MINIMUM BIAS ESTIMATION

by

Wm. Jackson Hall

University of North Carolina

This research was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. AF 49(600)-261. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Institute of Statistics
Mimeograph Series No. 213
December, 1958

TABLE OF CONTENTS

MINIMUM BIAS ESTIMATION[1,2]

Summary

This is an expository paper on the theory and application of minimum bias estimation, an approach to estimation theory which characterizes an estimator primarily according to its bias. Minimum risk theory, as developed by Wald and others, leads to estimators which minimize the average risk or the maximum risk. The place of bias in the risk theory is reinterpreted, and an alternative theory is developed with the role of the risk function replaced by a bias function (e.g., absolute bias, squared bias, or percentage absolute bias). When unbiased estimators do not exist, it is proposed to minimize either the average (in some sense) or the maximum of the bias function. This is accomplished by choosing an unbiased estimator of a suitable approximation to the function to be estimated. Relevant aspects of the theory of approximation are reviewed. Results bearing considerable conceptual similarity to Wald's results on minimax and Bayes estimators are thereby derived; for example, estimators which minimize the maximum absolute bias are shown to minimize also the average squared bias, the average being taken with respect to a least favorable prior distribution. The problem of estimating a root of the binomial parameter, arising in certain biological and military applications, is used to exemplify various aspects of the minimum bias theory.

---

1. Presented in part at the 118th Annual Meeting of The American Statistical Association, December 27, 1958, Chicago, Ill.

2. This research was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. AF 49(600)-261. Reproduction in whole or in part is permitted for any purpose of the United States Government.

1. A Review of Minimum Risk Estimation

We are concerned with some data which were generated by a probability distribution, whose form is completely known except for specification of a parameter θ lying in a space Ω. We denote the data simply by x, a point in a space $\mathcal{X}$. It is desired to estimate a numerical function g of θ; that is, we are to choose an estimator δ, a numerical function on $\mathcal{X}$, and identify the value $\delta(x)$ of δ at the observed value x with the unknown value $g(\theta)$ of the parametric function. We shall not consider randomized estimators, non-parametric estimation, nor sequential experiments, though these could be included with little alteration. The Neyman-Pearson theory of testing the hypothesis that θ is in Ω_0 versus θ in $\Omega - \Omega_0$ can be considered as the special case in which $g(\theta)$ is zero in Ω_0 and unity elsewhere and $\delta(x) = 0$ is equivalent to acceptance of θ in Ω_0; however, our primary concern will be genuine estimation problems.

Let $L(\delta(x), g(\theta))$ denote the monetary loss incurred by estimating $g(\theta)$ by $\delta(x)$. L is normally a non-negative function of the error of estimate $\delta - g$; e.g., squared error, percentage absolute error, or simple loss which is zero if the error is small and unity otherwise. Let $R(\delta, g(\theta))$ denote the expected value of the loss, called risk, when using the estimator δ and θ is the true parameter value.

What we shall refer to as minimum risk theory consists of those various approaches to estimation theory, developed by A. Wald and his followers (see, for example S. Wald (1950), E. L. Lehmann (1950), D. Blackwell and M. A. Girshick (1954), or D.A.S. Fraser (1957)),

but with foundations in the works of Gauss and Laplace, which evaluate an estimator primarily on the basis of its associated risk and consider estimators with small risk (in some sense) to be good estimators. Since the risk is a function of θ, its minimization must take this into account. And since uniform minimization is possible only in trivial problems, either some other type of minimization is required, or some conditions must first be imposed to reduce the class of estimators under consideration, or both. In any case, it is customary to restrict attention to admissible estimators--estimators which cannot be uniformly improved upon in terms of as small or smaller risk.

Two types of minimization commonly considered are given below; either of these criteria frequently yield unique admissible estimators.

(1) Minimization of average risk: This criterion leads to Bayes estimators, estimators chosen so that a weighted average (over the parameter space) of the risk function is a minimum. This is particularly appropriate if the parameter itself is a random variable with a known a priori probability distribution in which case the expected risk is thereby minimized. Alternatively, justification is sometimes made in terms of rational degrees of belief or concern about the parameter.

(2) Minimization of maximum risk: This criterion leads to minimax estimators, estimators chosen so that the maximum (over the parameter space) of the risk function is a minimum. This may be considered a conservative approach whereby, in absence of specific knowledge about the parameter, one guards against the least favorable eventuality.

Other possible criteria are: subject to the prior distribution belonging to some specified class, minimize the average risk, or, subject

perhaps to some global bounds on the risk, minimize the risk in some sense in some particular locality; these are not developed to any extent (except for the latter in the case of hypothesis testing).

Four different conditions, one or more of which are sometimes imposed to reduce the class of estimators under consideration, are given below; subject to such conditions, the risk is then minimized uniformly or otherwise. The first two will not be considered further in this paper.

(1) Restriction to linear estimators, the rationale usually being one of simplicity; its use is generally restricted to problems of estimating parameters in linear models in which case a normality assumption further justifies the linearity.

(2) Restriction to invariant estimators; for example, two statisticians using different units of measurement should obtain equivalent estimates.

(3) Restriction to estimators which are functions only of a sufficient statistic--in fact, a necessary and sufficient statistic (if existent). From the risk point of view, nothing is lost by this restriction so long as L is convex in δ (and this is not required if randomized estimators are allowed). Restriction to sufficient estimators is consistent with the Fisherian concept that a sufficient statistic contains all the relevant information.

(4) Restriction to unbiased estimators, estimators for which the expected value of the error of estimate is everywhere zero. Such a restriction is frequently offered in the guise of preliminary reduction of the class of estimators under consideration, after which risk is

minimized. However, as has been shown by Lehmann and Scheffé (1950), in a wide class of problems there is a unique unbiased estimator depending only on a necessary and sufficient statistic[3]. This condition prevails whenever there are no non-trivial unbiased estimators of zero depending only on a necessary and sufficient statistic[4], in which case the statistic is said to be complete (Lehmann and Scheffé (1950)). Apparently, few practical (non-sequential) problems fall outside this class. Thus, the condition of unbiasedness is such a strong one that its imposition really implies that risk is not even considered rather than that it is put in a position of secondary importance--to claim that un unbiased estimator has minimum risk (or minimum variance) is usually an empty claim except in comparison with insufficient estimators. Thus, restriction to unbiased estimators can be thought of as being outside the minimum risk theory.

It may be noted that only in rare circumstances are minimax or Bayes estimators unbiased--only in trivial situations if the loss is squared error--and sometimes no unbiased estimator is even admissible. Thus the criteria of unbiasedness and minimum risk are frequently incompatible as well.

2. An Introduction to Minimum Bias Estimation

An approach to estimation theory parallelling the minimum risk theory, with the role of the risk function filled by a bias function,

---

3. That the words "necessary and" can be inserted in these statements is due to Bahadur (1957), the terminology is originally due to Dynkin (1951).

4. Ibid.

is here developed. The bias function is some non-negative function of the bias, the expected error of estimate. Thus, the operations of taking expected value and of applying a non-negative function are interchanged: instead of minimizing the expected value of a non-negative function of the error of estimate as in risk theory, we minimize some non-negative function of the expected value of the error of estimate. A number of concepts and theorems completely analogous to those in risk theory can be stated for the bias theory, but the mechanics of obtaining minimum bias estimators are quite different.

In minimum bias theory, in contrast to minimum risk theory, uniform minimization is frequently possible, leading here to unbiased estimators. The theory of unbiased estimation is well-established[5]; we shall be concerned with a minimum bias theory applicable to situations when no unbiased estimator is available. For example, in (non-sequential) binomial problems, only polynomials of limited degree in the success probability admit unbiased estimators. If the restriction is made that the range of the estimator be within the range of the function to be estimated, then unbiased estimators are less frequently available (for an example, see below). If one desires the simplicity of linear estimators, then some bias may be unavoidable. In the hypothesis testing case, in which the range of g is only 0 and 1, unbiased estimators are never available (except in trivial cases). Thus, four situations can be delineated in which bias is frequently unavoidable: 1) the sample space

---

5. For a review, see Lehmann and Scheffé (1950), Lehmann (1950), or Fraser (1957).

is finite, 2) the range of the estimator is restricted, 3) the functional form of the estimator is specified, and 4) the parametric function to be estimated is discontinuous.

When uniform minimization of bias is not possible, we might look for estimators whose maximum bias (absolute, squared, or relative) or average bias is a minimum, or for estimators with locally small bias in some sense. Such minimum bias estimators will be considered in the sequel. A. Bhattacharyya (1954), in discussing binomial estimation problems, considered estimators with minimum average squared bias and estimators with locally small bias in the sense that all derivatives of the bias vanish at a specified parameter point. A. N. Kolmogorov (1950) considered a somewhat different approach to minimum bias estimation, which will be illustrated later in this paper; he suggested finding upper and lower estimators, the bias of the former being everywhere non-negative and of the latter non-positive. Thus, one obtains two estimators rather than one, but thereby obtains bounds on the bias of any estimator between the two. No theory was offered for obtaining such estimators in any optimal way, however. S. H. Siraždinov (1956) followed Kolmogorov's approach and found upper and lower estimators for a polynomial of degree n+1 in the binomial parameter which are optimum in a sense related to what we shall call the minimax sense.

Why should one be concerned with bias? The connotations of the words "bias" and "expected error of estimate" certainly make it appear undesirable to the practitioner. In this regard, an unbiased estimator does allow the sample to "speak for itself"; no other knowledge or opinion of the experimenter is allowed to influence the estimate.

(Other definitions of unbiasedness--e.g., in terms of medians rather than expectations--would have similar justifications.) Unbiased estimation thus seems to fit more naturally in a theory of statistical inference, in which there may be more reason to consider the estimator as a descriptive statistic, than in a theory of statistical decision in which all prior information and the consequences of the decision taken cannot easily be ignored. It is noteworthy in this regard that, in many problems, if the statistician wishes to limit the parameter space to some subset of its natural range and to limit the range of the estimator accordingly, then no unbiased estimator is available. For example, the success probability in n Bernoulli trials, if limited to any proper subset of the interval (0,1), does not admit an unbiased estimator with range similarly restricted.

This justification for requiring small bias, allowing the sample to "speak for itself", would seem to preclude the use of prior information--e.g., an a priori distribution of the parameter--in any minimum bias theory. Thus minimum bias estimators in the least squares sense, to be introduced below, may appear to be a contradiction in terms, but other justification for these estimators may overcome this objection.

Another justification for a requirement of small bias is possible when making repeated estimations; namely, that the average error of estimate will be arbitrarily small with high probability if repetitions are sufficiently numerous. It may be some consolation to know that the over-estimates in some trials tend to be compensated for by under-estimates in other trials, though if some consumer loses because of the statistician's over-estimate on one trial he is not likely to be consoled

by knowing that his losses are compensated for by some other consumer's gains; such consumers might be more concerned with small risk. The justification for requiring small risk is also founded largely on a long-run interpretation of expectation--applied to loss rather than error--and its justification other than in repeated experimentation presents similar problems.

An additional justification for unbiasedness may be that it usually eliminates the need for specifying a loss function since so frequently it leads to unique sufficient estimators; however, if unbiased estimators do not exist, specification of some function analogous to loss will be required in the theory developed here.

Whether or not satisfactory justification is available, it is a fact that statisticians frequently spend great efforts in "correcting for bias", and there seems to be only limited acceptance of any minimum risk estimators with large biases. For example, suppose x denotes the number of successes in n Bernoulli trials. Then, the minimax estimator (squared error loss) of the success probability θ is $(x + \frac{1}{2}\sqrt{n})/(n + \sqrt{n})$, and although for small samples its risk, $(2\sqrt{n} + 2)^{-2}$, is less than that of the unbiased estimator x/n over a wide range of parameter values, it is apparently difficult to persuade the experimenter of its superiority; this may be due to its large bias, $(1 - 2\theta)/(2\sqrt{n} + 2)$, for probabilities near 0 and 1 (or perhaps to the inappropriateness of squared error loss).

Perhaps some compromise between the risk and bias approach would be more readily accepted by the practitioners--estimators whose risk is minimized subject to the bias being within bounds, or, conversely,

whose bias is minimized subject to the risk being within bounds. The minimum bias theory developed herein is offered as a preliminary step to such developments. Only an example of such compromise approaches is offered here.

Suppose it is required to obtain a linear minimum risk (squared error loss) estimator in the minimax sense of the binomial success probability θ subject to the absolute bias being everywhere bounded by ρ. We shall assume that $\rho < (2\sqrt{n} + 2)^{-1}$, the maximum absolute bias of the (unrestricted) minimax estimator. Because of symmetry, it is easy to show that we only need consider estimators of the form $(x + \alpha)/(n + 2\alpha)$, $\alpha \geq 0$, (which, incidentally, includes all Bayes estimators relative to a symmetric beta prior distribution). Taking $\alpha = n\rho / (1 - 2\rho)$, the maximum absolute bias will be ρ and the maximum risk $(1 - 2\rho)^2 / 4n$, which cannot be further reduced. Thus $\delta = \frac{x}{n}(1 - 2\rho) + \rho$ is the required estimator.

Before considering minimum bias estimation in further detail, we shall review some relevant aspects of the theory of approximation.

3. Some Aspects of the Theory of Approximation

Let Φ represent a specified system of n (finite or infinite) linearly independent bounded and continuous functions $\phi_0, \phi_1, \ldots, \phi_n$ of θ where $\theta \in (\underline{\ }, \overline{\ })$, a subset of the real line (or some other metric space). Let $P = P(\Phi)$ denote the function space of all linear combinations $p = \sum_{i=0}^{n} a_i \phi_i$ where the a_i's are real constants. We call such functions p generalized polynomials of the system Φ. For example,

suppose $\underline{(\bar{\ })}$ is a finite interval and $\phi_i(\theta) = \theta^i$; then P is the space of polynomials of degree n on the interval. Let g be a bounded and continuous function on $\underline{(\bar{\ })}$ but not in P and λ a non-negative bounded and continuous function on $\underline{(\bar{\ })}$.[6]

DEFINITION 1: A function $p_o \in P$ is a best approximation to g in the minimax (Chebyshev) sense if

$$\sup_{\underline{(\bar{\ })}} \lambda|\, p_o - g| = \inf_{P} \sup_{\underline{(\bar{\ })}} \lambda|\, p - g|. \tag{1}$$

We shall assume in what follows that $\lambda \equiv 1$; if not, transform the problem by multiplying all other functions defined above by λ before proceeding.

Suppose now that $\underline{(\bar{\ })}$ is a closed finite or infinite interval and that any generalized polynomial other than the zero polynomial of the system Φ has at most n roots in $\underline{(\bar{\ })}$ (n finite) where a root at which the polynomial does not change sign is counted twice. Φ is then said to be a Chebyshev system of functions. An alternative characterization of a Chebyshev system is that

$$D(\theta) \equiv \begin{vmatrix} \phi_o(\theta) & \cdots & \phi_n(\theta) \\ \phi_o(\theta_1) & \cdots & \phi_n(\theta_1) \\ \cdots & \cdots & \cdots \\ \phi_o(\theta_n) & \cdots & \phi_n(\theta_n) \end{vmatrix}$$

vanishes only at n distinct points $\theta_1, \ldots, \theta_n$ in $\underline{(\bar{\ })}$ and that D changes sign in passing through successive θ_i's.

S. Bernstein (1926) has proved the following result as a generalization of Chebyshev's original work on polynomial approximation: if Φ

---

6. Some of the above restrictions can be relaxed in certain parts of the sequel.

is a Chebyshev system, there exists a best approximation to g in the minimax sense; moreover, p_o is unique, and a necessary and sufficient condition for $p = p_o$ is that the number of points where p-g attains its extremum, with alternating signs, be at least n+2.

Various extensions of this result and upper bounds on (1) are given by Bernstein, by C. de la Vallée Poussin (1919), and by others in more recent publications (many such results appear in N. I. Achieser (1956), J. Favard (1947), P. R. Clement (1953), and D. Jackson (1930)).

If the various functions are differentiable, the function p_o may be obtained as follows, though this method may be untractable analytically: let $b = p - g$ and let $\rho = \sup_\theta |b|$ when $p = p_o$. Let $\theta_0,\ldots,\theta_{n+1}$ be n+2 successive points in $\underline{(\bar{\ })}$ at which ρ is achieved with alternating signs by b, and let $p_o = \Sigma a_i \phi_i$. Then, denoting $b' = db/d\theta$,

$$(2) \qquad b(\theta_i) = \pm \rho(-1)^i, \qquad b'(\theta_i) = 0 \qquad (i = 0,\ldots,n+1)$$

gives 2n+4 equations in the 2n+4 unknowns $\theta_0,\ldots,\theta_{n+1}$, $a_0,\ldots,a_n$, and ρ. (If θ_0 or θ_{n+1} is an endpoint of $\underline{(\bar{\ })}$, then (2) need not hold at i=0 or n+1.)

In particular, if $\phi_i(\theta) = \theta^i$ $(i = 0,\ldots,n)$ and $g(\theta) = \theta^{n+1}$, it can be shown that Chebyshev polynomials can be used to obtain readily the best approximation to g in the minimax sense (see, for example, C. Lanczos (1952)). If, instead, g is any function with a series expansion throughout $\underline{(\bar{\ })}$, the expansion of g in Chebyshev polynomials, truncated after n+1 terms, will yield "almost" the best polynomial approximation to g. Even if g has no valid expansion, the tau-method of C. Lanczos (1952) may lead to an approximate solution, again utilizing Chebyshev polynomials.

A generalization of Bernstein's theorem for more general parameter spaces has been given by J. Bram (1958). He assumes that $\underline{(\bar{\ })}$ is a locally compact space. Then a necessary and sufficient condition that $\sup|b|$ be a minimum is that, for some $r \leq n$, there exist $r+2$ points $\theta_0,\ldots,\theta_{r+1}$ in $\underline{(\bar{\ })}$ such that the $(n+1)\times(r+2)$ matrix $\{\phi_i(\theta_j)\}$ has rank $r+1$ and such that if the subscripts are assigned so that the first $r+1$ rows are independent, and a_i is the sign of the cofactor of a_i in

$$\begin{pmatrix} a_0 & \cdots & a_{r+1} \\ \phi_0(\theta_0) & \cdots & \phi_0(\theta_{r+1}) \\ \cdots & \cdots & \cdots \\ \phi_r(\theta_0) & \cdots & \phi_r(\theta_{r+1}) \end{pmatrix}$$

then $b(\theta_i) = a_i \rho u$ for all i such that $a_i \neq 0$ where $u = \pm 1$.

DEFINITION 2: Let ξ be a probability measure on the Borel sets $\{w\}$ of $\underline{(\bar{\ })}$ and assume g and the ϕ_i's to be square-integrable. A function $p_\xi \in P$ is said to be a <u>best approximation to</u> g <u>in the least squares sense relative to</u> ξ if

$$N(\xi) \equiv \int \lambda^2 (p_\xi - g)^2 \, d\xi = \inf_P \int \lambda^2 (p-g)^2 \, d\xi \ .$$

Again, we shall assume for simplicity that $\lambda \equiv 1$. Analogous developments, using powers other than two, are also possible.

Let $p_0,\ldots,p_n$ constitute an orthonormal set in P w.r.t. the measure ξ (see G. Szegö (1939) or Achieser (1956), for example); i.e.,

$$\int p_i p_j \, d\xi = \delta_{ij} \qquad (\text{Kronecker } \delta_{ij}).$$

Such a set always exists and can be constructed from Φ. Denoting $c_i = \int p_i \, g \, d\xi$, then it is well-known that $p_\xi = \Sigma \, c_i \, p_i$ is a best

approximation to g in the least squares sense, and, moreover, that $N(\xi) = \int g^2 \, d\xi - \Sigma c_i^2$.

As a special case, suppose $\phi_i(\theta) = \theta^i$. Then $p_0,\ldots,p_n$ are the orthonormal polynomials associated with ξ, and p_ξ is the best polynomial approximation to g (in the sense of least squares).

At this point, we note the analogy in Definitions 1 and 2 with the minimax and Bayes solutions to problems in the theory of games, as treated by Wald (1950) and Blackwell and Girshick (1954), for example. We only need replace the role of the risk function or the expected pay-off in decision or game theory by $\lambda|p-g|$ or its square. We pursue this analogy further.

DEFINITION 3: A probability measure ξ_o on $\{w\}$ is said to be <u>least favorable</u> if $N(\xi_o) = \sup N(\xi)$ where the supremum is over all probability measures on $\{w\}$.

Moreover, the fundamental theorem of the theory of games is applicable, so that, with suitable compactness assumptions (Wald (1950)), it readily follows that $p_o = p_{\xi_o}$; i.e., that any best approximation in the minimax sense is also a best approximation in the least squares sense relative to a least favorable distribution.[7] It only needs to be noted that the operations of squaring and taking sup's (or inf's) can be interchanged when applied to the function $\lambda|p-g|$. Thus there exists a norm $\int \lambda^2(p-g)^2 \, d\xi_o$ which is minimized by the same p_o which minimizes the norm $\sup \lambda|p-g|$.

---

7. This result does not appear in the literature to the author's knowledge.

Other decision theory or game theory results will also carry over. For example, a sufficient condition for ξ to be least favorable is that it assign probability one to a subset of $\underline{(\bar{\ })}$ throughout which $\lambda|p_\xi - g| = \sup_{\underline{(\bar{\ })}} \lambda|p_\xi - g|$. Also, with weaker compactness assumptions, a sequence of least squares approximations relative respectively to a sequence of distributions having certain limit properties will yield a minimax approximation, analogously to Bayes solutions in the wide sense. Precise theorems have not been stated here because of the perfect analogy with those published elsewhere.

The possible relevance of the fundamental theorem here, analogous to its other applications, is that constructive methods for finding best approximations in the least squares sense are quite generally available whereas approximation in the minimax sense is usually more difficult; however, there seems to be no constructive, or even intuitive, way of finding least favorable distributions.

As indicated by Bernstein's theorem, ξ_o will frequently assign probability one to a finite point set in which case a result somewhat similar to that above was stated by la Vallée Poussin (1919) (see also J. L. Walsh (1956)).

Other definitions of best approximation are possible; e.g., one might choose p so that $\lambda|p-g|$ is minimized in the neighborhood of θ_o (in some sense). For example, if g possesses a valid series expansion in $\underline{(\bar{\ })}$ at θ_o, and P is the class of n^{th} degree polynomials, then one might approximate g by a truncated expansion at θ_o which lies in P. Alternatively, one may limit consideration to certain classes of polynomials which may have relevance in the particular problem and look

for approximations within this class. Examples of each of these will be mentioned below.

4. Theory of Minimum Bias Estimation

The problem we consider is that of estimating a numerical function g of a parameter θ which indexes the family of probability distributions assumed to generate the sample point x. Extensions to more general situations are possible.

We say a numerical parametric function is estimable if there exists an unbiased estimator of it. (Since we are concerned with situations in which no unbiased estimator is available, our subject is the anomalous one of estimating non-estimable functions!) Our approach is to approximate g by an estimable function and then estimate g by an unbiased estimator of its approximating function. Bounds on the error of approximation, derived in approximation theory, yield bounds on the bias of estimators.

Since all functions which are estimable are also estimable by functions of sufficient statistics, restriction to estimators depending only on sufficient statistics may be made, if desired. No increase in bias will obtain.

We consider a system Φ of estimable functions which generates a class $P(\Phi)$ of functions which are clearly also estimable. Let $\mathcal{D} = \mathcal{D}(\Phi)$ be the class of unbiased estimators of functions p in $P(\Phi)$. For $\delta \in \mathcal{D}$, we denote $\mathcal{E}_\theta \delta = p_\delta(\theta) \in P$. Then $p_\delta - g$ is the bias b of δ as an estimator of g.

An estimator δ_o is said to be a minimum bias estimator of g in the minimax sense if p_{δ_o} is the best approximation to g in the minimax sense. For example, for λ identically unity, δ_o minimizes the maximum absolute bias; for $\lambda = |g|^{-1}$, if finite, δ_o minimizes the maximum relative or percentage bias. The methods of the previous section are available for finding such estimators or approximations to them.

An estimator δ_ξ is said to be a minimum bias estimator of g in the least squares sense relative to ξ if p_ξ is the best approximation to g in the least squares sense relative to ξ. Thus, δ_ξ minimizes the expected quadratic bias $\lambda^2(p-g)^2$ relative to an a priori distribution over the parameter space. Averaging of other bias functions could be considered analogously.

If $p_0,\ldots,p_n$ constitute an orthonormal basis for P, then δ_ξ is equal to $\Sigma\, c_i\, \delta_i$ where the c_i's were defined previously and where the δ_i's are unbiased estimators of the p_i's. In the case of orthonormal polynomials associated with ξ, δ_ξ is an unbiased estimator of the best polynomial approximation to g (in the least squares sense).

Analogously to risk theory, minimum bias estimators in the minimax sense are also minimum bias estimators in the least squares sense relative to a least favorable prior distribution (under appropriate compactness asumptions). It is yet to be demonstrated, however, that this result is of any practical significance in minimum bias theory.

As noted previously, best approximations are frequently unique, and, if confined to functions of a necessary and sufficient statistic, the corresponding estimators of these approximations will frequently also be unique (whenever said statistic is complete). Thus, minimum

bias estimators will frequently be unique, and, as in the case of unbiased estimators, there is no further room for minimization of risk. Moreover, as will be demonstrated later by example, minimum bias estimators need not be admissible (in the risk sense).

Estimators with small local bias might also be considered. For example, in binomial estimation problems, Bhattacharyya (1954) considers estimators whose bias and all existing derivatives thereof up to order n vanish at a point θ_0, and shows that an unbiased estimator of the truncated Taylor expansion of g at θ_0 is such an estimator.

For reasons of convenience, or otherwise, one might restrict attention to certain types of polynomial approximations to g and use unbiased estimators of the approximating function. It is interesting to note that if x is binomially distributed with parameters n and θ, then restriction to Bernstein polynomial approximations (G. G. Lorentz (1953)) to the function g leads to maximum likelihood estimation of g; i.e., the maximum likelihood estimator of g is the unbiased estimator of the Bernstein approximation to g. Results concerning the error of approximation by Bernstein polynomials thus apply to the bias of maximum likelihood estimators.

Admissibility, in terms of bias rather than risk, could also be considered and various complete class theorems derived, in analogy with corresponding theorems in risk theory. Some asymptotic results are also possible. It should be noted that bias is reduced with increasing sample size only if the class of estimable functions increases correspondingly. Specifically, if the class of estimable functions are polynomials of degree m, then bias may be reduced with increasing

sample size n only if m increases with n; if m is fixed, bias is not asymptotically reduced.

The various aspects of minimum bias theory presented here will be illustrated in relation to an example in the Supplement.

5. Acknowledgements

The author would like to acknowledge the assistance of Professor Wassily Hoeffding in making several helpful comments and bringing several references to the author's attention.

References

Achieser, N. I. (1956), Theory of Approximation (translated by C. J. Hyman). New York: Frederick Ungar Publishing Company.

Bahadur, R. R. (1957), "On unbiased estimates of uniformly minimum variance," Sankhyā, 18, 211-224.

Bernstein, S. (1926), "Lecons sur les proprietes extremales," Borel Monograph Series. Paris: Gauthier-Villars.

Bhattacharyya, A. (1954), "Notes on the use of unbiased and biased statistics in the binomial population," Calcutta Statistical Assoc. Bull., 5, 149-164.

Blackwell, David, and Girshick, M. A. (1954), Theory of Games and Statistical Decisions. New York: John Wiley and Sons.

Bram, Joseph (1958), "Chebychev approximation in locally compact spaces," Proc. Amer. Math. Soc., 9, 133-136.

Clement, P. R. (1953), "The Chebyshev approximation method," Quart. Appl. Math., 11, 167-183.

Dynkin, E. B. (1951), "Necessary and sufficient statistics for a family of probability distributions," Uspehi Matem Nauk, 6, 68-90 (Russian: English translation by Statistical Laboratory, Cambridge).

Favard, J. (1947), "Sur l'approximation des fonctions d'une variable réelle," Analyse Harmonique. Paris: Centre National de la Recherche Scientifique.

Fraser, D. A. S. (1957), Nonparametric Methods in Statistics. New York: John Wiley and Sons.

Jackson, Dunham (1930), The Theory of Approximation. New York: Amer. Math. Soc.

Kolmogorov, A. N. (1950), "Unbiased estimates," Izvestiya Akad. Nauk SSR, Ser. Matem., 14, 303-326. (Amer. Math. Soc. Translation No. 98, 1953.)

Lanczos, C. (1952), Introduction to Tables of Chebyshev Polynomials $S_n(x)$ and $C_n(x)$, National Bureau of Standards Applied Mathematics Series, 9. Washington: Govt. Printing Office.

La Vallée Poussin, C. de (1919), "Leçons sur l'approximation des fonctions," Borel Monograph Series. Paris: Gauthier-Villars.

Lehmann, E. L. (1950), Notes on the Theory of Estimation (mimeographed). Berkeley: Associated Students Stores.

Lehmann, E. L., and Scheffé, Henry (1950), "Completeness, similar regions, and unbiased estimation -- part I," Sankhyā, 10, 305-340.

Lorentz, G. G. (1953), Bernstein Polynomials. Toronto: Univ. of Toronto Press.

Siraždinov, S. H. (1956), "Concerning estimations with minimum bias for a binomial distribution," Theory of Probability and Its Applications, Akad. Nauk SSSR, 1, 174-176 (Russian: English summary).

Szegö, G. (1939), Orthogonal Polynomials. New York: Amer. Math. Soc.

Wald, Abraham (1950), Statistical Decision Functions. New York: John Wiley and Sons.

Walsh, J. L. (1956), "Best-approximation polynomials of given degree," Proc. Symp. Appl. Math., 6, 213-218. New York: McGraw-Hill.