The Annals of Statistics
1975, Vol. 3, No. 5, 1038-1044

DESCRIPTIVE STATISTICS FOR NONPARAMETRIC MODELS I. INTRODUCTION

BY P. J. BICKEL[1] AND E. L. LEHMANN[2]

University of California, Berkeley

An overview is given of an approach to the definition of descriptive characteristics of populations and to their estimation. The emphasis is on the robustness and efficiency of the estimators. Detailed summaries will be found in successive papers of the series dealing with the problems of location, scale and kurtosis.

Introduction. Descriptive statistics deals with measures of different aspects of a population (or a distribution of population values). The population may be finite, as is the case for example when it consists of a set of data, or it may be infinite. Typical examples of descriptive measures are the mean and median as measures of location, the standard deviation or interquartile range as measures of scale, and the classical measures of skewness, kurtosis and correlation.

When defining such measures, one usually has in mind not a single population to which the measure is to be applied but a family of such populations. In particular, in statistics one is typically concerned with a family to which the given population is assumed to belong. The choice of suitable measures then depends strongly on the nature of this family. We shall distinguish a number of possibilities.

1. Parametric models. Here the distribution in question is characterized by a small number of (natural) parameters. Examples are the family of normal distributions, characterized by expectation and variance; the family of Pearson curves, characterized by the first four moments, and hence by the classical measures of location, scale, skewness and kurtosis; or the family of stationary, two-state Markov chains, characterized by initial and transition probabilities.

2. Nonparametric neighbourhood models. It has long been recognized that the validity of parametric models can at best be approximate, and that the discrepancies due to gross errors and other impurities can have a serious effect on the descriptive parameters and their estimates. This has led to the suggestion of models in which the main body of the observations come from a distribution belonging to a given parametric family (for example, the normal family) but a

Received April 1973; revised September 1974.

[1] This research was partially supported by the Office of Naval Research, Contract, NONR N00014-69-A-0200-1038.

[2] Work done while the author was Research Professor for the Adolph C. and Mary Sprague Miller Institute for Basic Research in Science, University of California, Berkeley.

AMS 1970 *subject classifications.* Primary 62G99; Secondary 62G05, 62G20, 62G35.

Key words and phrases. Descriptive statistics, nonparametric models, robustness, estimation, asymptotic efficiency, standardized variance, location, scale, kurtosis.

small proportion represent impurities about whose distribution no parametric assumptions are made. The result is a parametric model contaminated by a small nonparametric admixture, a model which one could describe as a nonparametric neighbourhood of a parametric family.

The parameters of interest in such a model are still the parameters of the parametric part of the model. However, in estimating or testing these parameters one must now guard against the nonparametric disturbances. An important and satisfactory body of inference in such models has been constructed by Huber (1964). This work is still in progress; a survey of the present state is given in Huber's Wald lecture (1972).

3. Nonparametric models with natural parameters. The above approach seems very well suited to the many situations in which extensive experience with the type of data under consideration makes an approximately parametric model reliable either for the measurements themselves or for suitable (known) transforms of these measurements. However, assumptions such as approximate normality will often be unwarranted, for example, in sociological or psychological investigations, where the observations by the nature of the situations are less well controlled. In such cases, one may prefer a totally nonparametric model, in which the observations may, for example, be assumed to be independently, identically distributed according to a distribution F, of which nothing is assumed except possibly certain smoothness or symmetry conditions.

However, even in such apparently totally nonparametric models there may exist natural parameters describing important aspects of the model, and the estimation or testing of such parameters has in fact been a very active field of study in recent years.

The case which has received the greatest attention is that in which F is a univariate distribution which is assumed to be symmetric. The natural location parameter for F is then its center of symmetry. A survey of the work concerning the estimation of this center, is given by Andrews *et al.* (1972). Another example is the two-sample shift problem where two distributions are assumed to differ only in location and where the difference in location then constitutes a natural shift parameter (see, for example, Høyland (1965)). More generally, there is a natural measure of contrasts in linear models, even when the (common) shape of the symmetric error distribution is unknown (see, for example, Lehmann (1963) and Spjøtvoll (1968)).

4. Neighbourhoods of nonparametric models with natural parameters. It seems natural to consider the extension of model 3. which parallels the extension 2. of model 1. A typical example would be to assume that we are dealing with a sample from an unknown symmetric distribution F which has a slight amount of asymmetric contamination; the parameter of interest would be the center of symmetry of F. Problems of this kind do not seem to have been considered in the literature.

The four types of models above share the important feature that there exist natural measures of the aspect of the model under consideration. There are many possible estimators of these measures, and a central problem is to choose among these.

Let us now turn to situations in which such natural measures do not exist. (The model may still possess natural parameters, but these do not correspond to the feature of interest. This is the case, for example, when we are concerned with the scale of a symmetric distribution.) It is here that the descriptive aspect of the problem becomes nontrivial. This problem has been pointed out and was discussed from a somewhat similar point of view by Takeuchi (1967).

How do we measure the location of a nonsymmetric distribution? How do we measure or describe its spread, its skewness or kurtosis? What do we even mean by kurtosis: does it describe the tail-behavior of a distribution, its peakedness, or, as has recently been argued (Darlington (1970), Chissom (1970)), its tendency to biomodality?

It is this kind of question, of course, with which we must begin the description of a particular aspect of a distribution, and we shall find it convenient to answer it by defining in each case when a distribution G (or a random variable Y with distribution G) possesses the attribute under consideration more strongly than a distribution F (or random variable X).

The literature contains many examples of such partial orderings. We may, for example, say that Y is "to the right" of X if Y is stochastically larger than X (a concept introduced by Mann and Whitney (1947) and studied further by Lehmann (1955)). A stronger definition is obtained by requiring monotone likelihood ratio. Partial orderings corresponding to the comparison of the skewness of two distributions, or of their kurtosis were introduced by van Zwet (1964) and by Barlow and Proschan (1966). Comparisons for dispersion were discussed by Z. W. Birnbaum (1948) and for the degree of positive association between the two components of a bivariate distribution by Lehmann (1966), Esary, Proschan and Walkup (1967), and Esary and Proschan (1972).

Suppose now that there has been defined a partial ordering, with $F \prec G$ meaning that G possesses the attribute under consideration more strongly than F. Then the first condition required of a measure θ of this attribute is that

$$(1) \qquad \theta(F) \leqq \theta(G) \qquad \text{whenever} \quad F \prec G\,.$$

A second condition characterizes the behaviour of $\theta(F)$ (which we shall also denote by $\theta(X)$ when X is a random variable with distribution F) under linear transformations. Thus, a measure of location should satisfy

$$(2) \qquad \theta(aX+b) = a\theta(X) + b \qquad \text{for all} \quad a\,, \quad b\,,$$

and a measure of scale

$$(3) \qquad \theta(aX+b) = |a|\theta(X) \qquad \text{for all} \quad a \neq 0 \quad \text{and all} \quad b\,.$$

A referee has pointed out that (3) is evidently desirable in the scale case if

$b = 0$ and $a > 0$. However he feels, and we agree, that if, for example, X is known to be positive, the constraints imposed by (3) for $b \neq 0$ and/or $a < 0$ are much less appealing. We here assume that no such restriction on X is given and that X must be considered totally unknown. In this case we feel that (3) is acceptable as stated.

To illustrate a slightly more complicated situation, suppose that the quantity in question is a standardized measure of location, of which $E(X)/[\mathrm{Var}\,(X)]^{\frac{1}{2}}$ is a typical example. We should characterize this as a ratio $\theta = \theta_1/\theta_2$ of a location parameter θ_1 satisfying (2), divided by a scale parameter θ_2 satisfying (3).

Another example of this kind is provided by a measure of kurtosis, which in a latter part of this paper we shall define as the ratio $\theta = \theta_1/\theta_2$ of two scale parameters which satisfies (1) when $F \prec G$ is the partial order introduced by Barlow and Proschan (1966) or the stronger ordering introduced by van Zwet (1964).

Once the conditions have been laid down which a measure θ is to satisfy, there typically will exist an infinity of measures satisfying these conditions. In the location case, for example, the functionals

$$\theta(F) = \int_0^1 F^{-1}(t)\, dK(t) \tag{4}$$

where K is any distribution function on (0, 1) which is symmetric with respect to $\frac{1}{2}$ defines a large (but by no means exhaustive) class of such measures.

How should one choose among this great variety of possibilities? An additional condition one might like to impose is that of robustness; that is, small changes in F should result in small changes of θ. An early discussion of the need for such a condition was given by Bahadur and Savage (1956). For a rigorous formulation of the condition, which is essentially a continuity requirement, and a detailed analysis of the problem see Hampel (1968, 1971). Such a condition would rule out, for example, expectation as a measure of location and standard deviation in the scale case. Even with this restriction, there will still be available an infinity of measures. The location measures (4), for example, will be robust provided K is continuous and assigns probability 1 to some interval (u, v) with $0 < u < v < 1$.

To make a choice among these measures let us now recall that it will be necessary to estimate the chosen $\theta(F)$ from the data, and let us compare different θ's in terms of the accuracy with which they can be estimated. The most natural estimator of $\theta(F)$ is $\theta(\hat{F})$ where $\hat{F}$ denotes the empirical cdf; that is, the values of θ for the whole population is estimated by its value for the subpopulation made up of the sample values. One might conjecture that in a very strong sense $\theta(\hat{F})$ is also the best estimator. This conjecture is invalidated by the phenomenon of superefficiency which we exemplify in BL II, Section 4. Nevertheless, the estimators $\theta(\hat{F})$ are the ones with which we shall be concerned.

At this point, a difficulty arises, which we have only been able to overcome in each case separately through an *ad hoc* argument adjusted to the case. This difficulty (to which reference is made also in Daniell (1920) and Tukey (1960)) concerns the role played by the functionals $h(\theta)$, where h is any strictly increasing function. In a sense, these functionals are all equivalent since each is determined by any other. However, they differ widely in the ease with which they can be estimated. Suppose now that we wish to compare two functionals θ_1 and θ_2 which are not equivalent in this sense, and suppose we use variance (or asymptotic variance) as our measure of accuracy. Then it will clearly make a rather essential difference whether we compare θ_1 with θ_2, or with $\theta_2/100$. In fact, we can make any given θ as accurate as we please by dividing it by a sufficiently large constant.

This problem is greatly simplified when θ is required to satisfy equivariance conditions such as (2) or (3). For if θ satisfies these conditions, then the only functions for which the condition can hold are linear functions. Let us now briefly indicate how the difficulty remaining after this reduction can be resolved in the three cases which will be studied in the later parts of this paper.

(i) *Location.* If θ is a functional satisfying (2), then the only functions h for which $h(\theta)$ satisfies (2) are

$$h(\theta) = \theta + b \tag{5}$$

where b is an arbitrary constant. Since our estimator of (5) is $\hat{\theta} + b$ and since addition of a constant does not affect the variance of the estimator, the choice of a member of (5) is immaterial.

(ii) *Scale.* If θ is a functional satisfying (3), the only functions h for which $h(\theta)$ satisfies (3) are

$$h(\theta) = a\theta\,, \quad a > 0\,. \tag{6}$$

For a given θ satisfying (3), we could thus choose instead $a\theta$, which would be estimated by $a\hat{\theta}$, and thereby multiply the variance by a^2. These choices will become equivalent if we measure the accuracy of $\hat{\theta}$ not by its variance (or asymptotic variance) but by its *standardized variance*

$$\operatorname{Var}(\hat{\theta})/\theta^2 \tag{7}$$

which is invariant under the transformations (6). With this definition of accuracy, it is now possible to compare different measures θ_1 and θ_2.

There is an alternative way of arriving at (7). If the distribution of X is symmetric about μ, attention can be restricted to the variables $|X - \mu|$. By taking logarithms, the scale problem is then reduced to the location problem (i). If θ is any measure of scale and $\log \theta$ the corresponding location measure, the latter is estimated by $\log \hat{\theta}$; the asymptotic variance of $\log \hat{\theta}$ under suitable regularity conditions is just the asymptotic variance of $\hat{\theta}$ divided by θ^2.

(iii) *Kurtosis.* As mentioned above, we shall define a measure of kurtosis as a suitable ratio of the form $\theta(F) = \tau_2(F)/\tau_1(F)$ where τ_1 and τ_2 are measures of scale and hence satisfy (3). If τ_1 and τ_2 are two such functionals, the only functions $h_1(\tau_1)$ and $h_2(\tau_2)$ with which they could be replaced are $h_i(\tau_i) = a_i\tau_i$; hence the only functions of θ with which we need to be concerned are those satisfying (6). The same argument as in (ii) shows that these choices become equivalent provided we use the standardized variance or asymptotic variance as our measure of accuracy.

At this point, we find ourselves in the following position. We have available a large class of measures of the characteristic in question. From this class we should like to choose a member which is robust and the estimator of which has good accuracy globally, i.e. for a large (nonparametric) family of distributions F. In the symmetric location case this turned out to be fairly easy. For estimating the center of symmetry of a symmetric distribution there exist a variety of estimators which are nearly as efficient as the mean for all F, and more efficient for many F. This is a piece of good luck, which one cannot expect in general. In the symmetric location case, a theoretically even more appealing possibility exists if one is willing to forego robustness. It was shown by Stein (1956), van Eeden (1970) and Takeuchi (1971) that in this case it is possible to find a fully efficient estimator, essentially by estimating the shape of F and adapting the estimator of the center to this estimated shape. The efficiency of this procedure depends heavily on the symmetry of F, and it seems unlikely that a similar approach would work in the cases considered here.

How should a measure be chosen if it cannot be simultaneously robust and estimated with high efficiency (either absolute or relative to a standard estimator) for all distributions satisfying the assumptions of the model? Two possibilities are the following.

(i) We can relax the condition of robustness. Instead of requiring the measure (and its estimator) to be totally robust, we may be satisfied if it is significantly more robust than the standard measure. (For details of such a comparative concept of robustness, see Section 4 of the following second part of this paper).

(ii) Alternatively, we can put additional restrictions on the model, which one could expect to be satisfied in practice, (for example, restrictions on the heaviness of the tails of the distributions). Such restrictions may make it easier for the estimators to be either robust or to have high efficiency for all distributions in the restricted model, or both.,

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimation of Location*. Princeton Univ. Press.

Bahadur, Raghu Raj and Savage, Leonard J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115-1122.

Barlow, Richard E. and Proschan, Frank (1966). Tolerance and confidence limits for classes of distributions based on failure rates. *Ann. Math. Statist.* **37** 1593-1601.

Birnbaum, Z. W. (1943). On random variables with comparable peakedness. *Ann. Math. Statist.* **19** 76-81.

CHISSOM, BRAD S. (1970). Interpretation of the kurtosis statistic. *Amer. Statist.* **24** (4) 19-23.

DANIELL, P. J. (1920). Observations weighted according to order. *Amer. J. Math.* **42** 222-336.

DARLINGTON, R. B. (1970). Is kurtosis really peakedness? *Amer. Statist.* **24** (2) 19-20.

ESARY, J. D. and PROSCHAN, F. (1972). Relationships among some concepts of bivariate dependence. *Ann. Math. Statist.* **43** 651-655.

ESARY, J. D., PROSCHAN, F. and WALKUP, D. W. (1967). Association of random variables with applications. *Ann. Math. Statist.* **38** 1466-1474.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph. D. Dissertation, Univ. of California, Berkeley.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887-1896.

HØYLAND, ARNLJOT (1965). Robustness of the Hodges-Lehmann estimator for shift. *Ann. Math. Statist.* **36** 174-197.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73-101.

HUBER, P. J. (1972). Robust statistics: A review. *Ann. Math. Statist.* **43** 1041-1067.

LEHMANN, E. L. (1955). Ordered families of distributions. *Ann. Math. Statist.* **26** 399-419.

LEHMANN, E. L. (1963). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34** 957-966.

LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137-1153.

MANN, HENRY B. and WHITNEY, D. R. (1947). On tests of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18** 50-60.

SPJØTVOLL, EMIL (1968). A note on robust estimation in analysis of variance. *Ann. Math. Statist.* **39** 1486-1492.

STEIN, CHARLES (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. on Math. Stat. and Prob.* 187-195, Univ. of California Press.

TAKEUCHI, K. (1967). Robust estimation and robust parameter. (Unpublished).

TAKEUCHI, K. (1971). A uniformly asymptotically efficient estimator of a location parameter. *J. Amer. Statist. Assoc.* **66** 292-301.

TUKEY, J. W. (1960). A survery of sampling from contaminated distributions. *Contributions to Probability and Statistics*. Ed. I. Olkin, Stanford Univ. Press.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1-67.

VAN EEDEN, CONSTANCE (1970). Efficiency—robust estimation of location. *Ann. Math. Statist.* **41** 172-181.

VAN ZWET, W. R. (1964). *Convex transformations of random variables.* Math. Centrum, Amsterdam.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720