On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin

Author(s): John Aitchison

Source: *Journal of the American Statistical Association*, Vol. 50, No. 271 (Sep., 1955), pp. 901-908

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: http://www.jstor.org/stable/2281175

Accessed: 14-02-2018 23:27 UTC

ON THE DISTRIBUTION OF A POSITIVE RANDOM VARIABLE HAVING A DISCRETE PROBABILITY MASS AT THE ORIGIN*

JOHN AITCHISON
University of Cambridge

In a number of situations we are faced with the problem of determining efficient estimates of the mean and variance of a distribution specified by (i) a non-zero probability that the variable assumes a zero value, together with (ii) a conditional distribution for the positive values of the variable. This estimation problem is analyzed and its implications for the Pearson type III, exponential, lognormal and Poisson series conditional distributions are investigated. Two simple examples are given.

1. THE PROBLEM

THE nature of the problem to be discussed is best introduced by examples. In a study of household expenditures it is often of interest to estimate, from a sample of household budgets, the mean expenditure per household on a certain commodity, say children's clothing. Over the period of the investigation it may well happen that a number of households in the sample spend nothing on children's clothing whereas the expenditures by the remainder of the households necessarily arise from the distribution of a positive variable, probably skew and possibly approximated by a lognormal curve. If such is the case, then clearly the correct procedure in any analysis is to recognize explicitly this dichotomy of the population into the categories, spender and non-spender. This type of situation is not, however, confined to the case of a continuous variable; it occurs also for discrete variables. For example, in a household composition study we may wish to investigate the distribution of the number of children in a household. This distribution is sometimes Poisson except that the number of households with no children is considerably larger than is suggested by Poisson theory. Again one solution of the difficulty is to assume that there is a proportion of households containing no children while the remainder is distributed as a truncated Poisson distribution.

Such problems lead us to consider a random variable X with the following properties. There is a non-zero probability θ that X is zero

* This paper is a development of some of the estimation problems discussed by Utting and Cole [5]. The author wishes to express his indebtedness to J. A. C. Brown of the Department of Applied Economics for helpful criticism and for suggesting the application of the Poisson series distribution to the analysis of household composition.

and hence a probability $1-\theta$, that X is non-zero; further the distribution of X conditional on $X \neq 0$ is some well-known distribution of a positive variable, either continuous or discrete. This we may write:

$$P\{X = 0\} = \theta, \tag{1}$$

$$P\{X > 0\} = 1 - \theta, \tag{2}$$

and, for the continuous case.

$$P\{X \subset (x, x + dx) \mid x > 0\} = g(x)dx, \tag{3}$$

where $g(x)$ is the conditional frequency function; and so

$$P\{X \subset (x, x + dx)\} = (1 - \theta)g(x)dx, \; x > 0. \tag{4}$$

If α and β are the mean and variance respectively of the $g(x)$ distribution and γ and δ are the corresponding parameters of X then

$$\gamma = (1 - \theta)\alpha \tag{5}$$

and

$$\delta = (1 - \theta)\beta + \theta(1 - \theta)\alpha^2. \tag{6}$$

We discuss in this paper the problem of efficient estimation of γ and δ.

2. EFFICIENT ESTIMATION

In this section we state some general results proved in the Appendix which allow us, under certain circumstances, to obtain best unbiased estimators of γ and δ; by the term *best unbiased estimator* we mean an unbiased estimator having minimum attainable variance (see, for example, Rao [3]). Let us suppose that, for the purpose of estimation, we have available a random sample S of size n from the population and that r of the sample values are zero while the remaining $(n-r)$ are $x_1, x_2, \cdots, x_{n-r}$. Then the following results hold.

(i) If, for a sample of size m from the $g(x)$ population a sufficient unbiased estimator of α, say $a_{(m)}$, exists then

$$c = \left(1 - \frac{r}{n}\right) a_{(n-r)}, \; r < n,$$

$$= 0, \; r = n, \tag{7}$$

is a best unbiased estimator of γ.

The twofold definition of c is necessary since $a_{(n-r)}$ is not defined for

$r=n$. If $a_{(m)}$ is the arithmetic mean of m sample values then c becomes the mean of the sample S (including zero values), namely

$$c = \frac{1}{n}\sum_{i=1}^{n-r} x_i \tag{8}$$

and the variance of the estimator in this case is

$$\text{var}\{c\} = \frac{\delta}{n}. \tag{9}$$

A result similar to (i) holds for δ provided that jointly sufficient estimators of α (and hence of α^2) and β exist.

(ii) If $e_{(m)}$ and $f_{(m)}$ are jointly sufficient unbiased estimtaors of α^2 and β respectively for a sample of size m, then

$$d = \left(1 - \frac{r}{n}\right) f_{(n-r)} + \frac{r}{n}\left(1 - \frac{r-1}{n-1}\right) e_{(n-r)},\ r < n,$$
$$= 0,\ r = n, \tag{10}$$

is a best unbiased estimator of δ.

It is seldom that such jointly efficient estimators of α and β occur. Often, however, β depends on α so that $a_{(m)}$ is sufficient for both α and β; an important case is $\beta = K\alpha^2$, for which we have the following property.

(iii) If $\beta = K\alpha^2$ and $a_{(m)}$, the sufficient unbiased estimator of α, is the sample mean then

$$\delta = (1 - \theta)(K + \theta)\alpha^2 \tag{11}$$

and

$$d = \left\{K + (1-K)\ \frac{r}{n} - \frac{r(r-1)}{n(n-1)}\right\} a_{(n-r)}{}^2 \Big/ \left\{1 + \frac{K}{n-r}\right\},\ r < n,$$
$$= 0,\ r = n, \tag{12}$$

is a best unbiased estimator of δ.

3. APPLICATION TO PARTICULAR DISTRIBUTIONS WITH EXAMPLES

It is interesting to apply the estimator procedure of the preceding section to a number of particular conditional distributions.

(i) *Pearson type III distribution*

For the Pearson type III distribution

$$g(x) = \left(\frac{p}{\alpha}\right)^p \frac{x^{p-1}e^{-px/\alpha}}{\Gamma(p)}, \quad x > 0, \tag{13}$$

where we assume p to be known. For the mean α of this distribution the sample mean is a sufficient unbiased estimator so that

$$c = \frac{1}{n}\sum_{i=1}^{n-r} x_i \tag{14}$$

is a best unbiased estimator of $\gamma = (1-\theta)\alpha$. Here $\beta = \alpha^2/p$ and so

$$\delta = (1-\theta)\left(\frac{1}{p} + \theta\right)\alpha^2, \tag{15}$$

and a best unbiased estimator of δ is given by (12) with $K = 1/p$.

(ii) *Exponential distribution*

For the exponential distribution,

$$g(x) = \frac{1}{\alpha} e^{-x/\alpha}, \; x > 0, \tag{16}$$

and this is the special case $p = 1$ of the Pearson type III distribution so that no new theory arises. The estimator of δ, however, simplifies to

$$d = \frac{1 + \dfrac{r-1}{n}}{1 - \dfrac{r-1}{n}} \frac{\left\{\sum_{i=1}^{n-r} x_i\right\}^2}{n(n-1)}, \; r < n,$$

$$= 0, \; r = n. \tag{17}$$

(iii) *Lognormal distribution*

If the conditional distribution is lognormal with parameters μ and σ^2 then

$$g(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\}, \; x > 0. \tag{18}$$

Here

$$\alpha = e^{\mu + \frac{1}{2}\sigma}, \tag{19}$$

so that

$$\beta = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1), \tag{20}$$

so that

$$\gamma = (1-\theta)e^{\mu+\frac{1}{2}\sigma^2}, \tag{21}$$

and

$$\delta = (1-\theta)\cdot e^{2\mu+\sigma^2}\left\{e^{\sigma^2} - (1-\theta)\right\}. \tag{22}$$

Finney [2] has obtained by considerations similar to those of this paper best unbiased estimators of α and β, and as an extension of his theory it may be shown that

$$c = \left(1 - \frac{r}{n}\right) e^{\bar{y}} \psi_{(n-r)}(\tfrac{1}{2}s^2), \quad r < n-1,$$
$$= \frac{x_1}{n}, \quad r = n-1,$$
$$= 0, \quad r = n, \tag{23}$$

and

$$d = \left(1 - \frac{r}{n}\right) e^{2\bar{y}} \left\{ \psi_{(n-r)}(2s^2) - \left(1 - \frac{r}{n-1}\right) \psi_{(n-r)}\left(\frac{n-r-2}{n-r-1} s^2\right) \right\}, \quad r < n-1,$$
$$= \frac{{x_1}^2}{n}, \quad r = n-1,$$
$$= 0, \quad r = n, \tag{24}$$

are best unbiased estimators of γ and δ respectively, where

$$y_i = \log x_i, \quad i = 1, 2, \cdots, n-r, \tag{25}$$

$$\bar{y} = \frac{1}{n-r} \sum_{i=1}^{n-r} y_i, \quad r < n-1, \tag{26}$$

$$s^2 = \frac{1}{n-r-1} \sum_{i=1}^{n-r} (y_i - \bar{y})^2, \quad r < n-1, \tag{27}$$

and[1]

$$\psi_m(t) = 1 + \frac{m-1}{m} t + \frac{(m-1)^3}{2!m^2(m+1)} t^2 + \frac{(m-1)^5}{3!m^3(m+1)(m+3)} t^3 + \cdots. \tag{28}$$

It can be shown that $1/n \sum_{i=1}^{n-r} x_i$ is less efficient than c in this case.

Example.[2] In a household expenditure inquiry carried out by the Ministry of Food in 1950, a sample of 1143 British households was

[1] A table of values of the function $\psi_m(t)$ will be published in a forthcoming monograph on the log-normal distribution by the Department of Applied Economics, University of Cambridge.

[2] The data for this and the following example have been obtained from The National Food Survey by courtesy of the Ministry of Food.

taken; one of the items in the classification gives expenditures in pence per household on sweet biscuits. Of the 1143 households only 512 bought this commodity and their expenditures appear to come from a lognormal population. The relevant sample data are:

$$n = 1143 \qquad \bar{y} = 3.664$$
$$r = 631 \qquad s^2 = 0.3721$$

The estimator c of γ obtained from (23) is then

$$c = \left(1 - \frac{631}{1143}\right) e^{3.664} \psi_{(512)}(0.1860)$$
$$= 20.95,$$

as compared with the value of 20.25 given by the ordinary sample mean. The value of d in this case is 4481 so that the estimated standard error of the sample mean is 1.97 and the standard error of c is necessarily less than this.

(iv) *Truncated Poisson distribution*

The truncated Poisson distribution[3] provides an application of the theory to a discrete variable. In this case

$$g(x) = P\{X = x \mid x > 0\}$$
$$= \frac{e^{-\mu}}{1 - e^{-\mu}} \frac{\mu^x}{x!}, \quad x = 1, 2, \cdots \tag{29}$$

and

$$\alpha = \frac{\mu}{1 - e^{-\mu}}. \tag{30}$$

The sample mean is again a sufficient unbiased estimator[4] of α and so c as given by (8) is a best unbiased estimator of γ where

$$\gamma = \frac{(1 - \theta)\mu}{1 - e^{-\mu}}. \tag{31}$$

It does not seem possible to find a simple expression for the estimator d in this case.

---

[3] See, for example, David and Johnson [1].

[4] See Tukey [4].

Example. The data of Table 1 analyze by number of children (under fourteen years of age) per household a sample of 4021 British households in 1950.

TABLE 1

NUMBER OF HOUSEHOLDS CONTAINING GIVEN NUMBER OF CHILDREN

No. of Children	0	1	2	3	4	5	6	7	8	9
(i) Observed	2303	831	565	212	67	23	15	3	1	1
(ii) Poisson	1856	1435	554	143	28	4	1	—	—	—
(iii) Truncated Poisson	2303	822	546	242	81	21	5	1	—	—

Rows (ii) and (iii) of the table show the results of fitting a complete Poisson distribution (estimated $\mu=0.773$) and a truncated Poisson distribution (estimated $\mu=1.33$) as described by (29). The complete Poisson distribution clearly does not give an adequate fit due to the extra large proportion of households with no children; the truncated distribution gives a better approximation and a best unbiased estimate of the mean number of children per household is simply the sample mean 0.773.

4. FURTHER CONSIDERATIONS

The limitations that the discrete probability mass is at the origin rather than at some other point, and that the conditional variable is essentially positive, may be removed without unduly complicating the theory; we have not thought it worth-while to develop this extension because of the lack of any obvious practical application. It would also have been interesting to compare the efficiency of other possible estimators with the estimators derived for the special distributions but this particular problem has also been left aside.

APPENDIX: DERIVATION OF THEORETICAL RESULTS

In this Appendix we derive the theoretical results of Section 2. The essential idea underlying the proofs is a property of jointly sufficient estimators: any function of jointly sufficient estimators is a best unbiased estimator of its expectation (cf. Rao [3, p. 149]).

As in Section 2 the random sample S consists of r zero values and $(n-r)$ other values $x_1, \cdots, x_{n-r}$. For the $g(x)$ population, $a_{(m)}$ is a sufficient unbiased

estimator of α for a sample of m values. If the distribution of X depends on parameters $\lambda, \cdots$ in addition to θ and α, then the likelihood function L of the sample may be written in the form

$$L(S \mid \theta, \alpha, \lambda, \cdots) = \binom{n}{r} \theta^r (1 - \theta)^{n-r} h(a_{(n-r)}, \alpha, \lambda, \cdots) k(S \mid \lambda, \cdots)$$

where h is a frequency function containing the sample values only in the form $a_{(n-r)}$, and k is a frequency function independent of θ and α. Hence

$$L = L_1\left(\frac{r}{n}, a_{(n-r)}, \theta, \alpha, \lambda \cdots\right) L_2(S \mid \lambda, \cdots)$$

where L_1 is a frequency function containing the sample values only in the forms r/n and $a_{(n-r)}$, and L_2 is a frequency function independent of θ and α. Thus r/n and $a_{(n-r)}$ are jointly sufficient estimators of θ and α. Consider the sample function:

$$\begin{aligned} c &= \left(1 - \frac{r}{n}\right) a_{(n-r)}, \quad r < n \\ &= O, \quad r = n. \end{aligned}$$

The expectation of the estimator is given by

$$\begin{aligned} E\{c\} &= P\{r = n\} E|_{r=n}\{c\} + P\{r < n\} E|_{r<n}\{c\} \\ &= P\{r < n\} E_{r|r<n}\left\{\left(1 - \frac{r}{n}\right) E|_{r=\text{const}<n}[a_{(n-r)}]\right\} \\ &= P\{r < n\} E_{r|r<n}\left\{\left(1 - \frac{r}{n}\right)\alpha\right\} \\ &= \alpha E\left(1 - \frac{r}{n}\right) \\ &= (1 - \theta)\alpha \\ &= \gamma \end{aligned}$$

so that, from the property of sufficient estimators referred to above, c is a best unbiased estimator of γ.

The proofs of the results (ii) and (iii) of Section 2 proceed in exactly the same manner and their details need not be reproduced here.

REFERENCES

[1] David, F. N., and Johnson, N. L., "The truncated Poisson," *Biometrics*, 8 (1952), 275–85.

[2] Finney, D. J., "On the distribution of a variate whose logarithm is normally distributed," *Journal of the Royal Statistical Society*, Series B, 7 (1941), 155–61.

[3] Rao, C. R., *Advanced Statistical Methods in Biometric Research*. New York: John Wiley and Sons, Inc., 1952.

[4] Tukey, John W., "Sufficiency, truncation and selection," *Annals of Mathematical Statistics*, 20 (1949), 309–11.

[5] Utting, J. E. G., and Cole, Dorothy, "Sample surveys for the social accounts of the household sector," *Bulletin of the Oxford Institute of Statistics*, 15 (1953), 1–24.