[*From* Biometrika, *Vol. XXXV. Parts III and IV, December* 1948.]

PRINTED IN GREAT BRITAIN

THE PRECISION OF OBSERVED VALUES OF SMALL FREQUENCIES

By J. B. S. HALDANE, F.R.S.

In recent genetical work numerous observers have recorded the frequencies of rare events, notably mutations. It has been realized that it is misleading to state the observed frequencies with their standard errors, since the distribution is decidedly skew. Various devices have been suggested to avoid this difficulty. But so far as I know it has not been pointed out that, when the frequency is small, its cube root is almost normally distributed. This will be proved and applied to actual observations.

Let a rare event be observed in a out of n trials, where n is much greater than a^2. Let x be the true value of the frequency, whose observed value is $p = a/n$. Let the *a priori* distribution of x be

$$dF = \phi(x)\,dx.$$

Let the probability distribution, after the observation has been made, be

$$dF = f(x)\,dx,$$

and let $x = y^3$.

Then for given values of n and x, the probability of a is

$$\binom{n}{a} x^a(1-x)^{n-a}.$$

Hence for given values of n and a, the distribution of x is

$$dF = f(x)\,dx = \frac{x^a(1-x)^{n-a}\phi(x)\,dx}{\int_0^1 x^a(1-x)^{n-a}\phi(x)\,dx}.$$

If we assume that all values of x are equiprobable, $\phi(x) = 1$, and

$$\bar{x} = \frac{(n+1)!}{a!\,(n-a)!}\int_0^1 x^{a+1}(1-x)^{n-a}\,dx = \frac{a+1}{n+2}.$$

This value should of course be a/n. As I have previously remarked (Haldane, 1932) and as Jeffreys (1948) has shown in greater detail, the assumption that $\phi(x) = 1$ introduces a bias. It is also contrary to common sense. If we are trying to estimate a mutation rate, we know *a priori* that it will almost certainly be less than 10^{-3} and greater than 10^{-20}. In a particular case we might perhaps guess that such a rate would be about as likely to lie between 10^{-5} and 10^{-6} as between 10^{-6} and 10^{-7}. In other words, when x is small it is more nearly true that all values of $\log x$ are equiprobable than that all values of x are equiprobable. This would imply that $\phi(x) = c/x$ in the region considered. However, this cannot continue to be true when x is sufficiently small. If we wished to state a plausible general form for the *a priori* distribution of x it might be somewhat as follows:

$$F = k \quad (x = 0),$$

$$dF = \frac{C}{(x+\epsilon)(1+\epsilon-x)} \quad (0 < x < 1),$$

$$F = k \quad (x = 1),$$

where k is some number less than $\frac{1}{2}$ expressing the possibility that x may prove to be zero or unity,

$$C = \frac{(1-2k)(1+2\epsilon)}{2\log(1+\epsilon^{-1})},$$

and ϵ is a very small number, perhaps of the order of 10^{-100}, expressing the fact that exceedingly rare events are relatively infrequent. If the universe is finite in space and in time, and if there is a minimum time in which an event can occur, it might imply that there is no sense in discussing events which have no appreciable probability of ever occurring.

For practical purposes, however, so long as we know that a exceeds zero, and is less than n, that is to say, that the event considered is possible and so is its converse, we can take $\phi(x) = \dfrac{C}{x(1-x)}$ without appreciable error. We then have

$$dF = \frac{(n-1)!}{(a-1)!\,(n-a-1)!}x^{a-1}(1-x)^{n-a-1}\,dx. \tag{1}$$

This is a Pearsonian Type I distribution, and

$$\overline{x^r} = \frac{(n-1)!\,(a+r-1)!}{(n+r-1)!\,(a-1)!}.$$

Thus $\bar{x} = a/n$, as it should be, $\overline{x^2} = \dfrac{a(a+1)}{n(n+1)}$, etc. When an^{-1} is small, this approximates very closely to the Type III distribution

$$dF = \frac{e^{-nx}x^{a-1}}{(a-1)!}dx. \tag{2}$$

Now Wilson & Hilferty (1931) showed that the cube root of χ^2 is almost normally distributed; and the same transformation will almost normalize many Type III distributions.

The standard form of this type, referred to its mode, is

$$dF = C\left(1+\frac{x}{a}\right)^{\gamma a}e^{-\gamma x}\,dx.$$

It is more convenient to change the origin to the point where the probability becomes zero, and write

$$dF = \frac{\gamma^c x^{c-1}\,dx}{\Gamma(c)\,e^{\gamma x}},$$

where

$$c = 1+p = 1+\gamma a = 4/\beta_1.$$

$\kappa_r = (r-1)!\,c\gamma^{-r}$, so the mean is $c\gamma^{-1}$ and the moments about it are

$$\mu_2 = c\gamma^{-2}, \quad \mu_4 = (3c^2+6c)\gamma^{-4}, \quad \mu_6 = (15c^3+130c^2+120c)\gamma^{-6}, \quad \mu_8 = (105c^4+\ldots)\gamma^{-8},$$
$$\mu_3 = 2c\gamma^{-3}, \quad \mu_5 = (20c^2+24c)\gamma^{-5}, \quad \mu_7 = (210c^3+924c^2+720c)\gamma^{-7}, \quad \mu_9 = (2520c^4+\ldots)\gamma^{-9}.$$

Let $x = c\gamma^{-1}+z$, so that $\overline{z^r} = \mu_r$ and $y = (\gamma x/c)^{\frac{1}{3}}$. Then

$$y = \left(1+\frac{\gamma z}{c}\right)^{\frac{1}{3}}$$

and

$$\overline{y^r} = 1+\tfrac{1}{3}r\frac{\gamma\bar{z}}{c}+\binom{\frac{1}{3}r}{2}\frac{\gamma^2\bar{z}^2}{c^2}+\binom{\frac{1}{3}r}{3}\frac{\gamma^3\bar{z}^3}{c^3}+\ldots$$

$$= 1+\frac{r(r-3)}{18c}+\frac{r(r-1)(r-3)(r-6)}{2(18c)^2}+\frac{r^2(r-3)^2(r-6)(r-9)}{6(18c)^3}$$
$$+\frac{r(r-3)(r-6)(r-9)(r-12)(5r^3-30r^2+15r+18)}{120(18c)^4}+O(c^{-5}).$$

Or, putting $t = \frac{1}{9c}$,

$$\begin{aligned}
\overline{y} &= 1 - t + \tfrac{10}{3}t^3 + \tfrac{11}{3}t^4 + O(t^5),\\
\overline{y^2} &= 1 - t + t^2 + \tfrac{7}{3}t^3 - \tfrac{28}{3}t^4 + O(t^5),\\
\overline{y^3} &= 1,\\
\overline{y^4} &= 1 + 2t - 3t^2 + \tfrac{10}{3}t^3 + \tfrac{41}{3}t^4 + O(t^5),\\
\overline{y^5} &= 1 + 5t - 5t^2 + \tfrac{25}{3}t^3 + \tfrac{14}{3}t^4 + O(t^5),\\
\overline{y^6} &= 1 + 9t.
\end{aligned}$$

Hence the cumulants of the distribution of y are

$$\left.\begin{aligned}
\kappa_1 &= 1 - t + \tfrac{10}{3}t^3 + \tfrac{11}{3}t^4 + O(t^5),\\
\kappa_2 &= t - \tfrac{13}{3}t^3 - 10t^4 + O(t^5),\\
\kappa_3 &= 4t^3 + 16t^4 + O(t^5),\\
\kappa_4 &= -2t^3 - 16t^4 + O(t^5),\\
\kappa_5 &= 8t^4 + O(t^5),\\
\kappa_6 &= -55t^4 + O(t^5),
\end{aligned}\right\} \qquad (3)$$

or

$$\sigma = \frac{1}{(9c)^{\frac{1}{2}}}\left[1 - \frac{13}{6(9c)^2} - \frac{5}{(9c)^3} + O(c^{-4})\right],$$

$$\gamma_1 = \frac{4}{(9c)^{\frac{1}{2}}}\left[1 + \frac{4}{9c} + O(c^{-2})\right],$$

$$\gamma_2 = \frac{-2}{9c}\left[1 + \frac{8}{9c} + O(c^{-2})\right],$$

$$\gamma_3 = \frac{8}{(9c)^{\frac{3}{2}}} + O(c^{-\frac{5}{2}}),$$

$$\gamma_4 = \frac{-55}{9c} + O(c^{-2}).$$

Thus provided 9c, or $9(1+p)$, is large, the approximation to normality, up to the sixth moment, is satisfactory. But it is of no value when p is negative, that is to say, the curve is J-shaped. (Here p is of course the parameter used in specifying Type III distributions, and not the observed frequency value.)

To apply these formulae to the distribution of y, we have only to put $a = c$, and to multiply κ_r by $p^{\frac{1}{2}r}$. We thus find

$$\left.\begin{aligned}
&\kappa_1 = p^{\frac{1}{2}}[1 - \tfrac{1}{9}a^{-1} + \tfrac{10}{2187}a^{-3} + O(a^{-4})], \quad \kappa_3 = p[\tfrac{4}{729}a^{-3} + O(a^{-4})],\\
&\kappa_2 = p^{\frac{1}{2}}[\tfrac{1}{9}a^{-1} - \tfrac{13}{2187}a^{-3} + O(a^{-4})], \quad \kappa_4 = p^{\frac{1}{2}}[\tfrac{2}{729}a^{-3} + O(a^{-4})], \quad \text{etc.}
\end{aligned}\right\} \qquad (4)$$

Thus $\sigma = p^{\frac{1}{2}}/3a^{\frac{1}{2}}$, $\gamma_1 = \frac{4}{27}a^{-\frac{3}{2}}$, $\gamma_2 = \frac{2}{9}a^{-1}$, all approximately.

The terms involving a^{-3} in the mean and standard error may be safely neglected in practice. Even when $a = 1$, the former is only 0·013 of the standard error. If we take (1) as our distribution of x, a term of order n^{-1} must be added to those of order a^{-3}. This also can be safely neglected.

Thus we find that y is almost normally distributed with mean $(1 - \frac{1}{9}a^{-1})\,p^{\frac{1}{2}}$, and standard deviation $p^{\frac{1}{2}}/3a^{\frac{1}{2}}$. For example, if $n = 1000$, $a = 8$, $p = 0{\cdot}008$, x is by no means normally distributed about 0·008, for $\beta_1 = 0{\cdot}5$ and $\beta_2 = 3{\cdot}75$. But y is very nearly normally distributed

about $0{\cdot}2 \times \frac{71}{72}$ or 0·1972 with standard deviation $\frac{1}{120}$ or 0·0083, with $\beta_1 = 0{\cdot}00004$, and $\beta_2 = 3{\cdot}028$. The method of Haldane (1938) would give an even better fit if $a > 10$.

Two examples will be given showing how the method can be actually used.

Muller (1928, p. 311) found 13 lethal genes in 1034 X-chromosomes of flies kept at 27° C., and 5 in 840 X-chromosomes of flies kept at 19·5° C. Thus corrected values of y_1 and y_2 are:—

$$y_1' = \left(\frac{13}{1034}\right)^{\frac{1}{3}}\left(\frac{1}{1-9\times 13}\right) = 0{\cdot}23054 \pm 0{\cdot}02150, \quad y_2' = 0{\cdot}17720 \pm 0{\cdot}02702.$$

$y_1' - y_2' = 0{\cdot}05334$, which is 1·55 times its standard error of 0·03452. The difference is therefore rather more significant than Muller, who used the usual formula, believed.

Again Muller (1940) obtained 7 translocations in 3366 flies with a dose of 375 r., and 56 in 2223 flies with a dose of 1500 r. The question at issue was as follows: 'the frequency may be proportional to the dosage, to its $\frac{3}{2}$th power or to its square. With which, if any, of these hypotheses are the observed results consistent?'

$$y_1' = 0{\cdot}125616 \pm 0{\cdot}016081, \quad y_2' = 0{\cdot}29256 \pm 0{\cdot}01306.$$

We therefore compare y_2' with

$$2^{\frac{1}{3}}y_1' = 0{\cdot}19940 \pm 0{\cdot}02559, \quad 2y_1' = 0{\cdot}25123 \pm 0{\cdot}03216, \quad 2^{\frac{4}{3}}y_1' = 0{\cdot}31653 \pm 0{\cdot}04052.$$

The differences are respectively 3·25, 1·19 and 0·56 times their standard errors, so either of the latter two hypotheses is admissible.

It is perhaps worth remarking that, if the emendation of the classical inverse probability distribution be rejected, and the calculation made according to Bayes's hypothesis, the cube root of the frequency is still almost normally distributed. It is also true that if the frequency of a rare event is estimated by the method described by Haldane (1945) when the observations cease when a fixed number m of rare events have occurred, the estimated frequency being $(m-1)/(n-1)$, where n is the total number of observations, the cube root of the estimate is almost normally distributed. Here too the cube root may be used with advantage in comparing different estimates.

I have to thank Prof. E. S. Pearson for valuable criticism.

Summary

When an event is rare, the distribution of the cube root of the frequency round the cube root of the estimate is much more nearly normal than the distribution of the true frequency round the estimate.

REFERENCES

Haldane, J. B. S. (1932). A note on inverse probability. *Proc. Camb. Phil. Soc.* **28**, 55–61.

Haldane, J. B. S. (1938). The approximate normalization of a class of frequency distributions. *Biometrika*, **29**, 392–404.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222–5.

Jeffreys, H. (1948). *Theory of Probability*. Oxford University Press.

Muller, H. J. (1928). The measurement of gene mutation rate in *Drosophila*, its high variability and its dependence on temperature. *Genetics*, **13**, 274–357.

Muller, H. J. (1940). An analysis of the process of structural change in chromosomes of *Drosophila*. *J. Genet.* **40**, 1–66.

Wilson, E. B. & Hilferty, M. M. (1931). The distribution of Chi-square. *Proc. Nat. Acad. Sci., Wash.*, **17**, 684–88.