This article was downloaded by: [New York University]

On: 27 July 2015, At: 19:14

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954

Registered office: 5 Howick Place, London, SW1P 1WG

Philosophical Magazine Series 6

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/tphm17>

XXVII. On the influence of past experience on future expectation

Karl Pearson F.R.S.

Published online: 16 Apr 2009.

To cite this article: Karl Pearson F.R.S. (1907) XXVII. On the influence of past experience on future expectation , Philosophical Magazine Series 6, 13:75, 365-378, DOI: [10.1080/14786440709463611](https://doi.org/10.1080/14786440709463611)

To link to this article: <http://dx.doi.org/10.1080/14786440709463611>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is

expressly forbidden. Terms & Conditions of access and use can be found at
<http://www.tandfonline.com/page/terms-and-conditions>

XXVII. *On the Influence of Past Experience on Future Expectation.* By KARL PEARSON, F.R.S.*

(1) **I**N the whole round of statistical investigation few principles are so much theoretically neglected at present and yet so largely and unconsciously appealed to as that theorem in probability by which our past experience is made a basis for future conduct. It is perfectly true that the logic of this theorem has been criticised and discussed by De Morgan, Boole, Venn, Edgeworth, and many others, but the extent to which it really lies at the basis of most state, municipal and individual actions is too often disregarded. It is not only the insurance companies, but the Chancellor of the Exchequer, the First Lord of the Admiralty, the local officer of health, the merchant, and the general practitioner, who almost daily express their faith in the approximate stability of statistical ratios. One and all, we act on the principle that the statistical ratio determined from our past experience will hold, at any rate approximately, for the near future. This category of the stability of statistical ratios is as all important not only in statistical theory but in practical conduct, as is from a second standpoint in physical theory and also in practical life the principle that the same causes will reproduce the same effects. Neither principle admits of an ultimate logical demonstration; both rest on the foundation of commonsense and the experience of what follows their disregard. Both need considerable care in their application, but what is quite clear is that practical life cannot progress without them. The object of the present paper is to put into a new form the mathematical process of applying the principle of the stability of statistical ratios, and to determine, on the basis of the generally accepted hypothesis, what is the extent of the influence which may be reasonably drawn from past experience.

The problem is of course not novel; it has been treated by a number of the greater writers on probability, but I believe the present investigation is original in character; it extends some of the usual conceptions and shows very serious limitations in the applications frequently made of the old results. Very crude applications of the principle are occasionally made in vital statistics, or even suggested in medico-statistical text-books. Thus past experience has shown that in a certain population a sample of n individuals shows a character in np individuals and an absence of this character in nq individuals ($p + q = 1$). These proportions are taken to hold good for future experience, and if a further sample m be

* Communicated by the Author.

expected, the result anticipated will be $m\bar{p}$ individuals of the given character with a probable error of $\cdot 67449\sqrt{m\bar{p}q}$. This result is often given without any regard to the relative magnitudes of m and n , or again of $\bar{p}$ and $\bar{q}$. Further, tables of the probability integral are applied to determine the probability of deviations, although many of the users would be much puzzled to cite where the justification for using the normal curve at all is to be found.

It seems therefore desirable to reinvestigate the whole subject. I start as most mathematical writers have done, with "the equal distribution of ignorance," or I assume the truth of Bayes' Theorem. I hold this theorem not as rigidly demonstrated, but I think with Edgeworth that the hypothesis of the equal distribution of ignorance is, within the limits of practical life, justified by our experience of statistical ratios, which *a priori* are unknown, *i. e.* such ratios do not tend to cluster markedly round any particular value. "Chances" lie between 0 and 1, but our experience does not indicate any tendency of actual chances to cluster round any particular value in this range. The ultimate basis of the theory of statistics is thus not mathematical but observational. Those who do not accept the hypothesis of the equal distribution of ignorance and its justification in observation, are compelled to produce definite evidence of the clustering of chances, or to drop all application of past experience to the judgment of probable future statistical ratios. It is perfectly easy to form new statistical algebras with other clustering of chances. We can form other geometries with varied axioms as to parallels. But as the usual axiom of parallels leads to results in close approximation to experience, so the "equal distribution of ignorance" leads us to results which accord reasonably with actual observation. As to the second alternative, even those who criticise Bayes' Theorem apply at every turn in practical life their past experience of statistical ratios to actual conduct.

With this prelude, which appears needful in the present state of opinion as to the basis of probability, I turn to the actual problem before me.

(2) Let the chance of a given event occurring be supposed to lie between x and $x + \delta x$, then if on $n = p + q$ trials an event has been observed to occur p times and fail q times, the probability that the true chance lies between x and $x + \delta x$ is, on the equal distribution of our ignorance,

$$P_x = \frac{x^p (1-x)^q dx}{\int_0^1 x^p (1-x)^q dx} \cdot \cdot \cdot \cdot \cdot \quad (i.)$$

This is Bayes' Theorem.

Now suppose that a second trial of $m = r + s$ instances be made, then the probability that the given event will occur r times and fail s , is on the *a priori* chance being between x and $x + \delta x$

$$= P_x \frac{m}{\begin{smallmatrix} r \\ s \end{smallmatrix}} x^r (1-x)^s,$$

and accordingly the total chance C_r , whatever x may be, of the event occurring r times in the second series, is

$$C_r = \frac{\begin{smallmatrix} m \\ r \end{smallmatrix} \int_0^1 x^{p+r} (1-x)^{q+s} dx}{\begin{smallmatrix} m \\ r \end{smallmatrix} \int_0^1 x^p (1-x)^q dx} \quad \dots \quad \text{(ii.)}$$

This is, with a slight correction, Laplace's extension of Bayes' Theorem.

A further evaluation of C_r is then usually made by expressing the integrals in β -functions, these again in γ -functions, then using Sterling's Theorem, and after a series of approximations, not always closely investigated, showing that C_r can be represented by the ordinates of a normal curve of errors. The present investigation is intended to replace this by a more rigid and non-approximative process.

We have :

$$C_r = \frac{\begin{smallmatrix} m \\ r \end{smallmatrix}}{\begin{smallmatrix} r \\ s \end{smallmatrix}} \frac{\Gamma(p+r+1) \Gamma(q+s+1) \Gamma(n+2)}{\Gamma(p+1) \Gamma(q+1) \Gamma(n+m+2)}.$$

Now write :

$$C_0 = \frac{\Gamma(q+m+1) \Gamma(n+2)}{\Gamma(q+1) \Gamma(n+m+2)}, \quad \dots \quad \text{(iii.)}$$

then the chances of the successive values of r , 0, 1, 2 m are the successive terms in the series :

$$C_0 \left\{ 1 + \frac{m}{1!} \frac{p+1}{q+m} + \frac{m(m+1)}{2!} \frac{(p+1)(p+2)}{(q+m)(q+m-1)} + \frac{m(m-1)(m-2)}{3!} \frac{(p+1)(p+2)(p+3)}{(q+m)(q+m-1)(q+m-2)} + \&c. \right\} \quad \text{(iv.)}$$

This is a simple hypergeometrical series, the momental properties of which have already been studied *.

* Pearson, "On Certain Properties of the Hypergeometrical Series, and on the Fitting of such Series to Observation Polygons on the Theory of Chance," Phil. Mag. Feb. 1899, pp. 236-246.

Now it is clear that the terms of this hypergeometrical series increase as long as

$$\frac{m-r+1}{r} \cdot \frac{p+r}{q+m-r+1} > 1,$$

or as long as

$$r < \frac{p(m+1)}{n}. \quad \dots \quad (\text{v.})$$

The nearest whole integer to $p(m+1)/n$ accordingly gives us the *modal* value, or maximum frequency to be expected in second trials. This result is wholly independent of the values of p , q , m , and n .

Let us write $\bar{p} = p/n$, $\bar{q} = q/n$ so that $\bar{p} + \bar{q} = 1$. Then p and q are the chances calculated from the first trial, and the modal or most probable value on the second trial is the nearest whole integer to $(m+1)\bar{p}$.

The value of the mean of the hypergeometrical series is given in Equ. (8) of the paper cited. Measured from $r=0$ we find :

$$\text{Mean} = \frac{m(p+1)}{n+2} = mp + \frac{m}{n+2}(q-p), \quad \dots \quad (\text{vi.})$$

whereas from the same origin

$$\text{Mode} = \text{greatest integer in } mp + \bar{p}. \quad \dots \quad (\text{vii.})$$

Hence for mode and mean to agree—the first condition for an approach to a Gaussian curve—it is necessary that p should not be very small, and further $m/(n+2)$ should be small. In other words, characters which occur in a very small percentage of the population, or when a relative large second sample is taken, cannot be expected to give on further sampling a *mean* simply determined by the experience of the first sample.

As illustration, let us take the following : 2 p. c. of a certain population, on the basis of a sample of 100, are found to be suffering from malaria. The *mean* value to be expected on the basis of this sample in future samples of 100 is 2.94 or nearly 3 and not 2 p. c. If the first sample were 1000, future samples of 100 might be expected to have an average percentage of 2.10 p. c. For both cases of first sampling the most probable frequency in second samples would be 2 p. c. *But it must be remembered that in dealing with the probable error, we generally take the probable error of the mean and not of the mode.* Hence in rare characteristics considerable caution must clearly be used in tacitly replacing the modes

by the means, from which they may differ, as just shown, by as much as 5 to 50 p. c. of their values.

I now pass to the standard deviation, σ , of the frequency, measuring it as usual about the mean. The value of $\mu_2 = \sigma^2$ is given in the paper referred to (p. 239, eqn. (13)) for a hypergeometrical series. We find, substituting the present values of the symbols, that :

$$\sigma^2 = \mu_2 = m \left(p + \frac{q-p}{n+2} \right) \left(q - \frac{q-p}{n+2} \right) \left(1 + \frac{m-1}{n+3} \right). \quad (\text{viii.})$$

It is clear that this can take values differing comparatively widely from the customary $\sigma^2 = m\bar{p}\bar{q}$. This latter carries on the face of it its unfitness to express the desired facts. The variability about the expected average must depend on the extent of the previous experience, but no sign of this is visible in $\sigma = \sqrt{m\bar{p}\bar{q}}$.

Suppose, for example, that the characteristic does not occur in a disproportionately large or small section of the population, then if the samples are considerable as they mostly will be in safe practice $(q-p)/(n+2)$ may be neglected and (viii.) gives :

$$\begin{aligned} \sigma^2 &= m\bar{p}\bar{q} \left(1 + \frac{m-1}{n+3} \right), \\ &= m\bar{p}\bar{q} \left(1 + \frac{m}{n} \right), \end{aligned}$$

if m and n are considerable. Hence a first sample n having given a percentage value of the characteristic p , a second sample of m individuals may be expected to give :

$$100p \pm 67.449 \sqrt{pq \left(\frac{1}{m} + \frac{1}{n} \right)} \text{ per cent.}^* \quad (\text{ix.})$$

This is the correct value, and not $100p \pm 67.449 \sqrt{\frac{pq}{m}}$ as is frequently adopted. To justify the latter the first sample must be indefinitely larger than the second, and in many cases this is not true. The first result given shows at once why a small first sample is quite useless in determining the expectancy in the case of a second large sample. The absence of the term $\frac{1}{n}$ in the second formula proves its unfitness.

* This value has been long used in my laboratory and by my statistical students.

Thus far we have seen that a sensible modification in the expected mean may be needful and most probably a still more appreciable deviation in the variability about that mean. It is necessary, however, to consider how far the distribution is really normal round the mean. Otherwise we have no justification for calculating the odds on the basis of tables of the probability integral.

The full values of the second, third and fourth moments about the mean for the hypergeometrical series are given in the paper already referred to (pp. 239-240), and with the changes requisite to suit our formula (iv.) are :

$$\mu_2 = \frac{m(p+1)(q+1)}{(n+2)^2} \left(1 + \frac{m-1}{n+3}\right), \quad \dots \dots \dots (x)$$

$$\mu_3 = \frac{m(p+1)(q+1)(q-p)}{(n+2)^3} \left(1 + \frac{m-1}{n+3}\right) \left(1 + \frac{2(m-1)}{n+4}\right), \quad \dots \dots \dots (xi)$$

$$\begin{aligned} \mu_4 = & \frac{3m(m-2)(p+1)^2(q+1)^2}{(n+2)^4} \left(1 + \frac{m-1}{n+3}\right) \left(1 + \frac{m-1}{n+4} \left(1 - \frac{8}{m-2} - \frac{9}{n+5}\right)\right) \\ & + \frac{m(p+1)(q+1)}{(n+2)^2} \left(1 + \frac{m-1}{n+3}\right) \left(1 + \frac{6(m-1)}{n+4} \left(1 + \frac{m-2}{n+5}\right)\right). \quad \dots (xii) \end{aligned}$$

In any actual case there is no difficulty in calculating μ_2 , μ_3 and μ_4 ; then, if $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$ we can find β_1 and β_2 , and these ought to be sensibly zero and 3 respectively, if the system is to have the symmetry and mesokurtosis, which characterize the Gaussian curve.

The values of β_1 and β_2 can, however, be found directly from the following formulæ :

$$\beta_1 = \frac{(q-p)^2}{m(p+1)(q+1)} \frac{\left(1 + 2\frac{m-1}{n+4}\right)^2}{1 + \frac{m-1}{n+3}}, \quad \dots \dots \dots (xiii.)$$

$$\begin{aligned} \beta_2 = & 3 \left(1 - \frac{2}{m}\right) \frac{1 + \frac{m-1}{n+4} \left(1 - \frac{8}{m-2} - \frac{9}{n+5}\right)}{1 + \frac{m-1}{n+3}} \\ & + \frac{(n+2)^2}{m(p+1)(q+1)} \frac{1 + 6\frac{m-1}{n+4} \left(1 + \frac{m-2}{n+5}\right)}{1 + \frac{m-1}{n+3}}. \quad \dots (xiv.) \end{aligned}$$

Or, if we write $\epsilon = (\bar{p} - \bar{q})/(n+2)$, we can throw β_1 and β_2 into a form which shows their approximation to the values 0 and 3 thus :

$$\beta_1 = \frac{(q - \bar{p})^2}{m(p + \epsilon)(q - \epsilon)} \frac{\left(1 - \frac{2}{n+2}\right)^2 \left(1 + \frac{2(m-1)}{n+4}\right)^2}{1 + \frac{m-1}{n+3}}, \quad \text{(xiii.)}^{bis}$$

$$\begin{aligned} \beta_2 = 3 \left(1 - \frac{2}{m}\right) \frac{1 + \frac{m-1}{n+4} \left(1 - \frac{8}{m-2} - \frac{9}{n+5}\right)}{1 + \frac{m-1}{n+3}} \\ + \frac{1}{m(p + \epsilon)(q - \epsilon)} \frac{1 + \frac{6(m-1)}{n+4} \left(1 + \frac{m-2}{n+5}\right)}{1 + \frac{m-1}{n+3}}. \quad \text{(xiv.)}^{bis} \end{aligned}$$

Hence if *both* m and n are large numbers :

$$\left. \begin{aligned} \beta_1 &= \frac{(q - \bar{p})^2}{m(p + \epsilon)(q - \epsilon)} \frac{\left(1 + 2\frac{m}{n}\right)^2}{1 + \frac{m}{n}} \\ \beta_2 &= 3 + \frac{1}{m(p + \epsilon)(q - \epsilon)} \frac{1 + 6\frac{m}{n} \left(1 + \frac{m}{n}\right)}{1 + \frac{m}{n}} \end{aligned} \right\} \quad \text{(xv.)}$$

If m be small relative to n , they become

$$\beta_1 = \frac{(q - \bar{p})^2}{m(p + \epsilon)(q - \epsilon)}, \text{ and } \beta_2 = 3 + \frac{1}{m(p + \epsilon)(q - \epsilon)}. \quad \text{(xvi.)}$$

While if n be small relative to m , we have

$$\beta_1 = \frac{4(\bar{q} - \bar{p})^2}{n(p + \epsilon)(q - \epsilon)}, \text{ and } \beta_2 = 3 + \frac{6}{n(p + \epsilon)(q - \epsilon)}. \quad \text{(xvii.)}$$

These again reduce to still simpler forms if $\epsilon = (\bar{q} - \bar{p})/(n+2)$ be not comparable with either $\bar{p}$ or $\bar{q}$, *i. e.* if neither $\bar{p}$ or $\bar{q}$ be very small. We then have for :

$$\text{(xvi.)}, \quad \beta_1 = (\bar{q} - \bar{p})^2/(m\bar{p}\bar{q}), \quad \beta_2 = 3 + 1/(m\bar{p}\bar{q});$$

$$\text{(xvii.)}, \quad \beta_1 = 4(\bar{q} - \bar{p})^2/(n\bar{p}\bar{q}), \quad \beta_2 = 3 + 6/(n\bar{p}\bar{q}).$$

Both forms result,—for n or m large and the product of either

with p and q not small,—in $\beta_1=0$ or $\beta_2=3$, *i. e.* in the symmetry and mesokurtosis, which are for practical purposes closely enough represented by the Gaussian curve. But if m and n be commensurable, and either p or q moderately small, this result by no means follows. Thus suppose we take:—

Illustration I. A sample of 100 of a population shows 10 p. c. affected with a certain disease. What percentage may be reasonably expected in a second sample of 100?

In this case $m=n=100$, $p=10$, $q=90$, and we easily find:

Mode=10. Mean=10.78. $\sigma=4.344$.

Probable error of mean=2.930.

Further: $\beta_1=.2749$, $\beta_2=3.3255$.

There is thus sensible skewness and sensible leptokurtosis.

If we suppose the probable error to still give a sufficiently definite meaning, we should put the answer in the form 10.784 ± 2.930 , or assert that the percentage of the character in the second sample was as likely to fall outside as inside the limits 7.85 to 13.71. If we used the value $m\bar{p} \pm .67449 \sqrt{m\bar{p}q}$, often adopted, we find the limits 7.98 to 12.02. The first result gives a very much more considerable chance of percentages arising in the second sample considerably beyond that shown by the first sample. It seems worth while investigating for this case, the actual errors in expectancy introduced by using the normal curve for our skew platykurtic frequency distribution. I have

accordingly calculated out series (iv.). I find $C = \frac{1}{2581}$

nearly, and obtain the distribution of frequency given in Table I. In this table the second column gives in a 1000 second samples of 100 each, the numbers which would show the percentage in the first column; the third column gives the numbers, which would show this percentage or any less percentage. This table proves at once the marked skewness of the distribution which with a modal value at 10 stretches from 0 on the defect to upwards of 30 on the excess side. If we interpolate to find the quartiles we have lower quartile at 7.65 and the upper quartile at 13.48. The median is at 10.40. Thus the lower quartile has 2.75 and the upper 3.08 for range; values in excess of the percentage observed are thus to be expected. Since $\frac{1}{2}(2.75 + 3.08) = 2.92$, we see that the "probable error" thus found is practically identical with the value found from the standard deviation. The range, however, as obtained from placing this probable error either

TABLE I.
Frequency of Percentages in a Second Sample
for Special Case.

(1) Percentage.	(2) Frequency.	(3) Sum of Frequency.	(1) Percentage.	(2) Frequency.	(3) Sum of Frequency.
0	4	4	17	29.5	927.7
1	2.2	2.6	18	22.0	949.7
2	7.1	9.7	19	16.0	965.7
3	15.9	25.6	20	11.4	977.1
4	28.9	54.5	21	7.9	985.0
5	44.8	99.3	22	5.4	990.4
6	61.3	160.6	23	3.6	994.0
7	76.1	236.7	24	2.3	996.3
8	87.0	323.7	25	1.5	997.8
9	92.8	416.5	26	.9	998.8
10	93.3	509.8	27	.6	999.3
11	89.1	598.9	28	.4	999.6
12	81.2	680.1	29	.2	999.8
13	71.0	751.1	30	.1	999.9
14	59.9	811.0	31	.0	1000
15	48.7	859.7			
16	34.5	898.2	Total ...	1000	1000

side of the mean (*i. e.* 7.85 to 13.1) is clearly too high. If we calculate the median by the rule of the third* and place the probable error either side of this theoretical median, we shall approximate closely to the actual range. The distance from mode to mean = $10.78 - 10 = .78$, and one-third of this is .26. Hence $10.78 - .26 = 10.52$ is the theoretical median. This gives 10.52 ± 2.93 , or a range of 7.59 to 13.45 as compared with the actual 7.65 to 13.48 within which half the number of second samples will fall. This is a far better result than the 7.98 to 12.02 of the usual method. The one amounts to saying that half † the second samples will fall between 7.5 and 13.5, which is sensibly in accord with actuality, while the other crowds them into the range between 8 and 12, which result is far in excess, only about 36 p. c. of the samples falling within this range.

This first illustration is, I think, sufficient to demonstrate that it is not possible in judging expectancy from past experience (*i.*) to neglect the relative sizes of the first and second samples, or (*ii.*) to neglect, even in characteristics which appear in 10 p. c. of the sample, the sensible deviation from the Gaussian distribution.

The above method of dealing with the problem is merely intended as a ready approximate method. It involves only

* Phil. Trans. vol. clxxvi. A. p. 375.

† More accurately 51.5 per cent.

the calculation of the mode, the mean and the S.D. from formulæ (vi.), (vii.), and (viii.). For more accurate work we must calculate the frequencies as I have done in Table I., a laborious process, or we can note that the values of β_1 and β_2 correspond to a skew curve of Type I. * We have in fact $\kappa_1 = 6 + 3\beta_1 - 2\beta_2 = .1737 > 0$. This leads to $r = 70.8325$, $\epsilon = 552.8582$, $m_1 = 7.93$, $m_2 = 60.90$,

and the equation to the curve is

$$y = 93.8 \left(1 + \frac{x}{12.78}\right)^{7.93} \left(1 - \frac{x}{98.13}\right)^{60.90}.$$

The origin is at 9.57 per cent. and the range from -3.2 to 107.7 per cent., extending somewhat beyond the possible limits, with, however, no sensible frequency. Considering that the raw moments have been taken to get the constants, the curve must be considered as giving very good results, and exhibiting in a manner amply sufficient for practical purposes the deviations to be expected in second samples from the Gaussian symmetry and mesokurtosis. The curves are shown in Diagram I.

Illustration II.—I proceed now to a more extreme case. Let us suppose that examination of 1000 members of a population has shown 20 affected with a certain characteristic. What is the expectancy if a second sample of 100 be examined? In this case $p=20$, $q=980$, $n=1000$, and $m=100$. We find :

$$\text{Mean} = 2.096, \quad \text{Mode} = 2, \quad \sigma = 1.5015,$$

and thus the probable error $= 1.013$.

There is not much difference between the mean and the mode, and the usual theory would give $\sigma = 1.40$, and state the answer to be expected from the second sample as $2 \pm .94$. The second sample is so much smaller than the first, that the standard deviation of second samples is fairly close to the true value when calculated by the erroneous formula. But the assumption of a Gaussian distribution falls wide of the mark. In Table II. are the calculated frequencies given by the series in (iv.) for a thousand samples †, the data being

* Phil. Trans. vol. clxxxvi. A. p. 369. This is an interesting illustration of the futility of Dr. K. Ranke's criticism of the worth of these curves on the ground that they may give negative and unintelligible values for the constants. In this particular case the n of Phil. Trans. vol. clxxxvi. A. p. 361, has become the *minus* ($n+2$) of this paper, and the pn and qn of that paper are here both negative. Yet we see that the n , p , and q of our present notation are perfectly intelligible, and that the skew curve follows at once from the series in (iv.).

† C_0 of (iv.) $= .1328$ nearly.

TABLE II.

Percentage.	Frequency.	Sum of Frequency.	Percentage.	Frequency.	Sum of Frequency.
0	132.8	132.8	6	16.8	992.0
1	258.2	391.0	7	5.7	997.7
2	260.6	651.6	8	1.7	999.4
3	181.6	833.2	9	0.5	999.9
4	98.2	931.4	10	0.1	1000.0
5	43.8	975.2	Total ...	1000.0	—

arranged as in Table I. We see at once that there is no approach to a Gaussian curve. If we proceed by simple interpolation we find the median at 1.919, the first quartile at .954, and the second at 3.042; these give a probable error in defect of .965 and in excess of 1.123, or a mean probable error of 1.044, somewhat greater than the value deduced from the standard deviation or 1.013, and considerably greater than that given by the usual theory. Here, as in the last illustration, we see a sensibly larger chance of the second sample exceeding than of its falling short of the percentage given by the long first sample. We might perhaps express this best by saying that the expectancy is 50 p. c. of samples with a percentage of the character between 0 and 1.92 and 50 p. c. with a percentage between 1.92 and 10.

To show the sort of erroneous inference that may be drawn, I suppose a percentage of 6 to have been observed in a second sample. Table II. shows that a percentage as large as this actually occurs with a frequency of 24.8 in the thousand, or its improbability is only measured by 39 to 1. On the other hand, if we adopt a Gaussian distribution we should only expect a deviation as large as 4 from the mean to occur in excess 3.5 times in the 1000 trials; or the odds are 285 to 1 against its occurrence*. Such long odds would reasonably lead us to suppose some modification in the population, since the first sample was drawn. The error of such a supposition really flows, however, from the assumption of a normal frequency distribution.

Looked at analytically we find:

$$\beta_1 = .5836, \quad \beta_2 = 3.6468;$$

or the curve is very sensibly skew and leptokurtic.

We find $\kappa_1 = .4572$, > 0 , or the curve is of Type I. We get $r = 27.0761$, $\epsilon = 87.3453$, and $m_1 = 2.7435$, $m_2 = 22.3326$;

* 475 to 1 against so large a deviation, if we adopt the old theory, for which such a deviation is 2.86 and not 2.70 times the standard deviation.

whence we deduce as equation to the curve :

$$y = 279.92 \left(1 + \frac{x}{2.5217} \right)^{274.35} \left(1 - \frac{x}{20.5272} \right)^{22.3326}$$

The origin is at 1.4308, and the range from -1.09 to +22.05. Considering that the moments have not been corrected, the fit is quite reasonable.

Diagram II. gives the curve set against the histogram of the hypergeometrical series, and the following table gives the

whole number frequencies, those for the curve being found by mechanical integration.

Percentages.	Frequencies.		Percentages.	Frequencies.	
	Hypergeo- metrical.	Curve.		Hypergeo- metrical.	Curve.
Less than .5	133	134	6	17	18
1	258	259	7	6	7
2	261	257	8	2	3
3	181	179	9	0	0
4	98	98	10	0	0
5	44	45	Sum.....	1030	1000

Conclusions.—The method often adopted for measuring the expectancy of a future sample from past experience is in certain respects defective, when (i.) the size of the future sample is at all comparable with the past experience, (ii.) the characteristic expected occurred only in a small percentage of cases in the first experience. In such cases the mean and probable error of the expectancy can be easily found, and depends upon the relative sizes of the past experience and of the future sample. The frequency of future samples is given by a certain hypergeometrical series, which is not at all closely represented by the Gaussian curve except when the past experience is very large as compared with the proposed sample, and further the characteristic expected does not occur in either a very large or very small percentage of the population.

XXVIII. *On the Radioactive Matter in the Earth and the Atmosphere.* By G. A. BLANC, *Ph.D. (Rome)* *.

IN a paper published in the Philosophical Magazine (September, 1906), A. S. EVE has given the very interesting results of the investigations made by him on the radioactive matter contained in the soil and atmosphere, concluding that the presence of radium in the earth and radium emanation (with its successive transformation products) in the atmosphere are the principal cause of the ionization which is observed in the open air and inside buildings.

Eve comes to the conclusion that a proportion of 1.8×10^{-11} gr. of radium bromide per c.cm. uniformly distributed in the

* Communicated by Lord Kelvin.