

Unit 1

A Statistical Model and Related Inference

Abstract

In this article a family of distributions is introduced by confining to a subset of the sample space of a well-known family of distributions, this in turn gives scope to widen the parameter space. Some distributional properties and related inference problems for the new family have been studied. Model fitting for real data sets have been studied. Simulation study has been carried out to study the performance of the estimator.

1. Introduction:

To describe the random behavior of a phenomenon experimenter can use various available statistical models. Some of the models can be easily discarded and from the other remaining, experimenter wishes to select 'best' model, which is 'well' described by the data sets.

However when models do not fit the data sets, one can think of describing new models. For this purpose, the knowledge that the experimenter gains on the basis of data sets helps him a lot to identify the added features that a new model should possess.

For example, if the variable of interest X is known to take values in a subset A of S , the sample space corresponding to a given well-known model $\underline{F}$, then a new model $\underline{G}$ for the variable of interest X can be developed by restricting to the set A . Here the form of the distribution remains the same, however the parameter space may become wider.

The natural parameter space of a given family of distributions depends on the convergence of the corresponding integral over the sample space. The integral may be improper because of (i) the sample space is infinite and/or (ii) the density function is unbounded. However, the existing indefiniteness of the improper integral can be removed by confining to a suitable subset A of the

sample space, so that the set A is finite and/or the defined density function is finite on the set A . Thus by using the same density and the sample space A , new models can be defined.

In the following, as simple illustrations we develop and study such models based on the Exponential and Geometric distribution and by considering the lower tails of the supports.

2. Development of Models:

Let $f(x, \theta)$ be the pdf/pmf of a r.v. X with an unbounded support S . Then Θ , the natural parameter space is given by

$$\Theta = \left\{ \theta : f(x, \theta) \geq 0 \forall x \in S, \int_S f(x, \theta) dx = 1 \right\}.$$

That is Θ is a set of values for which the improper integral $\int_S f(x, \theta) dx$ converges and is equal to one.

However by confining to a bounded subset A of S , define $\Theta_A = \{ \lambda : \text{for all } x \in A, f(x, \lambda) \text{ retains the same sign and } \int_A f_A(x, \lambda) dx \text{ converges to a non-zero finite number} \}$.

Then it is clear that $\Theta \subset \Theta_A$.

Define a function f_A from $A \times \Theta \rightarrow \mathbb{R}^+$ by

$$f_A(x, \lambda) = f(x, \lambda) I_A(x) \quad \dots\dots\dots (1)$$

where $I_A(x)$ is the indicator function.

For $x \in A$ and $\lambda \in \Theta_A$, let $g_A(x, \lambda)$ be an extension of $f_A(x, \lambda)$ 'extended' to the domain $A \times \Theta_A$. Let

$$b(\lambda) = \int_A f_A(x, \lambda) dx \quad \lambda \in \Theta_A.$$

$$\text{It is to be noted that } \{g(x, \lambda) = \frac{f_A(x, \lambda)}{b(\lambda)}, x \in A, \theta \in \Theta_A\}$$

is a family of density functions with Θ_A as the natural parameter space.

3. Examples:

(i) Let $f(x, \theta) = \theta e^{-\theta x}$, $x \in (0, \infty)$ $\theta \in (0, \infty)$ and zero otherwise, $A = [0, 1]$.

In this case

$\Theta_A = \{\lambda : \text{for all } x \in A \text{ } f(x, \lambda) \text{ retains the same sign, } \int_0^1 \lambda e^{-\lambda x} dx \text{ converges to non-zero finite number}\} = (-\infty, \infty) - \{0\}$.

Remark : Θ_A remains the same even if $A = [0, a]$. Since, for,

$\lambda = 0$, $\int_A \lambda e^{-\lambda x} dx = 0$, for $\lambda = 0$ we define $g(x, 0) = \lim_{\lambda \rightarrow 0} \frac{f_A(x, \lambda)}{b(\lambda)} (= 1)$.

Hence for $0 < x < 1$ we have ,

$$g(x, \lambda) = \begin{cases} \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda}} & \lambda \neq 0 \\ 1 & \lambda = 0 \end{cases} \quad \text{and} \quad G(x, \lambda) = \begin{cases} \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda}} & \lambda \neq 0 \\ x & \lambda = 0 \end{cases}.$$

Thus, an extended family obtained from the exponential distributions is given by the class of densities $\underline{g} = \{g(x, \lambda), x \in (0, 1), \lambda \in \mathbb{R}\}$. We have not seen the study of the class $\underline{g}$ in the existing literature. As the parameter space is $(-\infty, \infty)$ this class is different from the one obtained just by truncation of an exponential family of density functions.

In other well-known families like beta first kind, Pareto the parameter(s) appear in exponent. While in this model variable appears in the exponent.

(ii) Let $P(X = x) = (1 - q)q^x$, $S = \{0, 1, 2, \dots, \infty\}$. $0 \leq q \leq 1$,

Let $A = \{0, 1, 2, \dots, r\}$ then

$\Theta_A = \{q : \text{for all } x \in A, g(x, q) \text{ retains the same sign, } \sum_{x=0}^r (1 - q)q^x \text{ exists, non-zero and finite.}\}$

$= (0, \infty) - \{1\}$.

Hence for $x = 0, 1, \dots, r$ and $q \in (0, \infty)$ we have

$$g(x, q) = \begin{cases} \frac{(1-q)q^x}{1-q^{r+1}} & q \neq 1 \\ \frac{1}{r+1} & q = 1 \end{cases} \quad \text{and} \quad G(x, q) = \begin{cases} \frac{1-q^x}{1-q} & q \neq 1 \\ \frac{x}{r+1} & q = 1 \end{cases}$$

4. Distributional properties :

In the following we consider the model given in example (i). Similar results can be established for the model given in example (ii).

Let X be a r.v. having d.f. $F_x(x; \alpha) = \begin{cases} \frac{\alpha^x - 1}{\alpha - 1} & \alpha \neq 1 \\ x & \alpha = 1 \end{cases}$

The corresponding pdf is given by

$$f(x; \alpha) = \begin{cases} \frac{\log \alpha}{\alpha - 1} \alpha^x & \alpha \neq 1 \\ 1 & \alpha = 1 \end{cases}$$

for $0 < x < 1$ and zero otherwise.

The family $\{f(x, \alpha) : \alpha > 0\}$ forms one parameter exponential family and as such many related inference problems can be solved.

The following are some properties of F :-

1. F is convex if $\alpha < 1$ and concave if $\alpha > 1$.
2. Note that $f(x, \alpha) = f(1-x, 1/\alpha)$, $F(x, \alpha) = 1 - F(1-x, 1/\alpha)$. Thus, if X has d.f. $F(x, \alpha)$ then $1 - X$ has d.f. $F(x, 1/\alpha)$.
3. F is symmetric if and only if $\alpha = 1$, otherwise positively or negatively skewed according as $\alpha < 1$ or $\alpha > 1$.
4. For each fixed x ($0 < x < 1$), the function $F(x, \alpha)$ is decreasing in α . The mean of the distribution is given by

$$\mu(\alpha) = \frac{\alpha}{\alpha - 1} - \frac{1}{\log \alpha} \quad \text{and} \quad \lim_{\alpha \rightarrow 1} \mu(\alpha) = \frac{1}{2}.$$

5. Recurrence relation for moments:

$$\mu'_r = \frac{\alpha}{\alpha - 1} - \frac{r}{\log \alpha} \left\{ \frac{\alpha}{\alpha - 1} - \frac{(r-1)}{\log \alpha} \right\} \mu'_{r-2}$$

6. $V_\alpha(X)$, the variance of X is $\frac{1}{(\log \alpha)^2} - \frac{\alpha}{(\alpha-1)^2}$ and we note that

$V_\alpha(X) = V_{1/\alpha}(X)$. Also, $V_\alpha(X)$ is increasing in the interval $(0, 1)$ and

decreasing on $(1, \infty)$ and $\lim_{\alpha \rightarrow 1} V(\alpha) = 1/12$.

7. Characteristic function:

$$\phi_X(t) = \frac{\alpha e^{it} - 1}{\alpha - 1} \frac{\log \alpha}{\log \alpha + it}$$

8. Failure rate function : $r(t) = \frac{\alpha^t \log \alpha}{\alpha - \alpha^t}$ and hence F belongs to IFR

9. Fisher Information contained in single observation:

$$I(\alpha) = \left(\frac{1}{\alpha \ln(\alpha)} \right)^2 - \frac{1}{\alpha(\alpha-1)^2} = \frac{1}{\alpha^2} V_\alpha(X)$$

and it can be shown that $\alpha^2 I(\alpha) = \left(\frac{1}{\alpha} \right)^2 I\left(\frac{1}{\alpha} \right)$ and $\lim_{\alpha \rightarrow 1} I(\alpha) = 1/12$

The graph of the function $I(\alpha)$ is given appendix 1.4.

10. The above family can be extended by introducing the location parameter μ and the scale parameter σ . Let $Y = (X - \mu)/\sigma$ and

$$F(t) = \begin{cases} 0 & t < 0, \\ \frac{\alpha^t - 1}{\alpha - 1} & 0 \leq t \leq 1, \\ 1 & t \geq 1. \end{cases}$$

That is

$$F_{\alpha, \mu, \sigma}(u) = P[X \leq u] = \begin{cases} 0 & u < \mu, \\ \frac{\alpha^{\frac{u-\mu}{\sigma}} - 1}{\alpha - 1} & \mu \leq u \leq \mu + \sigma, \\ 1 & u \geq \mu + \sigma. \end{cases}$$

We note that for $\mu \leq a < b \leq \mu + \sigma$,

$$P[X \leq u \mid a < X < b]$$

$$= \begin{cases} 0 & u < a, \\ \frac{\alpha^{\frac{\mu-a}{\sigma}} - 1}{\alpha^{\frac{b-\mu}{\sigma}} - 1} & a < u < b, \\ 1 & u \geq b. \end{cases}$$

$$= \begin{cases} 0 & u < a, \\ \frac{\beta^{\frac{u-a}{b-a}} - 1}{\beta - 1} & a < u < b, \\ 1 & u \geq b, \end{cases}$$

where, $\beta = \alpha^{\frac{b-\mu}{\sigma}}$.

Thus, if $X \sim F_{\alpha, \mu, \sigma}$ then the conditional distribution of X given $a < X < b$ is $F_{\beta, a, b-a}$.

5. Related Inference of the models

5.1 Estimation:

Let $X_1, X_2, \dots, X_n$ be a random sample from d.f. F defined in 4.

Likelihood function is

$$L(\alpha) = \left(\frac{\log \alpha}{\alpha - 1} \right)^n \alpha^{\sum_{i=1}^n x_i}$$

To find the maximum likelihood estimator $\hat{\alpha}$ of α , we consider the likelihood equation

$$\frac{1}{\alpha \log \alpha} - \frac{1}{\alpha - 1} + \frac{\bar{x}}{\alpha} = 0.$$

Since the p.d.f. is of the form $c(\alpha) h(x) \exp\{x u(\alpha)\}$, m.l.e. coincides with the moment estimator. This equation can be solved by using method of scoring. The improved solution is obtained by

$$\alpha_{i+1} = \alpha_i + \frac{\frac{1}{\alpha_i \log \alpha_i} - \frac{1}{\alpha_i - 1} + \frac{\bar{x}}{\alpha_i}}{nl(\alpha)}$$

While using the above method sometimes next value may be negative, and hence we need to modify the method. One such modification is described below.

If next iterated value is negative then define $\alpha_{[i+1]} = \alpha_{[i]} / 2$ if numerator > 0 and denominator > 0.

Performance of the mle by simulation has been carried out by using a program in C language (given in Appendix 1.1) and the results are summarized in Appendix 1.2.

5.2 Testing of hypothesis:

For $0 < \alpha_1 < \alpha_2$, the ratio $\frac{f(x, \alpha_2)}{f(x, \alpha_1)}$ is increasing in x . Hence the UMP test of size α^* , for testing $H_0: \alpha \leq \alpha_0$ Vs $H_1: \alpha > \alpha_0$ is to reject H_0 if $x > k$ and k is such that $\frac{\alpha_0^k - 1}{\alpha_0 - 1} = 1 - \alpha^*$. The unique solution of it is given by

$$k = \frac{\log(\alpha_0 + \alpha^*(1 - \alpha_0))}{\log(\alpha_0)}.$$

The UMPU test of size α^* , for testing $H_0: \alpha = 1$ Vs $H_1: \alpha \neq 1$ is to reject H_0 if $x < c_1$ or $x > c_2$ and as the distribution of X under H_0 is symmetric about $1/2$ we have

$$c_1 = \alpha^*/2, c_2 = 1 - (\alpha^*/2).$$

The UMPU test for testing $H_0: \alpha = \alpha_0$ Vs $H_1: \alpha \neq \alpha_0$ is of the same form where the constants c_1 and c_2 are the solutions of the equations

$$\alpha_0^{c_1} - \alpha_0^{c_2} = (\alpha^* - 1)(\alpha_0 - 1)$$

$$c_1 \alpha_0^{c_1} - c_2 \alpha_0^{c_2} \frac{\alpha_0^{c_1} - \alpha_0^{c_2} + (\alpha_0 - 1)}{\log \alpha_0} + \alpha_0 = \alpha^* \left(\alpha_0 - \frac{\alpha_0 - 1}{\log \alpha_0} \right)$$

As for this family Cramer's regularity conditions are satisfied, one can obtain asymptotic confidence intervals, tests by using asymptotic normality of $\hat{\alpha}$.

5.3 Simulation study of m.l.e.

To study the performance of m.l.e., simulation study was carried out. The histograms of the normalized m.l.e. $\sqrt{nI(\alpha)}(\hat{\alpha} - \alpha) \sim N(0,1)$ for sample size 50 and 100 are given in the appendix-1.3. In fact $\hat{\alpha}$ is BAN for α with asymptotic variance $\frac{1}{nI(\alpha)}$ which is equal to the Cramer-Rao Lower Bound.

To study small, moderate and large sample behavior of m.l.e. simulation study was carried out by taking samples of sizes $N=10, 30, 50, 100, 500$ & generating 1000 ($= M$) samples each time. Estimators by taking the averages of the 1000 estimators ($\hat{\alpha}_1$) and that obtained by combining all the 1000 samples ($\hat{\alpha}_2$) are given in the table in Appendix 1.2. As expected m.l.e. does not perform well for large (small) values of α .

6. Data Analysis (Case Studies) :

A. In case of arrival of passengers to board a train it seems that the intensity of arrivals increases as the remaining time to departure of the train reduces. The same would be true in case of time to pay for the bills before the due date.

B. However, in the situations of getting benefits, the intensity is expected to be more in the earlier stages. For example, the purchase of very popular fresh-food-articles or the purchase of high discounted articles.

We considered some situations of type A and collected data related to

- (i) the arrival times of passengers to board a train (Data Set 6.1),
- (ii) the sign-in times of employees (Data Sets 6.2.1, 6.2.2 and 6.2.3) and
- (iii) the arrival times of persons to purchase cinema tickets (Data Sets 6.3.1, 6.3.2 and 6.3.3).

In all these cases, the arrival time T , observed during a fixed period t_1 to t_2 is converted to X ($0 < X < 1$) by using the transformation $X = (T - t_1)/(t_2 - t_1)$.

We shall describe in detail the method of data collection, estimation and goodness of fit carried out for the data set 6.1. Similar methods are used for other other data sets.

6.1 The arrival times of passangers to board the 6.40 p.m. train (Data Set 6.1) :

Data was collected over the period 6 p.m. ($= t_1$) and 6.40 p.m. (t_2 , the departure time of train) and arrivals were recorded in seconds counted from 6 p.m. . The first arrival was at 6 hrs. 11 min. 3 seconds and is being recorded as 663 seconds ($t = T - t_1$). As such X , the variable of interest is $X = t/2400$.

This particular train being a passenger train, many regular commuters are pass holders. Further, for Kolhapur Railway Station, passengers can enter the railway platform through many entry-points. Data was collected at the entrance gate nearest to the booking counter. As there is another train at 7.05 p.m., it was necessary to identity the passengers for the 6.40 p.m. passenger train. In all there were 128 arrivals.

DATA SET 6.1

The arrival times of passengers

Time in seconds (t)	No. of persons arrived	X	Time in seconds (t)	No. of persons arrived	X
663	1	0.27625	1580	1	0.658333
691	1	0.287917	1595	1	0.664583
720	2	0.3	1625	2	0.677083
760	1	0.316667	1640	1	0.683333
880	2	0.366667	329.9		0.137458
920	1	0.383333	1655	1	0.689583
930	3	0.3875	1665	2	0.69375
940	1	0.391667	1680	1	0.7
1015	1	0.422917	1685	4	0.702083
1020	1	0.425	1740	1	0.725
1040	2	0.433333	1770	5	0.7375
1050	4	0.4375	1800	1	0.75
1060	2	0.441667	1820	1	0.758333
1110	2	0.4625	1850	1	0.770833
1120	1	0.466667	1855	2	0.772917
1140	1	0.475	1875	2	0.78125
1145	2	0.477083	1905	5	0.79375
1200	3	0.5	1925	1	0.802083

1220	2	0.508333
1265	1	0.527083
1280	2	0.533333
1381	1	0.575417
1410	2	0.5875
1415	1	0.589583
1420	2	0.591667
1430	3	0.595833
1440	1	0.6
1460	1	0.608333
1470	1	0.6125
1475	1	0.614583
1495	1	0.622917
1505	2	0.627083
1540	1	0.641667
1535	2	0.639583
1565	5	0.652083

1935	2	0.80625
2000	1	0.833333
2025	1	0.84375
2030	3	0.845833
2045	4	0.852083
2070	3	0.8625
2075	1	0.864583
2115	6	0.88125
2175	2	0.90625
2190	2	0.9125
2265	1	0.94375
2300	1	0.958333
2310	2	0.9625
2320	1	0.966667
2380	5	0.991667
2390	1	0.995833

Assuming the model,

$$f(x; \alpha) = \begin{cases} \frac{\log \alpha}{\alpha - 1} \alpha^x & \alpha \neq 1 \\ 1 & \alpha = 1 \end{cases},$$

based on the data collected, $\hat{\alpha}$, the mle, is computed by using method of scoring. Iterative procedure was continued until the difference between two successive estimated values is less than 10^{-12} . In this case, $\hat{\alpha} = 9.39902$.

By using this estimated value, Goodness of fit test for the model is carried out. The observed and the expected frequencies for the various classes together with calculated χ^2 value are given in the following table :

TABLE FOR GOODNESS OF FIT

Class Intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square Statistics
0 - 0.1	0	0.029902	3.827445	3.82744486
0.1 - 0.2	0	0.037412	4.788695	4.78869526
0.2 - 0.3	2	0.046808	5.991361	2.65898857
0.3 - 0.4	10	0.058563	7.496071	0.83639257
0.4 - 0.5	16	0.073271	9.378685	4.6746216
0.5 - 0.6	17	0.091673	11.73411	2.36315933
0.6 - 0.7	23	0.114696	14.6811	4.71382774
0.7 - 0.8	23	0.143502	18.36821	1.16797004
0.8 - 0.9	22	0.179542	22.98132	0.04190326
0.9 - 1	15	0.224633	28.75301	6.57827498
	128	1	128	31.6512782

Here, Cal. $\chi^2_8 = 31.65731$ and Tab. $\chi^2_{8, 0.05} = 15.50731$. This implies that the proposed model does not fit for this data.

Remarks :- Let O_i and E_i be the observed and expected i^{th} class frequencies respectively.

1. Observe that O_1, O_2 are zero and O_3 is very small. As such, there is significant contributions to the χ^2 value from these 3 classes.
2. Moreover, $O_{10} < O_9$ while $E_{10} > E_9$ (since $\alpha > 1$). The difference ($O_{10} - E_{10}$) is also contributing significantly to the χ^2 value.

These facts indicate that the family of distributions has to be enlarged by considering additional parameters a, b ($0 < a < b < 1$), the lower and upper bounds for the variable X . The distribution of such a random variable is given in section 4. For each fixed α , m.l.e.s of a and b are $x_{(\min)}$ and $x_{(\max)}$ respectively.

In case of arrival of passengers to board a train, there is risk of missing the train, the passengers have the tendency of keeping a margin of time to avoid the risk of missing the train, to a very good extent. This margin of time is related with the amount of risk. As such the intensity does not increase over the entire period but it increases initially and then decreases. Because of this decreasing tendency (though on a small interval at the end), the proposed model does not give a very good fit.

There are two ways to overcome the drawback. The first being confine to the arrivals excluding the last risk period and the other is enlarge the class of distributions by introducing additional parameters or by using the technique of composite functions.

6.2 The arrival times of employees :

To study the pattern of arrival times of persons to the office, we collected data related to the arrival times of employees working at Main Building of Shivaji University, Kolhapur on three days.

Data was collected on 1st, 2nd and 3rd May 2003 during the period 10.00 a.m. to 10.25 a.m. During this period the register is kept at a particular place and the persons are expected to sign in the register during this period. Arrival times t (in seconds counted from 10.00 a.m. ($= t_1$)) to sign the register are recorded up to 10.25 a.m. ($= t_2$). As such, the variable of interest X ($0 < X < 1$) to be considered as $X = t/1500$.

Data Set 6.2.1

Data collected on 1st May 2003 :

Time (in sec.)	No. of persons	X
435	1	0.29
450	6	0.3
475	3	0.316667
485	1	0.323333
525	4	0.35
605	5	0.403333
615	5	0.41
630	4	0.42
640	1	0.426667
660	1	0.44
680	3	0.453333
685	1	0.456667
690	1	0.46
692	1	0.461333
700	2	0.466667
710	3	0.473333
715	1	0.476667
721	1	0.480667
725	1	0.483333
730	2	0.486667
745	1	0.496667
750	1	0.5
757	1	0.504667
760	1	0.506667
787	1	0.524667
790	1	0.526667
795	1	0.53
805	1	0.536667
810	1	0.54
820	1	0.546667
826	1	0.550667
827	4	0.551333
835	3	0.556667

Time (in sec.)	No. of persons	X
915	1	0.61
925	3	0.616667
930	1	0.62
940	1	0.626667
950	2	0.633333
955	4	0.636667
962	2	0.641333
967	2	0.644667
970	1	0.646667
995	1	0.663333
1015	1	0.676667
1025	1	0.683333
1030	2	0.686667
1032	2	0.688
1040	1	0.693333
1044	1	0.696
1045	1	0.696667
1047	2	0.698
1060	1	0.706667
1065	1	0.71
1075	1	0.716667
1078	1	0.718667
1085	1	0.723333
1090	4	0.726667
1093	8	0.728667
1095	1	0.73
1097	1	0.731333
1100	2	0.733333
1107	1	0.738
1110	3	0.74
1125	1	0.75
1135	1	0.756667
1143	4	0.762

Time (in sec.)	No. of persons	X
1230	1	0.82
1232	1	0.821333
1240	1	0.826667
1245	1	0.83
1257	1	0.838
1258	1	0.838667
1290	2	0.86
1315	2	0.876667
1323	1	0.882
1325	1	0.883333
1327	3	0.884667
1330	2	0.886667
1335	1	0.89
1338	4	0.892
1342	2	0.894667
1350	1	0.9
1352	2	0.901333
1355	5	0.903333
1360	2	0.906667
1365	1	0.91
1375	3	0.916667
1390	2	0.926667
1402	2	0.934667
1415	2	0.943333
1420	3	0.946667
1425	1	0.95
1435	2	0.956667
1443	1	0.962
1445	2	0.963333
1447	8	0.964667
1452	1	0.968
1455	1	0.97
1459	1	0.972667

855	1	0.57
860	1	0.573333
865	1	0.576667
870	3	0.58
872	3	0.581333
876	1	0.584
890	1	0.593333
910	3	0.606667

1152	1	0.768
1160	1	0.773333
1165	1	0.776667
1175	1	0.783333
1185	1	0.79
1205	1	0.803333
1212	2	0.808
1225	2	0.816667

1463	1	0.975333
1468	3	0.978667
1471	1	0.980667
1474	3	0.982667
1482	2	0.988
1485	5	0.99
1491	1	0.994
1495	3	0.996667

There were 232 number of arrival and MLE of the parameter ($\hat{\alpha}$) is 15.74629.

TABLE FOR GOODNESS OF FIT

Class intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square statistic
0 - 0.1	0	0.021524	4.993589	4.993589
0.1 - 0.2	0	0.028356	6.578557	6.578557
0.2 - 0.3	1	0.037356	8.666594	6.78198
0.3 - 0.4	14	0.049213	11.41738	0.584193
0.4 - 0.5	33	0.064833	15.04126	21.44212
0.5 - 0.6	28	0.085411	19.81536	3.380627
0.6 - 0.7	32	0.112521	26.10476	1.331321
0.7 - 0.8	36	0.148235	34.39043	0.075333
0.8 - 0.9	29	0.195284	45.30597	5.868644
0.9 - 1	59	0.257268	59.68611	0.007887
Total	232	1	232	51.04425

Here, Cal. $\chi^2_8 = 51.04425$, Tab. $\chi^2_{8, 0.05} = 15.50731$ and p-value is 2.57×10^{-8} .

Data Set 6.2.2

Data collected on 2nd May 2003 :

Time (in sec)	No. of persons	X
440	4	0.293333
445	2	0.296667
465	2	0.31
490	3	0.326667
520	2	0.346667
530	1	0.353333
550	3	0.366667

Time (in sec)	No. of persons	X
975	1	0.65
980	1	0.653333
998	4	0.665333
1005	1	0.67
1025	1	0.683333
1035	1	0.69
1040	3	0.693333

Time (in sec)	No. of persons	X
1248	2	0.832
1255	2	0.836667
1265	1	0.843333
1272	3	0.848
1275	2	0.85
1285	2	0.856667
1290	1	0.86

555	1	0.37
565	1	0.376667
570	1	0.38
605	1	0.403333
610	1	0.406667
620	1	0.413333
630	2	0.42
640	3	0.426667
650	1	0.433333
665	1	0.443333
685	1	0.456667
690	2	0.46
695	2	0.463333
702	1	0.468
735	3	0.49
750	2	0.5
760	2	0.506667
790	1	0.526667
800	2	0.533333
805	1	0.536667
830	1	0.553333
852	2	0.568
863	2	0.575333
874	2	0.582667
885	3	0.59
902	2	0.601333
945	1	0.63
953	1	0.635333
968	1	0.645333

1045	2	0.696667
1049	1	0.699333
1058	2	0.705333
1060	2	0.706667
1065	1	0.71
1075	1	0.716667
1085	3	0.723333
1095	2	0.73
1105	3	0.736667
1110	1	0.74
1125	2	0.75
1135	2	0.756667
1145	2	0.763333
1150	1	0.766667
1160	2	0.773333
1170	2	0.78
1180	1	0.786667
1190	3	0.793333
1195	2	0.796667
1198	1	0.798667
1202	1	0.801333
1210	2	0.806667
1218	3	0.812
1220	1	0.813333
1230	1	0.82
1237	1	0.824667
1240	4	0.826667
1243	1	0.828667
1245	3	0.83

1295	3	0.863333
1300	1	0.866667
1302	1	0.868
1310	1	0.873333
1325	4	0.883333
1365	3	0.91
1375	1	0.916667
1385	1	0.923333
1387	1	0.924667
1400	3	0.933333
1405	1	0.936667
1410	3	0.94
1420	1	0.946667
1425	1	0.95
1426	1	0.950667
1430	1	0.953333
1433	1	0.955333
1435	2	0.956667
1442	1	0.961333
1450	2	0.966667
1455	3	0.97
1460	4	0.973333
1465	2	0.976667
1470	1	0.98
1480	3	0.986667
1485	1	0.99
1490	3	0.993333
1495	1	0.996667

There were 191 number of arrivals. Here, $\hat{\alpha} = 17.8444$.

TABLE FOR GOODNESS OF FIT

Class intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square statistic
0 - 0.1	0	0.0198275	3.787051	3.787051
0.1 - 0.2	0	0.0264495	5.0518616	5.051862
0.2 - 0.3	6	0.0352832	6.7390974	0.081059
0.3 - 0.4	14	0.0470672	8.9898413	2.792228
0.4 - 0.5	19	0.0627869	11.992295	4.094956
0.5 - 0.6	18	0.0837566	15.997518	0.25066
0.6 - 0.7	20	0.1117299	21.340417	0.084193
0.7 - 0.8	33	0.1490458	28.467753	0.721563
0.8 - 0.9	40	0.1988246	37.975498	0.107928
0.9 - 1	41	0.2652286	50.658668	1.841538
Total	191	1	191	18.81304

Here, Cal. $\chi^2_8 = 18.81304$, Tab. $\chi^2_{8, 0.05} = 15.50731$ and p-value is 0.015892.

Data Set 6.2.3

Data collected on 3rd May 2003 :

Time (in sec)	No. of persons	X
5	2	0.003333
58	1	0.038667
70	1	0.046667
170	2	0.113333
190	1	0.126667
200	1	0.133333
244	1	0.162667
260	1	0.173333
345	1	0.23
380	1	0.253333
482	1	0.321333
488	2	0.325333
525	1	0.35
550	2	0.366667
570	4	0.38
585	1	0.39
605	1	0.403333
640	1	0.426667
645	1	0.43
658	2	0.438667
678	1	0.452
680	2	0.453333
690	1	0.46
695	2	0.463333
700	1	0.466667
710	2	0.473333
722	6	0.481333
750	1	0.5
755	1	0.503333
760	3	0.506667
805	2	0.536667
815	2	0.543333
817	1	0.544667
820	1	0.546667
835	3	0.556667
841	1	0.560667

Time (in sec)	No. of persons	X
847	1	0.564667
858	1	0.572
880	1	0.586667
885	1	0.59
890	1	0.593333
905	2	0.603333
935	4	0.623333
950	3	0.633333
964	3	0.642667
968	1	0.645333
970	2	0.646667
985	1	0.656667
990	1	0.66
992	1	0.661333
1000	2	0.666667
1015	1	0.676667
1027	1	0.684667
1040	2	0.693333
1050	1	0.7
1056	1	0.704
1075	2	0.716667
1078	2	0.718667
1095	5	0.73
1105	1	0.736667
1120	1	0.746667
1130	1	0.753333
1135	1	0.756667
1142	1	0.761333
1153	1	0.768667
1158	2	0.772
1165	1	0.776667
1175	2	0.783333
1180	1	0.786667
1195	3	0.796667
1210	10	0.806667
1225	1	0.816667

Time (in sec)	No. of persons	X
1230	1	0.82
1235	4	0.823333
1250	2	0.833333
1255	1	0.836667
1270	1	0.846667
1275	2	0.85
1283	1	0.855333
1290	1	0.86
1292	3	0.861333
1295	3	0.863333
1300	1	0.866667
1310	1	0.873333
1315	1	0.876667
1318	1	0.878667
1325	4	0.883333
1330	6	0.886667
1340	1	0.893333
1345	1	0.896667
1350	1	0.9
1352	5	0.901333
1370	2	0.913333
1375	4	0.916667
1385	1	0.923333
1390	2	0.926667
1400	1	0.933333
1405	2	0.936667
1415	3	0.943333
1435	2	0.956667
1445	2	0.963333
1455	1	0.97
1460	1	0.973333
1465	1	0.976667
1470	2	0.98
1475	1	0.983333
1480	2	0.986667
1495	2	0.996667

There were 194 number of arrivals. Here, $\hat{\alpha} = 11.05296$.

TABLE FOR GOODNESS OF FIT

Class intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square statistic
0 - 0.1	4	0.02701614	5.2411316	0.293907
0.1 - 0.2	6	0.03435352	6.6645822	0.066271
0.2 - 0.3	2	0.04368366	8.4746309	4.946628
0.3 - 0.4	11	0.05554781	10.776275	0.004645
0.4 - 0.5	20	0.07063416	13.703027	2.893657
0.5 - 0.6	20	0.08981785	17.424663	0.380631
0.6 - 0.7	24	0.11421168	22.157066	0.153288
0.7 - 0.8	26	0.14523068	28.174753	0.167865
0.8 - 0.9	46	0.18467421	35.826796	2.888734
0.9 - 1	35	0.23483029	45.557076	2.446422
Total	194	1	194	14.24205

Here, Cal. $\chi^2_8 = 14.24205$, Tab. $\chi^2_{8, 0.05} = 15.50731$ and p-value is 0.075671.

Observe that the proposed model does not fit to the data sets 6.2.1 and 6.2.2; but, it fits to the data set 6.2.3. By introducing the lower bound as a parameter the model can be enlarged. The data sets may give good fit for the enlarged models.

6.3 Arrival times of persons to a cinema hall :

Data was collected on 30th April 2003 during the period 2.30 p.m. to 3.15 p.m. and on 1st May 2003 during the period 5.30 p.m. to 6.15 p.m. at the entrance gate of Padma Theatre and on 1st May 2003 during the period 9.00 p.m. to 9.30 p.m. at the entrance gate of Uma Theatre. Arrival times t (in seconds counted from the beginning period ($= t_1$)) are recorded up to the starting time of the movie ($= t_2$). As such, the variable of interest X ($0 < X < 1$) in the first two cases is $t/2700$ and in the third case it is $t/1800$.

Data Set 6.3.1

Padma Theatre on 30th April 2003 (2.30p.m. to 3.15 p.m.)

Time (in sec)	No. of persons	X
9120	4	0.044444
9240	2	0.088889
9380	2	0.140741
9390	2	0.144444
9480	1	0.177778
9510	1	0.188889
9515	2	0.190741
9535	1	0.198148
9550	1	0.203704
9715	3	0.264815
9765	6	0.283333
9770	3	0.285185
9778	1	0.288148
9810	1	0.3
9820	2	0.303704
9825	1	0.305556
9830	1	0.307407
9838	1	0.31037
9865	1	0.32037
9880	1	0.325926
9905	2	0.335185
9910	1	0.337037
9935	2	0.346296
9950	2	0.351852
9990	1	0.366667
10010	2	0.374074
10015	4	0.375926
10135	2	0.42037
10140	2	0.422222
10215	2	0.45
10223	5	0.452963
10225	1	0.453704
10240	1	0.459259
10250	1	0.462963
10280	1	0.474074
10315	2	0.487037
10365	1	0.505556
10370	1	0.507407

Time (in sec)	No. of persons	X
10400	3	0.518519
10455	3	0.538889
10470	2	0.544444
10471	1	0.544815
10480	5	0.548148
10495	2	0.553704
10510	1	0.559259
10515	1	0.561111
10520	1	0.562963
10522	2	0.563704
10527	1	0.565556
10540	1	0.57037
10545	2	0.572222
10565	1	0.57963
10595	6	0.590741
10630	2	0.603704
10655	2	0.612963
10665	1	0.616667
10670	3	0.618519
10705	2	0.631481
10745	3	0.646296
10760	1	0.651852
10765	2	0.653704
10795	1	0.664815
10810	1	0.67037
10865	1	0.690741
10885	2	0.698148
10890	2	0.7
10895	1	0.701852
10900	2	0.703704
10935	1	0.716667
10953	1	0.723333
10960	2	0.725926
10974	2	0.731111
10990	3	0.737037
11015	2	0.746296
11030	1	0.751852

Time (in sec)	No. of persons	X
11045	1	0.757407
11060	2	0.762963
11075	1	0.768519
11080	1	0.77037
11082	1	0.771111
11083	1	0.771481
11085	1	0.772222
11093	3	0.775185
11105	1	0.77963
11115	1	0.783333
11155	2	0.798148
11162	1	0.800741
11185	2	0.809259
11200	1	0.814815
11230	2	0.825926
11260	1	0.837037
11285	2	0.846296
11290	4	0.848148
11305	2	0.853704
11310	1	0.855556
11320	1	0.859259
11345	1	0.868519
11380	4	0.881481
11385	4	0.883333
11395	4	0.887037
11405	1	0.890741
11440	1	0.903704
11445	1	0.905556
11465	1	0.912963
11500	1	0.925926
11505	3	0.927778
11555	1	0.946296
11560	1	0.948148
11565	2	0.95
11600	2	0.962963
11615	1	0.968519
11670	1	0.988889

There were 201 number of arrivals. Here, $\hat{\alpha} = 2.881696$.

TABLE FOR GOODNESS OF FIT

Class intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square statistic
0 - 0.1	6	0.05933034	11.925399	2.944166
0.1 - 0.2	9	0.06595408	13.25677	1.366856
0.2 - 0.3	14	0.0733173	14.736778	0.036836
0.3 - 0.4	22	0.08150257	16.382016	1.926609
0.4 - 0.5	17	0.09060165	18.210932	0.080521
0.5 - 0.6	34	0.10071657	20.244031	9.347283
0.6 - 0.7	21	0.11196074	22.504108	0.10053
0.7 - 0.8	32	0.12446022	25.016504	1.949482
0.8 - 0.9	31	0.13835517	27.809388	0.366064
0.9 - 1	15	0.15380137	30.914075	8.192313
Total	201	1	201	26.31066

Here, Cal. $\chi^2_9 = 26.31066$, Tab. $\chi^2_{9, 0.05} = 16.91896$ and p-value is 0.001817.

Data Set 6.3.2

Padma Theatre on 1st May 2003 (5.30p.m. to 6.15 p.m.)

Time (in sec)	No. of persons	X
225	4	0.083333
255	5	0.094444
270	1	0.1
275	2	0.101852
340	2	0.125926
405	1	0.15
445	1	0.164815
620	3	0.22963
750	1	0.277778
790	3	0.292593
805	2	0.298148
820	1	0.303704
880	1	0.325926
1025	3	0.37963
1230	7	0.455556
1240	1	0.459259
1280	3	0.474074
1285	1	0.475926
1555	3	0.575926
1570	2	0.581481
1590	2	0.588889
1655	1	0.612963
1670	2	0.618519
1685	1	0.624074

Time (in sec)	No. of persons	X
1690	1	0.625926
1695	1	0.627778
1850	1	0.685185
1890	1	0.7
1945	1	0.72037
1965	2	0.727778
2015	1	0.746296
2020	1	0.748148
2050	2	0.759259
2080	3	0.77037
2082	1	0.771111
2120	2	0.785185
2165	3	0.801852
2240	4	0.82963
2315	3	0.857407
2320	1	0.859259
2325	3	0.861111
2415	2	0.894444
2445	2	0.905556
2475	3	0.916667
2555	1	0.946296
2600	2	0.962963
2610	1	0.966667
2625	2	0.972222

There were 97 number of arrivals. Here, $\hat{\alpha} = 2.069178$.

TABLE FOR GOODNESS OF FIT

Class intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square statistic
0 - 0.1	9	0.07054405	6.8427733	0.68007912
0.1 - 0.2	7	0.07586478	7.3588836	0.01750231
0.2 - 0.3	9	0.08158682	7.9139212	0.14904965
0.3 - 0.4	5	0.08774043	8.510822	1.44825857
0.4 - 0.5	12	0.09435818	9.1527435	0.88573112
0.5 - 0.6	7	0.10147507	9.8430814	0.82119729
0.6 - 0.7	7	0.10912874	10.585487	1.21446653
0.7 - 0.8	14	0.11735968	11.383889	0.60120383
0.8 - 0.9	16	0.12621143	12.242509	1.15325532
0.9 - 1	11	0.13573082	13.16589	0.35630549
Total	97	1	97	7.32704923

Here, Cal. $\chi^2_9 = 7.327049$, Tab. $\chi^2_{9, 0.05} = 16.91896$ and p-value is 0.60311.

Data Set 6.3.3

Uma Theatre on 1st May 2003 (9.00 p.m. to 9.30 p.m.)

Time (in sec)	No. of persons	X
610	4	0.338889
735	3	0.408333
915	2	0.508333
1000	1	0.555556
1010	3	0.561111
1050	5	0.583333
1170	2	0.65
1270	2	0.705556
1300	2	0.722222
1335	2	0.741667
1355	1	0.752778
1365	2	0.758333
1530	3	0.85
1550	1	0.861111
1555	2	0.863889
1565	1	0.869444
1568	2	0.871111
1572	1	0.873333
1575	1	0.875

Time (in sec)	No. of persons	X
1632	1	0.906667
1635	2	0.908333
1640	1	0.911111
1645	1	0.913889
1650	2	0.916667
1655	4	0.919444
1665	1	0.925
1667	1	0.926111
1670	2	0.927778
1672	2	0.928889
1675	2	0.930556
1678	2	0.932222
1685	1	0.936111
1690	2	0.938889
1692	1	0.94
1700	5	0.944444
1705	1	0.947222
1710	1	0.95
1712	1	0.951111

1580	2	0.877778
1590	2	0.883333
1600	1	0.888889
1602	1	0.89
1615	2	0.897222
1618	1	0.898889
1630	2	0.905556

1720	2	0.955556
1725	2	0.958333
1738	2	0.965556
1755	1	0.975
1770	2	0.983333
1785	1	0.991667
1795	2	0.997222

There were 96 number of arrivals. Here, $\hat{\alpha} = 197.4751$.

TABLE FOR GOODNESS OF FIT

Class intervals	Obs. Freq.	Probability	Exp. Freq.	Chi-square statistic
0 - 0.1	0	0.00354493	0.3403129	0.340313
0.1 - 0.2	0	0.00601393	0.5773373	0.577337
0.2 - 0.3	0	0.01020257	0.9794468	0.979447
0.3 - 0.4	4	0.01730856	1.6616215	3.29077
0.4 - 0.5	3	0.02936379	2.8189239	0.011632
0.5 - 0.6	11	0.04981537	4.7822754	8.084039
0.6 - 0.7	2	0.08451126	8.1130813	4.606112
0.7 - 0.8	9	0.14337249	13.763759	1.648779
0.8 - 0.9	20	0.24322996	23.350076	0.480641
0.9 - 1	47	0.41263714	39.613165	1.377454
Total	96	1	96	21.39652

Here, Cal. $\chi^2_4 = 21.39652$, Tab. $\chi^2_{4, 0.05} = 9.487728$ and p-value is 0.0002642.

Remarks :

The model gives a good fit in case of data set 3.B while does not give good fit in cases of data sets 3.A and 3.C. In case of data set 3.C, there were no arrivals for the first 10 minutes, as such by introducing the lower limit of the variable as a parameter, the enlarged family of distributions may be considered. In this case, a large contribution to Chi-square value is from the class 0.5 to 0.6 and is due to very high observed value for this class and more over the next two classes have lesser frequencies. We have not made any attempts to make further possible reasons for such a great disturbance. By pulling the first five class intervals, the p-value for Chi-square increases from 0.0002642 to 0.0061454.

Appendix – 1.1

```

#include<stdio.h>
#include<math.h>
#include<stdlib.h>
#undef abs
#include<conio.h>
main()
{
    int i,j,k,m,n;
    float
        ialpha,norm,alpha,sum1,sum2,sums,mean1,mean,alpha1[10000]
        ,num,deno,sum3=0.0,mean2,galpha;
    double u,x[5000],delta=0.0000000000001,diff;
    FILE *fp;
    clrscr();
    randomize();
    fp=fopen("arati3.dat","w+");
    printf("Alpha=");
    scanf("%f",&alpha);
    printf("Number of observations=");
    scanf("%d",&n);
    printf("Number of iterations=");
    scanf("%d",&m);
    sums=0;
    ialpha=(1/(pow((alpha*log(alpha)),2)))-(1/(alpha*pow((alpha-
        1),2)));
    for(j=0;j<m;j++)
    {
        sum1=0;sum2=0;
        for(i=0;i<n;i=i+2)
        {
            u=(float)random(32000)/RAND_MAX;
            x[i]=log(1+(alpha-1)*u)/log(alpha);
            sum1=sum1+x[i];
        }
        for(i=1;i<n;i=i+2)
        {
            u=(float)random(32000)/RAND_MAX;
            u=1-u;
            x[i]=log(1+(alpha-1)*u)/log(alpha);
            sum2=sum2+x[i];
        }
        mean=(sum1+sum2)/n;
        sum3=sum3+sum1+sum2;
        alpha1[0]=mean+0.5;
        k=0;
        do
        {
            num=(1/log(alpha1[k]))-(alpha1[k]/(alpha1[k]-1))+mean;
            deno=(1/(alpha1[k]*pow(log(alpha1[k]),2)))-
                (1/pow((alpha1[k]-1),2));
            alpha[k+1]=alpha[k]+(num/(n*deno));
            if(alpha[k+1]<0)
            {

```

```

        if(num>0 && deno>0 || num<0 && deno<0)
            alphas[k+1]=alphas[k]/2;
        }
        k=k+1;
        diff=((alphas[k]-alphas[k-1])/alphas[k]);
        if(diff<0)
            diff=diff*(-1);
        }
        while(diff>delta);
        // printf("\n sol=%f\n", alphas[k]);
        norm=sqrt(n*ialpha)*(alphas[k]-alpha);
        fprintf(fp, "%f\n", norm);
        // fprintf(fp, "%f\n", alphas[k]);
        sums=sums+alphas[k];

    }
    mean1=sums/m;
    printf("simmean=%f\t", mean1);
    mean2=(sum3/m)/n;
    printf("xbarbar=%f\t", mean2);
    alphas[0]=mean2+0.5;
    k=0;
    do
    {
        num=(1/log(alphas[k]))-(alphas[k]/(alphas[k]-1))+mean2;
        deno=(1/(alphas[k]*pow(log(alphas[k]),2)))-
            (1/pow((alphas[k]-1),2));
        alphas[k+1]=alphas[k]+(num/(n*deno));
        if(alphas[k+1]<0)
        {
            if(num>0 && deno>0 || num<0 && deno<0)
                alphas[k+1]=alphas[k]/2;
        }
        k=k+1;
        diff=((alphas[k]-alphas[k-1])/alphas[k]);
        if(diff<0)
            diff=diff*(-1);
        }
        while(diff>delta);
        printf("sol=%f\n", alphas[k]);
        galphas=(alpha/(alpha-1))-(1/log(alpha));
        printf("\t galphas=%f", galphas);
        getch();
        fclose(fp);
    }
}

```

Appendix – 1.2

A Simulation study for the performance of m.l.e based on 1000 samples

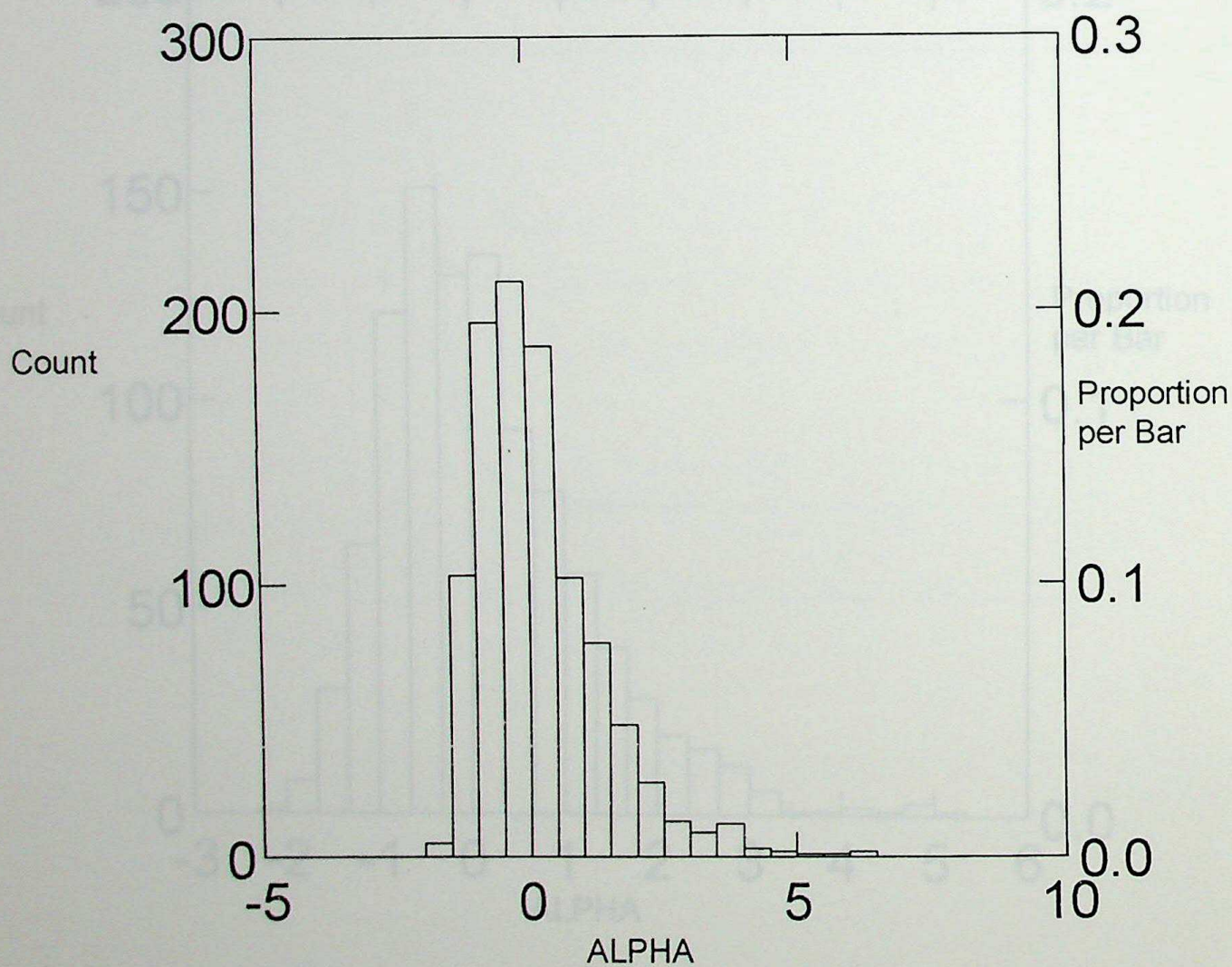
α	N	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$g(\alpha)$	$\bar{x}$
7	10	24.1480	7.4376	0.652768	0.656969
	30	10.2216	7.4783		0.657345
	50	8.7825	7.3055		0.655733
	100	8.0228	7.3948		0.656572
	500	7.5166	7.4056		0.656673
5	10	14.0719	4.9816	0.628665	0.628395
	30	6.5441	5.0868		0.629930
	50	6.2156	5.2266		0.631915
	100	5.5678	5.1533		0.630882
	500	5.2451	5.1692		0.631110
0.7	10	1.3585	0.7333	0.470340	0.474195
	30	0.8558	0.6993		0.470269
	50	0.7774	0.6959		0.469862
	100	0.7422	0.7004		0.470396
	500	0.7053	0.6975		0.470051

($\bar{x}$: mean of M sample means, $g(\alpha)$ - population mean)

Appendix – 1.3

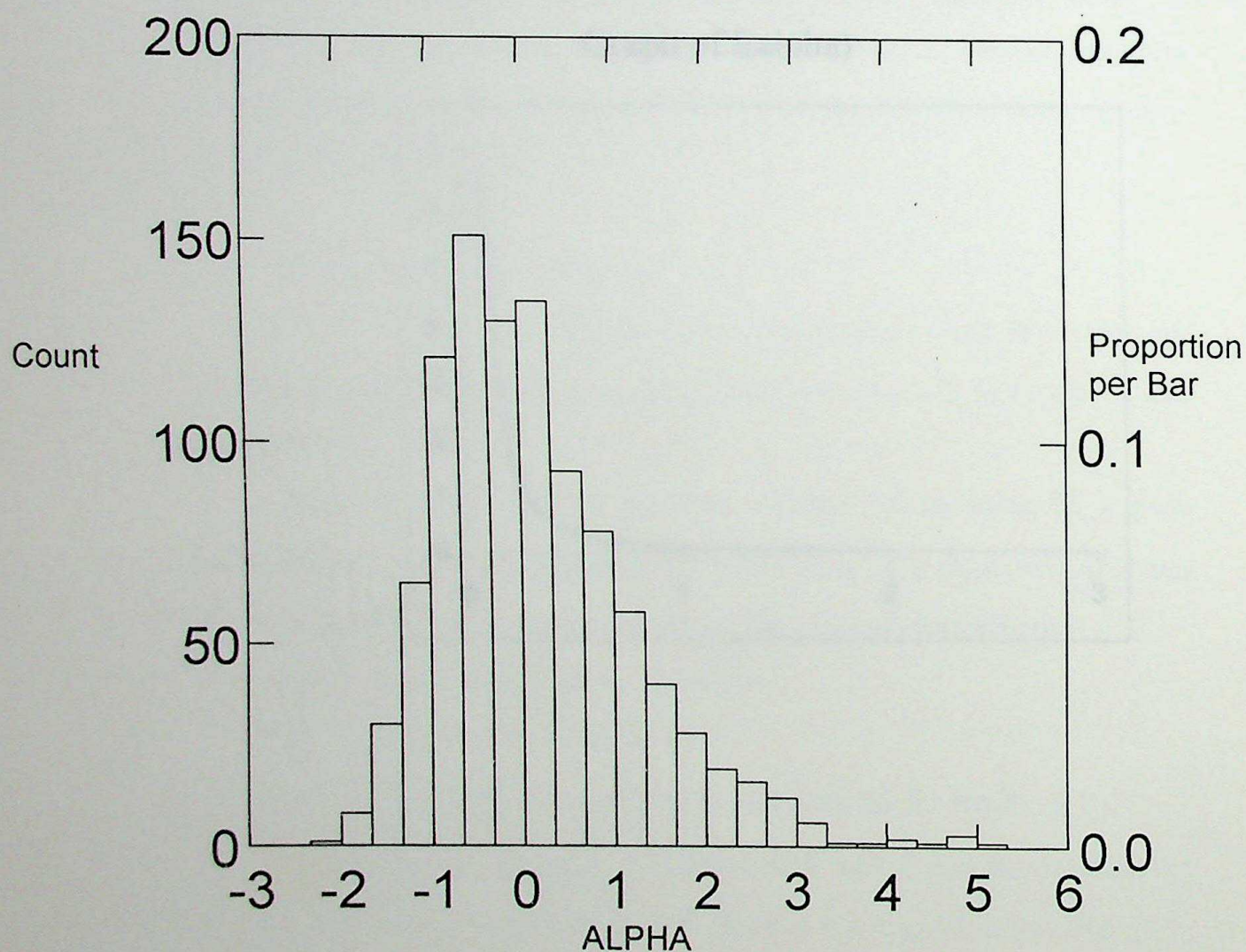
Frequency Distribution of normalised m.l.e.

$\text{Alpha} = 0.5, n = 50, m = 1000.$



Frequency Distribution of normalised m.l.e.

$\text{Alpha} = 0.5, n = 100, m = 1000.$



Unit 2

Models Based on Composite Functions

Abstract

Appendix – 1.4

Graph of $I(\alpha)$ 