-51.67822} \dots\dots\dots(\text{iii}).$$

The origin is at the opsonic index -1.4304 .

Diagram III gives this distribution reduced to a total of 1000 indices and set against Greenwood and White's experimental curve for 800 increased to 1000. This exaggerates the apparent deviations, but enables us to compare with the samples of 50 and 100. Considering the relative paucity of Greenwood and White's drawings, and the fact that we have fitted our curves to their non-replaced material, whereas a random sample would probably be better represented by the replaced material, the fits seem fairly close. The actual goodness of fit test was not applied, as at the time of writing the paper means of drawing the curves on a large scale and mechanically integrating them were not accessible.

(6) It is as well to look at a combined table of the frequency constants of the three distributions given above.

No. of Counts average based on	No. in Distribution of Means	Mean	Mode	σ	β_1	β_2
25	800	1.0356	.9330	.2776	.7060	4.3088
50	400	1.0205	.9611	.2072	.4837	4.1022
100	200	1.0116	.9774	.1562	.3123	3.9472

It will be seen that these values form a very consistent relatively smoothly altering system. But the approach to normality is very slow. Even with counts of 100 the distribution of the opsonic index is markedly skew and platykurtic, and it would not be safe to treat the distribution as a normal curve†. In all cases the

* Type VI is adjacent to Type IV and in this case the curve is almost on the boundary line—i.e. Type V: see Rhind's Diagram, *Biometrika*, Vol. VII. p. 389.

† The reduction of β_1 and β_2 to the Gaussian values 0 and 3 is, of course, not at the same rate as if we averaged 25, 50 and 100 opsonic indices. What we are doing here is to average the number of bacilli in 25, 50 and 100 cells on which the two factors of the index are based—a very different process.

modal or most probable value of the opsonic index of an individual tested by himself is less than unity and his average value greater than unity.

If we actually suppose the distributions normal and varying round the means with the standard deviations given above we have the following results:

Significance laid on an event which will not happen more frequently than	Number of Counts		
	25	50	100
Once in ten trials ...	·58—1·48	·68—1·36	·75—1·27
Once in eight trials ...	·61—1·46	·70—1·34	·77—1·25
Once in six trials ...	·65—1·42	·73—1·31	·80—1·23

This table is to be read in the following sense: If an opsonic index were based on a count of 50 cells, then once in eight trials an individual tested against himself would have an index lying outside the limits ·70 to 1·34. Or, again, with 100 counts once in six trials an individual tested against himself would have an index lying outside the limits ·80 to 1·23. It may be added that for most of the purposes of practical life or of exact science we should not consider an "improbability" which could happen once in ten trials as marking a significant differentiation. Much greater degrees of improbability would be required. In the case of medicine, however, much less certainty may be demanded of a judgment, and probably no weight would be given to an isolated opsonic determination in *diagnosis*. Still the matter gives ground for pause, the opsonic index of the same material tested by itself has a very wide range round unity.

The following table will show how closely Greenwood and White's experimental determinations agree with our theoretical evaluation of the constants:

Constant of Distribution	Samples of 25		Samples of 50		Samples of 100	
	G. + W.	P.	G. + W.	P.	G. + W.*	P.
Mean	1·0362	1·0356	1·0205	1·0205	1·0120	1·0116
Mode	·9432	·9330	·9536	·9611	·9689	·9774
Standard Deviation	·2585	·2776	·2204	·2072	·1553	·1562
β_1	·7099	·7060	·6952	·4837	·3654	·3132
β_2	4·3914	4·3088	4·8945	4·1022	3·7138	3·9472

The only substantial divergences are in Greenwood and White's values for β_1 and β_2 in the case of samples of 50. I have been through my results again and can find no error. Dr Greenwood has been through his figures and finds slight

* They have used the boundary curve Type V in the case of samples of 100.

slips (see footnote, p. 521) in the values of β_1 and β_2 . These are not sufficient to account for the divergence of their results from mine, and from the trend of their samples of 25 and 100. I am inclined to think the divergence is due to the presence of the outlying index 2.16 in their chance drawings—a result my curve shows to be exceedingly improbable not once in 3000 drawings, and the like of which does not occur in their samples of 50 (drawings replaced) (see p. 525). It would clearly be possible from their experimental method of drawing samples of 800, 400 and 200 indices to closely approach the corresponding theoretical distributions of 640,000, 160,000 and 40,000 represented in our curves. Our results then are in every way confirmatory of theirs, but place on a rather more satisfactory theoretical instead of experimental footing their deduction of index distributions. The general conclusion seems to be that except in the case of an extremely low or extremely high value of the opsonic index, little if any weight whatever ought to be placed on a *single* determination of this index. Hence the method would not be valid when applied to cases in which, owing to the evolution of a morbid process or the action of some drug, very few observations can be made under the same conditions, i.e. it must be of doubtful application in treatment. Further the concentration obtained by basing the index on a count of 100, rather than one of 50 or even 25, while sensible is not very rapid. It would require very large numbers—much beyond every-day practice—to reduce in a marked manner this variation of the opsonic index from unity, when an individual even is tested against himself. Generally the diagrams indicate that an extreme variation in excess is more likely to occur than an extreme variation in defect, but that the most probable index, when the individual is tested against himself, will be one somewhat less than unity.