The Annals of Statistics
1975, Vol. 3, No. 5, 1163–1174

A SIMPLE GENERAL APPROACH TO INFERENCE ABOUT THE TAIL OF A DISTRIBUTION

BY BRUCE M. HILL

University of Michigan

A simple general approach to inference about the tail behavior ot a distribution is proposed. It is not required to assume any global form for the distribution function, but merely the form of behavior in the tail where it is desired to draw inference. Results are particularly simple for distributions of the Zipf type, i.e., where $G(y) = 1 - Cy^{-\alpha}$ for large y. The methods of inference are based upon an evaluation of the conditional likelihood for the parameters describing the tail behavior, given the values of the extreme order statistics, and can be implemented from both Bayesian and frequentist viewpoints.

1. Introduction and summary. In certain situations it is of interest to draw inference about the behavior of a distribution function in the tails without assuming that a particular parametric form for the distribution function holds globally. The examples that gave rise to this article concerned a random sample $Z_1, \cdots, Z_k$ from a distribution F on the unit interval with $F(x) \sim Cx^\alpha$ as $x \to 0$, [1], [2], [3]. It was desired to draw inference about α without making assumptions about the form of F elsewhere. A similar situation occurs in connection with inference about the parameter α of a Zipf or Pareto Law $1 - F(x) \sim Cx^{-\alpha}$ as $x \to \infty$, where it is desirable to assume that the Zipf form holds only for large x, not globally [1], [2], [3]. A simple general form of inference, which can be implemented from either Bayesian or classical approaches, is proposed here for such situations.

Suppose that a sample $Y_1, \cdots, Y_k$, is drawn from a population with distribution G, and let $Y^{(1)} \geqq Y^{(2)} \geqq \cdots \geqq Y^{(k)}$ be the order statistics. (Here and throughout the usual definition of order statistics is reversed in order to simplify formulas in applications.) On the basis of theoretical arguments or previous data it is believed, or at least the hypothesis is tentatively entertained, that G has a known functional form, say, $G(y) = w(y; \theta)$, *for y sufficiently large*, where θ is a vector of parameters. The simplest case is that in which a number D is known such that for $y \geqq D$ this functional form is valid. Here, D need not be the smallest value for which this is true, and thus might be chosen quite conservatively in some situations. When the global form of G is unknown, so that ordinary parametric methods are unavailable, it is then perhaps intuitively plausible to base inference about θ on the values of those order statistics that exceed D, since it is only these that lie in the region where G is believed to have the specified form. Thus the values of such order statistics might be taken as a

Received February 1974; revised January 1975.

AMS 1970 *subject classification.* Primary 62F99.

Key words and phrases. Tail of distribution, Bayesian inference, order statistics.

conditioning event, or data, for the purpose of inference about θ. There is, of course, a certain degree of arbitrariness in the choice of such a conditioning event. For example, if $Y^{(r+1)} \geqq D > Y^{(r+2)}$, then it might be argued that it would be more appropriate to condition not only upon the values of the $r+1$ largest order statistics, but also upon the event $Y^{(r+2)} < D$. Here, as elsewhere in statistics, questions as to choice of conditioning events are subtle, and answers will be heavily colored by philosophical outlook. Such questions will be discussed in Section 4, where an attempt is made to justify the proposed procedures from a Bayesian viewpoint. To some extent, however, such questions are rather academic. Thus in typical applications D will not be known precisely, and it will be necessary to select a subset, consisting say of the $\hat{r}+1$ largest order statistics, on the basis of prior knowledge and a combination of various data analytic techniques. Such an $\hat{r}$ will often depend upon the data in a highly complicated way, and questions as to the precise form of conditioning event become of lesser importance than the choice of $\hat{r}$. The approach advocated here is to consider the inference based upon the values of the $r+1$ largest order statistics, for $r = 1, 2$, etc., until upon the basis of data analytic guides and prior knowledge a stopping point $\hat{r}$ is reached, beyond which it seems unwise to proceed. This approach is illustrated in Section 5, where the Zipf model is used for city size data.

In Section 2 conditional likelihood functions for θ are derived for inference about the lower tail behavior of a distribution F on the positive reals. This is done both when conditioning only upon the values of the $r+1$ smallest order statistics, and also when conditioning in addition upon the event $Y^{(k-r-1)} > d$. Such conditional likelihood functions form the basis for both Bayesian and classical inference. (Strictly speaking, all likelihood functions are conditional; however, the adjective is used here to emphasize that different subsets of the data are conditioned upon.) By means of such conditioning a possibly highly complex nonparametric problem is reduced to a relatively tractable parametric form. In the special cases $F(x) = Cx^{\alpha}$ as $x \to 0$, or $1 - F(x) = Cx^{-\alpha}$ as $x \to \infty$, the conditional likelihood functions and the conditional maximum likelihood estimates of α and C take on particularly simple forms for both types of conditioning events. In these cases it is seen that inference is rather insensitive to the precise nature of the conditioning event, provided that a modest number of order statistics fall in the region where the functional form is valid. Indeed, for inference about α alone, it would ordinarily suffice to condition merely upon ratios of consecutive order statistics.

In Section 3 it is shown how upper tail behavior of the Zipf form can be deduced from lower tail behavior of the type discussed in Section 2, the general form of inference for upper tails is obtained, and various other applications of the basic approach are suggested.

Recent related discussions of inference based upon extreme order statistics have been given by Johnson [4], who derives asymptotic properties of the likelihood function, and by Pickands [6], who proposes a method of inference using sample quantiles.

2. Lower tail inference. Let $Z_1, \cdots, Z_k$ be a sample from a continuous strictly increasing distribution F, with $F(0) = 0$, and let $Z^{(1)} \geqq Z^{(2)} \geqq \cdots \geqq Z^{(k)}$ be the order statistics. By the Renyi representation theorem [7]

$$(2.1) \qquad Z^{(i)} = F^{-1}\left[\exp - \left(\frac{e_1}{k} + \frac{e_2}{k-1} + \cdots + \frac{e_i}{k-i+1}\right)\right], \qquad \text{for} \quad i = 1, 2, \cdots, k\,,$$

where here and throughout the e_i are independent exponentially distributed random variables, each with expectation 1. Then

$$(2.2) \qquad e_j = (k-j+1)[\ln F(Z^{(j-1)}) - \ln F(Z^{(j)})]\,, \qquad \text{for} \quad j = 1, 2, \cdots, k\,,$$

where by definition $F(Z^{(0)}) = 1$.

Now suppose that $F(x) = w(x; \theta)$ for $x \leqq d$, where w is a specified function, d is known, and θ is an unknown parameter vector. From (2.2) it follows that for $Z^{(k-r)} \leqq d$,

$$(2.3) \qquad \begin{aligned} &(k-j+1)[\ln w(Z^{(j-1)}; \theta) - \ln w(Z^{(j)}; \theta)] = e_j\,, \\ &\qquad\qquad\qquad\qquad \text{for } j = k-r+1, k-r+2, \cdots, k\,, \\ &-k \ln w(Z^{(k-r)}; \theta) = k\left(\frac{e_1}{k} + \frac{e_2}{k-1} + \cdots + \frac{e_{k-r}}{r+1}\right), \end{aligned}$$

and so these equations determine the probability of any event for which $Z^{(k-r)} \leqq d$. If the observed values of the extreme order statistics are $Z^{(j)} = z^{(j)}$, for $j = k-r, k-r+1, \cdots, k$, with $z^{(k-r)} \leqq d$, then the conditional likelihood function for θ is

$$(2.4) \qquad \begin{aligned} L_1(\theta) \propto |J| \exp[-\textstyle\sum_{i=1}^{r} i(\ln w(z^{(k-i)}; \theta) - \ln w(z^{(k-i+1)}; \theta))] \\ \times\, p(-k \ln w(z^{(k-r)}; \theta))\,, \end{aligned}$$

where the Jacobian J is easily verified to be proportional to $\prod_{i=1}^{r+1} (d \ln w(z^{(k-i+1)}; \theta))/dz^{(k-i+1)}$, and p is the density function of $k(e_1/k + \cdots + e_{k-r}/(r+1))$, i.e., of a linear combination of independent exponentially distributed random variables. The density p can easily be evaluated as follows. Let $B(a, b)$ denote a random variable having the beta distribution with parameters a and b. If $F(x) = x$, for $0 \leqq x \leqq 1$, then it is well known that $Z^{(i)}$ is distributed like $B(k-i+1, i)$, so that from (2.1), p is in fact the density function of a random variable having the distribution of $-k \ln B(r+1, k-r)$. Hence

$$p(x) \propto \exp[-(r+1)x/k][1 - \exp(-x/k)]^{k-r-1} \qquad \text{for} \quad x > 0\,,$$

and the factor p in (2.4) reduces to

$$[w(z^{(k-r)}; \theta)]^{r+1}[1 - w(z^{(k-r)}; \theta)]^{k-r-1}\,.$$

It is easy to show that if the conditioning event used above to obtain $L_1(\theta)$ is supplemented by the additional condition $Z^{(k-r-1)} > d$, so that exactly $r+1$ of the order statistics are $\leqq d$, then the resulting conditional likelihood function is

$$(2.5)\qquad \begin{aligned} L_2(\theta) &\propto L_1(\theta)[(1-w(d;\theta))/(1-w(z^{(k-r)};\theta))]^{k-r-1} \\ &\propto |J|[1-w(d;\theta)]^{k-r-1}\textstyle\prod_{i=1}^{r+1} w(z^{(k-i+1)};\theta)\,. \end{aligned}$$

These likelihood functions can be used either to obtain conditional maximum likelihood estimates for θ, or, in conjunction with a prior distribution for θ, conditional posterior distributions for θ. The essential point is that by virtue of the conditioning, a highly complicated nonparametric problem has been reduced to parametric form. Although in general the conditional likelihood functions and maximum likelihood estimators may themselves be quite complicated, the reduction to the typically small number of parameters contained in θ would appear to be a substantial reduction in the scope of the inferential problem.

In the special case $w(x;\theta) = Cx^\alpha$ for $x \leqq d$, so $\theta = (\alpha, C)$, with $\alpha > 0$, $C > 0$, the above likelihood functions take on particularly simple forms. Indeed, suppose that we were to condition only upon $Z^{(k-r)} \leqq d$, and upon the values t_i of the random variables $T_i = i[\ln Z^{(k-i)} - \ln Z^{(k-i+1)}]$ for $i = 1, 2, \cdots, r$. It would follow immediately from (2.2) that the conditional likelihood function for α is then $L_0(\alpha) \propto \alpha^r \exp(-\alpha \sum_1^r t_i)$. In this case the conditional maximum-likelihood estimate of α is

$$\hat{\alpha}_0 = r(\textstyle\sum_{i=1}^r t_i)^{-1} = [\ln z^{(k-r)} - r^{-1}\sum_{i=0}^{r-1} \ln z^{(k-i)}]^{-1}\,.$$

The properties of $\hat{\alpha}_0$ as an estimator of α in the frequentist sense can easily be derived, conditional upon $Z^{(k-r)} \leqq d$. From (2.2), (2.3), and the independence of the e_i, it follows that the conditional distribution of $\hat{\alpha}_0$, given $Z^{(k-r)} \leqq d$, is that of $r/X(r)$, where $X(i)$ denotes a random variable having the gamma distribution with density proportional to $x^{i-1}\exp -x$, for $x > 0$. Hence

$$(2.6)\qquad \begin{aligned} &E(\hat{\alpha}_0 \mid Z^{(k-r)} \leqq d) = r\alpha/(r-1)\,, \quad \text{for} \quad r > 1\,, \qquad \text{and} \\ &\operatorname{Var}(\hat{\alpha}_0 \mid Z^{(k-r)} \leqq d) = (r\alpha)^2/[(r-1)^2(r-2)]\,, \quad \text{for} \quad r > 2\,. \end{aligned}$$

From a frequentist point of view, taking a conditional framework, the above conditional distribution of $\hat{\alpha}_0$, given $Z^{(k-r)} \leqq d$, might be regarded as an appropriate basis for inference about α, i.e., estimation, confidence intervals, hypothesis testing. This would be analogous, for example, to conditioning upon marginal totals for inference about contingency tables ([5], page 145). Alternatively, for fixed r and large k, Pr $[Z^{(k-r)} \leqq d]$ will be nearly 1, and the conditional moments should approximate the unconditional moments.

From a Bayesian or likelihood point of view it is not necessary even to consider Pr $[Z^{(k-r)} \leqq d]$, since conditioning upon $Z^{(k)}, \cdots, Z^{(k-r)}$, for example, with $Z^{(k-r)} \leqq d$, implies that Pr $[Z^{(k-r)} \leqq d \mid Z^{(k)}, \cdots, Z^{(k-r)}]$ is either 0 or 1, does not depend upon any unknown parameters, and thus cannot affect the likelihood function. However, it is necessary for a Bayesian to justify the ignoring of certain portions of the data that is implicit in basing inference upon the kinds of conditional events considered above. Such questions will be discussed in Section 4. Accepting for the time being that a conditional likelihood function such as $L_0(\alpha)$ can be justified, Bayesian inference in this case would consist of multiplying $L_0(\alpha)$ by a prior density function, and interpreting the product as proportional to the posterior density for α. Although the gamma family of prior densities would be particularly convenient, there is no need to restrict attention to this family. Finally, from a design point of view, a Bayesian might wish to take observations sequentially, stopping when there are sufficiently many observations $\leqq d$ for his needs. As usual, the stopping rule employed will have no effect upon Bayesian inference, provided that it depends only upon the data.

Consider next the conditional likelihood function $L_2(\alpha, C)$ given by (2.5), which can be viewed as resulting from conditioning upon $Z^{(k-r)} = z^{(k-r)} \leqq d$, and $Z^{(k-r-1)} > d$, in addition to the values of the T_i. Both maximum likelihood estimation and Bayesian inference are simplified by transforming the parameter C to $\beta = F(d) = Cd^{\alpha}$, where d is the fixed prespecified value below which the functional form is assumed valid. In terms of (α, β) the likelihood function is

$$(2.7) \qquad \begin{aligned} L_2(\alpha, \beta) &\propto \alpha^{r+1} \exp -\alpha[(r+1)\ln d - \textstyle\sum_{i=0}^{r} \ln z^{(k-i)}] \\ &\quad \times \beta^{r+1}(1-\beta)^{k-r-1}, \qquad \text{for} \quad \alpha > 0, \quad 0 < \beta < 1 . \end{aligned}$$

The conditional maximum likelihood estimates from (2.7) are then

$$(2.8) \qquad \begin{aligned} &\hat{\alpha}_2 = (r+1)/[(r+1)\ln d - \textstyle\sum_{i=0}^{r} \ln z^{(k-i)}] , \\ &\hat{\beta}_2 = (r+1)/k , \qquad \text{and} \qquad \hat{C}_2 = \hat{\beta}_2/d^{\hat{\alpha}_2} . \end{aligned}$$

The conditional distributions of these estimators are easily obtained.

For Bayesian inference a convenient and reasonably rich class of prior densities for (α, β) arises by letting α have a gamma distribution, while, conditional upon α, β has beta distribution $B(a(\alpha), b(\alpha))$, in which the parameters of the beta distribution are allowed to depend upon α. (Note that the parameter β could be replaced by the value of F at any fixed $x \leqq d$; however, in conjunction with (2.7), d is most convenient.) If the prior density for (α, β) is taken proportional to

$$[\alpha^{s-1} \exp -f\alpha]\beta^{a(\alpha)-1}[1-\beta]^{b(\alpha)-1}\Gamma(a(\alpha)+b(\alpha))[\Gamma(a(\alpha))\Gamma(b(\alpha))]^{-1} ,$$

where Γ is the usual gamma function, then multiplying by $L_2(\alpha, \beta)$, it is seen that the posterior density is of a similar form. Particularly simple results are obtained when $a(\alpha) = a$ and $b(\alpha) = b$ do not depend upon α, so that a priori α and β are independent. In this case they are also independent a posteriori, and the posterior expectations of α, β, and C, i.e., the Bayes estimates for squared error loss, are

$$(2.9) \qquad \begin{aligned} &\hat{\alpha} = (r+s+1)/[(r+1)\ln d - \textstyle\sum_{i=0}^{r} \ln z^{(k-i)} + f] , \\ &\hat{\beta} = (r+a+1)/(a+b+k) , \qquad \text{and} \\ &\hat{C} = \hat{\beta}[1 + \hat{\alpha}\ln d/(r+s+1)]^{-(r+s+1)} . \end{aligned}$$

A similar Bayesian analysis is possible using $L_1(\alpha, C)$ from (2.4). The maximum likelihood estimates in this case are

$$\hat{\alpha}_1 = [(r+1)/r]\hat{\alpha}_0 \quad \text{and} \quad \hat{C}_1 = [(r+1)/k][z^{(k-r)}]^{-\hat{\alpha}_1}. \tag{2.10}$$

It is to be noted that when a, b, f, $s-1$, and $z^{(k-r-1)} - z^{(k-r)}$ are small, with $z^{(k-r)} \leqq d \leqq z^{(k-r-1)}$, and if r is not too small, then both the Bayes estimates (2.9), and the maximum likelihood estimates obtained under the various forms of conditioning, are all approximately equal. Thus a certain degree of robustness, both to the parameters of the prior distribution, and to the type of conditioning event, can often be anticipated. Although the prior distribution upon which (2.9) is based will sometimes be reasonable as an approximation, it was used here primarily to illustrate the form which Bayesian inference takes in this problem.

3. Upper tail inference and applications.

Consider first a sample $Y_1, \cdots, Y_k$ of positive random variables with distribution function $G(y) = 1 - Cy^{-\alpha}$ for $y \geqq D$, where D is known. The simplest way to draw inference about α and C is to observe that if $Z_i = Y_i^{-1}$, then $\Pr[Z_i \leqq x] = \Pr[Y_i \geqq x^{-1}] = Cx^{\alpha}$, if $x \leqq d = D^{-1}$, so that the theory developed in the preceding section is directly applicable. Conditioning upon the values $Y^{(i)} = y^{(i)}$, $i = 1, 2, \cdots, r+1$, of the $r+1$ largest observations, where $y^{(r+1)} \geqq D$, yields

$$\begin{aligned} \hat{\alpha}_1 &= (r+1)/[\textstyle\sum_{i=1}^{r} \ln y^{(i)} - r \ln y^{(r+1)}] \quad \text{and} \\ \hat{C}_1 &= [y^{(r+1)}]^{\hat{\alpha}_1}(r+1)/k\,, \end{aligned} \tag{3.1}$$

as the maximum likelihood estimates of α and C, based upon (2.10) for the z_i, and expressed in terms of the $y^{(i)}$.

This simple method of inference for an upper tail of the Zipf form was obtained by noticing a suitable transformation of the data in order to reduce the problem to the form considered earlier. As another illustration of such a transformation, suppose that the random variable X has a distribution F of the Weibull form, $1 - F(x) = \exp[-(rx)^s]$ for large x. Transforming to $Y = \exp X^s$ yields $\Pr[Y > y] = y^{-r^s}$ for large y, which thus reduces to Zipf's Law with $\alpha = r^s$. These examples illustrate how the results of Section 2 for lower tails can be exploited by means of transformations of the variables to deal with upper tails. An alternative general approach following the lines of that in Section 2 for lower tails is to assume that $G(y)$ has a known form in the upper tail, say, $G(y) = w(y; \theta)$, if $y \geqq D$. Then, conditional upon $Y^{(r+1)} \geqq D$, from (2.2),

$$\begin{aligned} e_i &= (k - i + 1)[\ln w(Y^{(i-1)}; \theta) - \ln w(Y^{(i)}; \theta)]\,, \\ &\qquad \text{for} \quad i = 2, \cdots, r+1\,, \quad \text{and} \\ e_1 &= -k \ln w(Y^{(1)}; \theta)\,. \end{aligned} \tag{3.2}$$

It is worth noting that (3.2) is not quite symmetrical to the corresponding (2.3) for lower tails. Conditioning upon $Y^{(i)} = y^{(i)}$, $i = 1, \cdots, r+1$, as data, where

$y^{(r+1)} \geqq D$, the conditional likelihood function for θ is

$$L_1(\theta) \propto |J| \exp[k \ln w(y^{(1)}; \theta) - \textstyle\sum_{i=1}^{r} (k - i) \ln [w(y^{(i)}; \theta)/w(y^{(i+1)}; \theta)]] \,, \tag{3.3}$$

where the Jacobian J is proportional to $\prod_{i=1}^{r+1} (d \ln w(y^{(i)}; \theta))/dy^{(i)}$.

It is interesting to observe that the identities which can be obtained, on the one hand by using (3.3), and on the other hand by making use of transformations as illustrated above, are often by no means obvious.

For other applications of the basic approach in this article, observe that although the discussion so far has concerned the tails of distributions, the relationships given by (2.3) and (3.2) can in fact be employed for any pair of consecutive order statistics. Thus if the functional form $F(x) = w(x; \theta)$ were believed valid in some interval containing a middle subset of the order statistics, for example, if F were believed to be approximately normal *except in the tails*, then a slight modification of the approach presented here would allow inference about location and scale parameters.

Still another application arises when F is a distribution on the real line and both the upper and lower tail behavior of F are of interest. The above approach would enable symmetry of F (at least in the tails) to be tested; or, if F was known to be symmetric, it would lead to pooling of the information from the upper and lower order statistics, to yield more precise inference about the tail behavior of F.

Finally, the method of inference suggested here can be of use even in cases where some global specification of the model seems appropriate on theoretical grounds or on the basis of experience. Thus if it were thought that the underlying distribution was a stable law with index α, then a comparison of the estimate of α based upon the global stable law model with the estimates based upon the much weaker assumptions of the present approach, could be used in part to decide upon the adequacy of the global model. Even when the latter is appropriate, the present simple estimates of α might be used as initial values to aid in the difficult computational task of obtaining maximum likelihood estimates under the stable law model.

4. Bayesian comments.

Maximum likelihood estimators such as those derived above seem to have some justification from a classical view of inference, both in conditional and unconditional frameworks. However, from a strict Bayesian viewpoint, the posterior distribution should be based upon all the data. Thus, in the example of Section 2, the posterior distribution for α using $L_0(\alpha)$ as likelihood function would be justified only if it were known to yield a satisfactory approximation to the marginal posterior distribution of α, given all the data. However, in order to evaluate such a marginal posterior distribution, it would be necessary to specify F globally, and thus to choose a parametric form for F, say, in terms of additional (nuisance) parameters λ and finally to integrate out these nuisance parameters in the joint posterior distribution of α and λ. Except in some rather special situations, where a particular global specification seems appropriate, there is at present no satisfactory way of carrying out such a program. Hence, reduction of the data to the values of the T_i or $Z^{(k-i+1)}$, although undesirable, seems necessary in order to deal with the problem at all. Furthermore, there are reasons to anticipate that when prior knowledge of the global form of F is vague, then the posterior distribution for α based upon $L_0(\alpha)$ should approximate the "true" marginal posterior distribution for α. Thus, suppose $p(\alpha, \lambda) = p_1(\alpha)p_2(\lambda)$ is the prior density for (α, λ). Then the likelihood function for (α, λ), given all the data, could be written as $L(\alpha, \lambda) \propto L_0(\alpha)L_0{}^*(\alpha, \lambda)$, where $L_0{}^*(\alpha, \lambda)$ is the likelihood function based upon the data $z^{(k-r)}, \cdots, z^{(1)}$. The marginal posterior distribution for α, given all the data, would then be proportional to

$$p_1(\alpha)L_0(\alpha) \int L_0{}^*(\alpha, \lambda)p_2(\lambda)\, d\lambda\ .$$

But with $p_2(\lambda)$ of a diffuse nature, the integral will typically be a gentle function of α relative to $L_0(\alpha)$, and thus via a stable estimation argument, one anticipates that $p_1(\alpha)L_0(\alpha)$ may provide a satisfactory approximation. This argument of course depends crucially upon the assumed knowledge of the functional form of F for $x \leqq d$. Needless to say, a similar argument could equally well have been made for basing inference upon $L_1(\alpha, C)$ or $L_2(\alpha, C)$, so this discussion should not be construed as a justification for using $L_0(\alpha)$ in preference to the other conditional likelihood functions. Rather, it is an attempt to make it clear that any actual use of Bayesian methods of inference is implicitly based upon precisely such a stable estimation argument, together with a hopefully judicious choice of just what to include as data for the analysis. Although in principle Bayesian methods condition upon *all* available data, so that the only data that can be ignored in connection with a specific problem of inference or decision-making is that which itself, together with all of its underlying parameters, is regarded as independent of the parameters of interest, in practice such a full analysis is impossible, and any actual Bayesian analysis of data makes the necessary approximations. Whether it is wise to ignore certain aspects of the data in a particular application must be judged in the light of what can be done with, as opposed to without, formal analysis of such aspects. When such a formal analysis simply cannot be made, or even when it is merely very difficult and of dubious validity, then there is little choice but to condition upon that part of the data that can be effectively dealt with, and rely upon some form of the stable estimation argument. Such information as is lost by this procedure (and there is some reason to expect it is relatively little) could not, in any case, have been utilized satisfactorily.

In so far as choice between $L_0(\alpha)$, $L_1(\alpha, C)$, and $L_2(\alpha, C)$, is concerned, one would anticipate from the general argument given above, that it would make little difference which of these is employed for inference about α, or which of the last two is employed for inference about C, provided that d is reasonably well know a priori, and that there are a substantial number of observations $\leqq d$. This is borne out by the close relationships between the maximum likelihood estimators derived from these various likelihood functions. On the other hand, when d is only vaguely known, or when there are very few order statistics for which one can be reasonably sure that the relationship $F(x) = Cx^{\alpha}$ holds, then it is necessary to employ certain crude but effective data analytic techniques as a guide to the choice of an appropriate value of r.

My own preference is for $L_1(\alpha, C)$, and thus conditioning upon the values of the $r + 1$ smallest order statistics, but not upon the event $Z^{(k-r-1)} > d$. This is based upon the following considerations. In practice not only will d usually be unknown, but the relation $F(x) = Cx^{\alpha}$, will be at best an approximation for any x. The crucial question becomes whether or not the approximation can be justified for a particular set of order statistics, say, $z^{(k)}, \cdots, z^{(k-r)}$, and such a question could be answered, for example, on the basis of the usual bias versus precision considerations. But in such a context d loses the precise (but highly artificial) meaning it has heretofore held. Indeed, it should be recalled that even before this question of approximations was brought up, d was not assumed to be the largest value for which the functional form was valid. Given this indeterminacy as to the precise meaning of d, which could be resolved, but only in a highly artificial manner, it seems ordinarily preferable that d should not appear explicitly in the inferential formulae. In the next section it is shown how inference can be made using $L_1(\alpha, C)$, for increasing values of r, until on the basis of certain guides, a particular value $\hat{r}$ is chosen.

5. Data analysis.

The methods proposed above depend upon a subjective choice of d or D. In situations where such a choice is difficult, or for other reasons deemed inappropriate there are a variety of data-analytic techniques which can be useful in the choice of r upon which to base $L_1(\alpha, \beta)$. Consider again the case of a sample $Y_1, \cdots, Y_k$ from a distribution G with $G(y) = 1 - Cy^{-\alpha}$ for $y \geqq D$, but no longer assume that D is known. Let $V_i = \ln [Y^{(i)}/Y^{(i+1)}]$, for $i = 1, \cdots, r$. From (2.2) it follows that, conditional upon $Y^{(r+1)} \geqq D$, $\alpha i V_i = e_{k-i+1}$, for $i = 1, \cdots, r$, where the e_i are again independent exponentially distributed random variables, each with expectation 1. Hence if r has been chosen sufficiently small so that $y^{(r+1)}$ is in fact $\geqq D$, or to put it another way, so that the approximation of $G(y)$ by $1 - Cy^{-\alpha}$ is satisfactory for $y \geqq y^{(r+1)}$, then the iV_i should in all respects behave like a random sample from an exponential distribution with parameter α, at least for $i = 1, \cdots, r$. On the other hand, if r has been chosen too large, so that $y^{(r+1)}$ and perhaps other of the larger order statistics have values where the approximation is poor, then the iV_i should exhibit certain systematic discrepancies from their known behavior under the exponential distribution. This observation forms the basis for a variety of methods for choosing an appropriate r. Thus from a frequentist view one can simply test the hypothesis that the iV_i have an exponential distribution for $i = 1, \cdots, r$, using any of the standard test procedures, for example, the chi-square goodness of fit test. If the hypothesis is accepted for a particular r, then one can increase r step by step, until eventually the hypothesis is rejected. Such a procedure would be analogous to the practice of fitting an nth degree polynomial to a set of data in a regression analysis, and then testing, step by step, whether the degree can be reduced. Some other interesting test statistics, which can alternatively be viewed simply as measures of discrepancy, are

$$H^{(r)} = \hat{\alpha}_0^2 \sum_{i=1}^{r} (iV_i - \hat{\alpha}_0^{-1})^2 \quad \text{and}$$
$$K^{(r)} = \sum_{i=1}^{r} [\ln (iV_i) - r^{-1} \sum_{i=1}^{r} \ln (iV_i)]^2 ,$$

where $\hat{\alpha}_0 = \hat{\alpha}_0(r) = r[\sum_{i=1}^{r} iV_i]^{-1} = r[\sum_{i=1}^{r} \ln Y^{(i)} - r \ln Y^{(r+1)}]^{-1}$ is just the $\hat{\alpha}_0$ of Section 2 expressed in terms of the $Y^{(i)}$. The distributions of such statistics, conditional upon $Y^{(r+1)} \geqq D$, are easily derived. For example, the conditional distribution of $H^{(r)}$ is that of $\sum_{i=1}^{r} (e_i - \bar{e})^2/\bar{e}^2$, where $\bar{e} = (\sum_{i=1}^{r} e_i)/r$. It can be shown that the conditional expectation and variance of $H^{(r)}$ are approximately $r - 1$ and $8(r - 1)$, respectively.

From a Bayesian point of view it would be preferable to incorporate prior knowledge about the departures G might exhibit from its presumed form in the upper tail, and on the basis of such considerations, for example, to estimate α by a weighted average of the form $[\sum_{i=1}^{k-1} is_i V_i]^{-1}$, where $s_i \geqq 0$, $\sum_{i=1}^{k-1} s_i = 1$, and with the weights decreasing with i and nearly 0 for i near $k - 1$, so that most weight is given to the largest observations, where it is more likely that the model is appropriate. Such estimators, compromising between bias and precision, are of course very natural from a risk function point of view. It is not clear, however, just how much would be gained by such a more refined analysis, as opposed to merely selecting a value $\hat{r}$ by inspection of the iV_i or the measure of discrepancy $H^{(r)}$, and using this value $\hat{r}$ to form $L_1(\alpha, \beta)$.

The approach discussed in this article was developed as a means of examining the adequacy of a theoretical model for Zipf's Law proposed in [1], [2], and [3]. In that model the parameter α plays a distinguished role and therefore it was of importance to draw reliable inference about α under minimal assumptions about the global form of the distribution G. In [2] the model was used to describe city sizes, this being an area for which Zipf's Law is generally regarded as appropriate. Table 1 gives the sizes of the 30 largest cities in the United States in 1940, where the cities are defined by political boundaries. Several of the statistics proposed in this article are also displayed as functions of r for $r = 1, 2, \cdots, 29$.

It is interesting to note that the standard deviation of $\hat{\alpha}_0(29)$, as given by (2.6), is $.20\alpha$, and the fluctuations of $\hat{\alpha}(r)$ seem, if anything, too slight. This is also reflected in $H^{(r)}$, which stays somewhat surprisingly close to $r - 1$, considering the magnitude of its variance. The fact that $H^{(r)}$ is almost always smaller than $r - 1$, with a substantial discrepancy for r near 19, is perhaps indicative of a correlation amongst the iV_i. A chi-square goodness of fit test on the 29 values of the iV_i using the class intervals 0 to .5, .5 to 1, and > 1, yields a value of .3 for chi-square with 1 degree of freedom. On the whole the fit of the model seems embarrassingly good.

TABLE 1[1]

Rank	Size × 10^{-3}	iV_i	$\hat{\alpha}_0(r)$	$\hat{\alpha}_1(r)$	$H^{(r)}$	$k\hat{C}_1(r)$
1	7,455	.79	1.27	2.54		8.3×10^{16}
2	3,397	1.13	1.04	1.57	.06	2.1×10^{10}
3	1,931	.52	1.23	1.64	.28	6.3×10^{10}
4	1,623	.30	1.46	1.82	.81	9.2×10^{11}
5	1,504	2.69	.92	1.10	3.05	2.2×10^{7}
6	878	.13	1.08	1.26	5.07	2.0×10^{8}
7	859	.36	1.18	1.35	6.47	7.7×10^{8}
8	816	.45	1.25	1.41	7.51	1.8×10^{9}
9	771	1.24	1.18	1.31	6.90	4.5×10^{8}
10	672	.13	1.29	1.42	8.99	2.0×10^{9}
11	663	.47	1.34	1.46	9.80	3.5×10^{9}
12	635	.94	1.31	1.42	9.45	2.0×10^{9}
13	587	.25	1.38	1.49	10.99	5.2×10^{9}
14	576	2.12	1.21	1.30	11.16	3.8×10^{8}
15	495	.09	1.29	1.38	13.45	1.1×10^{9}
16	492	1.22	1.25	1.32	12.82	5.3×10^{8}
17	456	1.00	1.23	1.30	12.52	3.8×10^{8}
18	430	1.35	1.19	1.25	12.03	1.9×10^{8}
19	399	.58	1.21	1.27	12.53	2.4×10^{8}
20	387	.10	1.26	1.32	14.50	5.2×10^{8}
21	385	.95	1.25	1.31	14.27	4.2×10^{8}
22	368	2.73	1.13	1.18	16.10	7.0×10^{7}
23	325	.21	1.16	1.21	17.81	1.2×10^{8}
24	322	.22	1.20	1.25	19.52	1.9×10^{8}
25	319	1.04	1.19	1.24	19.19	1.6×10^{8}
26	306	.09	1.23	1.28	21.42	2.8×10^{8}
27	305	.24	1.26	1.31	23.09	4.3×10^{8}
28	302.3	.01	1.31	1.36	25.81	7.9×10^{8}
29	302.2	.10	1.35	1.40	28.23	1.4×10^{9}
30	301.2					

[1] The data is taken from Statistical Abstract of the United States, 1950, page 57.

Although it would be possible to make some arbitrary definition of the population from which these city sizes were drawn, and of the sample size k, it is perhaps preferable not to become involved with such delicate questions in this article. The values of $k\hat{C}_1(r)$ are presented for illustrative purposes.

REFERENCES

[1] Hill, B. M. (1970). Zipf's Law and prior distributions for the composition of a population. *J. Amer. Statist. Assoc.* **65** 1220–1232.

[2] Hill, B. M. (1974). The rank frequency form of Zipf's Law. *J. Amer. Statist. Assoc.* **69** 1017–1026.

[3] Hill, B. M. and Woodroofe, M. (1975). Sronger forms of Zipf's Law. *J. Amer. Statst. Assoc.* **70** 212–219.

[4] Johnson, R. A. (1974). Asymptotic results for inference procedures based on the r smallest observations. *Ann. Statist.* **2** 1138–1151.

[5] Lehman, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[6] PICKANDS, J. (1975). Statisuical inference using extreme order statistics. *Ann. Statist.* **3** 1–13.

[7] RENYI, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hungar* **4** 191–232.

DEPARTMENT OF STATISTICS
THE UNIVERSITY OF MICHIGAN
1447 MASON HALL
ANN ARBOR, MICHIGAN 48104