ESTIMATION OF PARAMETERS IN TRUNCATED PEARSON FREQUENCY DISTRIBUTIONS

By A. C. Cohen, Jr.

The University of Georgia

1. Introduction and summary. A method based on higher moments is presented in this paper by which the type of a univariate Pearson frequency distribution (population) can be determined and its parameters estimated from truncated samples with known points of truncation and an unknown number of missing observations. Estimating equations applicable to the four-parameter distributions involve the first six moments of a doubly truncated sample or the first five moments of a singly truncated sample. When the number of parameters to be estimated is reduced, there is a corresponding reduction in the order of the sample moments required. A sample is described as singly or doubly truncated according to whether one or both "tails" are missing. Estimates obtained by the method of this paper enjoy the property of being consistent and they are relatively simple to calculate in practice. They should be satisfactory for (a) rough estimation, (b) graduation, and (c) first approximations on which to base iterations to maximum likelihood estimates.

Previous investigations of truncated univariate distributions include studies of truncated normal distributions by Pearson and Lee [1], [2], Fisher [3], Stevens [4], Cochran [5], Ipsen [6], Hald [7], and this writer [8], [9]. In addition, the truncated binomial distribution has been studied by Finney [10], and the truncated Type III distribution by this writer [11].

2. Complete distributions. The Pearson system of frequency curves has its genesis in the differential equation

$$\frac{1}{f(x)} \frac{df(x)}{dx} = \frac{a - x}{b_0 + b_1 x + b_2 x^2}, \tag{1}$$

where the origin is arbitrarily taken. Since we are concerned with truncated distributions, it is convenient to take the origin at the left terminus. In the derivations which follow we regard a, b_0, b_1, and b_2 as primary characterizing parameters of the distributions studied. The mean, standard deviation, α_3, and α_4 are expressed as functions of these quantities. To obtain a moment recursion formula for the general Pearson (complete) function, $f(x)$, we separate the variables of (1), multiply both sides of the resulting equation by x^k, and integrate over the full range of permissible values of x, i.e., $r \leq x \leq s$. Thereby we obtain

$$\int_r^s (b_0 + b_1 x + b_2 x^2) x^k \, df(x) = \int_r^s (a - x) x^k f(x) \, dx. \tag{2}$$

Let the kth moment of the complete (population) distribution function, $f(x)$, about the origin selected (i.e., about the left terminus) be designated as

$$\mu_k' = \int_r^s x^k f(x)\, dx, \tag{3}$$

and the right member of (2) becomes

$$a\mu_k' - \mu_{k+1}'.$$

Upon integrating the left member of (2) by parts, we obtain

$$[(b_0 + b_1x + b_2x^2)x^k f(x)]_r^s - kb_0\mu_{k-1}' - (k-1)b_1\mu_k' - (k+2)\mu_{k+1}'.$$

The Pearson system includes only those solutions of (1) for which $f(r) = f(s) = 0$, and moreover only those for which the left member of the above expression vanishes at both limits. As a consequence of these restrictions, we may combine the left and right members above, to obtain the following recursion formula for moments of the complete distribution about the origin:

$$h\mu_k' + b_0k\mu_{k-1}' + b_1k\mu_k' + b_2(k+2)\mu_{k+1}' = \mu_{k+1}', \tag{4}$$

where we have written

$$h = a + b_1. \tag{5}$$

If $f(x)$ is normalized so that $\mu_0' = 1$, and we let $k = 0, 1, 2$, and 3, successively, in (4), the resulting system of equations may be written as

$$\begin{aligned}
&(2b_2 - 1)\mu_1' &&= -h,\\
&(b_1 + h)\mu_1' + (3b_2 - 1)\mu_2' &&= -b_0,\\
&2b_0\mu_1' + (2b_1 + h)\mu_2' + (4b_2 - 1)\mu_3' &&= 0,\\
&3b_0\mu_2' + (3b_1 + h)\mu_3' + (5b_2 - 1)\mu_4' &&= 0.
\end{aligned} \tag{6}$$

On solving (6) for moments of the complete distribution we obtain

$$\begin{aligned}
\mu_1' &= h/(1 - 2b_2),\\
\mu_2' &= [b_0 + (b_1 + h)\mu_1']/(1 - 3b_2),\\
\mu_3' &= [2b_0\mu_1' + (2b_1 + h)\mu_2']/(1 - 4b_2),\\
\mu_4' &= [3b_0\mu_2' + (3b_1 + h)\mu_3']/(1 - 5b_2).
\end{aligned} \tag{7}$$

With h, b_0, b_1, and b_2 known, it is a simple matter to determine μ_1', μ_2', μ_3', and μ_4' from equations (7) in the order named. These equations might, of course, be rewritten with μ_1', μ_2', and μ_3' entirely eliminated from the right members. However, when this is done they become too complex in structure to be of practical value. After calculating μ_1', μ_2', μ_3', and μ_4' from (7), corresponding central moments can be determined from the well known translation formula

$$\mu_k = \sum_{i=0}^{k} \binom{k}{i} \mu_{k-i}'(\mu_1')^i, \tag{8}$$

and the standard moments from

$$\alpha_k = \mu_k/\sigma^k, \tag{9}$$

where $\sigma^2 = \mu_2$. The second central moment then becomes

$$\mu_2 = \frac{1}{1 - 3b_2}\left[b_0 + \frac{h}{1 - 2b_2}\left\{b_1 + \frac{hb_2}{1 - 2b_2}\right\}\right]. \tag{10}$$

Similar formulas can also be written for μ_3 and μ_4, but they are too unwieldy to be useful. For each practical application, it seems preferable to compute noncentral moments about the left terminus from (7). Central and standard moments, as required, can then be obtained from (8) and (9).

If we designate the left truncation point in standard units of the complete distribution by ξ', we have $\xi' = (0 - \mu_1')/\sigma$ and thus

$$\mu_1' = -\sigma\xi'. \tag{11}$$

Although formulas expressing the mean, standard deviation, α_3, and α_4 explicitly as functions of a, b_0, b_1, and b_2 are unduly complex for the four-parameter distributions, as shown below they become relatively simple for Type III and Normal distributions.

Type III distribution. In this case $b_2 = 0$, and we have

$$\begin{aligned} \mu_1' &= h, & h &= -\sigma\xi', \\ \sigma &= \sqrt{b_0 + b_1 h}, & b_1 &= \sigma\alpha_3/2, \\ \alpha_3 &= 2b_1/\sqrt{b_0 + b_1 h}, & b_0 &= \sigma^2[1 + \xi'\alpha_3/2]. \end{aligned} \tag{12}$$

Normal distribution. In this case $b_1 = b_2 = 0$, and

$$\begin{aligned} \mu_1' &= h, & h &= -\sigma\xi', \\ \sigma &= \sqrt{b_0}, & b_0 &= \sigma^2. \end{aligned} \tag{13}$$

3. Recursion formula for moments of incomplete distributions. If the limits of integration in equation (2) are reduced to include only the truncated range $0 \leq x \leq d$, where $r \leq 0$ and $d \leq s$, we have

$$\int_0^d (b_0 + b_1 x + b_2 x^2)x^k \, df(x) = \int_0^d (a - x)x^k f(x)\, dx. \tag{14}$$

Define the kth moment of the truncated distribution about the left terminus as

$$m_k = \int_0^d x^k f(x)\, dx, \tag{15}$$

with $m_0 = 1$, and the right member of (14) becomes

$$am_k - m_{k+1}.$$

On integrating the left member of (14) by parts we obtain

$$[(b_0 + b_1 x + b_2 x^2)x^k f(x)]_0^d - kb_0 m_{k-1} - (k + 1)b_1 m_k.$$

Since we are not integrating over the full range of x, the first term of the above expression does not vanish as it did with the complete distribution. However, if we define

$$F_1 = f(0)b_0^*, \qquad (16)$$
$$F = f(d)[b_0 + db_1 + d^2 b_2],$$

and then combine left and right members above, we obtain the following recursion formula for moments of the truncated distribution:

$$hm_k + b_0 k m_{k-1} + b_1 k m_k + b_2(k+2)m_{k+1} - d^k F = m_{k+1} \qquad (k \geq 1). \qquad (17)$$

If we let $k = 0$ in (14) prior to integrating, and then proceed as outlined above, we obtain

$$h + 2m_1 b_2 + F_1 - F = m_1, \qquad (18)$$

which may be regarded as a companion equation to (17) for the case $k = 0$.

4. Estimating h, b_0, b_1, and b_2 from doubly truncated samples. To obtain estimates by equating observed sample moments to population moments, we substitute the observed sample moments ν_k for the m_k in (17), simultaneously replacing the population parameters h, b_j, $\cdots$, and F by their estimates h^*, b_j^*, $\cdots$, F^*. Setting $k = 1, 2, \cdots, 5$, successively, we find the estimating equations

$$
\begin{aligned}
\nu_1 h^* + b_0^* + \nu_1 b_1^* + 3\nu_2 b_2^* - dF^* &= \nu_2, \\
\nu_2 h^* + 2\nu_1 b_0^* + 2\nu_2 b_1^* + 4\nu_3 b_2^* - d^2F^* &= \nu_3, \\
\nu_3 h^* + 3\nu_2 b_0^* + 3\nu_3 b_1^* + 5\nu_4 b_2^* - d^3F^* &= \nu_4, \qquad (19) \\
\nu_4 h^* + 4\nu_3 b_0^* + 4\nu_4 b_1^* + 6\nu_5 b_2^* - d^4F^* &= \nu_5, \\
\nu_5 h^* + 5\nu_4 b_0^* + 5\nu_5 b_1^* + 7\nu_6 b_2^* - d^5F^* &= \nu_6.
\end{aligned}
$$

These constitute a linear system of five equations in the five estimates, h^*, b_0^*, b_1^*, b_2^*, and F^*, which can be solved by any of the standard methods applicable to such systems. For practical applications, the writer suggests using either the method of "single division" or "multiplication and subtraction" as described by Dwyer [12]. With estimates h^*, b_0^*, b_1^*, and b_2^* thus calculated, we substitute these values in (7) to estimate moments of the complete distribution, and subsequently compute the required estimates of population (complete distribution) parameters with the aid of (8) and (9). F_1^* can be computed from (18) upon replacing parameters by their estimates and m_1 by ν_1. It will be noted that these estimates are consistent since if they should be calculated from the entire population they would obviously equal the required parameters.

Although neither F_1^* nor F^* is required in estimating moments of the complete distribution, a comparison of their values found on solving (18) and (19) with corresponding values computed from the finally fitted curve with the aid of (16) affords a check on agreement between the fitted curve and observed sample data.

It should be noted here that estimates are distinguished from parameters throughout this paper by starring (*) the estimates.

5. Determining type of distribution. With estimates of μ_1', σ, α_3, and α_4 computed as indicated in Section 4, the type of the distribution involved can be established from the original Pearson criteria, an excellent exposition of which has been given by Elderton [13], or from the Carver-Craig criteria [14]. In the present instance, however, since estimates of b_0, b_1, and b_2 must of necessity be computed before estimates of the population parameters can be obtained, it seems more appropriate to determine the type directly from the quadratic equation

$$b_0 + b_1 x + b_2 x^2 = 0. \tag{20}$$

The general solution of the differential equation (1) can be written as

$$f(x) = C(x - r_1)^{m_1}(r_2 - x)^{m_2}, \tag{21}$$

where r_1 and r_2 are roots of (20) (cf., for example, [14]). The nature of these roots determines the type of the distribution. If we let D designate the discriminant, $D = b_1^2 - 4b_0b_2$, the principal Pearson curves[1] may be classified as follows:

Type I	$r_1 - \mu_1' < 0 < r_2 - \mu_1'$,	$D > 0$;
II	$(r_1 - \mu_1') = -(r_2 - \mu_1')$, $b_1 = 2b_2\mu_1'$,	$D > 0$;
III	$b_2 = 0$	
IV	r_1 and r_2 imaginary, $b_1 \neq 2b_2\mu_1'$,	$D < 0$;
V	$(r_1 - \mu_1')$, $(r_2 - \mu_1')$ of the same sign,	$D > 0$;
VI	r_1 and r_2 imaginary, $b_1 = 2b_2\mu_1'$,	$D < 0$;
Normal	$b_1 = b_2 = 0$.	

It can be shown that a necessary condition for the odd central moments to equal zero (i.e., for $f(x)$ to be symmetrical about is mean) is that

$$b_1 = 2b_2\mu_1'.$$

6. Singly truncated samples. If only the left tail is omitted, then F vanishes, and we can drop from (19) the last equation, which would otherwise be required, after placing $F^* = 0$ in the remaining equations. If only the right tail is missing, then $F_1 = 0$, and by changing the variable from x to $d - x$ we can translate the origin to the truncation point on the right, set $F_1^* = 0$, and again drop the last equation otherwise required in (19). As an alternate and frequently preferable

[1] The numbering of the types followed here is that of Craig [14].

procedure when some origin other than the truncation point of a singly truncated sample has been employed, we might substitute (18) for the last equation of (19) after replacing parameters by their estimates. In both instances, the order of the highest order sample moment required is reduced by one from the requirements for doubly truncated samples.

In practical applications, finding either F_1^* or F^* equal or almost equal to zero from a sample that is represented as being doubly truncated, suggests that perhaps the sample was in fact only singly truncated. In this case, either the sample terminus is the terminus of the complete distribution or the absence of lower sample values is due to the small probability associated with their occurrence. Finding both F_1^* and F^* equal or nearly equal to zero suggests that the sample was not truncated after all, and that the necessary estimates should be computed from estimating equations applicable to complete samples.

When the sample terminus is employed as an estimate of the corresponding population terminus, an additional equation may be dropped from (19) since in this case we are estimating one less parameter from the moment equations. To illustrate, consider a Type III distribution for which the left sample terminus (origin) is considered as an appropriate estimate of the population lower limit. We then have

$$h = 2\sigma/\alpha_3 ,$$

and from (12)

$$h = (b_0 + b_1 h)/b_1 .$$

Consequently it follows that $b_0 = 0$, and the system of estimating equations to be solved consists of the first two equations of (19) plus (18) with $b_0^* = b_2^* = 0$. The parameters appearing in (18) are of course replaced by their estimates.

7. Type III and normal distributions. When it is desired to estimate parameters of a Type III distribution (for which $b_2 = 0$) from a doubly truncated sample, we need calculate only the first five sample moments and solve the first four equations of (19) after placing $b_2^* = 0$. With singly truncated samples from which the left tail is missing, we require only the first four sample moments and need solve only the first three equations of (19) after setting $F^* = 0$.

To estimate parameters of a normal distribution (for which $b_1 = b_2 = 0$) from doubly truncated samples, we calculate the first four sample moments and solve the first three equations of (19) after setting $b_1^* = b_2^* = 0$. With singly truncated samples from which the left tail is missing, we require only the first three sample moments and need solve only the first two equations of (19) after setting $F^* = 0$.

8. A numerical example. To illustrate the application of results obtained in this paper to practical problems, we consider an example given by Miss Shook [15] on the weights of 1000 female students (cf. Table 1). Miss Shook considered her data as a complete (untruncated) sample from a Pearson Type III popula-

TABLE 1

Weights of 1000 female students

Weight in pounds	Observed fre-quency	Graduated frequencies based on Type III distribution			
		Complete sample	Truncated on right	Limit at sample terminus	Doubly truncated
70– 79.9	2	0	0.2	0.0	0.2
80– 89.9	16	4	12.8	8.4	12.7
90– 99.9	82	102	94.0	97.5	94.0
100–109.9	231	238	213.7	223.6	214.0
110–119.9	248	250	254.1	247.9	253.9
120–129.9	196	184	200.9	191.7	200.8
130–139.9	122	111	120.7	118.6	120.6
140–149.9	63	59	59.5	63.2	59.4
150–159.9	23	29	25.2	30.2	25.2
160–169.9	5	13	9.5	13.3	9.5
170–179.9	7	6	3.3	5.5	3.3
180–189.9	1	3	1.0	2.2	1.0
190–199.9	2	1	0.3	0.8	0.3
200–209.9	1	0	0.1	0.0	0.1
210–219.9	1	0	0.0	0.0	0.0
Total.............	1000	1000	995.3	1002.9	995.0
Total frequency in truncated range..	981	977	980.9	981.1	980.6
M^* (lbs.)		118.74	118.55	119.14	118.56
σ^* (lbs.)		16.9175	16.027	16.958	16.024
α_3^*		0.976424	0.655	0.865	0.657
Lower limit (lbs.)..........		84.09	69.61	79.95	69.77
F_1^* (from sample moments).............			0	0	−0.002
F_1^* (from fitted curve)..................			0.006	0	0.005
F^* (from sample moments).............			0.767	1.358	0.769
F^* (from fitted curve)..................			0.732	1.183	0.733

Truncated sample obtained by truncating the complete sample on the left at 79.95 lbs. and on the right at 159.95 lbs.

tion, and employed the method of moments to estimate population parameters. Using these estimates, she then graduated the observed sample data.

For our purposes, we truncate Miss Shook's sample on the left at 79.95 lbs. and on the right at 159.95 lbs., thus eliminating the first and the last six cells of the grouped data. The retained (truncated) sample then consists of 981 observations, all of which are within the range 79.95 to 159.95 lbs. We disregard all prior knowledge about the type of the population, and accordingly compute the first six sample moments about the lower terminus. In order to compensate for moment errors due to grouping, we apply Sheppard's[2] corrections for non-central moments. Both sets of moments are given below.

Uncorrected moments	*Corrected moments*
$n_1 = (7.56676860)\ 5$	$\nu_1 = (7.56676860)\ 5$
$n_2 = (66.4026504)\ 5^2$	$\nu_2 = (66.0693171)\ 5^2$
$n_3 = (649.817533)\ 5^3$	$\nu_3 = (642.250764)\ 5^3$
$n_4 = (6913.71764)\ 5^4$	$\nu_4 = (6781.37901)\ 5^4$
$n_5 = (78479.9827)\ 5^5$	$\nu_5 = (76331.5834)\ 5^5$
$n_6 = (937015.638)\ 5^6$	$\nu_6 = (902910.393)\ 5^6$

We substitute these values in (19) and solve the system by Dwyer's method of multiplication and subtraction to obtain $h^* = 44.973178$, $b_0^* = -53.5929$, $b_1^* = 12.339508$, $b_2^* = -0.084107$, and $F^* = 0.578321$. From (18) we then obtain $F_1^* = -0.196817$. The small negative value thus computed suggests that perhaps F_1 actually has the value zero and that there was no truncation on the left.

Considering the sample as being truncated on the right only, and rather than translate our origin to the right sample terminus, we substitute (18) with $F_1^* = 0$ for the last equation of (19) to obtain a new system of five equations in the same five unknown estimates as before, but involving only the first five sample moments. On solving the new set of equations, we obtain $h^* = 38.530928$, $b_0^* = 54.83444$, $b_1^* = 5.179891$, $b_2^* = 0.000986$, and $F^* = 0.771707$.

The small values obtained for b_2^* in both the above cases lead us to conjecture that b_2 actually has the value zero, and that our sample came from a Type III population.

With the sample considered as coming from a Type III population and as being truncated on the right only, we solve the system consisting of the first three equations of (19) plus (18) with $b_2^* = F_1^* = 0$, and obtain $h^* = 38.600670$, $b_0^* = 54.1194$, $b_1^* = 5.247727$, and $F^* = 0.766827$. On substituting these values in (12) we have $\mu_1'^* = 38.60$ lbs., $\sigma^* = 16.027$ lbs., and $\alpha_3^* = 0.655$. The mean referred to zero as an origin is estimated as $M^* = \mu_1'^* + 79.95$ lbs. $= 118.55$ lbs. The corresponding estimate of the lower limit is 69.61 lbs. A graduation of the sample data using these estimates and carried out with the aid of Salvosa's

[2] See for example reference [16], formula 27.9.3, page 361.

tables [17] is given in Table 1, along with Miss Shook's original graduation which was based on estimates from the complete sample.

To provide additional comparisons, we compute further estimates with the sample assumed to be doubly truncated from a Type III population. Accordingly, we solve the first four equations of (19) with $b_2^* = 0$ to obtain $h^* = 38.605540$, $b_0^* = 53.5835$, $b_1^* = 5.262710$, and $F^* = 0.769439$. From (18) we find $F_1^* = -0.002258$. Similarly, we calculate an additional set of estimates under the assumption that the sample was truncated on the right only but with the left sample terminus being the lower limit of the complete distribution. In this case, the system of three equations consisting of the first two equations of (19) plus (18) with $b_0^* = b_2^* = 0$ yields the solutions $h^* = 39.191957$, $b_1^* = 7.337336$, and $F^* = 1.358114$. Estimates of the basic population parameters for each of the above cases, along with graduations over the complete sample range, are also included in Table 1.

The agreement between observed and graduated frequencies is found to be much better for estimates based on the truncated sample than for estimates based on the complete sample. The improved results obtained with the truncated sample suggest that perhaps some of the extreme observations in Miss Shook's original data came from a different population than that which accounted for the main body of her data. It makes little difference whether the truncated sample is considered as being singly or doubly truncated or whether the left sample terminus is used as an estimate of the population lower limit or not. It will be also noted that the values of F_1^* and F^* as computed from the finally fitted curves with the aid of (16) are in substantially close agreement with the corresponding values found on solving the moment equations. In each case the graduations are very nearly equal throughout the entire sample range, and any one of the truncated sample graduations would be considered as a satisfactory fit to the observed data over the truncated range. Certainly any one of the three sets of estimates would, for this example, be adequate as first approximations on which to base iterations to maximum likelihood estimates in a manner similar to that previously employed by Koshal [18] in improving moment estimates from complete samples. The writer hopes to give further consideration in a subsequent paper to the problems of such iterations when samples are truncated.

REFERENCES

[1] KARL PEARSON AND ALICE LEE, "On the generalized probable error in multiple normal correlation," *Biometrika*, Vol. 6 (1908), pp. 59–68.

[2] ALICE LEE, "Table of Gaussian 'tail' functions when the 'tail' is larger than the body," *Biometrika*, Vol. 10 (1915), pp. 208–215.

[3] R. A. FISHER, "Properties and applications of H_h functions," *Mathematical Tables*, Vol. 1, pp. xxvi–xxxv, British Association for the Advancement of Science, 1931.

[4] W. L. STEVENS, "The truncated normal distribution," Appendix to "The calculation of the time-mortality curve" by C. I. Bliss, *Annals of Applied Biology*, Vol. 24 (1937), pp. 815–852.

[5] W. G. Cochran, "Use of IBM equipment in an investigation of the truncated normal problem," *Proc. Research Forum*, International Business Machines Corp., 1946, pp. 40–43.

[6] Johannes Ipsen, Jr., "A practical method of estimating the mean and standard deviation of truncated normal distributions," *Human Biology*, Vol. 21 (1949), pp. 1–16.

[7] A. Hald, "Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point," *Skandinavisk Aktuarietidskrift*, Haft 3–4 (1949), pp. 119–134.

[8] A. C. Cohen, Jr., "On estimating the mean and standard deviation of truncated normal distributions," *Jour. Am. Stat. Assn.*, Vol. 44 (1949), pp. 518–525.

[9] A. C. Cohen, Jr., "Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 557–569.

[10] D. J. Finney, "The truncated binomial distribution," *Annals of Eugenics*, Vol. 14 (1949), pp. 319–328.

[11] A. C. Cohen, Jr., "Estimating parameters of Pearson Type III populations from truncated samples," *Jour. Am. Stat. Assn.*, Vol. 45 (1950), pp. 411–423.

[12] P. S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, Vol. 6 (1941), pp. 101–129.

[13] W. P. Elderton, *Frequency Curves and Correlation*, 3rd ed., Cambridge University Press, 1938.

[14] C. C. Craig, "A new exposition and chart for the Pearson system of frequency curves," *Annals of Math. Stat.*, Vol. 7 (1936), pp. 16–28.

[15] B. L. Shook, "Synopsis of elementary mathematical statistics," *Annals of Math. Stat.*, Vol. 1 (1930), pp. 14–41.

[16] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[17] Luis R. Salvosa, "Tables of Pearson's Type III function," *Annals of Math. Stat.*, Vol. 1 (1930), appended.

[18] R. S. Koshal, "Application of the method of maximum likelihood to the improvement of curves fitted by the method of moments," *Jour. Roy. Stat. Soc.*, Vol. 96 (1933), pp. 303–13.