Estimation of The Extreme-Value Distribution Parameters from Grouped and Censored Observations

Mohamed Tawfik El-Bolkiny   Yaser Mohamed El-Adl

Faculty of Commerce, Mansoura University.

Abstract

This paper discusses the problem of estimating the parameters of the truncated Extreme-value distribution when individual sample data are not available and one can only work with grouped relative frequency data. Considering the individual observation falling in a particular group follows a truncated Extreme-value distribution with scale and location parameters. Three procedures are presented here to estimate the scale and location parameters; the maximum likelihood, the least squares and the minimum Chi-square.

Key words: **Truncated Extreme-Value Distribution, Actuarial Data, Grouped and Censored Data**.

1- Introduction

Actuarial and mortality data has some characteristics not found in the other applications (Grenander; 1956). These characteristics attained for different reasons. First, for the differing in insurance ages, the observations may consist of lives whose ages at death are independent but not identically distributed. Second, since a person life is only observed to the age at which the period of observation expires, the ages at deaths may not be observed for all sample observations. Hence, the distribution has different truncation points. Third, the data may be grouped with only frequencies of deaths recorded.

Data with such characteristics are found in the study of income distributions by (Salem & Mount 1974). The studies of wildlife population subject to migration. The studies based on clinical trials with treatments beginning at different ages and ending before death. The incomplete follow up studies to determine ages at death occurring after the trial period by (Bowers et. al.; 1997).

Usually group frequency data that available for the mortality data, measured by its location parameters (the midpoint of that group). But using the midpoints of the groups would therefore tend to concentrate the sample data within each group but disperse the group mean more than would be theoretically desirable.

(El-Bolkiny 1989 & 1990) considered the case of the initial ages s_i are equals, i.e. $s_i = s$ for $i=1,2,3\ldots k$. where $s=t_0<t_1<t_2<\ldots t_k < \infty$ delimit k+1 age intervals. The generalization of this paper allows for different s_i's as well as withdrawals (and censoring) prior to death. The ages; $s_i=s$ for $i=1,2,3,\ldots k$. are restricted to set of group limits, $t_0<t_1<t_2< \ldots t_k$, which they are considered the withdrawal ages. This is a reasonable restriction in life insurance and annuity applications, since insureds and annuitants are generally assigned as "insuring age", which is typically integral at the inception of the contract. If t_i's are integers; then the integral "insuring ages" scale is integer. Records are generally based on "insuring age" and a policy year. So these records lend themselves naturally to the analysis described in this paper.

Considering the underlying distribution of life at birth is extreme-value distribution with distribution function; F(x):

$$F(x) = 1- \exp[- \exp [-(x-\mu)/\sigma]; \qquad (1.1)$$

where; $-\infty < \mu < \infty$ and $\sigma>0$ are location and scale parameters; respectively.

In this paper; section 2 deals with the maximum likelihood estimator procedure using grouped and censored data is first formulated. It is difficult to find solution for the parameter estimates in closed form, thus numerical method is used to find a solution. Alternative procedure based on least-squares estimates and its generalization is presented in section 3. A minimum Chi-square is given in section 4. Finally, numerical examples are given in the last section before the summary and concluding remarks.

2- Maximum Likelihood Estimation

Choi and Wette (1969) discussed the problem of the maximum likelihood estimation MLE in the case of Gamma distribution. However; the Gamma distribution in Johnson and Kots (1970) belongs to the exponential family distribution with limited type, and the Extreme-value distribution belongs to the same family with unlimited type.

Let e_i denote the number of new insureds (or new entrants) at age t_i, i=0,2,3, ...k. Also, let n_i denote the number of deaths occurs in the i^{th} interval, i=1,2,3, ...k+1. Let w_i the number of withdrawals at t_i, i=0,1,2, ...k. Of course $w_o=0$. The likelihood function is the product of two factors depends on the sample member. For a sample member initially observed at age t_e (age at contracting) and withdrawn at age t_w (age at withdrawal) before death, the factor is: $[1-F(t_w)]/[1-F(t_e)]$. For a sample member initially observed at age t_e, for whom death accurs in the $d^{\underline{th}}$ interval, the factor is: $[F(t_d\ )-F(t_{d-1})]/[1-F(t_e)]$. Hence; the natural logarithms of the two factors are:

$$\ln[1-F(t_w)]-[\ln(1\text{-}F(t_e)] \quad \text{and} \quad \ln[F(t_d)-F(t_{d-1})]-\ln[1\text{-}F(t_e)].$$

Summing over all members of the sample results; then the likelihood function is;

$$L=\prod_{i=1}^{k}\left(\frac{1-F(t_o)}{1-F(t_e)}\right)^{w_i}\left(\frac{F(t_d)-F(t_{d-1})}{1-F(t_e)}\right)^{n_i} \qquad (2.1)$$

The natural logarithm of the likelihood function lnL=l in (2.1) will be:

$$\ell = \sum_{i=1}^{k} w_i \ln[1-F(t_i)] + \sum_{i=1}^{k} n_i \ln[F(t_i)-F(t_{i-1})]$$

$$-\sum_{i=1}^{k} e_i \ln[1-F(t_i)] + cons\tan t \qquad (2.2$$

$$l = \sum_{i=1}^{k}(w_i - e_i)\ln[1-F(t_i)] + \sum_{i=1}^{k} n_i \ln[F(t_i)-F(t_{i-1})]$$

$$+ \text{ constant.} \qquad (2.3)$$

Using the F(x) in (1.1) then l in (2.2) will be:

$$l = -c\sum_{i=0}^{k}(w_i - e_i)(g_i - h_i/2) - c\sum_{i=1}^{k+1} n_i g_i + \sum_{i=1}^{k+1} n_i \ln(2\text{shin}(ch_i/2)$$

$$+ \text{ constant.} \qquad (2.4)$$

where; $c = \exp(-\mu/\sigma)$, $g_i = [\exp(t_i/\sigma) + \exp(t_{i-1}/\sigma)]/2$ and $h_i = \exp(t_i/\sigma) - \exp(t_{i-1}/\sigma)$; ι=1,2,...κ+1. Differentiate (2.3) with respect to c and σ instead of μ and σ then;

$$\frac{\partial l}{\partial c} = -\sum_{i=0}^{k+1}(n_i + w_i - e_i)g_i - \sum_{i=1}^{k}(w_i - e_i)h_i/2 + \sum_{i=1}^{k} n_i h_i \coth(ch_i/2)/2$$

$$(2.5)$$

$$\frac{\partial l}{\partial \sigma}=\frac{c}{\sigma^2}[\sum_{i=0}^{k+1}(n_i+w_i-e_i)v_i+\sum_{i=1}^{k}(w_i-e_i)m_i/2-\sum_{i=1}^{k}n_im_i\coth(ch_i/2)/2]$$

(2.6)

where;

$v_i = [t_i\exp(t_i/\sigma)+t_{i-1}\exp(t_{i-1}/\sigma)]/2$

and $m_i = t_i\exp(t_i/\sigma) - t_{i-1}\exp(t_{i-1}/\sigma)$. for interval time $t_{(-1)}=0$.

To obtain the maximum likelihood estimators (MLE) for the parameters μ and σ, we have to solve equations (2.4) and (2.5) jointly for the c and σ; respectively.

3- Least-squares estimation

The least squares estimators (LSE) are minimizes the sum of the squared deviations S, for;

$$S=\sum_{i=1}^{k}(\frac{n_i}{e_i}-p_i)^2 \qquad (3.1)$$

where e_i represents the number of lives at the beginning of the i^{th} interval, for i = 1,2,3,.......k_{k+1} and;

$$p_i = \int_{t_{i-1}}^{t_i} f(x_i)dx = [\exp[\exp(\frac{x_i-\mu}{\sigma})-\exp(\frac{s-\mu}{\sigma})]]_{t_{i-1}}^{t_i}$$

$$=\exp[\exp(\frac{s-\mu}{\sigma})-\exp(\frac{t_{i-1}-\mu}{\sigma})]-\exp[\exp(\frac{s-\mu}{\sigma})-\exp(\frac{t_i-\mu}{\sigma})] \quad (3.2)$$

A numerical method may be used to search for μ and σ that minimizes (3.1). Consider c = exp(-μ/σ) and d = exp(s/σ); then the two conditions for a minimum Chi-square are;

$$\frac{\partial S}{\partial c} = 2\sum_{i=1}^{k} (\frac{n_i}{e_i} - p_i)(-\frac{\partial p_i}{\partial c}) = 0 \tag{3.3}$$

$$\frac{\partial S}{\partial \sigma} = 2\sum_{i=1}^{k} (\frac{n_i}{e_i} - p_i)(-\frac{\partial p_i}{\partial \sigma}) = 0 \tag{3.4}$$

Using p_i in (3.2) then;

$$\frac{\partial p_i}{\partial c} = p_i(d - g_i) + \exp\{c(d - g_i)\}\{2\cosh(ch_i/2)\}\{h_i/2\} \tag{3.5}$$

$$\frac{\partial p_i}{\partial \sigma} = \exp\{c(d - g_i)\}\{2\cosh(ch_i/2)\}\{c'h_i + ch_i'\}/2 \tag{3.6}$$

Since each group under truncated Extreme-value distribution then the above result may not have a common variance. We may improve the efficiency of the above result by the generalized least-squares GLS.

Rewrite (3.1) in matrix notation;

$$S = (Y - P)`\ (Y - P) \tag{3.7}$$

where;

$Y = (n_1/e_1, n_2/e_2, n_3/e_3, \ldots . n_{k+1}/e_{k+1}) = (Y_1, Y_2, Y_3 \ldots . Y_{k+1})$ and $P = (p_1, p_2, p_3, \ldots . p_{k+1})$ for $\sum_{i=1}^{k+1} p_i = 1$. We notice that;

$$Y = P + U; \tag{3.8}$$

where $U = (u_1, u_2, u_3, \ldots\ldots, u_{k+1})$ is the vector of errors. We assume that;

$$E\ (U) = 0; \tag{3.9}$$

$$E\ (UU`) = \Omega \tag{3.10}$$

and each u_i is multinomially distributed. The covariance structure is then:

$$\Omega_{k+1xk+1} = \begin{bmatrix} \frac{p_1(1-p_1)}{n} & \frac{-p_1 p_2}{e_2} & '' \; '' & \frac{-p_1 p_{k+1}}{e_{k+1}} \\ \frac{-p_2 p_1}{n} & \frac{p_2(1-p_2)}{e_2} & '' \; '' & \frac{-p_2 p_{k+1}}{e_{k+1}} \\ '' & '' & '' \; '' & '' \\ '' & '' & '' \; '' & '' \\ '' & '' & '' \; '' & '' \\ \frac{-p_{k+1} p_1}{n} & \frac{-p_{k+1} p_2}{e_2} & & \frac{p_{k+1}(1-p_{k+1})}{e_{k+1}} \end{bmatrix} \tag{3.11}$$

Due to the fact that $\sum_{i=1}^{k+1} p_i = 1$, the covariance matrix $\Omega_{(k+1)x(K+1)}$ is singular and its inverse does not exist. Thus for generalized least-squares, we minimize the sum of squares S_*;

$$S_* = (Y_* - P_*)` \, \Omega_*^{-1} \, (Y_* - P_*) \tag{3.12}$$

where;

$$\Omega^{-1}_{*\,kxk} = \begin{bmatrix} (\frac{n}{p_1} + \frac{e_{k+1}}{p_{k+1}}) & \frac{e_{k+1}}{p_{k+1}} & \cdots\cdots\cdots & \frac{e_{k+1}}{p_{k+1}} \\ \frac{e_{k+1}}{p_{k+1}} & (\frac{e_2}{p_2} + \frac{e_{k+1}}{p_{k+1}})\cdots\cdots\cdots & & \frac{e_{k+1}}{p_{k+1}} \\ '' & \cdots\cdots\cdots & \cdots\cdots & \cdots\cdots\cdots \\ '' & \cdots\cdots\cdots & \cdots\cdots & \cdots\cdots\cdots \\ \frac{e_{k+1}}{p_{k+1}} & \frac{e_{k+1}}{p_{k+1}} & \cdots\cdots\cdots & (\frac{e_k}{p_k} + \frac{e_{k+1}}{p_{k+1}}) \end{bmatrix} \tag{3.13}$$

for Y_* is Y with the last element deleted, P_* is the mean vector and Ω_* is the dispersion matrix; the asterisks denote the last dimension of P and Ω deleted (Salem & Mount; 1974). The expand of S_* may be written as follows;

$$S_* = n\sum_{i=1}^{k}\ \sum_{j=1}^{k} (y_i - p_i)\left(\frac{\delta_{ij}}{p_i} + \frac{1}{p_{k+1}}\right)(y_j - p_j) \tag{3.14}$$

where δij is the kronecker delta (Judje et. al. ; 1988), which is equal to 1 when i=j and zero otherwise. S_* it can be further reduced to the following:

$$S_* = n[\sum_{i=1}^{k} (y_1 - p_i)\frac{1}{p_i}\sum_{j=1}^{k}\delta_{ij}(y_j - p_j) + \sum_{i=1}^{k}(y_i - p_i)\frac{1}{p_{k+1}}\sum_{j=1}^{k}(y_j - p_j)]$$

$$= n[\sum_{i=1}^{k} (y_i - p_i)\frac{1}{p_i}(y_j - p_j) + (y_{k+1} - p_{k+1})\frac{1}{p_{k+1}}(y_{k+1} - p_{k+1})]$$

$$= n\sum_{i=1}^{k+1} (y_i - p_i)^2 \ / p_i . \tag{3.15}$$

4- Minimum chi-square Estimation

Since the chi-square goodness of fit test is often employed to test the closeness of the estimated distribution to a given sample distribution, we may directly use minimum Chi-square as a decision criterion. To test the fittness of the observed and theoritical frequencies n_i and $e_i p_i$, for i=1,2,3, ...k+1. The following Chi-square statistic may be used;

$$\chi^2 = \sum_{i=1}^{k+1} (n_i - e_i\, p_i)^2 / e_i\, p_i . \tag{4.1}$$

The statistic χ^2 in (4.1) has a Chi-square distribution with k degrees of freedom. The above χ^2 expression in (4.1) can be written as;

$$\chi^2 = \sum_{i=1}^{k+1} (\frac{n_i}{e_i} - p_i)^2 \ e_i/p_i . \tag{4.2}$$

Clearly; (4.2) is identical to (3.15) since $y_i = \frac{n_i}{e_i}$.

In short the generalized least square estimates of μ and σ are also the minimum Chi-square estimators. The Chi-square estimators can be obtained by solving the following two conditions:

$$\sum_{i=1}^{k+1} (\frac{n_i^2}{e_i \ p_i} - e_i \ p_i)(\frac{\partial \ p_i}{\partial \ c}) = 0 \tag{4.3}$$

and

$$\sum_{i=1}^{k+1} (\frac{n_i^2}{e_i \ p_i} - e_i \ p_i)(\frac{\partial \ p_i}{\partial \ \sigma}) = 0 . \tag{4.4}$$

where $\frac{\partial \ p_i}{\partial \ c}$ and $\frac{\partial \ p_i}{\partial \ \sigma}$ are defined in (3.5) and (3.6); respectively.

Again it is difficult to solve (4.3) and (4.4) in closed form, thus a numerical solution method is recommended.

5- Numerical Results of Mortality Distribution

Using data shown in Table (1) and a procedure of numerical analysis given by (Robston; 1965), the root of the equations (2.4) and (2.5) can only be found numerically. The desired equations involves simultaneous solutions in two unknowns μ and σ. Thus, the likelihood equations may be represented by f = f (μ,σ) = 0 and g = g (μ,σ) = 0. Let f_1, f_2, g_1, g_2 be the first and second order derivatives of f and g with respect to μ and σ respectively, and let μ_j , σ_j be the values of the j^{th} approximation to (μ,σ). A solution of $f(\mu,\sigma)=0$ and $g(\mu,\sigma)=0$ may be obtained using the Newton-Raphson method in two variables.

$$\begin{bmatrix}\mu_{j+1}\\ \sigma_{j+1}\end{bmatrix}=\begin{bmatrix}\mu_j\\ \sigma_j\end{bmatrix}-\frac{\begin{bmatrix} g_2 & -f_2\\ -g_1 & f_1\end{bmatrix}\cdot\begin{bmatrix} f\\ g\end{bmatrix}}{\begin{vmatrix} f_1 & f_2\\ g_1 & g_2\end{vmatrix}}\Bigg|_{\mu_j,\ \sigma_j}$$

If the value of the first approximation (μ_o,σ_o) is not sufficiently close to the roots, the procedure may diverge. A first approximation may be obtained by treating all deaths occurring in the intervals as having occurred at the midpoint of the interval and using the methods for the complete sample.

Table (1): Observed Deaths by Age Groups*

Groups	Observed Deaths	Groups	Observed Deaths
10-15	2	50-55	40
15-20	4	55-60	56
20-25	6	60-65	79
25-30	6	65-70	103
30-35	8	70-75	137
35-40	11	75-80	148
40-45	17	80-85	150
45-50	16	85 and over	185

* Source: Complete life table for U.S.A. 1951-1959.

If the initial values $\mu_0 = 70$ and $\sigma_0 = 5$ the procedure converges after 12 iterations to the mean $\hat{\mu} = 83.47745$ and $\sigma_0 = 15.151314$ correct to 5 decimal places. However, if the initial values $\mu_0 = 120$ and $\sigma_0 = 50$, the procedure converges after 16 interactions to the roots correct to 5 decimal places. The Newton-Raphson method requires a lot of computations per iteration, but it is known that if the initial values μ_0 and σ_0 are sufficiently close to the roots then the Newton-Raphson method will converge easily and quickly.

The numerical method also used to find the roots of the equations (3.12) and (3.13) also equations (4.3) and (4.4). The results, are given in Table(2), the GLS and MLE are absolutely better than the results of LSE in terms of minimum chi-square. The estimated mortality rates are very close to the observed rates for GLS and MLE than for LSE.

Table (2)

Estimates	LSE	GLS = MLE
μ	83.75755	83.47765
σ	15.59312	15.151314
χ^2	4.59195	1.08357

A surprising result is that the estimates of μ and σ for GLS and MLE are equals. Calculating p_i using the estimated values $\hat{\mu}$ and $\hat{\sigma}$, the estimated number of mortality rates in each interval are given in table (3). These results confirm our theoretical expectations.

Table (3): Observed and Estimated Mortality Rates

Groups	Observed rates	Estimated Mortality Rates	
		LS	GLS = MLE
10-15	0.00221	0.00399	0.00232
15-20	0.0456	0.0648	0.00459
20-25	0.00618	0.00813	0.00615
25-30	0.0641	0.0843	0.00742
30-35	0.00802	0.01003	0.00898
35-40	0.01147	0.01343	0.01195
40-45	0.01812	0.02007	0.01899
45-50	0.02869	0.03065	0.02889
50-55	0.04557	0.04751	0.04552
55-60	0.06663	0.06867	0.06665
60-65	0.10017	0.10225	0.10321
65-70	0.14463	0.14654	0.14559
70-75	0.20847	0.21051	0.20890
75-80	0.30297	0.30499	0.30309
80-85	0.44776	0.44981	0.44880
85 and over	1.00000	0.99999	0.99959

6- Summary and Concluding Remarks.

This paper dealt with the problem of estimating the parameters of the truncated Extreme-value distribution. When individual sample data are not available and can only work with the grouped relative frequency data. The maximum likelihood function is formulated assuming that the distribution of individuals falling in a particular group is a truncated Extreme-value distribution. It is difficult to find a solution in closed form for the explicit estimators, and an appropriate numerical solution method is recommended. The generalized least-squares procedure is shown to have the same objective function as minimum chi-squares. The resulting optimality conditions are not different from those of MLE and LSE.

References

Bowers, J., Gerber, H., Hickman, J., Jones, D. and Nesbitt, C. (1997). "Actuarial Mathematics" The society of Actuaries, Itasca, Illinois, U.S.A.

Choi, S.C. and Wette R. (1969). "Maximum Likelihood Estimation of the Parameters of Gamma Distribution and Their Bias". Technometrics, Vol. II, No. 4, pp. 683-690.

El-Bolkiny, M.T. (1989). "The existence and uniqueness of the MLE of the Truncated Extreme-value Distribution". 14th International Conference for Statistics, Computer Science, Social and Demographic Research. Ain Shams University, Cairo, Egypt.

El-Bolkiny, M.T. (1990). "Estimation of the parameters of Extreme-Value-Distribution from grouped observation". The Egyptian Journal for Commercial Studies. Vol. 14, No. 1, pp. 1-13.

Grenander, ULF (1956). "On the theory of mortality measurement". Skand, Aktuirietidskt. 39, Vol. I, pp. 70-96, and Vol. II, pp.125-153.

Johnson, N.L. and Kotz, S. (1970). "Continuous Univariate Distribution". Vol. II., Hougbron Miffin Company, New York.

Judje, G. Hill, R., Gviffiths, W., Lutkepohl, H. and Lee. T. (1988). "Introduction to theory and practice of econometrics" John Wiley & sons, New York, U.S.A.

Robston, A. (1965). A First Course in Numerical Analysis, Mc Graw-Hill, inc., New York, U. S. A.

Salem, A. B. and Mount,T. D.(1974)."A convenient Descriptive Model of Income Distributions: the Gamma Density", Econometrics, Vol. 42. No. 6, pp. 1115-1127.