OLD AND NEW METHODS OF ESTIMATION AND

THE PARETO DISTRIBUTION

by

Richard E. Quandt

OLD AND NEW METHODS OF ESTIMATION AND
THE PARETO DISTRIBUTION[*]

1. Introduction

Recently much attention has been paid to the statistical distribution of certain socio-economic quantities such as personal incomes, the assets of firms, the sizes of cities and the number of firms in various industries.[1] Some authors posit that the quantities in question are generated by stochastic processes which have as limiting distribution either the Pareto distribution, or the lognormal distribution, or some other distribution strongly skewed to the right. In view of the fact that some of the competing distributions are fairly similar, it becomes relevant to test the hypothesis that data have been generated by a particular distribution against the alternative hypothesis that some particular other distribution(s) is (are) responsible for generating the data. Since many of the candidate hypotheses closely resemble each other, the problem of estimating the parameters of the various distributions is by no means trivial.[2]

---

1. See [2], [9], [11], [13], [14]. The intellectual antecedents of these studies can be found in the works of Pareto, Gibrat and others. See [7], [10].

2. For a more detailed discussion of discriminating between rival hypotheses, see [11].

The objective of this paper is to discuss various methods of estimating the Pareto distribution which has been one of the most distinguished candidates for the honor of explaining the distribution of incomes, assets, etc. Some of the methods discussed are traditional; one is probably novel and appears to be sufficiently promising as to be generally usable in problems of estimating the parameters of distributions. Section 2 is devoted to a discussion of various traditional methods of estimating the Pareto distribution. Section 3 presents the results of some sampling experiments with these methods. Section 4 discusses a new approach to estimation and analyzes the results of pertinent sampling experiments. Section 5 contains some concluding remarks.

2. Estimation of the Pareto Distribution

We distinguish between the distribution function $F(x)$ and the density function $f(x)$ of a random variable x where $F(x) = \int_{-\infty}^{x} f(\xi)d\xi$ is the probability that the random variable assumes a value $\leq x$. The Pareto distribution[3] is given by

$$F(x) = 1 - \left(\frac{k}{x}\right)^a \qquad (2\text{-}1)$$

$k > 0$, $a > 0$ and $x \geq k$. Its parameters k and a (where k is the lower bound of the random variable x) can be estimated by a variety of methods. These methods are discussed and some theorems about the

---

3. More properly called the Pareto distribution of the first kind since Pareto himself proposes three distributions. The Pareto distribution of the second kind is $F(x) = 1 - \dfrac{K}{(x + c)^a}$ and that of the third kind is $F(x) = 1 - Ke^{-bx}/(x + c)^a$. See [10].

properties of the estimates are proved below. Specifically, we shall be concerned with the consistency of the estimates on the grounds that the convergence in probability of the estimates to the true values represents a minimum standard of acceptability.

The Method of Moments. Provided that $a > 1$, the mean of the Pareto distribution exists and is given by

$$E(x) = \int x dF(x) = \int_k^\infty x \frac{ak^a}{x^{a+1}} dx = \frac{ak}{a-1} \qquad (2\text{-}2)[4]$$

We estimate a by equating (2-2) to the sample mean $\bar{x}$, yielding

$$\hat{a} = \frac{\bar{x}}{\bar{x} - \hat{k}} \qquad (2\text{-}3)$$

where $\hat{k}$ is some estimate of k.

The estimation of k from samples of n observations is accomplished as follows: the probability of an observation greater than x is $(k/x)^a$ from (2-1). Hence, the probability that all n sample values $x_1, \ldots, x_n$ are greater than x is $(k/x)^{an}$. This is, therefore, also the probability that the lowest sample value is greater than x. Thus the probability distribution of the lowest sample value is

$$G(x) = 1 - \left(\frac{k}{x}\right)^{an} . \qquad (2\text{-}4)$$

The corresponding density function is

$$g(x) = \frac{ank^{an}}{x^{an+1}}$$

and the expected value of the lowest sample observation is

---

4. It is well known that for $a < 2$ the variance does not exist.

$$\int_{k}^{\infty} \frac{ank^{an}}{x^{an}} dx = \frac{ank}{an - 1} \ .$$ (2-5)

Equating the lowest sample value, x_o, to the expected value,[5] we obtain

$$\hat{k} = \frac{(an - 1)x_o}{an}$$ (2-6)

and therefore

$$\hat{a} = \frac{\overline{x}}{\overline{x} - \frac{\hat{a}n - 1}{\hat{a}n}x_o}$$

yielding

$$\hat{a} = \frac{n\overline{x} - x_o}{n(\overline{x} - x_o)} \ .$$ (2-7)

The estimators are thus given by (2-6) and (2-7).

<u>Theorem 1.</u> The method of moments yields consistent

estimates.

<u>Proof</u>: Since $\text{plim}\, x_o = k$, and since the sample mean is a consist-

ent estimator of the population mean,[6] (2-6) yields immediately

$$\text{plim}\, \hat{k} = k \ .$$ (2-8)

Taking probability limits in (2-7) we obtain

$$\text{plim}\, \hat{a} = \text{plim} \frac{n\overline{x} - x_o}{n(\overline{x} - x_o)} =$$

$$= \text{plim} \frac{n\frac{ak}{a - 1} - k}{n\left(\frac{ak}{a - 1} - k\right)} = \text{plim}\left(a + \frac{1 - a}{n}\right) = a$$ (2-9)

5. By the subscript o , we denote the least of the n values $x_1, \dots, x_n$.

6. Without having to require that the variance of the random variable
   exists. See [4], pp. 228-233.

as asserted.

The Method of Maximum Likelihood. The likelihood function for a sample $(x_1, \ldots, x_n)$ is

$$\mathcal{L} = \frac{a^n k^{an}}{\left(\prod_i x_i \right)^{a+1}}$$

and taking logarithms,

$$L = n \log a + an \log k - (a + 1) \sum_i \log x_i .$$

Hence

$$\frac{\partial L}{\partial a} = \frac{n}{a} + n \log k - \sum_i \log x_i = 0$$

yielding for a the estimate

$$\hat{a} = \frac{n}{\sum_i \log \frac{x_i}{\hat{k}}} . \tag{2-10}$$

A maximum likelihood estimate cannot be obtained for k by differentiating L with respect to k since L is unbounded with respect to k. But since k is the lower bound of the random variable x, we may maximize L subject to the constraint

$$\hat{k} \leqq \min_i x_i . \tag{2-11}$$

Clearly $\mathcal{L}$ is maximized with respect to k subject to (2-11) when

$$\hat{k} = \min_i x_i \tag{2-12}$$

which is, therefore, the maximum likelihood estimate for k.

Since the partial derivatives of the likelihood function do not all vanish at the maximum, we convince ourselves of the consistency of

the maximum likelihood estimates by the following argument.

Theorem 2. The maximum likelihood estimates are consistent.

Proof: We first observe that $\text{plim}\left|k - \min_i x_i\right| = 0$, and hence k
is a consistent estimator. Rewriting (2-10) we have

$$\hat{a} = \frac{1}{\dfrac{\sum_i \log x_i}{n} - \log \hat{k}} \quad . \tag{2-13}$$

The consistency of $\hat{a}$ can be established if we can show that $\text{plim} \dfrac{\sum_i \log x_i}{n} = \log k + \dfrac{1}{a}$. But $\left(\sum_i \log x_i\right)/n$ is the arithmetic mean of a random variable

$$y = \log x \; . \tag{2-14}$$

Transforming the Pareto density according to the transformation (2-14) we obtain

$$f(y)dy = ak^a e^{ay}dy \; . \tag{2-15}$$

Since the sample arithmetic mean is a consistent estimator of the mean of the distribution, we require $E(y)$. But

$$E(y) = \int_{\log k}^{\infty} ak^a ye^{-ay}dy = \log k + \frac{1}{a} \; .$$

Then

$$\text{plim}\,\hat{a} = \text{plim} \frac{1}{\dfrac{\sum_i \log x_i}{n} - \log \hat{k}} = \frac{1}{\log k + \dfrac{1}{a} - \log k} = a$$

as asserted.

Quantile Methods. Choose two probability levels P_1 and P_2 and determine the corresponding quantiles x_1 and x_2 from

$$P_1 = 1 - \left(\frac{k}{x_1}\right)^a \tag{2-16}$$

and

$$P_2 = 1 - \left(\frac{k}{x_2}\right)^a .$$
(2-17)

Then we obtain an estimate for a by solving the above equations to
yield

$$\hat{a} = \frac{\log \dfrac{1 - P_1}{1 - P_2}}{\log \dfrac{x_2}{x_1}}$$
(2-18)

which, when substituted into (2-16) or (2-17) yields the corresponding
estimate for k .

Theorem 3. The quantile estimates are consistent.

Proof: We merely have to observe that sample quantiles are con-
sistent estimators of the population quantiles.

Least Squares Estimates. The (cumulative) distribution function
can be rewritten as

$$1 - F(x) = \left(\frac{k}{x}\right)^a$$

and taking logarithms on both sides

$$\log(1 - F(x)) = a \log k - a \log x .$$
(2-19)

The parameters of equation (2-19) are estimated by least squares where
the dependent variable is the logarithm of 1 minus the sample cumulative
distribution. For the same reason as in Theorem 3, the estimates are con-
sistent.

In addition to mixed methods by which several of the above methods
are combined, we may single out for special mention a class of methods
which could best be designated as

Qualitative Methods. These methods have as their purpose not so much the precise estimation of the parameters k and a but rather the verification that the sample is generated by the Pareto distribution at all.

Such a qualitative method consists of examining the Lorenz curve of the sample. The Lorenz curve is frequently employed in studies of income distribution and is a locus of points such that the ordinate of each point represents the fraction of income accruing to that fraction of recipients which is the abscissa of that point. We can define the Lorenz curve parametrically as follows: letting $F(x)$ be the distribution of the random variable x we have

$$\text{for abscissa:} \quad F(x)$$

$$\text{for ordinate:} \quad F_1(x) = \frac{\int_k^x \xi \, dF(\xi)}{\int_k^\infty x \, dF(x)}$$

where $F_i(x)$ is referred to as the ith moment distribution function corresponding to $F(x)$ and where k is the lower bound of the random variable x .[7] As a measure of the inequality of distribution we use the coefficient defined by

$$L = 1 - \int_k^\infty F_1(x) \, dF(x)$$

i.e., 1 minus the area under the Lorenz curve. Clearly the Lorenz measure and curve are not defined when $E(x)$ does not exist; hence we restrict ourselves now to cases in which $a > 1$.

---

7. See [1].

<u>Theorem 4</u>. The Lorenz measure for the Pareto distribution

is $a/(2a - 1)$.

<u>Proof</u>: The first moment distribution of the Pareto distribution

is

$$F_1(x) = \frac{1}{ak/(a - 1)} \int_k^x \frac{ak^a}{\xi^{a+1}} \xi \, d\xi = 1 - \left(\frac{k}{x}\right)^{a-1}.$$

This provides values of the ordinate of the Lorenz curve; the abscissae

are given by the cumulative distribution $1 - (k/x)^a$. The Lorenz meas-

ure therefore is

$$L = 1 - \int_k^\infty \left(1 - \frac{k^{a-1}}{x^{a-1}}\right) d\left(1 - \frac{k^a}{x^a}\right) = \frac{a}{2a - 1}$$

as asserted.

For purposes of comparison with other distributions, it is of

interest to examine the symmetry of the Lorenz curve arising from the

Pareto distribution. We shall establish that the Lorenz curve is not

symmetric about the 45 degree line perpendicular to the line of equal

distribution and that the point at which the slope of the curve equals

unity occurs above the line which is perpendicular to the line of equal

distribution. Hence, possibly, we may determine whether a sample has

been generated by the Pareto distribution by examining the sample Lorenz

curve. We shall refer to the line perpendicular to the line of equal

distribution as the <u>alternate diagonal</u>.

<u>Theorem 5</u>. The slope of the Lorenz curve equals unity at

the value $x = E(x)$.

<u>Proof</u>: Denoting the Lorenz curve in the parametrized form

$$y = 1 - \left(\frac{k}{x}\right)^{a-1} \tag{2-20}$$

$$z = 1 - \left(\frac{k}{x}\right)^a \qquad (2\text{-}21)$$

where y and z are the ordinate and abscissa respectively, we obtain the slope

$$\frac{dy}{dz} = \frac{dy/dx}{dz/dx} = \frac{(a-1)x}{ak}$$

which equals unity when $x = ak/(a-1)$ which is $E(x)$ by (2-2).

Theorem 6. The point on the Lorenz curve corresponding to $x = ak/(a-1)$ is above the alternate diagonal.

Proof: By the definition of the Lorenz curve, the intersection of the alternate diagonal with the Lorenz curve occurs at the point given by

$$F(x) = 1 - F_1(x) \qquad (2\text{-}22)$$

and substituting from (2-20) and (2-21) we obtain

$$1 - \left(\frac{k}{x}\right)^a = \left(\frac{k}{x}\right)^{a-1}$$

or

$$1 = \frac{k^{a-1}(k+x)}{x^a} . \qquad (2\text{-}23)$$

We can rewrite (2-23) by considering the right hand side as a function of x as

$$\phi = \frac{k^{a-1}(k+x)}{x^a} \qquad (2\text{-}24)$$

and evaluate ϕ at the point $x = ak/(a-1)$. Substituting directly into (2-24)

$$\phi = \frac{1 + \dfrac{a}{a-1}}{\dfrac{a^a}{(a-1)^a}} =$$

$$= \left[\frac{a-1}{a}\right]^{a-1}\left[\frac{2a-1}{a}\right] \qquad (2\text{-}25)$$

$$= (a - 1)^{a-1} \left(\frac{1}{a}\right)^{a} (2a - 1) \ . \qquad (2\text{-}26)$$

Clearly as $a \to 1$, $\phi \to 1$ since $\lim_{u \to 0} u^{u} = 1.$ [8]   Moreover, for values

of a close to unity,

$$\frac{1}{\phi} \frac{d\phi}{da} = \log\left(1 - \frac{1}{a}\right) + \frac{2}{2a - 1} < 0 \ . \qquad (2\text{-}27)$$

We also observe from (2-25) that as $a \to \infty$, $\phi \to 2e^{-1}$ which is less than

unity. Since ϕ is continuous and differentiable for $a > 1$, it can be-

come greater than 1 only if it has at least two extreme points in the

range $a > 1$, one of which must be a maximum. However

$$\frac{1}{\phi} \frac{d^{2}\phi}{da^{2}} = \frac{1}{\phi^{2}} \left(\frac{d\phi}{da}\right)^{2} + \frac{1}{a(a - 1)(2a - 1)^{2}}$$

and since $\phi > 0$ for $a > 1$, $d^{2}\phi/da^{2} > 0$ for all $a > 1$ and hence ϕ

can have no maximum. Consequently $\phi < 1$ for all $a > 1$ and the value

$E(x)$ does not satisfy the intersection of the Lorenz curve and the alter-

nate diagonal. We finally note that ϕ diminishes as x increases.

Thus the value of x at which the slope of the Lorenz curve is unity is

greater than the value at which it intersects the alternate diagonal and

this point is therefore above the alternate diagonal, as asserted.

Such qualitative considerations for establishing that a sample

has been generated from the Pareto distribution are fairly crude. The

property derived in Theorem 6 distinguishes the Lorenz curve of the

Pareto distribution from the Lorenz curve of, say, the exponential dis-

tribution but not from the Lorenz curve of the three-parameter family

of lognormal distributions. Another qualitative device, the measurement

---

8. This is immediate from l'Hôpital's rule.

of the moments of sequentially increasing samples, may distinguish the Pareto distribution (for which the sample moments may have a clear tendency to become unbounded)[9] from the lognormal distribution but not from other distributions with infinite means and/or variances. Although these methods are not without interest one obviously cannot place excessive reliance on them.

3. Sampling Experiments

For purposes of comparing the various methods by sampling experiments the Pareto distribution with $k = 1.0$ and $a = 1.5$ was chosen.[10] Pareto distributed samples were generated by generating uniformly distributed pseudo-random deviates between 0 and 1. For each such deviate u_i we determined the corresponding Pareto deviate x_i by solving $u_i = 1 - 1/x_i^{1.5}$ for x_i.[11] The sample size N took on the values 25, 50, 100, 300, 500, 1000, 2000. For each sample size 100 samples were generated and k, a, and the value of the Lorenz coefficient were estimated for each sample by the method of moments, the method of least squares, the method of maximum likelihood and four quantile methods. These latter differ from each other only in that they are based on different quantiles. Quantile methods 1, 2, 3, and 4 respectively use deciles 1 and 9, 2 and 8, 3 and 7, and 4 and 6. Table 1 displays the mean estimates over one

---

9. See the sampling experiments discussed in [9].

10. If a is less than 1, the method of moments will not yield consistent estimates.

11. By the well known proposition that, for any distribution $F(x)$, the values of $F(x)$ itself are uniformly distributed on the $[0,1]$ interval. The u_i play a role only in generating the samples and are not known for estimating purposes.

hundred samples; Table 2 contains the root mean square error, and Table 3 the predicted Lorenz coefficients. Since the true value of a is 1.5, the theoretical value of the Lorenz coefficient is $L = .75$ by Theorem 4.

TABLE 1. Mean Estimates of Parameters

	N = 25		N = 50		N = 100		N = 300		N = 500		N = 1000		N = 2000	
	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$
Method of Moments	1.009	1.721	1.001	1.663	1.000	1.617	1.000	1.570	1.000	1.550	1.000	1.533	1.000	1.522
Method of Least Squares	1.013	1.680	1.012	1.612	1.013	1.592	1.007	1.544	1.005	1.527	1.002	1.511	1.001	1.500
Maximum Likelihood	1.033	1.626	1.014	1.558	1.007	1.540	1.002	1.517	1.001	1.509	1.001	1.503	1.000	1.497
Quantile Method 1	.997	1.511	.995	1.432	1.000	1.498	1.001	1.494	1.000	1.492	.999	1.492	1.000	1.492
Quantile Method 2	.999	1.464	1.004	1.493	1.001	1.504	1.001	1.501	.999	1.499	.998	1.496	1.000	1.496
Quantile Method 3	.984	1.545	1.003	1.520	1.006	1.531	1.003	1.516	1.002	1.512	.999	1.497	1.001	1.500
Quantile Method 4	.991	1.628	.985	1.519	.986	1.491	1.001	1.522	.999	1.509	.999	1.504	.999	1.497

TABLE 2. Root Mean Square Errors

	N = 25		N = 50		N = 100		N = 300		N = 500		N = 1000		N = 2000	
	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$	$\hat{k}$	$\hat{a}$
Method of Moments	.037	.386	.013	.288	.007	.217	.002	.133	.002	.098	.001	.081	.000	.065
Method of Least Squares	.093	.430	.071	.311	.049	.222	.032	.124	.026	.096	.019	.065	.014	.045
Maximum Likelihood	.049	.356	.019	.240	.009	.166	.003	.087	.002	.068	.001	.048	.000	.032
Quantile Method 1	.060	.412	.032	.294	.023	.211	.014	.127	.011	.090	.008	.065	.005	.044
Quantile Method 2	.124	.363	.063	.281	.043	.209	.026	.133	.020	.105	.012	.063	.009	.041
Quantile Method 3	.172	.621	.078	.305	.058	.214	.032	.121	.028	.106	.018	.076	.013	.053
Quantile Method 4	.069	.895	.117	.450	.077	.290	.047	.179	.041	.149	.027	.104	.018	.071

TABLE 3. Mean Lorenz Coefficients

	N = 25	N = 50	N=100	N=300	N=500	N=1000	N=2000
Method of Moments	.719	.725	.730	.736	.740	.743	.746
Method of Least Squares	.736	.742	.737	.742	.745	.748	.750
Maximum Likelihood	.741	.749	.746	.748	.749	.750	.751
Quantile Method 1	.783	.793	.762	.756	.754	.753	.753
Quantile Method 2	.796	.774	.761	.754	.753	.752	.752
Quantile Method 3	.796	.768	.754	.750	.750	.752	.751
Quantile Method 4	.796	.783	.773	.752	.753	.752	.752

We shall consider five individual performance criteria: (i) the mean bias in $\hat{k}$, (ii) the mean bias in $\hat{a}$; (iii) the root mean square error of $\hat{k}$; (iv) the root mean square error of $\hat{a}$; (v) the mean bias in L. The tables reveal the following:

(1) By all criteria Quantile Methods 1 and 2 perform better than Methods 3 and 4. The values .15 and .85 seem therefore reasonable values to use for P_1 and P_2 in the Quantile Method.

(2) Since Quantile Methods 1 and 2 are comparable with each other in performance, we shall choose (arbitrarily) Quantile Method 1 to represent the best of the quantile methods. For each of the five criteria, we obtained a table of rankings in which the four methods[12] are ranked by each of the seven sets of samples corresponding to the seven values of N. For each of these tables of rankings Kendall's coefficient of concordance W was calculated. The W values are displayed in Table 4.

---

12. Method of Moments, Least Squares, Maximum Likelihood, Quantiles-1.

TABLE 4.   W Statistic for Ranking of Estimating Methods

.	Bias $\hat{k}$	Bias $\hat{a}$	RMSE $\hat{k}$	RMSE $\hat{a}$	L
W	.718	.755	.974	.683	.698

These values are all significant on the .01 level and allow one to reject the null hypothesis that the rankings of methods according to different sample sizes are random.

(3) If rank totals are obtained for each method and criterion, one can rerank the methods (on the basis of the rank totals) by each of the five criteria. Kendall's W calculated from the resulting table of rankings is .362 which is not significant on the .05 level and does not allow one to reject the null hypothesis that the rankings of the different methods by the five criteria are random. The rankings of the four methods by the five criteria are displayed in Table 5.

TABLE 5.   Rankings of Methods

Method	Criterion				
	Bias $\hat{k}$	Bias $\hat{a}$	RMSE $\hat{k}$	RMSE $\hat{a}$	L Bias
Moments	1	4	1	4	4
Least Squares	4	3	4	3	2
Maximum Likelihood	3	2	2	1	1
Quantiles	2	1	3	2	3

(4) The related Friedman two-way analysis of variance test yields a χ^2 value (with 3 degrees of freedom) of 3.48 which does not permit one to reject the null hypothesis that the rank totals are not significantly

different. By inspection of Table 5 one would single out the maximum likelihood and quantile estimates as yielding best results but such a distinction is casual and does not rest on a probabilistic basis.[13]

(5) The performances of the methods as measured by the bias and by the root mean square error of a given parameter are highly correlated. Also, the performances of the methods as measured by (the biases or root mean square errors of) k and a are substantially divergent.

(6) Since L depends only on a , it is not surprising that the ranking of methods according to L is substantially more similar to the rankings according to $\hat{a}$ than to those according to $\hat{k}$.

4. A New Method of Fitting

The previous sections discussed various traditional methods of fitting distributions to samples with reference to the Pareto distribution. No sharp differences in performance were found among the various methods but informally the quantile and maximum likelihood methods seemed to have the edge.

All of the traditional methods, however, suffer from a distinct disadvantage. Broadly speaking, this disadvantage is that traditional methods of fitting do not allow one to discriminate statistically among competing but mathematically closely related alternative hypotheses.

Suppose, for example, that one hypothesizes that a given sample was generated by the Pareto distribution and that the alternative hypothesis is that the sample comes from, say, the lognormal distribution. It

---

13. One must note the difficulty of evaluating methods such as the Method of Moments which is best by some and worst by other criteria.

is easily possible and in fact frequently the case that two or more conflicting hypotheses appear to yield good fits.[14] In such instances one is typically dissatisfied with the nature of the criteria of goodness of fit as well as with the inconclusive results to which these criteria lead.

Standard Goodness of Fit Tests. Some of the goodness of fit tests that might be suggested as appropriate are the following.

(1) The χ^2 goodness of fit test. Accordingly the observations are grouped and the statistic

$$G = \sum_i \frac{(e_i - f_i)^2}{e_i}$$

is obtained, where e_i and f_i are the expected and actual frequencies in the ith group respectively. Under the null hypothesis that the parent of the sample is the distribution in question, G has approximately the χ^2 distribution with $r - k$ degrees of freedom, where r is the number of groups and k the number of parameters fitted. According to this approach one would declare that one of all competing hypotheses to be the winner which yields the G statistic representing the highest significance level. But this approach has several undesirable features: (a) the grouping of observations is arbitrary and if all candidate distributions fit fairly well, even small alterations in the method of grouping will tend to alter the resulting ranking of rival hypotheses; (b) the validity of the χ^2 test does not rest on any specific alternative hypothesis; therefore it is not strictly proper for evaluating the fit from distribution X_1 when the only alternatives are, say, distributions X_2 and X_3; (c) in any particular study interest

14. See [11].

may be focused on behavior in the right tail of the distribution where expected frequencies are small; small expected frequencies are, however, a violation of the conditions that must be fulfilled for a valid application of the test;[15] (d) the χ^2 test essentially ignores small but systematic deviations of the sample from the theoretical distribution.

(2) The Kolmogorov-Smirnov Test. Given a sample of n observations $x_1, \ldots, x_n$ and empirical and theoretical cumulative distribution functions $S(x)$ and $F(x)$, the statistic is

$$D = \max_i \left| S(x_i) - F(x_i) \right|$$

and measures the distance between the empirical and theoretical distributions. Accordingly the distribution yielding the smallest D statistic for a given sample would be declared to fit that sample best. This method also has serious disadvantages: (a) it shares difficulties (b) and (d) above with the χ^2 test; (b) critical values of the D statistic cannot be obtained when the parameters of the distribution have been estimated from the sample. Thus, even though the Kolmogorov-Smirnov test is probably more powerful than the χ^2 test in cases in which they can both be validly applied,[16] there is probably little reason for believing that either test is appropriate for present purposes.

A Widely Applicable Fitting Procedure. Denote by $F(x)$ the distribution to be estimated and $F(x_i)$ its value at the ith sample point. Let the ordered sample be $(x_1, \ldots, x_n)$ and let there be two fictitious points x_o and x_{n+1} such that $F(x_o) = 0$ and $F(x_{n+1}) = 1$. The quantity $F(x_i) - F(x_{i-1})$ has expected value of $1/(n+1)$ for all values of

15. See [3].

16. See [12], p. 51.

$i = 1, \ldots, n+1$ since each of the intervals $F(x_i) - F(x_{i-1})$ is identically distributed. The proposed procedure is to estimate the parameters of $F(x)$ by minimizing

$$S = \sum_{i=1}^{n+1} \left(F(x_i) - F(x_{i-1}) - \frac{1}{n+1} \right)^2 \qquad (4\text{-}1)$$

with respect to the parameters of the distribution. The minimization of (4-1) is generally feasible by gradient or other numerical methods if $F(x)$ is twice differentiable. The resulting estimates have the property of consistency as shown in the following

Theorem 7. If (1) $F(x)$ is a member of a k - parameter family of distributions, (2) if the parameters are continuous and single valued functions of the cumulative probability in the sense that k sample points x_i, $i = 1, \ldots, k$, $x_i \neq x_j$ for all i and j, are sufficient to determine unique values of the parameters, (3) if the null hypothesis that the x_i were generated by $F(x)$ is true, then the estimates resulting from minimizing $S = \sum_{i=1}^{n+1} \left(F(x_i) - F(x_{i-1}) - \frac{1}{n+1} \right)^2$ are consistent.

Proof: Consider the quantities $c_i = F(x_i) - F(x_{i-1})$, $i = 1, \ldots, n+1$, called the coverages corresponding to the order statistics $F(x_i)$. We shall need the probability distributions of the c_i. Since the $F(x_i)$ are uniformly distributed, so are the c_i and this distribution is identical for all c_i.[17] The first coverage c_1 is given by $c_1 = F(x_1) - 0$ and hence the required probability distribution is that of $F(x_1)$. Now

17. See [6], p. 151.

the probability that the first order statistic $F(x_1)$ is greater than or equal to some quantity z is

$$\Pr[F(x_1) \geq z] = (1 - z)^n .$$

Then

$$\Pr[F(x_1) \leq z] = 1 - (1 - z)^n$$

and

$$\Pr[c_i \leq z] = 1 - (1 - z)^n .$$

The corresponding density function is

$$f(z) = n(1 - z)^{n-1} .$$

Now let $y = nz$ and thus

$$f(y) = \left(1 - \frac{y}{n}\right)^{n-1}$$

and clearly $\lim_{n \to \infty} f(y) = e^{-y}$. It follows that, for any $\epsilon > 0$ and a particular c_i,

$$\lim_{n \to \infty} \Pr\left[\left|c_i - \frac{1}{n+1}\right| > \frac{\epsilon}{\sqrt{n+1}}\right] = \lim_{n \to \infty} \Pr\left[\left|(n+1)c_i - 1\right| > \epsilon\sqrt{n+1}\right] =$$

$$= \lim_{n \to \infty} \Pr\left[\left|y + c_i - 1\right| > \epsilon\sqrt{n+1}\right] = \lim_{n \to \infty} \Pr\left[y > -c_i + 1 + \epsilon\sqrt{n+1}\right] =$$

$$= \lim_{n \to \infty} \left(e^{c_i - 1 - \epsilon\sqrt{n+1}}\right) = 0 .$$

Then $\lim_{n \to \infty} \Pr\left[\text{Any}\left|c_i - \frac{1}{n+1}\right| > \epsilon\sqrt{n+1}\right] = \lim_{n \to \infty} n e^{c_i - 1 - \epsilon\sqrt{n+1}} = 0$

and $\lim_{n \to \infty} \Pr\left[\max\left|c_i - \frac{1}{n+1}\right| \leq \epsilon\sqrt{n+1}\right] = 1 .$

It further follows that

$$\lim_{n \to \infty} \Pr\left[\sum_{i=1}^{n+1} \left(c_i - \frac{1}{n+1}\right)^2 \leq \epsilon^2\right] = 1$$

and the quantity S converges in probability to zero for the true values

of the parameters. It is also clear that for values of the parameters

other than the true values S does not have zero as its limit and S

asymptotically possesses a minimum at the true values. Thus the method

of choosing estimates for the parameters by minimizing S yields, for

$n \rightarrow \infty$, a sequence of parameter estimates converging in probability to

the true values.

Testing Goodness of Fit. The current procedure employs two

notions of the goodness of fit. The first of these is called the close-

ness of the fit and is measured by the value of S at the minimum. The

second is the randomness of the fit and may be measured in several ways.

The approach rests on the notion that a good fit is characterized by two

circumstances: (a) the residuals $F(x_i) - F(x_{i-1}) - \frac{1}{n+1}$,

$F(x_{i+1}) - F(x_i) - \frac{1}{n+1}$, etc., are small; (b) the residuals are random.

In general a distribution will be considered to give a bad fit if it

fails by either criterion. Closeness is a fairly natural criterion of

goodness of fit and is related to the general notion of the distance be-

tween the sample and the fitted distribution. Randomness of the residu-

als is desirable since, if the null hypothesis is true, we would expect

the increments in cumulative probability $F(x_i) - F(x_{i-1})$ — associated

with going from the $(i-1)^{th}$ to the ith sample point, as estimated

from the fitted distribution — sometimes to exceed and sometimes to fall

short of their mathematical expectation, in no predictable manner.

Since estimates are obtained by minimizing S , the value of S

at the minimum is a natural measure of closeness. The randomness of the

residuals can be measured in several ways. Three particular methods are

discussed here.

1. A run test on the number of runs of positive and negative residuals. On the hypothesis that the permutations of positive and negative residuals are randomly generated, the number of runs in large samples is approximately normally distributed with mean

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

and standard deviation

$$\sigma = \left[\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \right]^{1/2}$$

where n_1 and n_2 are the number of positive and negative residuals. For small samples exact tables are available for testing the hypothesis of randomness.[18]

2. The reduction in the sum of the squares of residuals due to fitting to the residuals orthogonal polynomials up through the kth degree.[19] The value of k should be a number small relative to the total number of observations but high enough to fit well fairly high frequency oscillations. In the experiments described below k was chosen to be 15. According to this method a random series of residuals will yield a low reduction in the sum of squares. If there are low frequency oscillations in the residuals indicating systematic deviations of the sample from the fitted distribution, the reduction in the sum of squares will be considerable. Systematic very high frequency deviations which are also a sign of nonrandomness will also yield very small reduction in the

---

18. See [12].

19. See [5] and [8].

sum of squares but this is not likely to occur with unimodal densities.

3. The spectral density of the residuals according to which we consider the series of residuals ordered by the subscript i as a time series and display the density of frequencies generating the series. The more the spectrum resembles that of white noise the better the fit is considered from this point of view.

Sampling Experiments. A separate set of sampling experiments similar to those described in Section 3 was performed. Sample sizes of 25, 50, 100, 300, 500 and 1000 were employed and 100 samples of each size were generated from the Pareto distribution with $k = 1.0$ and $a = 1.5$. The parameters were fitted by the method described in this section and the various goodness of fit statistics were calculated.

Estimated critical values for the S statistic from 100 samples are shown in Table 6. As an illustration, we would reject the hypothesis

TABLE 6. Estimated Critical Values for the S-Statistic

N	Significance Level		
	.20	.10	.05
25	.0362	.0401	.0438
50	.0198	.0218	.0239
100	.0106	.0112	.0123
300	.0036	.0038	.0038
500	.0021	.0022	.0022
1000	.0010	.0011	.0011

of a good fit on the .05 level if, for example, a sample of 100 observations yielded an S value in excess of .0123. The mean parameter estimates and the root mean square errors are displayed in Table 7, and compare favorably with other methods of estimation. For large values of N

TABLE 7. Parameter Estimates and Root Mean Square Errors

N	$\hat{a}$	$\hat{k}$	RMSE $\hat{a}$	RMSE $\hat{k}$
25	1.435	.981	.100	.021
50	1.383	.993	.125	.008
100	1.440	1.003	.071	.004
300	1.480	1.008	.033	.008
500	1.490	1.009	.028	.009
1000	1.493	1.009	.019	.009

the mean bias of $\hat{k}$ is slightly larger than for other estimators. For $\hat{a}$ the mean bias tends to be somewhat larger than for the quantile method and for maximum likelihood estimates, but is generally of the same magnitude. The root mean square errors for $\hat{k}$ are larger and those for $\hat{a}$ smaller than with the best of the alternative methods. On the basis of the apparent properties of the present estimating method, it seems to be a reasonable alternative to the others.

The several methods discussed above of testing for randomness of fit were applied to the residuals arising out of fitting the Pareto distribution to the data. Since the data were generated from the Pareto distribution, the null hypothesis is known to be true.

Since the distribution of runs (under the null hypothesis) is known, the run test was performed for only 36 samples (6 for each value of N). In 3 out of the 36 cases we rejected the null hypothesis on the .05 level of significance. Considering the situation to be a binomial one with probability $p = .95$ of success, the probability of three or more failures in 36 tries is .268 and we cannot reject the hypothesis that .95 is a correct estimate of the probability of success.

The fitting of orthogonal polynomials up to and including the

fifteenth degree permits us to estimate empirically the percentage reduction in the total sum of squares of residuals due to fitting the first 15 degrees. The critical values of the percentage reduction are displayed in Table 8. In a concrete case we would select a significance

TABLE 8. Estimated Critical Values for the Percentage Reduction in the Sum of Squares Resulting from 15th Degree Orthogonal Polynomials

N	Significance Level							
	.975	.950	.900	.800	.200	.100	.050	.025
25	.821	.777	.693	.661	.487	.415	.365	.316
50	.462	.442	.392	.351	.212	.177	.155	.145
100	.219	.205	.188	.174	.102	.090	.079	.066
300	.085	.077	.066	.060	.034	.028	.026	.024
500	.062	.051	.044	.038	.019	.016	.015	.015
1000	.028	.026	.024	.020	.010	.008	.008	.006

level, say .05, and compare the empirically calculated percentage reduction with the critical values for the appropriate value of N. If the empirically calculated figure is outside the interval specified by Table 8, we reject the hypothesis of randomness. These tables are therefore suitable for testing against both alternatives of very low or very high frequency oscillations.

Finally we display in Figures 1-6 the spectral densities of the residuals for 36 cases (6 for each value of N). These may serve as a standard of comparison in cases in which the null hypothesis is not known to be true. Although the spectral densities displayed are not very meaningful for small values of N, they generally behave like the spectrum of white noise.

The applicability of the various measures suggested for testing

goodness of fit to a variety of other distributions is affected by the fact that tables of critical values for S (measuring closeness) and tables of critical values for the percentage reduction in the sum of squares of residuals due to fitting orthogonal polynomials were derived from sampling experiments based on a two-parameter family of distributions. Clearly with distributions with a different number of parameters to be estimated, our estimates in Tables 6 and 8 are not fully valid, those being based on cases with the wrong degrees of freedom. It appears unlikely, however, that this will make a great deal of difference when the number of observations is large.[20]

5. Conclusion

Four standard methods of estimating the parameters of the Pareto distribution have been discussed in some detail. These are the method of moments, the method of maximum likelihood, the method of least squares and the method of quantiles. In addition, some more qualitative methods of judging whether a sample was generated by the Pareto distribution have been analyzed, with particular reference to the properties of the Lorenz curve and the properties of sequential samples. Sampling experiments were used to obtain experimental evidence concerning the goodness of the various (nonqualitative) methods. Strictly no great differences were found among the four methods; more informally the methods of maximum likelihood and of quantiles performed best.

---

20. Initial application of these techniques seems to yield finer discrimination among alternative hypotheses than could be achieved with standard methods. See [11].

General dissatisfaction with some existing methods of judging the goodness of a fit has led to the formulation of a new method of estimation. This method involves the minimization of the criterion function

$$S = \sum_{i=1}^{n} \left(F(x_i) - F(x_{i-1}) - \frac{1}{n+1} \right)^2.$$

When a distribution has been fitted by minimizing S we judge the goodness of fit on the basis of two criteria: (a) the closeness of the fit as measured by the value of S at the minimum, and (b) the randomness of the fit as measured by (i) the number of runs of positive and negative residuals $F(x_i) - F(x_{i-1}) - \frac{1}{n+1}$, (ii) the percentage reduction in the total sum of squares due to fitting orthogonal polynomials to the residuals, (iii) the spectral density of the residuals. This method of estimating the parameters of a distribution seems to yield results comparable with those obtained by standard methods as judged by root mean square errors of estimates and similar criteria, and seems superior to standard methods with regard to goodness of fit problems in providing finer discrimination among alternative hypotheses.

BIBLIOGRAPHY

[1]   Aitcheson, J. and J. A. C. Brown, The Lognormal Distribution, Cam-
         bridge University Press, 1957.

[2]   Champernowne, D. G., "A Model of Income Distribution," Economic Jour-
         nal, LXIII (1953), 318-351.

[3]   Cochran, W. G., "Some Methods for Strengthening χ^2 Tests,"
         Biometrics, 10 (1954), 417-451.

[4]   Feller, W., An Introduction to Probability Theory and Its Applica-
         tions (2nd ed., John Wiley and Sons, 1957).

[5]   Fisher, R. A., Statistical Methods for Research Workers (12th ed.,
         Oliver and Boyd, 1954).

[6]   Fraser, D. A. S., Nonparametric Methods in Statistics (John Wiley and
         Sons, 1957).

[7]   Gibrat, R., Les inégalités économiques, Paris, Sirey, 1931.

[8]   Goulden, C. H., Methods of Statistical Analysis (2nd ed., John Wiley
         and Sons, 1952).

[9]   Mandelbrot, B., "New Methods in Statistical Economics," Journal of
         Political Economy, LXXI (1963), 421-440.

[10] Pareto, V., Corso di Economia Politica, ed. by G. Einaudi, 1949.

[11] Quandt, R. E., "Statistical Discrimination Among Alternative Hypothe-
         ses and Some Economic Regularities," Proceedings of the Regional
         Science Association, forthcoming.

[12] Siegel, S., Nonparametric Statistics, McGraw-Hill, 1956.

[13] Simon, H. A., "On a Class of Skew Distribution Functions," Biometrika,
         Vol. 42, Parts 1 and 2 (1955), 425-440.

[14] Simon, H. A. and C. P. Bonini, "The Size Distribution of Business
         Firms," American Economic Review, XLVIII (1958), 607-617.

[15] Kalecki, M., "On the Gibrat Distribution," Econometrica, 13 (1945),
         161-170.

[16] Zipf, G. K., Human Behavior and the Principle of Least Effort,
         Addison-Wesley Press, Cambridge, 1949.
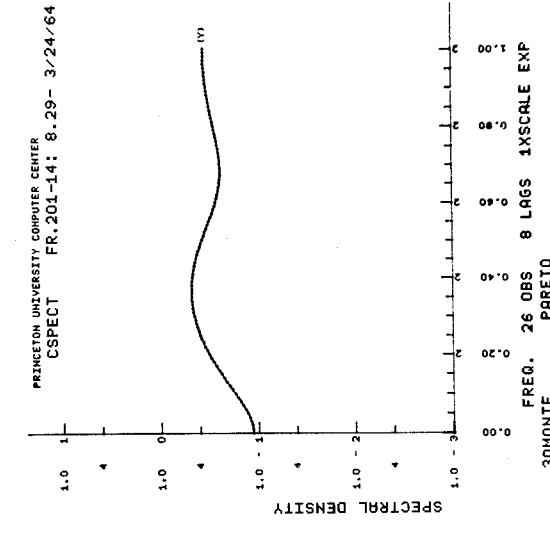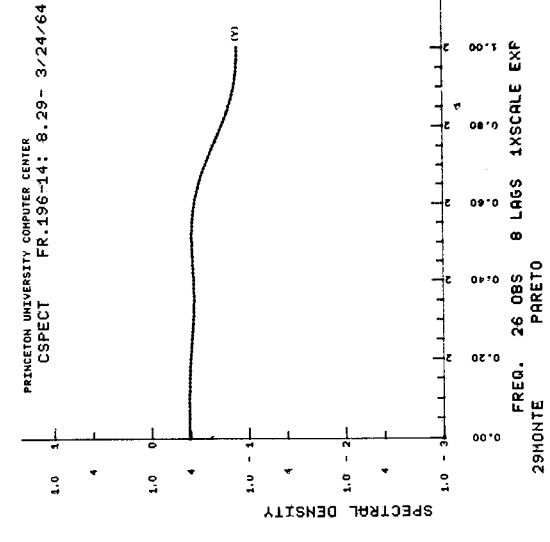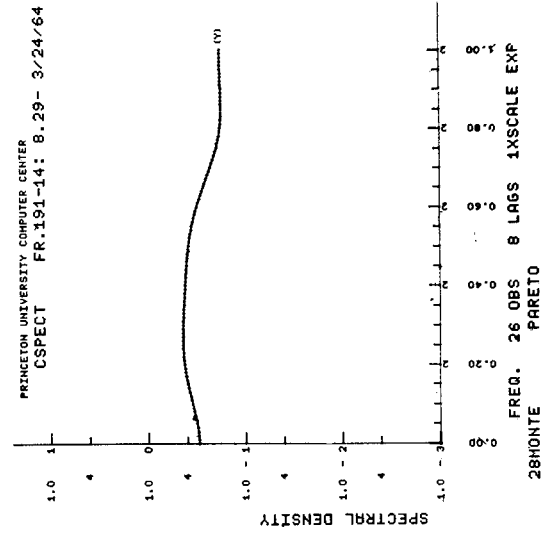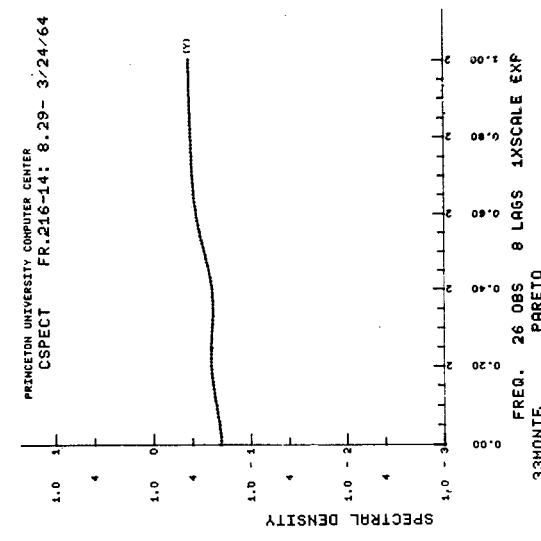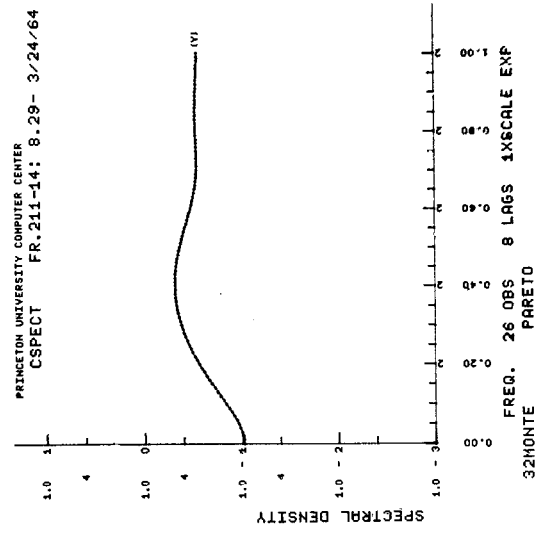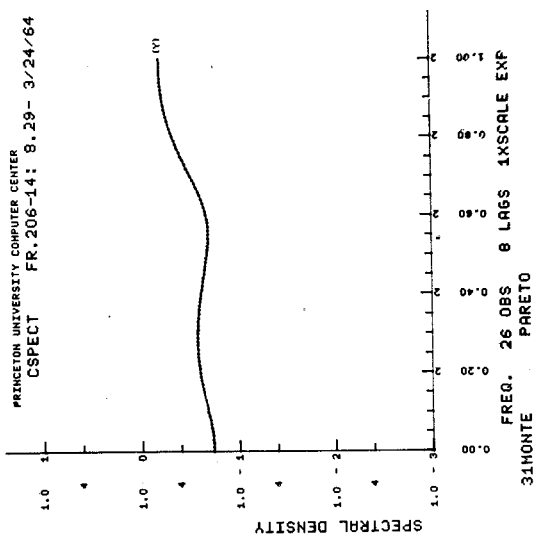
Figure 1
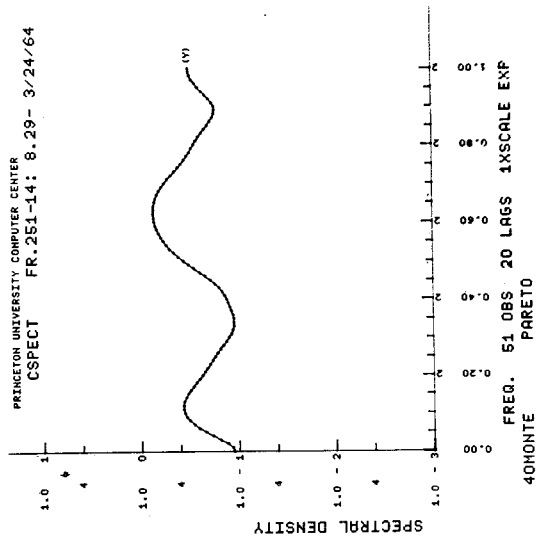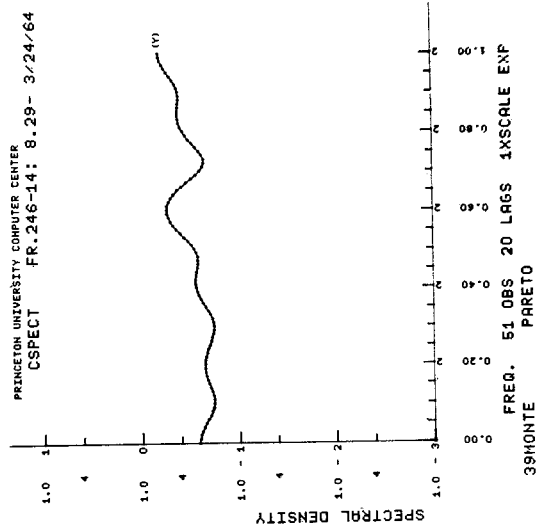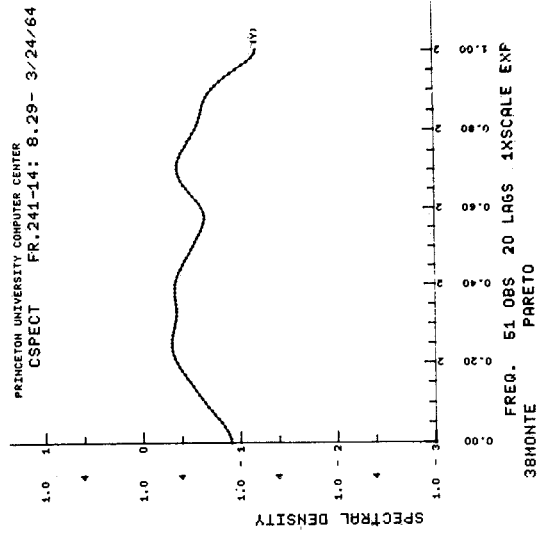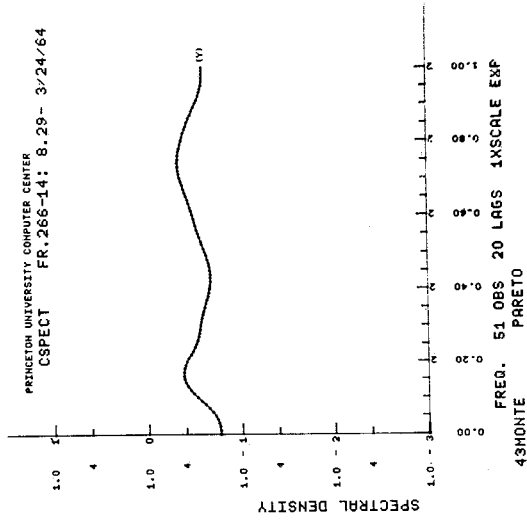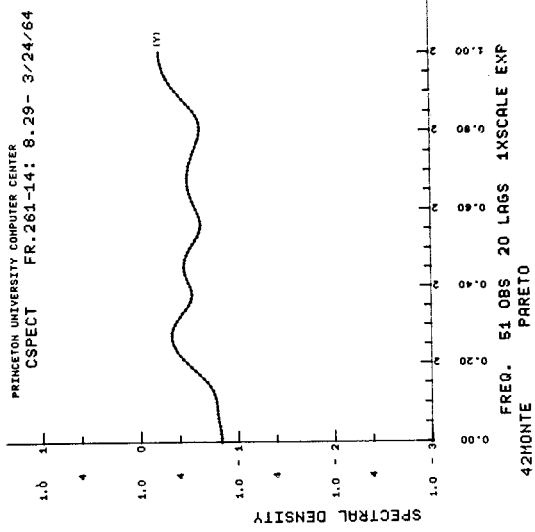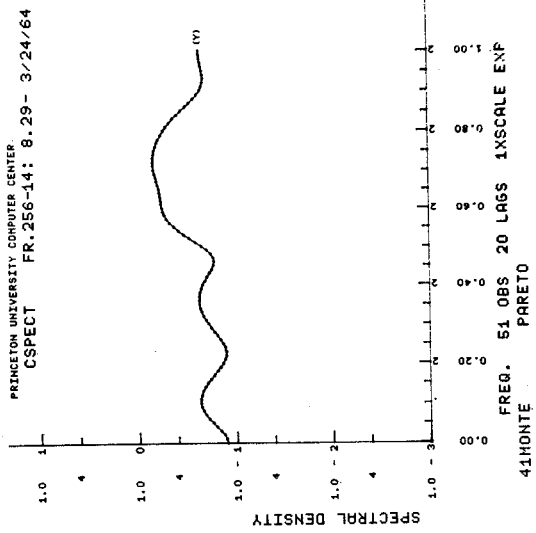
Figure 2

Figure 4

Figure 6