Simple formula for the distortions in a Gaussian representation of a Poisson distribution

L. J. Curtis
Department of Physics and Astronomy
University of Toledo
Toledo, Ohio 43606
(Received 2 April 1975; revised 7 May 1975)

There are many examples in nature of random processes which have a constant probability of occurrence.[1] The fluctuations in the number of events arising from such a process during a fixed time interval follow the Poisson frequency distribution[2]

$$P(\mu, n) = \exp(-\mu)\mu^n/n! \,, \tag{1}$$

where μ is the average number of events obtained over many repeated trial intervals, n is the number of events which occur during a particular trial interval, and P is the fraction of repeated trial intervals which will yield a value n. It is well known that this distribution is skewed so that the median[3] and most probable[4] values lie below the mean. However, if the time interval of a trial is made sufficiently long that μ becomes a "large number," the Poisson distribution asymptotically approaches a Gaussian density function of the same mean and variance as Eq. (1), given by

$$G(\mu, n) = (2\pi\mu)^{-1/2} \exp[-(n-\mu)^2/2\mu]. \tag{2}$$

Many data analysis techniques, such as the method of least squares, the χ^2 test, the F test, the Student t test, etc., are based in part on the assumption of a normal Gaussian distribution and are applicable to a Poisson distribution only if the asymptotic limit of Eq. (2) is appropriately achieved. For example, if atomic or nuclear counting data are to be fitted to a theoretical form, the desired fit is *not* rigorously obtained by minimizing the mean square deviations from the mean, since the Poisson distribution is always skewed below the mean. When this involves a decay curve measurement, the problem is further compounded by the fact that the average number of counts detected may decrease by several orders of magnitude over the decay curve so that the degree to which it approaches the asymptotic density is not uniform. For such cases it would be valuable to develop criteria for the Gaussian asymptotic limit to the Poisson distribution in terms of μ. For this reason we have examined the relative difference between Eqs. (1) and (2) near the peak and found that it has a very simple polynomial form which is valid for essentially any value of μ. This provides a pictorial model for the errors in a Gaussian representation of a Poisson distribution, exposes the positions of maximum and minimum (zero) error which characterize the asym-

metries of the distribution, and is useful both in the treatment of experimental data and as a pedagogic device to display the basic relationships between these two important distributions.

To facilitate calculation, Eq. (1) can be conveniently rewritten by using Stirling's approximation for the factorial,

$$n! \simeq (2\pi n)^{1/2}(n/e)^n, \tag{3}$$

where the error in the approximation is less than 8% for $n = 1$ and vanishes as $1/12\,n$ for large n. Thus Eq. (1) becomes, to this approximation,

$$P(\mu, n) = (2\pi\mu)^{-1/2} \exp(n-\mu)(\mu/n)^{n+1/2}. \tag{4}$$

It is well known[5] that, in the limit of large μ, Eq. (4) approaches Eq. (2). Corrections to this approximation have been computed by Fry,[6] but direct expansion of Eq. (4) leads to tedious algebra and cumbersome expressions which are difficult to interpret. We have obtained a more tractable expansion for these corrections through consideration fo the logarithmic ratio of Eqs. (4) and (2), with n reexpressed with respect to the mean by $\delta \equiv n - \mu$, which is given by

$$\ln(G/P) = -\delta - \delta^2/2\mu + (\mu + \delta + \tfrac{1}{2})\ln(1 + \delta/\mu). \tag{5}$$

A neatly ordered expansion can be obtained if this expression is refactored into the form

$$\ln(G/P) = \tfrac{1}{2}\ln(1+\delta/\mu) + \delta[\ln(1+\delta/\mu) - \delta/\mu] + \mu[\ln(1+\delta/\mu) - \delta/\mu + \delta^2/2\mu^2]. \tag{6}$$

Expanding the logarithms on both sides of this equation for $\delta^2 < \mu^2$,

$$\left(\frac{G}{P}-1\right) - \frac{1}{2}\left(\frac{G}{P}-1\right)^2 + \cdots = \frac{\delta}{2\mu}\left[1 - \frac{1}{2}\left(\frac{\delta}{\mu}\right) + \frac{1}{3}\left(\frac{\delta}{\mu}\right)^2 - \cdots\right] + \frac{\delta^3}{6\mu^2}\left[1 - \frac{1}{2}\left(\frac{\delta}{\mu}\right) + \frac{3}{10}\left(\frac{\delta}{\mu}\right)^2 - \cdots\right]. \tag{7}$$

Retaining only the dominant terms (it is important to note that δ^2 is of order μ near the peak), we obtain

$$\frac{G-P}{P} \simeq \frac{\delta - \delta^3/3\mu}{2\mu}, \tag{8}$$

which describes the relative error with high reliability in the vicinity of the peak even for low values of μ. Notice that Eq. (8) has zeros for $\delta = 0, \pm(3\mu)^{1/2}$, which correspond to the crossings of the two curves, and has extrema for $\delta = \pm(\mu)^{1/2}$, which correspond to the maximum relative errors in the peak region (larger relative errors occur on the tails). The latter are

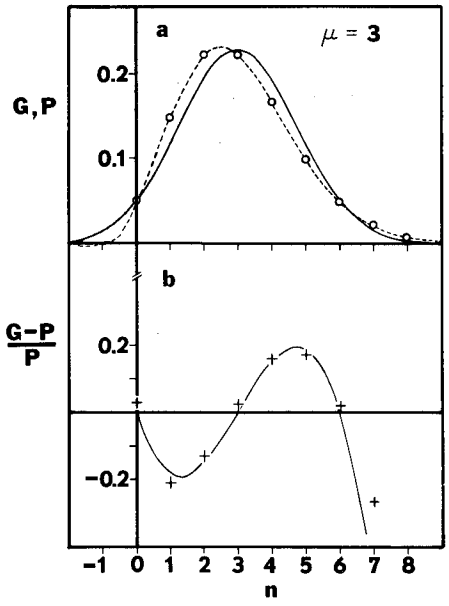

Fig. 1. (a) Gaussian (solid curve) and Poisson (circles) distributions for $\mu = 3$. The Poisson distribution is correctly a histogram, but is computed for noninteger n (dashed curve) for comparison purposes. (b) Relative error in the Gaussian distribution. Exactly calculated values (crosses) are compared with the approximation of Eq. (8) (solid curve).

$$[(G-P)/P]_{\max} = \pm(3\sqrt{\mu})^{-1}. \tag{9}$$

These five characteristic points are useful in visualizing the dependence. The expansion restriction $\delta^2 < \mu^2$ is satisfied within the peak region between the off-mean crossings ($\delta^2 < 3\mu$) for values $\mu \geqslant 3$. The exactly calculated Poisson and Gaussian curves for $\mu = 3$ are compared in Fig. 1(a), and subtracted differences relative to the Poisson curve, together with the predictions for these values given by Eq. (8), are shown in Fig. 1(b). The crossings and extrema agree well with the predictions here, and quantitative agreement occurs for even smaller values of μ (e.g., $\mu = 2$, 1, ½, etc.), although there the lower crossings occur in the nonphysical continuation of the curves into the region of negative n. For higher μ, Eq. (8) gives increasingly precise results.

The standard measures of the asymmetry of a distribution (coefficient of skewness, cumulative probability between the median and mean, etc.) all indicate that the Poisson distribution is skewed below the mean by an amount proportional to $(1/\mu)^{1/2}$, exactly as would be deduced by consideration of Eq. (9) alone. Thus it is clear that the majority of the skewness arises from the regions between the zero error crossings, which is only partially offset by contributions from the tails, and correct qualitative inferences can be drawn by consideration of this region only. It is interesting that the measures of the skewness of the Poisson distribution approach zero as $(1/\mu)^{1/2}$, just as does the ratio of the standard deviation to the mean. Thus it could be said with some justification that the Gaussian limit is approached no faster than the delta function limit, since the skewness becomes negligible only when the scatter about the mean also becomes negligible. Clearly, there are aspects of the Gaussian assumption which are not particularly sensitive to a slight skewness error, but the persistence of this Poisson skewness for large μ is not emphasized in most statistics textbooks.

The approach of the Poisson distribution to the Gaussian form can be confusing to students and researchers alike. For example, one can compare the statement of one author[7] that "the normal approximation to the Poisson distribution is quite adequate for values of $\mu \gtrsim 8$" with that of another author[8] that "the Poisson distribution is always skewed about the mean and rather more so for counts below about 100." Both of these statements are correct in a proper context, which includes a careful assessment of acceptable inaccuracies. These considerations become especially important if nonstatistical uncertainties are reduced to a small fraction of 1%, in which case small deviations from a Gaussian distribution can contribute significant errors (which are not accounted for in many standardized computer programs). The formulation presented here provides a convenient model for comparing the Poisson corrections to the Gaussian assumption with other known sources of inaccuracy.

[1]Among the more illustrious examples are the distribution in time of Prussian Cavalry soldiers killed by the kick of a horse [cf. L. von Bortkiewicz, *Das Gesetz der kleinen Zahlen* (Teubner, Leipzig, 1898)], outbreaks of war [cf. L. F. Richardson, J. R. Stat. Soc. **107,** 242 (1945); and in *The World of Mathematics,* edited by J. R. Newman (Simon and Schuster, New York, 1956), pp. 1254–1263], telephone trunk traffic [cf. E. C. Molina, *Poisson's Exponential Binomial Limit* (Van Nostrand, New York, 1942)], and atomic and nuclear decay processes [cf. P. F. Hinrichsen, Am. J. Phys. **42,** 231 (1974)]. Similar examples can also be drawn from spatial distributions, for example, the number of yeast cells visible in the field of a microscope at a given moment [cf. W. S. Gossett (pseud. "Student"), Biometrika **5,** 351 (1907)].

[2]S. D. Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile* (Bachelier, Paris, 1837). The work grew out of a study of the application of probability theory to the decisions of juries.

[3]It can be readily demonstrated numerically that the median value occurs at approximately $\mu - 1/6$, through consideration of the cumulative probability $C(\mu,N) = \sum_{n \le N} P(\mu,n)$. This provides an interesting student exercise on a small computer, and can be achieved either by choosing various values for the integer N and showing that $C = \frac{1}{2}$ when $\mu = N + \frac{1}{3}$ (since the Nth bin is centered at N, the median is then at $N + \frac{1}{2}$), or by choosing various values for μ and linearly interpolating C between integer N values to locate $C = \frac{1}{2}$.

[4]The most probable value (or mode) has been shown to be at approximately $\mu - \frac{1}{2}$ by J. R. Priest [Am. J. Phys. **38,** 658 (1970)].

[5]The "Gaussian" form of the binomial expansion was first obtained by A. De Moivre, *Approximatio ad Summam Terminorum Binomii* $(a+b)^n$ *in Seriem Expansi* (London, 1733). An English translation of this article is given in D. E. Smith's *Source Book in Mathematics* (McGraw-Hill, New York, 1929), pp. 566–575. This paper predates the work of J. Stirling and of C. F. Gauss by many years, and both Eqs. (2) and (3) should properly be credited to De Moivre. References to many of the original papers on probability are given by M. G. Bulmer, *Principles of Statistics* (MIT, Cambridge, MA, 1965).

[6]T. C. Fry, *Probability and its Engineering Uses* (Van Nostrand, New York, 1928), p. 238.

[7]B. R. Martin, *Statistics for Physicists* (Academic, London, 1971), p. 41.

[8]P. H. R. Orth, W. R. Falk, and G. Jones, Nucl. Instrum. Methods **65,** 301 (1968).