DRUG DEVELOPMENT AND INDUSTRIAL PHARMACY, 18(13), 1379-1393 (1992)

ON TRUNCATED POISSON DISTRIBUTION FOR DETERMINING MEAN NUMBER OF DEFECTIVES IN PHARMACEUTICAL PRODUCTS

N. R. Bohidar

Philadelphia College of Pharmacy and Science
Villanova University
and

Nicholas R. Bohidar

Villanova University

ABSTRACT

The Poisson distribution plays a dominant role in the determination of the mean value of a distribution of the number of defective units (e.g. tablets, capsules) per sample, based on several samples of same size. If, however, the data emanates from samples with at least one defective unit in each sample, involving the absence of the zero-defective category, then the formula of the Poisson distribution as well as of the mean number of defective units are no longer tenable. In this presentation, appropriate formula for the Poisson distribution, called the truncated Poisson distribution, and for the mean, θ, are developed. The maximum likelihood method of estimation of the parameter θ by employing numerical (iterative) analysis methods is depicted, in detail. The procedure for conducting the chi-square test of goodness of fit of the experimental data to the truncated Poisson distribution is demonstrated. The results of the analyses of two recent experiments based on the methods described above are presented and appropriately interpreted.

INTRODUCTION

The Poisson distribution has been profusely used in determining the average number of defectives (1) associated with an equipment or a production process in which the number of units, such as tablets, capsules and vials, to be examined, is sufficiently large ($n > 50$) and the probability of the occurrence of a defective unit in a sample is considerably small ($p < 0.05$). For the purpose of determining the mean value of a distribution of the number of defective units per sample, the total data is grouped into several discrete categories and is presented in the following format: zero or no defective units in n_1 samples of R units each, one defective unit in n_2 samples of R units each, two defective units in n_3 samples of R units each, and so on. Consider that there are K categories, then the average number of defective units (X^*) is calculated (1) as,

$X^* = [(0xn_1) + (1xn_2) + (2xn_3) + ----(K-1)n_k/ n_1 + n_2 + n_3 + ---n_k]$, which is the appropriate formula for the mean of the Poisson distribution, symbolically expressed (1) as, $\exp(-\Theta)\Theta^X/X!$, where X= 0, 1, 2,---,

Θ = mean of the distribution (average number of defectives per sample) and $X! = X.(X-1).(X-2)..2.1$. However, this simple formula for X^* is no longer tenable if in an experiment the data for the first (zero) category is non-existent due to the specific nature of the process. Indeed, the formula for the Poisson distribution, given above, is also no longer appropriate. The truncated Poisson distribution primarily pertains to a Poisson distribution in which the zero category is absent. This happens when each sample drawn from an on-going process, has at least one defective unit present. If one uses erroneously the simple formula of X^*, given above, for estimating the mean of a truncated Poisson distribution, the value of the mean will not only be incorrect but also will yield a higher rate of defectives than the rate actually exists. This error could be serious if one is engaged in establishing specification limits and/or quality control limits for the purpose of complying with the compendial and regulatory requirements.

The primary purpose of this paper is to derive the exact distribution function for the truncated Poisson distribution, to depict, in detail, the procedure for estimating correctly the mean of the truncated Poisson distribution and to provide a numerical illustration based on the data taken from two recent experiments.

THEORY

The explicit expression of the probability distribution function (PDF) of the Poisson distribution is as (2) follows:

$f(X,\Theta) = \exp(-\Theta)\Theta^X/X!$ where $X = 0,1,2,---$. Using the above PDF, one can find the probability of obtaining a sample with no defective units to be $\exp(-\Theta)$, since in this case $X = 0$. In the same way, one can find the probability of obtaining a sample with at least one defective unit to be $[1 - \exp(-\Theta)]$, since $\Sigma\exp(-\Theta)\Theta^X/X!$, summed over $X = 0$ to $X = \infty$, is equal to one. (Note that $\Sigma\Theta^X/X! = 1 + \Theta + \Theta^2/2! + \Theta^3/3! + --- = \exp(+\Theta)$ and therefore, $\exp(-\Theta)\Sigma\Theta^X/X! = \exp(-\Theta).\exp(+\Theta) = 1$). Now, from the above derivation, one has the expression for the probability of obtaining a sample with at least one defective unit as follows:

$$[1 - \exp(-\Theta)] = \Sigma\exp(-\Theta)\Theta^X/X!$$

where, the right hand side is summed over $X=1$ to $X=\infty$. Now divide both sides of the above equation by $[1-\exp(-\Theta)]$ and obtain the equation as follows:

$$1 = [1-\exp(-\Theta)]/[1-\exp(-\Theta)] = \Sigma\exp(-\Theta)\Theta^X/X![1-\exp(-\Theta)]$$

The above equation clearly shows that the right hand expression adds to one and therefore the explicit expression of the PDF of the truncated Poisson distribution is as follows:

$$f(X,\theta) = \exp(-\theta)\theta^X/X![1 - \exp(-\theta)]$$

Maximum Likelihood Estimator of Parameter θ(2,3):

Consider that there are n observations, X_1, X_2,--- X_n from a truncated Poisson distribution. The likelihood function, by definition, is the joint probability function obtained as the product of the individual probability functions, as follows:

$$L(\theta) = \pi f(X_i,\theta) = f(X_1,\theta),\ f(X_2,\theta)\ \text{----}\ f(X_n,\theta)$$

which is explicitly,

$$L(\theta) = \exp(-n\theta)\theta^{\Sigma X}/\pi X_i![1 - \exp(-\theta)]^n$$

The log likelihood function has the following expression

$$LL(\Theta) = Log_e L = -n\Theta + \Sigma X Log_e\Theta - nLog_e[1-\exp(-\Theta)] - \Sigma Log_e X_i!$$

The interest here is to find that value of the parameter θ, as a function of the observations, which maximizes the likelihood function (or, equivalently the log likelihood function). Accordingly, one obtains the first derivative of [LL(Θ)] with respect to Θ, sets the derivative, dLL(Θ)/dΘ, equal to zero, and then solves for Θ from the resulting equation (2,3). The expression for the derivative is as follows:

$$LL'(\Theta) = dLL(\Theta)/d\Theta = -n + (\Sigma x/\Theta) - n\exp(-\theta)/[1 - \exp(-\theta)]$$

By setting the above derivative to zero, one finds that the resulting equation, unfortunately, has no explicit solution for θ. Consequently, one has to take recourse to one of the numerical analysis techniques for the solution. Since the first two derivatives of LL(Θ), LL'(Θ) and LL"(Θ) are not too complicated functions of θ, the most obvious method is the Newton-Raphson algorithm (4,5). Let θ_0 denote the initial value of Θ and h denote the correction to be applied to obtain the correct value of Θ which is expressed as $LL(\theta_0 + h) = 0$.

Expanding this by the Taylor series expansion method, one obtains,

$$LL(\theta_0 + h) = LL(\theta_0) + hLL'(\theta_0) + (h^2/2)LL''(\theta_0 + ph)$$

where, $0 < p < 1$. If h is sufficiently small, the terms containing h^2 could be ignored. Thus the equation becomes, $LL(\theta_0) + hLL'(\theta_0) = 0$. Solving for h one obtains the correction factor, $h = LL(\theta_0)/LL'(\theta_0)$. It should be noted here that, in the context of the likelihood function, however, the correction factor becomes, $h = LL'(\theta_0)/LL''(\theta_0)$, since here, one is seeking the solution of θ from the first derivative function of $LL(\theta)$. Now that the correction factor has been defined, the subsequent iterations have the following recurrence relationship,

$$\theta_1 = \theta_0 - LL'(\theta_0)/LL''(\theta_0)$$
$$\theta_2 = \theta_1 - LL'(\theta_1)/LL''(\theta_1)$$

$$\theta_{i+1} = \theta_i - LL'(\theta_i)/LL''(\theta_i)$$

The last equation depicts the general structure of the recurrence relationship. When the difference between two successive values of θ, θ_i and θ_{i+1}, is very small (.001) the process is said to have converged and the θ-value thus obtained is indeed the maximum likelihood estimate of the population θ. The explicit expression of $LL''(\theta)$ is as follows:

$$LL''(\theta) = -\Sigma X/\theta^2 + n\exp(-\theta)/[1-\exp(-\theta)]^2$$

If the initial value of θ, θ_0, is not close to the true value of θ and the value of $LL''(\theta)$, in the denominator, tends to approach zero during the iterative process, then the Newton-Raphson procedure may not yield an acceptable solution. For this reason, Fisher's moment (scoring) algorithm (2,3,6) which is a modification of the Newton-Raphson algorithm has been developed primarily for statistical applications. The basic step of the method is as follows:

$$\theta_{i+1} = \theta_i - [(dL/d\theta)/E(d^2L/d\theta^2)]$$

Note that in this algorithm, $(d^2L/d\theta^2)$ is replaced by its expected value (E). The derivation of the expected value is as follows:

$$E[LL''(\theta)] = E(-\Sigma x/\theta^2) + n\exp(-\theta)/[1-\exp(-\theta)]^2$$

now, $$E(x) = \Sigma x\exp(-\theta)\theta^x/x![1 - \exp(-\theta)]$$
$$= [\theta/(1 - \exp(-\theta))]\Sigma\exp(-\theta)\theta^{x-1}/(x-1)!$$

After a minor manipulation, it can be shown that the expected value of x is $E(x) = \theta/[1-\exp(-\theta)]$. When this value is inserted into the $LL''(\theta)$ expression, one obtains,

$$LL''(\theta)F = -n/\theta[1-\exp(-\theta)] + n\exp(-\theta)/[1-\exp(-\theta)]^2$$

Both methods will be numerically illustrated in the next section.

<u>Chi-square Test of Goodness of Fit of Truncated Poisson Distribution:</u>

After obtaining the maximum likelihood estimate of θ, θ^*, it is incumbent upon the analyst to show that the observed data is indeed from a truncated Poisson distribution. For this purpose, one must carry out a chi-square goodness of fit test for the distribution by comparing the observed frequency with the expected (theoretical) frequency for each category, as follows:

$$X_c^2 \text{ (chi-square)} = \Sigma[(O_i - E_i)^2/E_i$$

where, $i = 1,2,---k$ (k = number of categories), O_i denotes the observed frequency (number of samples) associated with the ith category, E_i denotes the expected (theoretical) frequency for the ith category, and $X^2{}_c$ has (k-2) degrees of freedom (7) to be used for obtaining the tabulated value (7) of $X^2{}_c$ for the test. Now the expected frequency of each category (i = 1,2,3,---k) is calculated in two steps: (i) computation of theoretical proportion P(x) for each category and (ii) computation of the quantity [nP(x)] for each category, which constitutes the value of the expected frequency (7). The computational formula for P(x) for each category is as follows:

$$P(x=1) = \exp(-\theta)\theta/[1-\exp(-\theta)]$$

$$P(x=2) = \exp(-\theta)\theta^2/2![1-\exp(-\theta)]$$

$$P(x=3) = \exp(-\theta)\theta^3/3![1-\exp(-\theta)]$$

$$P(x=k) = \exp(-\theta)\theta^k/k![1-\exp(-\theta)]$$

To accomplish the above computations, θ^*, the maximum likelihood estimate (the convergent value of θ) of the parameter θ of the truncated Poisson distribution, should be used. Now to obtain the expected frequency for each category, the theoretical proportion P(x) for each category must be multiplied by n (7), the total number of samples in the study.

PROCESS VALIDATION STUDY: RESULTS AND DISCUSSION

During an interlaboratory process validation investigation, it was discovered that a tablet compressing equipment (Model-M) in laboratory A (Lab-A) and a tablet compressing equipment (Model-F) in laboratory B (Lab-B) were performing unsatisfactorily with respect to production of defective tablets. The available experimental data was analyzed to estimate the mean number of defective tablets per sample for constructing appropriate control limits for each equipment. The data consisted of the number of defective tablets in each sample of 100 tablets for a total of 200 samples collected from each equipment. The samples were collected at specific intervals and were subjected to a careful inspection for visual defects. TABLE -I contains the observed data for both laboratories, organized according to the format presented in the introduction section.

The data shows that there is no zero-category (x=0) in either of the two sets of data and therefore, it is proposed to estimate the mean number of defective tablets for each laboratory by using the maximum

TABLE -I

DISTRIBUTION OF NUMBER OF SAMPLES CONTAINING ONE OR MORE DEFECTIVE TABLETS FOR LAB-A AND LAB-B*

LAB-A

X =	1	2	3	4	5
N=	120	55	20	4	1

LAB-B

X =	1	2	3	4	5	6	7
N=	58	65	43	20	8	4	2

Legend: X = number of defective tablets, N= number of samples containing exactly X number of defective tablets.

*Note that, for LAB-A, $\Sigma N = 200$ and $\Sigma XN = 311$; and for LAB-B, $\Sigma N = 200$ and $\Sigma XN = 475$

TABLE -II

MAXIMUM LIKELIHOOD ESTIMATION OF θ FOR LAB-A USING NEWTON-RAPHSON ALGORITHM

ITERATION	θ-VALUE	LL′(θ)	LL″(θ)	LL(θ)
0	1.555	-53.546283	-60.735058	-126.2553
1	0.673363	53.705446	-261.102978	-114.9979
2	0.879049	11.807175	-159.688805	-108.6128
3	0.952988	0.827907	-138.157829	-108.1555
4	0.958981	0.004603	-136.626240	-108.1530
5*	0.959015	0.000000	-136.617840	-108.1530
6**	0.959015	0.000000	-136.617840	-108.1530

LEGEND: LL(θ) = Log Likelihood Function, LL′(θ) = First Derivative of LL(θ), LL″(θ) = Second Derivative of LL(θ), * = The θ-value for the 5th iteration is 0.959014519 without rounding and $LL'(\theta_5) = 8.3 \times 10^{-8}$ which is virtually zero, ** = The θ-value for this iteration is the convergent θ-value and the ML estimate.

likelihood (ML) iterative method as described in the previous section.

LAB-A(MODEL-M): The detailed results of the analysis are presented in TABLE-II. The first thing to notice is that the initial value of θ, θ_0, is taken as 1.555 which is calculated by using $\Sigma XN/\Sigma N$ (=311/200 = 1.555) to get the iterative process started. Then, using the value of ΣXN and of ΣN (see TABLE-I), one can calculate $LL'(\theta)$, $LL''(\theta)$ and $L(\theta)$ for each iteration based on the θ-value of the previous iteration. For instance, θ_1 is obtained from θ_0 by computing 1.555 - (-53.546283/-60.735058) = 0.673363(=θ_1) and now this value of $\theta(\theta_1)$ is utilized to generate the required quantities for the second iteration and so on. The convergence criterion considered here is 0.00001 which implies that when the absolute difference $(\theta_{i+1} - \theta_i)$ is equal to or less than 0.00001, then the iterative process comes to an end and the value of θ, θ^*, thus obtained at the last iteration becomes the appropriate ML estimate of the mean number of defectives, which is for LAB-A (See TABLE-II) is equal to 0.9590 or to 1.0 (when rounded to the nearest integer). It should be pointed out here that, if one would have used X* formula instead, the value of the mean would be 1.6 ≈ 2.0 (when rounded to the nearest integer) showing wrongly that the mean is twice as large as the mean number of defectives actually exists in the sample. This clearly demonstrates that the ML estimation procedure is indispensible for attaining the correct results. The contents of TABLE-II clearly show that the iterative process does indeed meet all the function maximization criteria at the convergent value of θ, since (i) $LL'(\theta^*)$ is very close to zero (8.3×10^{-8}), (ii) $LL''(\theta^*)$ is negative (-136.6262) and (iii) $L(\theta^*)$ does attain a maximum at -108.1530, starting at -126.2553.

TABLE -III

MAXIMUM LIKELIHOOD ESTIMATION OF θ FOR LAB-A USING FISHER'S SCORING ALGORITHM

ITERATION	θ-VALUE	LL'(θ)	LL"(θ)F	LL(θ)
0	1.555	-53.546283	-95.169967	-126.2553
1	0.992362	-4.419716	-133.046115	-108.2275
2	0.959142	-0.017430	-136.603582	-108.1530
3	0.959015	-0.000000	-136.617840	-108.1530

LEGEND: LL(θ) = Log Likelihood Function, LL'(θ) = First Derivative of LL(θ), LL"(θ)F = Expected Value (E) of LL"(θ), the Second Derivative of LL(θ)

TABLE -IV

CHI-SQUARE TEST OF GOODNESS OF FIT FOR TRUNCATED POISSON DISTRIBUTION ASSOCIATED WITH LAB-A

X	O	P(x)	E	$(O-E)^2/E$
1	120	0.59599	119.1971	0.005408
2	55	0.28578	57.1559	0.081320
3	20	0.09136	18.2711	0.163597
4	4	0.02190	4.3806	0.033068
5	1	0.00420	0.8402	0.030393
TOTAL	200	0.99923	199.8449	0.312458

LEGEND: X = Number of defective tablets, O = Observed number of samples containing exactly x number of defective tablets, P(x) = Theoretical proportion of the total number of samples associated with category x, E = Expected (theoretical) number of samples containing exactly x number of defective tablets, and the sum of the entries of the last column, $\Sigma(O-E)^2/E$, is the test statistic for the chi-square test of goodness of fit.

TABLE-III depicts the results of the iterative process associated with Fisher's Moment (Scoring) Algorithm for LAB-A. It is demonstrated that there is a considerable reduction in the number of iterations with the same end results as shown in TABLE-II. It is recommended that both methods be carried out concurrently, for confirmation and accuracy. The detailed results of the test of goodness of fit for the truncated Poisson distribution are depicted in TABLE-IV. A cursory examination of the contents of the table indicates that the magnitudes of the theoretical expected values (E) are very close to that of the experimentally observed value (O) confirming the fact that the process indeed follows a truncated Poisson distribution. (Note that this would not be the case if an ordinary Poisson distribution would have been applied).

Since the calculated X^2_C value of 7.81 for 3 (= 5-2) degrees of freedom at a 5% level of significance, the test statistic is not significant ($p > 0.05$), implying the fact that the distribution is indeed a truncated Poisson distribution, and the formulations and computations associated with the analysis are absolutely appropriate. A by-product of this analysis is to notice that 59.6%, 28.6% and 9.1% of the total number of samples (=200) pertain to the first three categories, respectively.

LAB-B (Model -F): The experimental data for LAB-B is presented in TABLE I. The contents of TABLE-V pertain to the statistical results of the ML iterative estimation method using the Newton-Raphson algorithm. The initial value of θ, θ_O (=2.375) is calculated by using the formula, $\Sigma XN/\Sigma N$, (=475/200). The successive iterations are generated by the same procedure given above. The convergent value of θ, θ^* is 2.0 (rounded) which represents the mean number of defective units for

TABLE - V

MAXIMUM LIKELIHOOD ESTIMATION OF Θ FOR LAB-B USING NEWTON-RAPHSON ALGORITHM

ITERATION	Θ-VALUE	LL'(Θ)	LL"(Θ)	LL(Θ)
0	2.375	-20.510689	-61.596395	-44.6005
1	2.042015	2.789893	-79.642957	-41.4840
2	2.077045	0.040491	-77.349824	-41.4347
3	2.077568	0.000009	-77.316395	-41.4347
4*	2.077568	0.000000	-77.316395	-41.4347

LEGEND: LL(Θ) = Log Likelihood Function, LL'(Θ) = First Derivative of LL(Θ), LL"(Θ) = Second Derivative of LL(Θ), * = The convergent value of Θ

TABLE - VI

MAXIMUM LIKELIHOOD ESTIMATION OF Θ FOR LAB-B USING FISHER'S SCORING ALGORITHM

ITERATION	Θ-VALUE	LL'(Θ)	LL"(Θ)F	LL(Θ)
0	2.375	-20.510689	-70.232475	-44.6005
1	2.082960	-0.415938	-77.173170	-41.4358
2	2.077570	-0.000154	-77.316335	-41.4347
3*	2.077568	-0.000000	-77.316395	-41.4347

LEGEND: LL(Θ) = Log Likelihood Function, LL'(Θ) = First Derivative of LL(Θ), LL"(Θ)F = Expected Value (E) of LL"(Θ), the Second Derivative of LL(Θ), and * = convergent value of Θ

TABLE -VII

CHI-SQUARE TEST OF GOODNESS OF FIT FOR TRUNCATED POISSON DISTRIBUTION ASSOCIATED WITH LAB-B

X	O	P(x)	E	$(O-E)^2/E$
1	58	0.29743	59.486	0.03712
2	65	0.30897	61.793	0.16644
3	43	0.21397	42.793	0.00100
4	20	0.11113	22.227	0.22305
5	8	0.04618	9.235	0.16526
6	4	0.01599	3.198	0.20124
7	2	0.00475	0.949	1.16340
TOTAL	200	0.99842	199.681	1.95751

LEGEND: X = Number of defective tablets, O = Observed number of samples containing exactly x number of defective tablets, P(x) = Theoretical proportion of the total number of samples associated with category x, E = Expected (theoretical) number of samples containing exactly x number of defective tablets, and the sum of the entries of the last column, $\Sigma(O-E)^2/E$, is the test statistic for the Chi-square test of goodness of fit.

LAB-B. The iterative results do indeed confirm to the maximization criteria described in conjunction with the contents of TABLE-II. TABLE-VI depicts the results of the iterative process using Fisher's scoring algorithm. The end results are the same as given in TABLE-V, involving less number of iterations. The results of the chi-square test of goodness of fit are presented in TABLE-VII.

It is found that the theoretical expected values (E) are very close to their experimentally observed counterparts (O), substantiating the fact that the process indeed follows a truncated Poisson distribution. The calculated X^2c value of 1.96 is less than the tabulated X^2c value (7) of 11.7 for 5 (=7-2) degrees of freedom at a 5% level of significance indicating that the distribution is indeed a truncated Poisson distribution ($P > 0.05$). It is interesting to note that 29.7%, 30.9%, 21.4% and 11.1% of the total number of samples (=200) pertain respectively to the first four categories of the distribution.

It should be noted here that the standard deviation associated with the mean number of defective units is the square-root of the convergent θ^*-value to be used in establishing the appropriate control limits.

(Note: All calculations in this study are accomplished by using a pocket calculator (TI-60, Texas Instrument) with ten data memory locations. Note also that for each entry, several decimal places are displayed in the tables to facilitate reproduction of results with substantial numerical precision.)

ACKNOWLEDGEMENTS

Grateful acknowledgement is due to Mr. F.J Mark of Merck and Co. for providing timely computational assistance using SAS™. Special thanks are due to Mrs. B.J. Tomlinson for accomplishing the task of typing the manuscript with great efficiency.

REFERENCES

1. E.L. Grant and R.S. Leavenworth, "Statistical Quality Control", 5th ed. Chapter 6, pg. 203-205. McGraw-Hill Co. New York (1980).

2. A.M. Mood, "Introduction to Theory of Statistics," 1st ed. Chapter 3, Chapter 10, pgs. 208-209. McGraw-Hill Co. New York (1950).

3. R.L. Anderson and T.A. Bancroft, "Statistical Theory in Research", Chapter 8, pg. 96-97 McGraw-Hill Co. New York (1952).

4. N.R. Bohidar, Estimation of an Apparent Half-Life in the Presence of Extensive Enterohepatic Circulation. Proceedings of American Statistical Association, Biopharmaceutical Section, pg. 91-97 (1986).

5. R.H. Pennington, "Introductory Computer Methods and Numerical Analysis". 2nd ed. Chapter 8, pg. 286-287. The Macmillan Company, New York (1970).

6. S.Y Lee and R.I. Jenrich, A Study of Algorithms for Covariance Structure Analysis with Specific Comparisons Using Factor Analysis. Psychometrika, Vol. 44, pg. 99-113 (1979).

7. G.W. Snedecor and W.G. Cochran, "Statistical Methods" 6th ed. Chapter 7, pg. 130-133, 470. Iowa State University Press, Ames, Iowa (1980).