SUPPLEMENTARY MATERIAL to:

Confidence Intervals for Asbestos Fiber Counts: Approximate Negative-Binomial Distribution

David Bartley, James Slaven, and Martin Harper

Corresponding author: Address correspondence to David Bartley, 3904 Pocahontas Avenue, Cincinnati, Ohio 45227, United States
E-mail address: bartleydavy@gmail.com

Appendix A. NEGATIVE-BINOMIAL APPROXIMATION

Parameterization for counting particles

A simple and yet accurate closed-form approximation to quantiles of the negative-binomial distribution for an important range of parameters is derived here.
The probability distribution function $P[n;p,r]$ for random integer n distributed according to the negative-binomial distribution is given by:

$$P[n;p,r] = \frac{(n+r-1)!}{(r-1)!\ n!}\,(1-p)^r\,p^n\,. \tag{A1}$$

The parameters r and p may be any real numbers with $0 < p < 1$ and $r > 0$.

The mean and variance are given by:

$$\begin{aligned} Mean[n] &= pr/(1-p) \\ Variance[n] &= pr/(1-p)^2 \quad . \end{aligned} \tag{A2}$$

An alternative set of useful parameters N and s may be defined by:

$$\begin{aligned} p &= (Ns/\sigma)^2 \\ r &= s^{-2} \quad , \end{aligned} \tag{A3}$$

where σ^2 is the variance:

$$\sigma^2 \equiv Variance[n] = N + s^2 N^2\text{, and} \tag{A4a}$$

$$Mean[n] \quad = N\,. \tag{A4b}$$

The parameter *s* therefore represents a *true relative standard deviation* (relative to the mean *N*) in the limit *N* >> 1. Also, the approach to the Poisson distribution in the limit $s \to 0$ ($r \to \infty$) is manifest. Application is therefore suggested, for example, in describing counts *n* of mean *N* particles per given area, with counter variability expressed by *s*.

The range, *s* < 100%, is of practical importance. For example, Ogden, 1982, presented data on intra-lab asbestos counter variability with *s* of the order of 20%. Similarly, NIOSH 7400 assumes a value *s* ~ 45% for covering inter-lab variability.

Therefore, analysis of the negative-binomial distribution with *r* >> 1 is of interest. Also, the range with counts *n* >>1 is important. This suggests treating all the factorial arguments in equation (A1) asymptotically. By focusing on these ranges, upper single-sided quantiles may be accurately approximated for *N* as small as 1.00, and, surprisingly, lower quantiles for *N* > 3 at *s* < 60%.

Asymptotic representation of the distribution function

The approach taken here is to use Stirling's approximation for the factorials in equation (A1):

$$Log[z!] \approx z\,Log[z] - z + \tfrac{1}{2}Log[2\pi z] \quad for\ z >> 1\ . \tag{A5}$$

Sums over discrete values of *n* for defining the quantiles are approximated by integrals, adjusting end-points trapezoidally. Finally, perturbation from quantiles of the normal distribution is easily calculated.

Therefore, equation (A1) is approximated as:

$$\begin{aligned} Log[P[n;p,r]] = {} & (n+r-1)Log[n+r-1] - (n+r-1) + \tfrac{1}{2}Log[2\pi(n+r-1)] \\ & -n\,Log[n] \qquad +n \qquad -\tfrac{1}{2}Log[2\pi n] \\ & -Log[(r-1)!] \\ & +r\,Log[1-p] \ + \ n\,Log[p] \quad . \end{aligned} \tag{A6}$$

The expression $Log[P[n;p,r]]$ is now expanded as a Taylor's series about its maximum, the peak of $P[n;p,r]$ itself. The first derivative is

$$\frac{d}{dn}Log[P[n;p,r]] = Log[n+r-1] + \tfrac{1}{2}(n+r-1)^{-1} - Log[n] - \tfrac{1}{2}n^{-1} + Log[p]\ . \tag{A7}$$

Treating *n* as continuous, $P[n;p,r]$ therefore has a maximum at $n \approx n_0$ given, after collecting terms and exponentiation, by:

$$p\bullet(n_0 + r - 1) / n_0 = Exp[\tfrac{1}{2} n_0^{-1} - \tfrac{1}{2}(n_0 + r - 1)^{-1}] . \qquad \text{(A8)}$$

At fixed r, as $n_0 \to \infty$ the argument of the exponential function $\to 0$ rapidly, proportional to n_0^{-2}, and the right-hand-side of Equation (A8) correspondingly $\to 1$. Therefore in an asymptotic limit, to lowest order, n_0 can be approximated by n_{00} given by:

$$p\bullet(n_{00} + r - 1) / n_{00}] = 1 , \qquad \text{(A9)}$$

and therefore

$$\begin{aligned} n_{00} &= p(r-1)/(1-p) \\ &= N(1 - s^2) \ . \end{aligned} \qquad \text{(A10)}$$

Iteration of equation (A8) then gives n_0 approximately according to:

$$\begin{aligned} p(n_0 + r - 1)/n_0] &= 1 + \tfrac{1}{2} n_{00}^{-1} - \tfrac{1}{2}(n_{00} + r - 1)^{-1} \\ &= 1 + \tfrac{1}{2}(1-p)/n_{00} \ . \end{aligned} \qquad \text{(A11)}$$

Thus, n_0 is approximated as simply:

$$\begin{aligned} n_0 &\approx n_{00} - \tfrac{1}{2} \\ &= N(1-s^2) - \tfrac{1}{2}. \end{aligned} \qquad \text{(A12)}$$

With a view to expanding $Log[P[n;p,r]]$ as a Taylor's series about n_0, the 2nd and 3rd derivatives are:

$$\begin{aligned} \frac{d^2}{dn^2} Log[P[n;p,r]] &= (n + r - 1)^{-1} - n^{-1} - \tfrac{1}{2}(n + r - 1)^{-2} + \tfrac{1}{2} n^{-2} \\ &\approx -(1-p)/n_{00} \quad at \ \ n = n_0 \\ &\approx -\sigma^{-2} \ , \end{aligned} \qquad \text{(A13)}$$

where to leading order in s, σ^2 is again:

$$\sigma^2 \equiv N + s^2 N^2 \ , \qquad \text{(A14)}$$

i.e., the variance in equation (A4). Also, to leading order,

$$\frac{d^3}{dn^3} Log[P[n;p,r]] = -(n+r-1)^{-2} + n^{-2}$$
$$\approx (1-p^2)\, N^{-2} \;\; (n = n_0) \qquad \text{(A15)}$$
$$= (1+2Ns^2)/\sigma^4 .$$

Finally, expanding as a Taylor's series we have:

$$Log[P[n;p,r]] \approx \; constant - \tfrac{1}{2\sigma^2}(n-n_0)^2 + \tfrac{1}{6}(1+2Ns^2)\,\sigma^{-4}(n-n_0)^3 . \qquad \text{(A16)}$$

$$P[n;p,r] \approx \frac{1}{\sqrt{2\pi}\sigma} Exp[-\tfrac{1}{2\sigma^2}(n-n_0)^2] \;\; (1 \; + \; \tfrac{1}{6}(1+2Ns^2)\,\sigma^{-4}(n-n_0)^3) \;\; , \qquad \text{(A17)}$$

where exponentiation of the cubic term has been expanded as perturbation. $P[n;p,r]$ has been normalized to unity:

$$\int_{-\infty}^{+\infty} P[n;p,r] \; dn \; = 1 , \qquad \text{(A18)}$$

within the accuracy of equation (A6) at $n = n_0$, noting that the cubic term in equation (A17) does not contribute.

Comparison can now be made between both the negative-binomial distribution and its approximate equation (A17) as well as the real data reported by Ogden, 1982, by computing cumulative distributions of the function pivot[*n*,*N*] (equation (2)) as in Fig. 1 which summarizes the experimental data.

Plotted in Fig. A1 is the cumulative distribution of pivot[*n*,*N*] at s = 20%. The discrete points refer to exact computation according to equation (A1) for N equal to 5, 15, and 25. The approximation (equation (A17)) is seen in Fig. A1 as virtually a single curve for N = 5, 15, and 25.

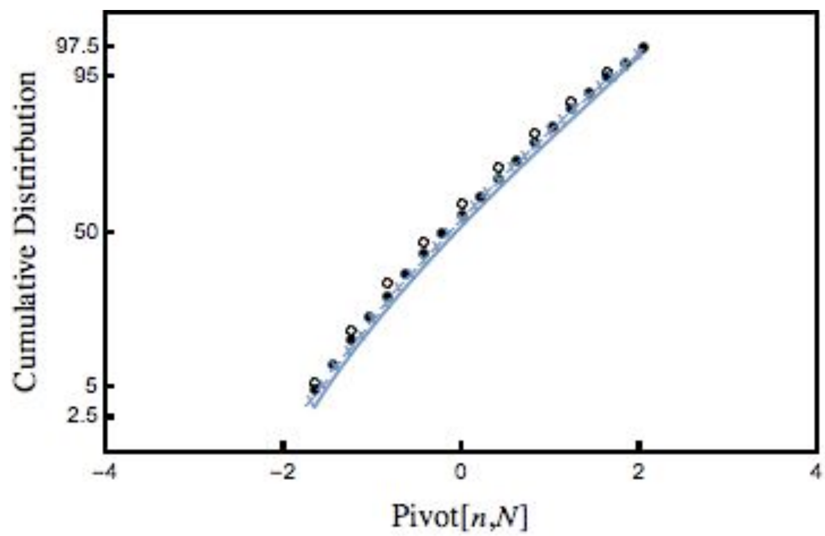

Fig. A1. Cumulative distribution of pivot[*n*,*N*] at *s* = 20%. Open circles refer to *N* = 5, closed to *N* = 15, and "x" to *N* = 25, all computed directly according to the negative-binomial distribution (equation (A1)). The solid curve was computed from the approximation (A17).

The closeness of the results from the approximate equation (A17) and the exact negative binomial distribution (1) is apparent as is also the near *N*-independence in the distribution of pivot[*n*,*N*] as assumed and found by Ogden, 1982. Also, the results are consistent with the real data of Figure 1, considering that the large tail in the experimental data is subject to large uncertainty due to the small number of data points in this region.

Quantile estimates

We are now in position for computing quantiles. The quantile at level β (e.g., 5%) is found by solving for Δ in:

$$\beta = \sum_{n=0}^{n_0+z_\beta\sigma+\Delta} P[n;p,r] \qquad (A19)$$

$$\approx \{\int_{-\infty}^{z_\beta} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z'^2}\,dz' + \tfrac{1}{2}\tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_\beta^2}\} + \Delta/\sigma \times \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_\beta^2} + \tfrac{1}{6}(1+2Ns^2)/\sigma \int_{-\infty}^{z_\beta} \tfrac{1}{\sqrt{2\pi}} z'^3 e^{-\frac{1}{2}z'^2}\,dz' \ ,$$

where z_β is the normal-distribution quantile (e.g., $z_{0.05} = -1.645$ and $z_{0.95} = +1.645$).

The term in brackets is $\sum_{n=0}^{n_0+z_\beta\sigma} \frac{1}{\sqrt{2\pi}\sigma} Exp[-\frac{1}{2}z^2]$, approximated as an integral by adding in ½ of the end point---the inverse of trapezoidal integration. The 2nd term is the first-order term in an expansion of $\int_{-\infty}^{z_\beta+\Delta/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z'^2}\, dz'$ in Δ/σ. The 3rd term is $\int_{-\infty}^{z_\beta+\Delta/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z'^2} z^3\ \frac{1}{6}(1+2Ns^2)/\sigma\, dz' \approx \int_{-\infty}^{z_\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z'^2} z^3\ \frac{1}{6}(1+2Ns^2)/\sigma\, dz'$ to lowest (0th) order in Δ/σ. Now the integral inside the bracket equals β, which therefore cancels β on the left-hand side. Secondly, the cubic integral of the 3rd term is explicitly:

$$\int_{-\infty}^{z_\beta} \frac{1}{\sqrt{2\pi}} z'^3 e^{-\frac{1}{2}z'^2}\, dz' = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_\beta^2}(2+z_\beta^2),$$

regardless of the sign of z_β. Therefore, finally, the above equation can be solved for Δ as:

$$\Delta = -\tfrac{1}{2} + \tfrac{1}{6}(1+2Ns^2)(2+z_\beta^2). \qquad \text{(A20)}$$

The quantile n_β is then:

$$n_\beta = 1 + IntegerPart[n_0 + z_\beta\sigma + \Delta], \qquad \text{(A21)}$$

$$n_0 = N(1-s^2) - \tfrac{1}{2} \quad \text{(see equation (A12) above)}$$

where the function *IntegerPart* (which truncates fractions) expresses the step-function character of the quantiles. Equation (A21) implies that continuous (averaging) quantiles may be defined as:

$$\begin{aligned} \bar{n}_\beta &= \tfrac{1}{2} + n_0 + z_\beta\sigma + \Delta \\ &= N + z_\beta\sigma + \tfrac{1}{6}(z_\beta^2 - 1)(1+2Ns^2)\,, \end{aligned} \qquad \text{(A22)}$$

as illustrated in Figures (2-4).

Poisson distribution (s = 0)

As is easily seen in Figure 5, $(\overline{n}_\beta - N)/\sigma$ is significantly non-constant for s < 20%. This reflects the fact that $(\overline{n}_\beta - N)/\sigma$ is not analytic at the point $s \to 0$ (Poisson) and $N^{-1/2} \to 0$. For dealing with the Poisson distribution, rather than using the pivot approximation, equation (12) can be solved directly for N, as setting s equal to zero results in a quadratic equation in $\sqrt{N}$. Then in this special case:

$$\sqrt{N[n_\beta, z_\beta]} = -\tfrac{1}{2} z_\beta + \tfrac{1}{2}\sqrt{\tfrac{1}{3} z_\beta^2 + \tfrac{2}{3} + 4 n_\beta}\,. \qquad \text{(A23)}$$

Confidence limits (neglecting the fluctuations (illustrated below) from the discreteness of n) on N are then given in terms of the function $N[n, z_\beta]$:

$$\Pr[N[n, z_\beta] < N < N[n, z_{1-\beta}]] = 2\beta - 1\,. \qquad \text{(A24)}$$

Equation (A24) can be compared to the more accurate traditional approximate confidence limits given by a relation of Poisson and chi-square quantiles (Johnson et al., 1993):

$$\Pr[\chi^2_{2n\ 1-\beta} < N < \chi^2_{2(n+1)\ \beta}] \approx 2\beta - 1\,. \qquad \text{(A25)}$$

In order to judge the accuracy of the above expressions when applied to discrete random variables, many simulations were carried out. At each of mean number N = 5.1, 5.1, ..., 30, a random number generator produced 10,000 Poisson-distributed values for checking the above inequalities. Results are given in Figures (A2, A3). The analogue to equation (18) is:

$$\lim[N] \approx \beta \pm \frac{1/2}{(\sqrt{N} + \tfrac{1}{3} z_\beta)\,\sqrt{2\pi}\ \ Exp[z_\beta^2/2]}\ . \qquad \text{(A26)}$$

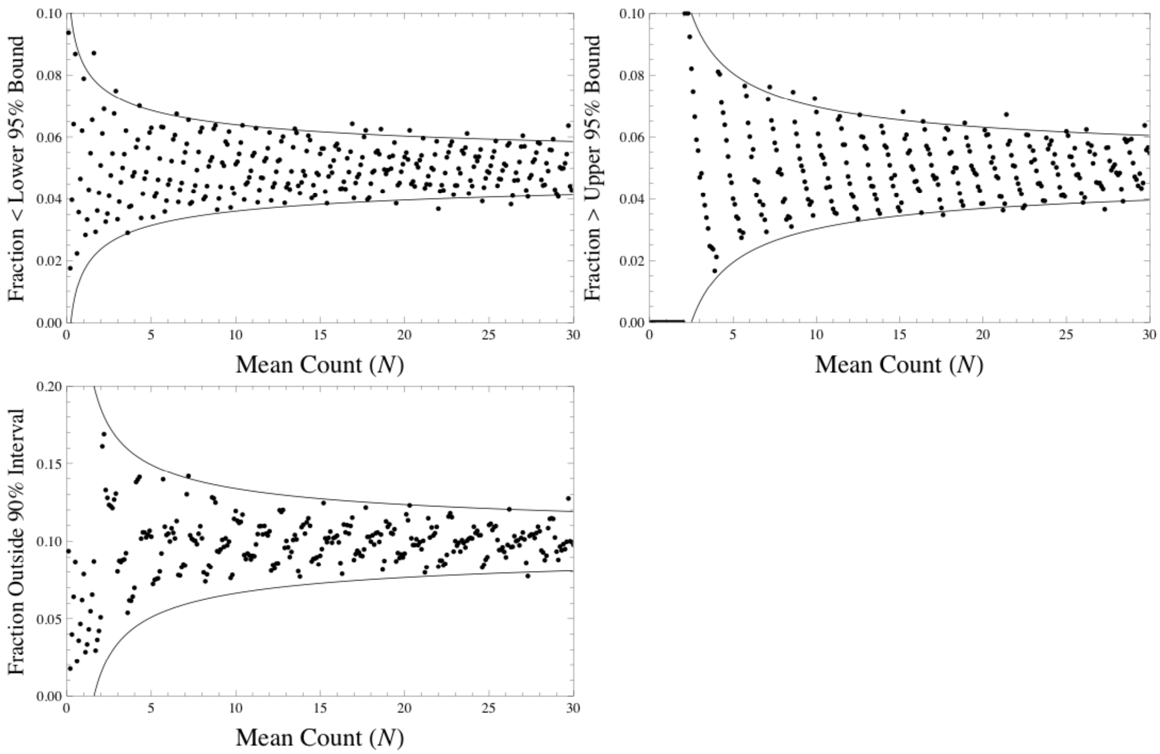

Fig. A2 Coverage probabilities with intervals computed following Stirling's approximation. Solid curves are approximate limits on the discrete variation (equation (A26)). Each point represents 10,000 simulated Poisson-distributed values.

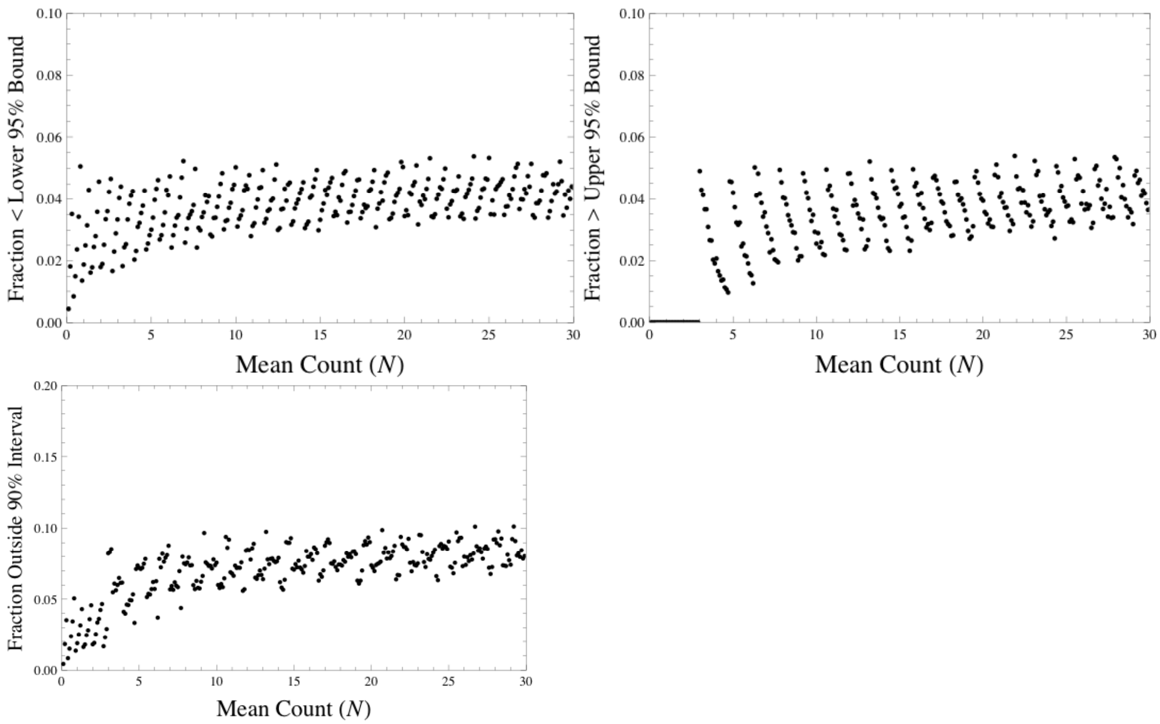

Fig. A3 Coverage probabilities with chi-square intervals. Each point represents 10,000 simulated Poisson-distributed values.

Note that the chi-square results are biased in a way that the probability of *N* falling outside the intended (5% or 10%) confidence intervals is controlled conservatively. On the contrary, the Stirling approximation leads naturally to unbiased limits controlled in the mean. Of course, since the discrete scatter is strictly limited, conservative control could be set up if needed.