A UNIFIED THEORY OF ESTIMATION, I.*

BY

Allan Birnbaum

Columbia University

*This research was sponsored in part by the Office of Naval Research under Contract Number Nonr-266(33), Project Number NR 042-034. Reproduction in whole or in part is permitted for any purpose of the United States Government.

1. General introduction.

The next sections introduce a treatment of problems of statistical estimation in which some previously developed theories are extended and unified.

The formulation of the estimation problem adopted here is of the general kind used in the theory of estimation and testing hypotheses due to Neyman and Pearson: estimation methods are evaluated in terms of probabilities of various possible errors. For statistical inference or decision problems generally, the most general possible formulation of this kind is evidently that of Lindley (1); there, for the problem of deciding among k simple hypotheses, one considers in a logically symmetrical way each of the k(k-1) error probabilities p_{ij} = probability of adopting (when using a given inference rule) hypothesis i when hypothesis j is true, $i \neq j$. The formulation adopted here is equally general, but introduces also the particular ordering among hypotheses which represents the structure of any given parametric estimation problem. For example, in a one-parameter estimation problem, the whole distribution function of an estimator, under each possible value of the parameter, is dealt with rather explicitly and not "summarized" by some functional such as mean squared error.

To this formulation, the concept of admissibility and the decision-theoretic methods due to Wald are applied to obtain characterizations of complete classes of admissible estimators for many problems. Because of the broad definition of admissibility used, these complete classes include all estimators which are admissible with respect to particular loss functions such as mean squared error. Thus the present theory has theoretical and technical

relevance for estimation theories based on different formulations, and in a formal sense it embraces such theories. (Most sections of this series of papers will be readable by statisticians who are not decision theorists.)

Fisher's principle of estimation by maximum likelihood is generalized, given exact (non-asymptotic) justification, and unified with the theory of confidence regions and tests of Neyman and Pearson. These parts of the theory utilize primarily the methods of the latter theory, and yield some new practical techniques of estimation.

The preceding remarks apply primarily to point estimation problems, but the treatment adopted unifies point and confidence-region estimation methods, and incorporates both in an "omnibus" definition of "an estimate" which is proposed and illustrated as a practically useful device as well as a theoretically useful notion. For problems of estimation of a real-valued parameter, such an estimate is called a "confidence curve" and is a representation of the family of upper and lower confidence limits for the parameter at each possible confidence coefficient (thus differing at most in form from the "confidence distributions" described by Cox ((2), p. 363)).

Some problems of sequential estimation are treated. Relations between asymptotic and non-asymptotic theories of estimation are developed in some detail.

2. On various formulations of the problem of point estimation.

We consider here problems of estimation with reference to a specified experiment E, leaving aside for the time being all questions of experimental design including those of choice of a

...ple size or possibly a sequential sampling rule; some definite ...pling rule is assumed to be specified as a part of E. Let $S = \{x\}$ ...note the sample space of possible outcomes x of the experiment. ...t $p(x,\theta)$ denote one of the elementary probability functions on S ...ich are specified as possibly true, and let $\Omega = \{\theta\}$ denote the ...ecified parameter space, so that for each θ in Ω and for each ...beat A of S, the probability that E yields an outcome x in A is ...iven by

$$\text{Prob}\left\{X \in A|\theta\right\} = \int_A p(x,\theta)\, d\mu(x);$$

...ere μ is a specified measure on S. (We assume tacitly here and ...low that consideration is appropriately restricted to measurable ...ets and functions only.)

Let $Y = Y(\theta)$ be any function defined on Ω (an important ...se is $Y(\theta) \equiv \theta$), and let Γ denote its range. Then a point ...timator of Y is a function $g=g(x)$ defined on S and taking ...lues in Γ. (The function g(.) is called an estimator; given a ...ecific observed outcome x, the specific value g(x) is called an ...timate.)

The following definition, containing as it does a number of ...ther vague terms, seems to render as precisely as possible the ...ull intuitive meaning of "the estimation problem": A problem of ...oint-estimation is a problem of choosing a good estimator, that is, ...n estimator g which tends to take values close to the true but ...nknown value of Y.

For purposes of comparing alternative estimators, the only ...precise interpretation which this definition allows is evidently ...he extremely limited one in which, under each hypothesis θ, the

:rrect value $g = \hat{\gamma}(\theta)$ is considered closer to the true value
: an is any other value $g' \neq \hat{\gamma}(\theta)$, but no definite comparisons as
: closeness are possible among different incorrect values. (Note
:at no ordering or parametric structure in Ω has been assumed as
:et.) In Lindley's formulation described above, this definition
:eads to a necessary preference for one estimator with error
:robabilities p_{ij} over a second one with error probabilities p'_{ij}
:ly if $p_{ij} \leq p'_{ij}$ for all i and $j \neq i$ and also $p_{ij} < p'_{ij}$ for some
: and some $j \neq i$. However, in most estimation problems the event
:hat an estimate is precisely correct has typically negligible
:robability and is of little interest.

In order to obtain a definite basis for comparing estimators,
:t is necessary to specify further the "closeness" of the various
: correct values of an estimator under each hypothesis. On the
:her hand, any proposed quantitative measure of closeness, or of
:st or loss due to errors, is necessarily open to the criticism
:hat it is partly subjective or arbitrary and hence is not a
:oroughly satisfactory formulation of the goal of "the estimation
:roblem" as stated above. Nevertheless, beginning with Laplace and
:auss (cf. Neyman ([3], pp. 9-14) for an interesting brief sketch
of the history of the theory
:f the history of estimation, and references), useful formulations
: the estimation problem have been made in terms of specific
:unctions adopted to represent the loss due to each possible error.
:or estimation of a real-valued parameter, the use of the absolute
:alue of the error, introduced by Laplace, was replaced by Gauss
:y the squared error because the latter proved more treatable
:athematically and provided an equally reasonable definite formula-
:ion of the problem. The use of loss functions in estimation and

other statistical problems was given particular emphasis in the decision-theoretic formulations due to Wald.

The criticism that any particular loss function adopted may be somewhat arbitrary seems to be answered for many practical and theoretical purposes by noting (a) that the large bodies of statistical theory and techniques to which these formulations have led include many estimation methods which can be evaluated directly on any basis desired, without necessary reference to the particular loss function adopted originally; (b) that on various such evaluations these methods are found very useful; and (c) that substantial parts of this body of theory and techniques have been found less sensitive than might be anticipated to moderate changes in the specification of the loss function, so that in many problems a number of different "reasonable looking" loss functions will lead to substantially the same estimation methods.

Such formulations dispose of the vagueness of "the estimation problem" by representing the individual who must make inferences by a formal model of a "rational economic man" who has (at least implicit in his patterns of inductive behavior) a utility function defined over the possible outcomes of the inference situation. While these formulations, and the theories and techniques to which they lead, are extremely useful and illuminating, they do not seem to embrace as much of the intuitive content of "the estimation problem" as one might wish, and they introduce elements one might wish to avoid or postpone using as far as is possible in the development of a general and useful theory of estimation. This seems true particularly with regard to the situation of the scientific research worker concerned not with helping to make

immediate decisions but rather with developing knowledge and using estimation methods in his analysis, interpretation, and reporting of research data; here measurement of utilities of various outcomes is rather hypothetical, as is, in important cases, the possibility of depicting his inference-making situation in terms of more definite decision problems.

Even when the vagueness of the estimation problem is dealt with by adoption of a loss function, typically a large class of estimators remain admissible for reasonable consideration, and selecting one of these requires adoption of a further formal criterion of optimality or an informal comparison, judgment and choice. For estimation problems in scientific research, the adoption of any such formal criterion to define an optimum expected loss function of an estimator seems to compound the somewhat arbitrary or subjective choice of the loss function itself; while informal comparisons of estimators can often be made in at least as simple and satisfying a way by direct consideration of error-probabilities of estimators as by consideration of expected values of loss functions.

In general it seems worthwhile to distinguish as far as possible informative inference, as a basic function served by statistical methods in scientific research, from other functions served by statistical methods including those represented in theories of rational decision-making. This distinction is not rendered a purely formal one by the consideration that most statistical theories and techniques may prove useful for both functions; nor even by the fact that the nature of the former function can often usefully be explicated (analyzed and interpreted) by use of the concepts and

formulations of theories of decision-making. The latter fact does not imply that such interpretations exhaust the nature and meaning of the informative inference function. This is not the place for a detailed discussion of this very general methodological question; however, it may be noted that there are problems in which quite different formal theories and practical statistical techniques are appropriate, depending upon whether an informative or a more specific utilitation goal is adopted. A striking example is provided by the problem of estimation of the mean of a multivariate normal distribution, and the formally similar problems of the classical theory of linear estimation (assuming normality of errors). For these problems, the use of the classical estimators for each component or each parametric function seems generally appropriate for purposes of informative inference, and their use is indicated by various formal criteria for estimators, including that of mean-squared error. (Such justification of these classical estimators is discussed in more detail in Section 5 below.) It has been proved by Stein (4), however, that for the problem of estimating simultaneously three or more components of the mean of a multivariate normal distribution (or three or more independent parametric functions), if one adopts as a loss function the sum of squared errors of the respective estimates, then the classical estimators are inadmissible. It seems clear that this result, striking as it is, does not at all detract from the reasonableness of using the classical estimators for informative inference, especially since information about components or parametric functions individually is often of primary interest. The result does point up the need for caution in connection with use of loss functions to represent the goal of informative

nferences about parameters. Stein's result depends crucially on the precise quantitative form specified for the loss function, together with the use of compounding (by adding their loss functions) of distinguishable inference problems which, for purposes of informative inference, there is no clear reason to consider in combination. The usefulness of over-all error rates for a family of confidence interval estimates on parametric functions is not inconsistent with the fact that the most useful families of interval estimates, for informative inferences, consist of confidence intervals each based on a classical estimate.) Stein's result points a direction for development of new statistical techniques for purposes of successful prediction, techniques which may differ strikingly from the classical estimators; and helps to make clear that the different functions of informative inference on the one hand and successful decision-making on the other will sometimes be served best by quite distinct theoretical formulations and practical techniques.

Such considerations lead us to develop as far as possible a theory of estimation in which (a) the notion of closeness to the correct value, of various incorrect values of estimators, is expressed in a precise and useful way without introduction of loss functions or similar devices, and (b) estimators are evaluated in terms of probabilities of errors of various sorts in the style of the Neyman-Pearson theory. For problems of estimation of a real-valued parameter θ, such comparisons as to closeness are available in an obvious way between every two incorrect values each less than the correct value, and between every two incorrect values each greater than the correct value; this leaves open only the comparison of incorrect values on opposite sides of the correct value. We are

this led to develop a theory of estimation based on direct consideration of the whole distribution function of each estimator under each hypothesis. On these terms, one estimator can often but not always be judged necessarily better than a second one. Savage's discussion ((5), pp. 224-225) of various criteria for estimators is unusual in its inclusion of a criterion of admissibility based in this way upon the whole c.d.f. of an estimator. Systematic use of such a criterion seems not to have been made previously.

Apart from the goal of developing a theory of estimation in a form which seems appropriate for purposes of informative inference, the development of the present theory has theoretical and technical relevance for theories of estimation based on other formulations, since the present theory evidently includes all other theories in a formal sense. A knowledge of the admissible class in the present broad sense can be helpful when other approaches are to be applied. For example, every estimate which is admissible with respect to a mean-squared-error loss function must be admissible in the present sense; hence the search for good m.s.e.-estimates can be restricted, without loss, to the admissible class (or a complete class) in the present sense. (In this way, a hierarchy of definitions of admissibility (cf. (6)) leads to a corresponding nested hierarchy of admissible or complete classes.) Again, a criterion of estimation may be shown to lead, for certain problems, to estimates which are not admissible in the present sense; such results facilitate our appraisal of such criteria; e.g. the criterion of unbiasedness has been shown to lead to inadmissible estimates in this sense in some problems of interest. A criterion for choice of an estimator

may be shown to lead, in some classes of problems, to estimates which are always admissible; when this occurs, the reasonableness of the criterion is confirmed, and furthermore any a priori intuitive attractiveness of the criterion, together with this confirmation, may be a satisfying practical basis for resolving the sometimes-difficult problem of choosing an estimator from a large admissible class; under fairly general conditions, maximum likelihood estimators can be given this kind of justification, as will be indicated in a later section. Finally, a criterion which is a priori attractive may lead to an estimator which seems unreasonable; evidently this indicates that some of the intuitive content of the estimation problem is not faithfully expressed by the given criterion, as discussed above in connection with Stein's result. Use of a broad criterion of admissibility can resolve such apparently-incongruous situations, and can exhibit a variety of admissible estimators from among which a satisfying choice may be possible.

2. Admissible estimators.

The present section deals with the concept of admissibility of an estimator of a real-valued parameter θ. (Extensions to multiparameter problems, problems involving nuisance parameters, and problems with other parametric structures will be given in subsequent sections.) Let Ω be any subset of the real line (finite, or countably or uncountably infinite). The family of elementary probability functions $p(x,\theta)$ under consideration has a parametric structure only in that each $p(x,\theta)$ is labeled by a different real number θ. Let $\theta^* = \theta^*(x)$ be any (measurable) estimator; it is useful (as will appear below) to allow the range of θ^* to be

any subset of the closure $\overline{\Omega}$ of Ω (e.g. if Ω is the whole real line, θ^* may take the values $\pm\infty$). For each possible true value θ, the use of an estimator θ^* leads to certain probabilities of over-estimation, represented by the function

$$a(u,\theta,\theta^*) = \begin{cases} \text{Prob}(\theta^*(X) \leq u|\theta) & \text{for each real } u < \theta, \\ \text{Prob}(\theta^*(X) \geq u|\theta) & \text{for each real } u > \theta. \end{cases}$$

The function $a(u,\theta,\theta^*)$ is thus defined at each $\theta\in\Omega$ and at each $u\neq\theta$. For a fixed θ, $a(u,\theta,\theta^*)$ will be called the <u>risk curve</u> of the estimator θ^* at θ.

<u>Definitions</u>. For a given estimation problem, an estimator θ^* is called <u>at least as good as</u> an estimator θ^{**} if $a(u,\theta,\theta^*) \leq a(u,\theta,\theta^{**})$ for all $\theta\in\Omega$ and all $u\neq\theta$. If θ^* <u>and</u> θ^{**} are each at least as good as the other, then $a(u,\theta,\theta^*) \equiv a(u,\theta,\theta^{**})$, and the estimators are called <u>equivalent</u>. If neither of θ^*, θ^{**} is at least as good the other, the two estimators are called <u>not comparable</u>. If θ^* is at least as good as θ^{**} and if $a(u,\theta,\theta^*) < a(u,\theta,\theta^{**})$ for some $\theta\in\Omega$ and some $u\neq\theta$, θ^* is called <u>better than</u> θ^{**}. An estimator θ^* is called <u>admissible</u> if no other estimator is better than θ^*. The class of admissible estimators is called <u>the admissible class</u>. A class of estimators is called <u>complete</u> if, for each estimator outside the class, there is a better one in the class. The <u>minimal</u> (smallest) <u>complete class</u>, if one exists, coincides with the admissible class. A class of estimators is called <u>essentially complete</u> if, for each estimator not in the class, there is one at least as good in the class. A <u>minimal essentially complete class</u>, if one exists, is a subclass of the admissible class.

It is useful to define also, for each θ and θ^*,

$$a(\theta-,\theta,\theta^*) = \text{Prob}\left\{\theta^*(X) < \theta|\theta\right\} = \lim_{\substack{\epsilon \to 0, \\ \epsilon > 0}} a(\theta-\epsilon;\theta,\theta^*) ,$$

$$a(\theta+,\theta,\theta^*) = \text{Prob}\left\{\theta^*(X) > \theta|\theta\right\} = \lim_{\substack{\epsilon \to 0, \\ \epsilon > 0}} a(\theta+\epsilon;\theta,\theta^*) ,$$

for each θ. When reference to a given estimator θ^* is understood, we shall write simply $a(u,\theta)$, $a(\theta-,\theta)$, or $a(\theta+,\theta)$. The functions $a(\theta-,\theta)$ and $a(\theta+,\theta)$ of θ admit useful interpretations, as will be seen below; they will be called, respectively, the lower location function and the upper location function of θ^*.

In many problems, estimators for which $\text{Prob}\left\{\theta^*(X) = \theta|\theta\right\} > 0$ for some θ will be found not useful. Then for the remaining estimators we have $a(\theta-,\theta) \equiv 1-a(\theta+,\theta)$ as the value of the c.d.f. of $\theta^*(X)$ at θ when θ is true. No two such estimators, having different location functions, can be comparable; for $a(\theta-,\theta,\theta^*) < a(\theta-,\theta,\theta^{**})$ is equivalent to $a(\theta+,\theta,\theta^*) > a(\theta+,\theta,\theta^{**})$; this shows that neither estimator is at least as good as the other.

The very broad and "weak" definition of admissibility adopted above leads to very large admissible classes in typical problems. However, it does not seem unreasonable to conceive of the problem of point estimation as one in which the investigator chooses an estimator on the basis of consideration of the risk curves of all estimators in some essentially complete class. In principle this consideration should be complete, but of course the practical counterpart of this can be at least a more or less extensive familiarity with the essentially complete class, developed by study of a variety of particular estimators, possibly strengthened by

some general theoretical considerations, and perhaps also by reference to one or several loss functions and criteria of optimality which may seem more or less appropriate in particular problems. Such an approach is indeed necessary if the undesirable features of formulations based primarily on loss functions are to be avoided.

It may be feared that such an approach entails awkward and formidable tasks of comparisons of risk curves, and is impractical to carry out. It will be seen below that this is not the case.

In most decision-theoretic formulations of statistical problems, a real-valued risk function $r(\theta,\theta^*)$ is defined for each parameter point and each decision function. In the present formulation, we associate with each pair θ,θ^* a set of error-probabilities $a(u,\theta,\theta^*)$, $u \neq \theta$. These respective error-probabilities, for each fixed θ and θ^*, may be regarded as components of a vector denoted by $r(\theta,\theta^*) = \left\{ a(u,\theta,\theta^*) \right\}$, the components $a(u,\theta,\theta^*)$ having index u. (We could define $a(\theta,\theta,\theta^*) = 0$ for formal convenience.) Then $r(\theta,\theta^*)$ is an example of a vector-valued risk function (6).

Example. Let X be normally distributed with unknown mean θ and variance 1, with $\Omega = \left\{ \theta \mid -\infty < \theta < \infty \right\}$. The classical estimator, based on one observation, is $\hat{\theta}(x) = x$. The error-probabilities of this estimator, when $\theta = 1$, are

$$a(u,1,\hat{\theta}) = \begin{cases} \Phi(u-1) & \text{for } u < 1 , \\ 1-\Phi(u-1) & \text{for } u > 1 . \end{cases}$$

This risk curve is graphed in Figure 3.1.

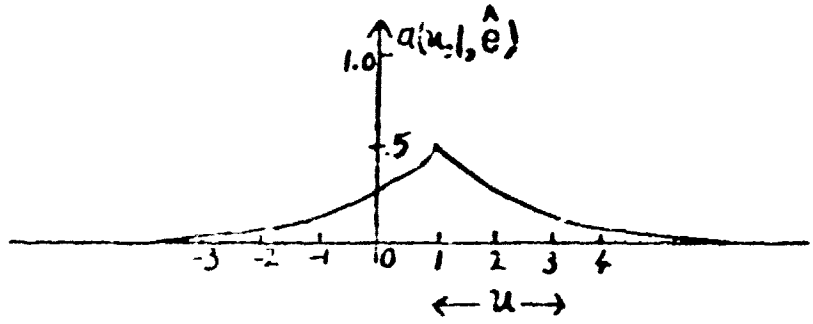


Figure 3.1

A second estimator, $\theta^*(x) = x + 1$, has, when $\theta = 1$, the following error-probabilities, illustrated in Figure 3.2:

$$a(u,1,\theta^*) = \begin{cases} \frac{1}{2}(u-2) & \text{for } u < 1 , \\ 1-\frac{1}{2}(u-2) & \text{for } u > 1 . \end{cases}$$

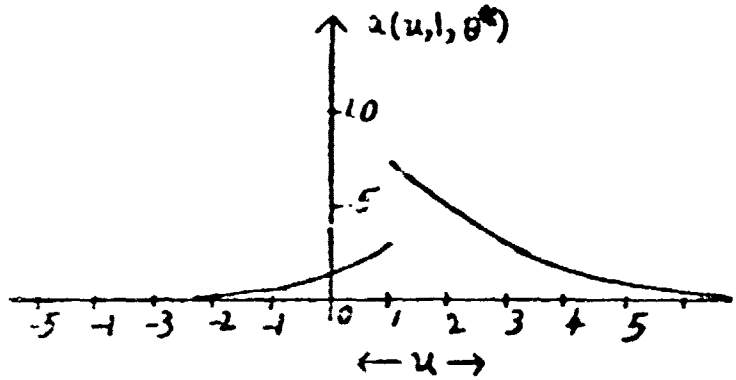


Figure 3.2

Our wishful intuitive goal in choosing an estimator would be to minimize simultaneously all ordinates of such curves (for all θ, and for all u/θ) since each ordinate is the probability of an error

which we wish to avoid; of course this goal cannot be re lized in non-trivial problems. The estimator θ^* is superior to $\hat{\theta}$ with respect to all errors of under-estimation, but is correspondingly worse with respect to errors of over-estimation. From this standpoint neither can be called better than the other; they are not comparable.

A third estimator is the apparently trivial one, $\theta^{**}(x) \equiv +\infty$, whose error-probabilities, when $\theta = 1$, are illustrated in Figure 3.3:

$$a(u,1,\theta^{**}) = \begin{cases} 0 & \text{for } u < 1 , \\ 1 & \text{for } u > 1 . \end{cases}$$

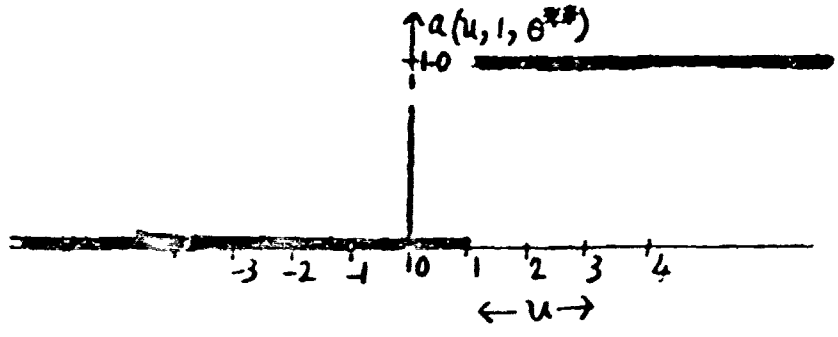


Figure 3.3

This estimator (but no "smaller" one) is perfect in avoiding errors of under-estimation, but is as bad as possible with respect to over-estimation.

This example illustrates the technical usefulness of our slight generalization of the usual definition of an estimator, which allows an estimator to take values not only in Ω but throughout its closure $\overline{\Omega}$.

It will be shown below (a) that each of these three estimators is admissible: (b) that indeed every non-decreasing function $\theta^*(x)$ of x is an admissible estimator; (c) that a minimal essentially complete class of estimators is constituted by the right-continuous non-decreasing functions $\theta^*(x)$ of x; and (d) that among estimators which are "median-unbiased" (that is, which over-estimate and under-estimate with equal probabilities, for each θ), the classical estimator is uniformly best (that is, $\hat{\theta}$ minimizes $a(u,\theta)$ simultaneously for all θ and all $u \neq \theta$).

4. Elementary theory of admissible estimators.

A useful part of the theory of admissible estimators of a real-valued parameter can be developed conveniently by use of the theory of one-sided tests of hypotheses due to Neyman and Pearson. In fact, for problems having a simple structure, the complete theory can be developed in this way. In this section, some useful methods and results are introduced under simplifying assumptions; later sections will contain more general results, such as complete classes of admissible estimators, for some of the problems considered here.

4.1. Relations between estimation and one-sided testing problems.

Any given estimator $\theta^*(x)$ of a real-valued parameter θ can be used in the following way to define a test of the hypothesis $H: \theta < \theta_0$, where θ_0 is given, against $H': \theta \geq \theta_0$:

If $\theta^*(x) < \theta_0$ is observed, infer that $\theta < \theta_0$ ("accept H");

if $\theta^*(x) \geq \theta_0$ is observed, infer that $\theta \geq \theta_0$ ("reject H").

Similarly, θ^* can be used to define a test of $H: \theta \leq \theta_0$ against $H': \theta > \theta_0$ as follows:

If $\theta^*(x) \leq \theta_0$, infer $\theta \leq \theta_0$;

if $\theta^*(x) > \theta_0$, infer $\theta > \theta_0$.

If θ^* is such that

(1) $$\text{Prob}\left\{\theta^*(x) = \theta_0 \middle| \theta\right\} = 0 \quad \text{for all} \quad \theta \in \Omega,$$

then the two tests are equivalent for testing either hypothesis H.

For any estimator θ^* satisfying (1), let $A_{\theta_0} = \left\{x \middle| \theta^*(x) \leq \theta_0\right\}$, the acceptance region of such a test. The Type I errors of such a test have probabilities

$$1 - \text{Prob}\left\{A_{\theta_0} \middle| \theta\right\} = a(\theta_0, \theta, \theta^*) \quad \text{for each} \quad \theta < \theta_0;$$

the Type II errors have probabilities

$$\text{Prob}\left\{A_{\theta_0} \middle| \theta\right\} = a(\theta_0, \theta, \theta^*) \quad \text{for each} \quad \theta > \theta_0.$$

For testing $H: \theta < \theta_0$, the Type II error when $\theta = \theta_0$ has probability

$$\text{Prob}\left\{A_{\theta_0} \middle| \theta_0\right\} = a(\theta_0-, \theta_0, \theta^*) = 1-a(\theta_0+, \theta_0, \theta^*).$$

For testing $H: \theta \leq \theta_0$, the Type I error when $\theta = \theta_0$ has probability

$$1 - \text{Prob}\left\{A_{\theta_0} \middle| \theta_0\right\} = a(\theta_0+, \theta_0, \theta^*) = 1 - a(\theta_0-, \theta_0, \theta^*).$$

In this way, for any estimator θ^* which satisfies (1) and which takes values only in Ω, each of the error-probabilities of θ^* as an estimator can be interpreted as an error-probability

of a one-sided test based on θ^*. Such interpretations give the following useful partial answer to the question whether any given estimator is admissible:

Lemma 1. If, for a specified family of densities $f(x,\theta)$, $\theta \in \Omega$, $\theta^*(x)$ is any estimator satisfying

(a) $\text{Prob}\left\{\theta^*(X) = \theta_0 \middle| \theta\right\} = 0$ for all θ_0 and $\theta \in \Omega$,

(b) $\theta^*(x)$ takes values in Ω only,

(c) $A_{\theta_0} = \left\{x \mid \theta^*(x) \leq \theta_0\right\}$ is the acceptance region of a test which is admissible for testing $H: \theta \leq \theta_0$ against $H': \theta > \theta_0$, and for testing $H: \theta < \theta_0$ against $H': \theta \geq \theta_0$, for each $\theta_0 \in \Omega$,

then θ^* is an admissible estimator.

Proof: If θ^* satisfies (a)-(c) and if θ^{**} is better than θ^*, then for some $\theta_0 \in \Omega$ and some $\theta' \in \Omega$, $\theta' \neq \theta_0$, we have $a(\theta_0, \theta', \theta^{**}) < a(\theta_0, \theta', \theta^*)$, while for all $\theta \in \Omega$, $\theta \neq \theta_0$, we have $a(\theta_0, \theta, \theta^{**}) \leq a(\theta_0, \theta, \theta^*)$. But then $\left\{x \middle| \theta^{**}(x) \leq \theta_0\right\}$ is the acceptance region of a test of $H: \theta \leq \theta_0$ which is better than $\left\{x \middle| \theta^*(x) \leq \theta_0\right\}$, contradicting the assumed admissibility of the latter test. (A test is called admissible if no other test has all error-probabilities at least as small, with at least one strictly smaller.)

For any estimator θ^*, since $\theta_1 < \theta_2$ implies $\left\{x \middle| \theta^*(x) \leq \theta_1\right\} \subset \left\{x \middle| \theta^*(x) \leq \theta_2\right\}$, the acceptance regions $\left\{A_\theta\right\}$ defined as above constitute a non-decreasing sequence of sets with index $\theta \in \Omega$.

Many admissible estimators can be constructed by the device

of determining, for each $\theta_0 \in \Omega$, an admissible acceptance region A_{θ_0} for the hypothesis $H: \theta \leq \theta_0$, in such a way that the sequence $\{A_{\theta_0}\}$ is non-decreasing in θ_0. A simple method for such constructions is given by the following

Corollary 1. For a specified family of densities $f(x,\theta)$, $\theta \in \Omega$, let $v(x,\theta)$ be any function satisfying

(a) for each $\theta \in \Omega$, $v(x,\theta)$ is measurable and $v(X,\theta)$ has a continuous c.d.f.;

(b) for each $x \in S$, $v(x,\theta)$ is decreasing in θ and $v(x,\theta) = 0$ has a (unique) solution $\theta \in \Omega$;

(c) for each $\theta_0 \in \Omega$, $\{x | v(x,\theta_0) \leq 0\}$ is the acceptance region of a test which is admissible for testing $H: \theta \leq \theta_0$ and $H: \theta < \theta_0$.

For each $x \in S$, let $\theta^*(x)$ be the solution θ of $v(x,\theta) = 0$. Then θ^* is an admissible estimator.

Proof: Since $\{x | v(x,\theta_0) \leq 0\} = \{x | \theta^*(x) \leq \theta_0\}$, the corollary follows from Lemma 1.

4.2 Monotone likelihood ratio condition.

Suppose the family $f(x,\theta)$, $\theta \in \Omega$, admits a real-valued sufficient statistics $t = t(x)$ satisfying

(A) t has, for each $\theta \in \Omega$, a density function $h(t,\theta)$ with respect to some measure $\nu = \nu(t)$; and $\theta_1 < \theta_2$ implies that $h(t,\theta_2)/h(t,\theta_1)$ is increasing in t.

If we add the assumption that t has a continuous distribution for each θ, we can define $t(\theta,\alpha)$ so as to satisfy $\alpha = \text{Prob}\{t(X) \leq t(\theta,\alpha) | \theta\}$, for each $\theta \in \Omega$ and each α, $0 \leq \alpha \leq 1$. Let $\alpha(\theta)$ be any function, $0 < \alpha(\theta) < 1$, such that $t(\theta,\alpha(\theta))$ is continuous and strictly increasing in θ. Then the conditions of Corollary 1 are met by $v(x,\theta) = t(x) - t(\theta,\alpha(\theta))$. In fact, the acceptance region (for testing $H: \theta \leq \theta_0$ or $H: \theta < \theta_0$)

$A_{\theta_0} = \left\{x|v(x,\theta_0) \leq 0\right\} = \left\{x|t(x) \leq t(\theta,\alpha(\theta))\right\}$ is well known to minimize simultaneously the probabilities of errors of Types I and II subject to the condition $\text{Prob}\left\{A_{\theta_0}|\theta_0\right\} = \alpha(\theta_0)$. It follows from Corollary 1 that the estimator $\theta^*(x)$, defined for each $x\in S$ as the solution θ of $t(\theta,\alpha(\theta)) = t(x)$, is an admissible estimator; and further, among all estimators with the same location functions

$$a(\theta -, \theta, \theta^*) = 1-a(\theta +, \theta, \theta^*) = \alpha(\theta),$$

$\theta^*(x)$ is uniformly best (that is, θ^* minimizes simultaneously all error probabilities $a(u,\theta)$, $u \neq \theta$, $\theta\in\Omega$). (Equivalently, every increasing function $\theta^*(t)$ taking values in Ω is an admissible estimator.)

Taking $\alpha(\theta) \equiv \alpha$, $0 < \alpha < 1$, $\theta^*(x)$ is familiar as an upper confidence limit with confidence coefficient $1-\alpha$ (and/or a lower confidence limit, with coefficient α).

4.3 Generalized maximum-likelihood estimators.

For a given family $f(x,\theta)$, $\theta\in\Omega$, let $\Delta(\theta)$ and $\Delta'(\theta)$ be any two functions satisfying $\Delta(\theta) > 0$, $\Delta'(\theta) \geq 0$, $\theta + \Delta(\theta) \in \Omega$, and $\theta - \Delta'(\theta) \in \Omega$, for each $\theta\in\Omega$. Assume $f(x,\theta) > 0$ for each $x\in S$ and each $\theta\in\Omega$, and let

$$v(x,\theta) = \frac{\log f[x,\theta + \Delta(\theta)] - \log f(x,\theta - \Delta'(\theta))}{\Delta(\theta) + \Delta'(\theta)} - G(\theta, \alpha(\theta)),$$

where $\alpha(\theta)$ satisfies $0 \leq \alpha(\theta) \leq 1$ and $G(\theta, \alpha(\theta))$ satisfies

$$\alpha(\theta) = \text{Prob}\left\{v(X,\theta) \leq 0|\theta\right\}$$

for each $\theta\in\Omega$. In many problems these conditions can be satisfied by suitable choices of $\Delta(\theta)$, $\Delta'(\theta)$, and $\alpha(\theta)$, and the resulting

function $v(x,\theta)$ will also satisfy the conditions of Corollary 1, which we now assume.

To see that a set $A_{\theta_o} = \left\{x | v(x,\theta_o) \leq 0\right\}$ is necessarily acceptance region of the admissible test of $H: \theta \leq \theta_o$ and of $H: \theta < \theta_o$, note that $v(x,\theta_o)$ is a monotone function of the likelihood ratio $f(x,\theta_o + \Delta(\theta_o))/f(x,\theta_o - \Delta'(\theta_o))$; it follows, by the Neyman-Pearson lemma, that A_{θ_o} gives a test of the simple hypothesis $H_1: \theta = \theta_o - \Delta'(\theta_o)$ against the simple alternative $H_2: \theta = \theta_o + \Delta(\theta_o)$, with minimum Type II error among all tests with the same or smaller Type I error. Since $\text{Prob}\left\{v(X,\theta_o) = 0 | \theta\right\} = 0$ for each $\theta \in \Omega$, this best test is determined essentially uniquely; hence A_{θ_o} represents a test which is admissible for testing $H: \theta \leq \theta_o$ or $H: \theta < \theta_o$.

It follows by Corollary 1 that an admissible estimator is given by $\theta^*(x)$ defined, for each $x \in S$, as the solution θ of $v(x,\theta) = 0$.

If in the above definition of $v(x,\theta)$ we consider for each θ a sequence of choices of $\Delta(\theta)$ and $\Delta'(\theta)$ such that $\Delta(\theta) + \Delta'(\theta) \longrightarrow 0$, we obtain, as a formal limit of the first term of $v(x,\theta)$, $\frac{\partial}{\partial\theta} \log f(x,\theta)$, provided that, for each $x \in S$ and $\theta \in \Omega$, this derivative exists. Leaving aside this limiting process, it is convenient to consider directly an alternative definition of $v(x,\theta)$ as follows:

$$v(x,\theta) = \frac{\partial}{\partial\theta} \log f(x,\theta) - G(\alpha, \alpha(\theta)) \equiv \frac{f'(x,\theta)}{f(x,\theta)} - G(\theta, \alpha(\theta)),$$

assuming existence of the derivative. Here $G(\theta,\alpha(\theta))$ is defined as before so that, for each θ, $\alpha(\theta) = \text{Prob}\left\{v(X,\theta) \leq 0 | \theta\right\}$. The acceptance region $A_{\theta_o} = \left\{x | v(x,\theta_o) \leq 0\right\}$ is clearly equivalent to one based on the inequality $f'(x,\theta_o)/f(x,\theta_o) \leq G(\theta_o, \alpha(\theta_o))$.

This is well known to be the acceptance region of the locally-best test, of size $1-\alpha(\theta_0)$, of the hypothesis H: $\theta = \theta_0$ against H': $\theta > \theta_0$ (and to be the rejection region of the locally-best test, of size $\alpha(\theta_0)$, of H: $\theta = \theta_0$ against H': $\theta < \theta_0$), provided that $f(x,\theta)$ satisfies $\frac{\partial}{\partial\theta}\int_W f(x,\theta)\,d\mu = \int_W \frac{\partial}{\partial\theta} f(x,\theta)\,d\mu$ for each (measurable) WCS. If $v(x,\theta)$ satisfies the conditions of Corollary 1 above, then the following estimator is admissible: For each x, let $\theta^*(x)$ be the solution θ of $v(x,\theta) = 0$, that is, of $\frac{\partial}{\partial\theta}\log f(x,\theta) = 0(\theta,\alpha(\theta))$. Among all estimators with the same location functions, such estimators minimize error-probabilities $a(u,\theta,\theta^*)$ for u in the neighborhood of θ, for each $\theta\in\Omega$.

The estimators obtained in this section are generalizations of the maximum-likelihood estimator $\hat\theta(x)$ which is defined as the solution of the equation $\frac{\partial}{\partial\theta}\log f(x,\theta) = 0$. Thus, under the conditions mentioned, the maximum-likelihood estimator is admissible. It has the location functions $a(\theta-, \theta, \hat\theta) \equiv 1-a(\theta+, \theta, \hat\theta) \equiv \alpha(\theta)$, where $\alpha(\theta) = \text{Prob}\left\{f'(X,\theta)/f(X,\theta) \le 0 \mid \theta\right\}$, where $f'(x,\theta) = \frac{\partial}{\partial\theta} f(x,\theta)$. If $f'(X,\theta)/f(X,\theta)$ has a symmetrical distribution, its median will coincide with its mean, which is $E(f'(X,\theta)/f(X,\theta)\mid\theta) \equiv 0$, giving $\alpha(\theta)\equiv .5$. More generally, if x is a sample of independent observations, the normal approximation (based on the Central Limit Theorem) will often apply relatively well for moderate sample sizes to give $\alpha(\theta) \doteq .5$ as an approximation to the location function of $\hat\theta$.

4.4 Examples.

The first two examples below illustrate that the method of Section 4.2 can often be applied conveniently as a case of the methods of Section 4.3.

Example1. Normal mean. Let $x = (y_1, \ldots y_n)$ be a sample of n independent observations from a normal distribution with known variance, say $\sigma^2 = 1$, and unknown mean θ, $-\infty < \theta < \infty$. Then

$$f(x,\theta) = \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2} .$$

Let

$$v(x,\theta) = \frac{\partial \log f(x,\theta)}{\partial \theta} \sim \mathcal{O}(\theta, \alpha(\theta)), \text{ where } \alpha(\theta) \text{ is}$$

a given function. Then

$$v(x,\theta) = n(\bar{y} - \theta) \sim \mathcal{O}(\theta, \alpha(\theta)) = n\bar{y} - n\theta \sim \sqrt{n}\, \Phi^{-1}(\alpha(\theta)) ,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\Phi(u)$ is the standard normal c.d.f. Then $v(x,\theta)$ clearly satisfies conditions (a) and (c) of Corollary 1. Condition (b) is met if $\alpha(\theta)$ is such that $\theta + \frac{1}{\sqrt{n}} \Phi^{-1}(\alpha(\theta))$ is increasing in θ; as n increases, this condition becomes a less restrictive one on $\alpha(\theta)$; (b) is obviously satisfied if $\alpha(\theta) \equiv \alpha$, $0 \leq \alpha \leq 1$ for all θ. For each function $\alpha(\theta)$ satisfying (b), an admissible estimator $\theta^*(x)$ is defined as the solution θ of $v(x,\theta) = 0$, that is, of

$$\theta + \frac{1}{\sqrt{n}} \Phi^{-1}(\alpha(\theta)) = \bar{y} .$$

Denoting the solution by $Q(\bar{y})$, this gives $\theta^*(x) = Q(\bar{y})$; $Q(\bar{y})$ may be any increasing function of $\bar{y}$ if $\alpha(\theta)$ is suitably chosen. For $\alpha(\theta) \equiv \alpha$, this becomes

$$\theta^*(x) = \bar{y} - \frac{1}{\sqrt{n}} \Phi^{-1}(\alpha) ,$$

an upper confidence limit of confidence coefficient $1-\alpha$ (and/or a

lower confidence limit of coefficient α). Since the function $v(x,\theta)$ in each of these cases also meets the conditions of Section 4.2, each of these estimators is uniformly best among all estimators with the same location functions $a(\theta -,\theta) \geq 1-a(\theta +, \theta) \geq \alpha(\theta)$. Taking $\alpha(\theta) \equiv \frac{1}{2}$ gives $\hat{\theta}^*(x) = \hat{\theta}(x) = \bar{y}$; for this location function, the estimator obtained is independent of the particular value assumed for σ^2 ; hence the classical (maximum likelihood) estimator is uniformly best among all "median-unbiased" estimators of θ even if σ^2 is not known.

Example 2. Normal variance. Let $x = (y_1,\ldots y_n)$ be a sample of n independent observations from a normal distribution with known mean, say $\mu = 0$, and unknown standard deviation $\theta = \sigma$, $0 < \sigma < \infty$. Then

$$f(x,\theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2} \quad . \text{ Let}$$

$v(x,\theta) = \frac{\partial}{\partial\theta} \log f(x,\theta) - G(\theta,\alpha(\theta))$, where $\alpha(\theta)$ is a given function. Then

$$v(x,\theta) = \frac{n}{\sigma}\left(\frac{s^2}{\sigma^2} - 1\right) - G(\sigma,\alpha(\sigma)),$$

where $s^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$.

For a given σ , $\frac{ns^2}{\sigma^2}$ has the Chi-square distribution with n degrees of freedom; hence $G(\sigma, \alpha(\sigma)) = \frac{1}{\sigma}\left(\chi^2_{n,\alpha(\sigma)} - n\right)$, where $\chi^2_{n,\alpha}$ is the lower α-point of the Chi-square distribution with n degrees of freedom. Thus $v(x,\theta) = \frac{1}{\sigma}\left(\frac{ns^2}{\sigma^2} - \chi^2_{n,\alpha(\sigma)}\right)$. If, for example, $\alpha(\sigma) \equiv \alpha$, then $v(x,\theta)$ satisfies conditions (a)-(e) of Corollary 1. Taking $\alpha(\sigma) \equiv \frac{1}{2}$ gives, as the solution of $v(x,\theta) = 0$, the

median-unbiased estimator of σ,

$$\tilde{\sigma} = \tilde{\sigma}(x) = s\sqrt{n/\chi^2_{n,.5}} \ .$$

Similarly, the median-unbiased estimator of σ^2 is

$$\tilde{\sigma}^2 = (\tilde{\sigma})^2 = s^2 n / \chi^2_{n,.5} \ .$$

When n is not small, $n/\chi^2_{n,.5} \doteq 1$, and $\tilde{\sigma} \doteq s$ and $\tilde{\sigma}^2 \doteq s^2$, where s^2 is the classical unbiased estimate. (These relations are considered in more detail in Section 5.11 below.) In each of these cases, the acceptance regions $A_{\theta_o} = \left\{x \middle| v(x,\theta_o) \leq 0\right\}$ are equivalent to those obtainable as in Section 4.2, $A_{\theta_o} = \left\{x \middle| s^2 \leq \sigma_o^2 \chi^2_{n,\alpha}(\sigma_o)/n\right\}$.

Hence each of the estimators considered is uniformly best among all estimators with the same location functions.

The following two examples illustrate the method of Section 4.3.

Example 3. Logistic mean. Let $x = (y_1, \ldots y_n)$ be a sample of n independent observations from a logistic distribution with unknown mean θ: $\text{Prob}\left\{Y \leq y \middle| \theta\right\} = \Psi(y - \theta) = 1/\left(1 + e^{-(y-\theta)}\right)$, $-\infty < y < \infty$, $-\infty < \theta < \infty$; Y has the density function $\psi(y-\theta) = e^{-(y-\theta)}/\left(1+e^{-(y-\theta)}\right)^2$. Taking $\Delta(\theta) \equiv \Delta'(\theta) \equiv \Delta > 0$ and $\alpha(\theta) \equiv \alpha$, $0 < \alpha < 1$, let

$$v(x,\theta) = \frac{1}{2\Delta}\left[\log f(x,\theta+\Delta) - \log f(x,\theta-\Delta)\right] - G(\theta,\alpha)$$

$$= \frac{1}{2\Delta}\left[\sum_{i=1}^{n} (\log \psi(y_i - \theta - \Delta) - \log \psi(y_i - \theta + \Delta)\right] - G(\theta,\alpha) \ .$$

Condition (a) of Corollary 1 is satisfied, since $\psi(Y-\theta-\Delta)/\psi(Y-\theta+\Delta)$

is a positive random variable with a continuous c.d.f. Condition (b) is satisfied, since $\psi(y_i - \theta - \Delta) / \psi(y_i - \theta + \Delta)$ is decreasing in θ from $+\infty$ to $-\infty$, and $G(\theta, \alpha)$ is independent of θ. Condition (c) is satisfied (since $\left\{ x \middle| v(x,\theta_o) \leq 0 \right\}$ is the acceptance region of an essentially-uniquely determined best test, of $H_1: \theta = \theta_o - \Delta$ against $H_2: \theta = \theta_o + \Delta$). Hence an admissible estimator $\theta^*(x)$ is given by the solution θ of $v(x,\theta) = 0$. We have by symmetry that $G(\theta, \frac{1}{2}) = 0$, so that an admissible median-unbiased estimator is $\tilde{\theta}(x)$, the solution θ of

$$\sum_{i=1}^{n} \left[\log \psi(y_i - \theta - \Delta) - \log \psi(y_i - \theta + \Delta) \right] = 0 .$$

This estimator and its risk curves depend upon the particular value Δ chosen; the error-probabilities $a(\theta - \Delta, \theta, \theta^*) = a(\theta + \Delta, \theta, \theta^*)$ have a minimized common value for all θ.

Taking the "limiting case $\Delta \longrightarrow 0$," we define

$$v(x,\theta) = \frac{\partial}{\partial \theta} f(x,\theta) - G(\theta,\alpha)$$

$$= 2 \sum_{i=1}^{n} \Psi(y_i - \theta) - n - G(\theta,\alpha) .$$

The conditions of Corollary 1 can again be verified. To determine $G(\theta,\alpha)$, note that when θ is true, $\Psi(Y - \theta)$ has the unit rectangular distribution; the c.d.f. of $\sum_{i=1}^{n} \Psi(Y_i - \theta)$ can be calculated, or approximated closely, except for extreme values of α and very small n, by a normal distribution, as shown in [7, pp. 244-246]. Since $E(Y - \theta) = \frac{1}{2}$ and $\mathrm{Var}(\Psi(Y - \theta)) = \frac{1}{12}$, we have by the normal approximation $G(\theta,\alpha) \doteq \sqrt{\frac{n}{12}}\, \Phi^{-1}(\alpha)$. Hence the locally-best

estimator $\theta^*(x)$ with location functions $a(\theta-,\theta) \equiv 1-a(\theta+,\theta) \equiv \alpha$ is the solution θ of

$$\sum_{i=1}^{n} \Psi(y_i-\theta) = \frac{n}{2} + \frac{1}{2} G(\theta,\alpha) \doteq \frac{n}{2} + \frac{1}{4}\sqrt{\frac{n}{3}}\,\Phi^{-1}(\alpha) .$$

By symmetry, $G(\theta,\frac{1}{2}) = 0$, so that the locally-best median-unbiased estimator coincides with the maximum likelihood estimator, and is the solution of

$$\sum_{i=1}^{n} \Psi(y_i - \theta) = \frac{n}{2} .$$

Solutions of the latter two equations numerically are easily obtained by use of tables of the functions $\Psi(u)$ [3] .

Example 4. Laplacean mean. Let $x = (y_1,\ldots y_n)$ be a sample of n independent observations from a Laplacean (double exponential) distribution with unknown mean θ, $-\infty < \theta < \infty$, with density function

$$h(y,\theta) = \tfrac{1}{2} e^{-|y-\theta|} , \quad -\infty < y < \infty .$$

For any $\Delta > 0$, let

$$v(x,\theta) = \frac{1}{2\Delta}\left[\log f(x,\theta+\Delta) - \log f(x,\theta-\Delta)\right] - G(\theta,\alpha(\theta))$$

$$= \frac{1}{2\Delta}\left[\sum_{i=1}^{n}\left(|y_i-\theta-\Delta|-|y_i-\theta+\Delta|\right)\right] - G(\theta,\alpha(\theta)).$$

We note that

$$|y-\theta-\Delta|-|y-\theta+\Delta| = \begin{cases} 2\Delta & \text{if } \theta \leq y-\Delta , \\ 2(y-\theta) & \text{if } y-\Delta \leq \theta \leq y+\Delta , \\ -2\Delta & \text{if } y+\Delta \leq \theta , \end{cases}$$

and hence $-2\Delta n \leq \sum_{i=1}^{n}(|y_i-\theta-\Delta|-|y_i-\theta+\Delta|) \leq 2\Delta n$ for all x.

Since $\text{Prob}\left\{X \leq \theta-\Delta \mid \theta\right\} = \frac{1}{2}e^{-\Delta}$, the c.d.f. of $\sum_{i=1}^{n}(|X_i-\theta-\Delta|-|X_i-\theta+\Delta|)$ has a jump of $(\frac{1}{2}e^{-\Delta})^n$ at each end of its range, and is continuously increasing between these jumps. Hence $G(\theta,\mu(\theta))$ is well-defined if $\mu(\theta)$ satisfies $(\frac{1}{2}e^{-\Delta})^n < \mu(\theta) < 1-(\frac{1}{2}e^{-\Delta})^n$ for all θ; if also $\mu(\theta)\equiv\mu$, then $G(\theta,\mu)$ is independent of θ, since the distribution of $X-\theta$ is independent of θ; by symmetry, $G(\theta,\frac{1}{2}) = 0$. A simple computation gives

$$\text{Var}(|Y-\theta-\Delta|-|Y-\theta+\Delta|) = 8(1-e^{-\Delta}-\Delta e^{-\Delta}), = v, \text{ say;}$$

then for n not very small and μ not extreme, the normal approximation to the distribution of $v(X,\theta)$ gives

$$G(\theta,\mu) \doteq \sqrt{nv}\ \Phi^{-1}(\mu).$$

Condition (a) of Corollary 1 is not satisfied by $v(x,\theta)$, but we can resort to Lemma 1 to prove the admissibility of the estimator defined by use of $v(x,\theta)$: For any $\mu(\theta)$ bounded as above, and such that condition (b) of Corollary 1 is satisfied by $v(x,\theta)$, it is easily verified that the estimation $\theta^*(x)$, defined as the solution θ of $v(x,\theta) = 0$, satisfied the conditions of Lemma 1, and hence is admissible.

The median-unbiased estimator $\tilde{\theta}(x)$ obtained as the solution of

$$\sum_{i=1}^{n}\left|(y_i-\Delta)-\theta\right| = \sum_{i=1}^{n}\left|(y_i+\Delta)-\theta\right|$$

(which is easily solved numerically), depends upon the particular

value Δ chosen; the error-probabilities $\alpha(\theta-\Delta,\theta,\tilde{\varphi}) = \alpha(\theta+\Delta,\theta,\tilde{\varphi})$ have a minimized common value for all θ.

Taking the "limiting case $\Delta \longrightarrow 0$," we obtain

$$v(x,\theta) = \sum_{i=1}^{n} I(y_i > \theta) - \sum_{i=1}^{n} I(y_i < \theta) - \sigma(\theta,\mu(\theta)) ,$$

where, for any relation R, the indicator-function I(R) is defined by I(R) = 1 if R is true and I(R) = 0 if R is false. Thus $\sum_{i=1}^{n} I(y_i > \theta) - \sum_{i=1}^{n} I(y_i < \theta)$ is the number of observations y_i exceeding θ minus the number of observations less than θ; with probability one, the observations y_i have n distinct values, and may be ordered, $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. Then

$$\sum_i I(y_i > \theta) - \sum_i I(y_i < \theta) = \begin{cases} n, & \text{if } \theta < y_{(1)} , \\ n-1, & \text{if } \theta = y_{(1)} , \\ n-2, & \text{if } y_{(1)} < \theta < y_{(2)} , \\ \vdots & \\ -n+1, & \text{if } \theta = y_{(n)} , \text{ and} \\ -n, & \text{if } \theta > y_{(n)} . \end{cases}$$

Let r be any integer, $0 < r < n$. It is easily seen that the locally-best test of $H: \theta = \theta_0$ against $H': \theta > \theta_0$, of size $1 - \alpha(\theta_0) = \frac{1}{2^n} \sum_{u=0}^{n-r} \binom{n}{r}$ is given by the acceptance region $A_{\theta_0} = \left\{ x \mid y_{(r)} \leq \theta_0 \right\}$. Taking

$$\alpha(\theta) \equiv 1 - \frac{1}{2^n} \sum_{u=0}^{n-r} \binom{n}{r} , \quad \sigma(\theta,\mu(\theta)) \equiv n + 1 - 2r ,$$

we have $v(x,\theta) = \sum_{i=1}^{n} I(y_i > \theta) - \sum_{i=1}^{n} I(y_i < \theta) - (n + 1 - 2r)$;

with probability one, $v(x,\theta) = 0$ will have a unique solution θ, namely $\theta^*(x) = y_{(r)}$. This estimator satisfies the conditions of Lemma 1, and so is admissible; among all estimators with the same location functions, it minimizes the error-probabilities $a(u,\theta)$ in the neighborhood of θ, for every θ. If n is odd, $\tilde{\theta}(x) = \tilde{y} = y_{(\frac{n+1}{2})}$, the sample median, is an admissible median-unbiased estimator.

5. Median-unbiased estimators.

A part of the admittedly vague goal in "the estimation problem" is expressed in the notion that for each θ the distribution of an estimator should have θ as a "typical" or "central" value in some sense. Past discussions of the use of statistics to describe distributions make clear the simple point that there is no index (parameter) of the location of a distribution which is uniquely reasonable for general descriptive purposes. For the estimation problem as considered here, the use of the criterion (proposed by Brown [3]) of median-unbiased (exact or approximate), to specify suitable centering of a point-estimator's distributions, seems at least as satisfactory as any alternative: An estimator $\theta^*(x)$ of a real-valued parameter θ is called median-unbiased if for each θ its probabilities of overestimation and underestimation are equal:

$$a(\theta +, \theta, \theta^*) = a(\theta -, \theta, \theta^*) \text{ for each } \theta \in \Omega .$$

An advantage of median-unbiased estimators, which seems particularly appropriate for our primary purpose of conveying information about parameters, is a property of invariance: If $\theta^*(x)$ is any median-unbiased estimator of θ, then each strictly monotone of function $\gamma = \gamma(\theta)$ has the median-unbiased estimator $\gamma^* = \gamma(\theta^*(x))$.

The most widely used criterion of location for estimators is of course mean-unbiasedness; the latter term will be used, when necessary to avoid possible terminological confusion, in place of the usual term unbiasedness. Unbiased estimators lack the property of invariance described above (except in the restricted case that $\gamma(\theta)$ is any linear function). This seems to be a disadvantage, at least in principle, for applications to models of experiments in which one parameter has a single significance for the structure of the model, but where it is natural to consider several different functions of the parameter, each of which has a distinct meaningful interpretation. A simple example is the standard deviation σ of a normal distribution; σ and the variance σ^2 have distinct interpretations, and it is sometimes desired to consider estimates of each in the same context. For such purposes it is desirable to define, if possible, good estimators of each parameter which are consistent with the mathematical relation between the parameters.

A second reason for considering alternatives to the mean-unbiasedness criterion for our present purpose is that this criterion seems somewhat akin to criteria based on loss functions, use of which we have set out to avoid as far as possible.

A third and perhaps strongest reason is that this criterion sometimes forces the use of an estimator which could otherwise be strictly improved upon from the standpoint of admissibility. A familiar example is the problem of estimating a component of variance. Here the obvious improvement of replacing a negative estimate by the value zero can be made only if the criterion of unbiasedness is dropped. On the other hand, in problems where

some θ^* satisfies $a(\theta,\theta^*) = a(\theta',\theta^*) < 1/2$ on, for example, $\Omega = \{\theta | 0 < \theta < \infty\}$, θ^* may be redefined as zero on any x for which it is negative, resulting in general in an improvement without altering the median-unbiasedness of θ^*.

The most important use made of the principle of mean-unbiasedness seems to be in the development of the theory and techniques of linear estimation. For many purposes the basic assumptions of that theory, which concern only the means and covariances of observational errors, are supplemented by a normality assumption. Whenever the latter assumption is added, we can (as indicated in Example 1 of Section 4 above) dispense with the justification usually given for the classical estimators of linear regression theory which includes the mean-unbiasedness criterion, and instead recommend the classical estimators on the following basis, which seems superior from the present standpoint for both theoretical and practical purposes: the classical estimators are (a) median-unbiased (and hence have the invariance property described above), and (b) among all such estimators, they minimize simultaneously all probabilities of over-estimation and under-estimation of every parametric function which is "estimable" (as defined in that theory). If this justification for the classical estimators is adopted, the use of mean-unbiasedness properties remains essential at the level of technical development of the theory and techniques, but does not play any role at the level of basic justifying criteria. The theory of linear estimation without normality assumptions is not strictly comparable with the present discussion, since it is based on an incompletely specified model for the observational errors (only first and second moments being specified) while the present

approach assumes fully specified probability models $p(x,\theta)$; however, the preceding comments on the important special case of normally distributed errors seem of some relevance for the more general theory.

A second extensive area of use of the mean-unbiasedness criterion is in asymptotic theories of estimation. However this use is typically made with reference only to asymptotically normally distributed estimators. Here, as with the linear estimators just discussed, and indeed wherever a problem has essentially the structure of the problem of estimation of the mean of a normal distribution, the criteria of mean and median-unbiasedness are generally satisfied by the same estimators; it may then be considered a matter of choice which (if either) criterion is preferred and adopted as a justifying criterion, and which is to be considered an entailed property. (A more detailed discussion of some aspects of asymptotic estimation theory will be given in a later section.)

An important and distinct part of the estimation theory developed in the following sections concerns median-unbiased estimators, particularly admissible ones. While this section has described certain advantages of median-unbiasedness as a criterion for point-estimators, it will be recalled that our program is not to use a theory of point-estimators as a proposed solution of "the estimation problem", but to use satisfactory point-estimators and confidence limits at various confidence levels jointly to form omnibus estimators called "confidence curves" which will be proposed as relatively satisfactory solutions for many estimation problems. An important additional advantage of median-unbiased point estimators is that they fit well into this program.

Concerning the development of theory and techniques of median-unbiased point estimation, it is useful to note the following: Every upper confidence limit with confidence coefficient .5 (which is also a lower confidence limit with the same coefficient) is a median-unbiased estimator (with minor qualifications). This simple observation leads us to adopt for our purposes, first of all, the considerable body of theory and techniques of estimation by confidence limits, and more generally to consider in principle all admissible estimators with location functions $a(\theta-,\theta) \equiv a(\theta+,\theta) \equiv .5$.

The following comments complement the definition of median-unbiasedness given above:

If any estimator θ^* satisfies $\text{Prob}\left\{\theta^* = \theta \mid \theta\right\} = 0$ for all θ, then θ^* is median-unbiased if and only if

$$a(\theta-;\theta,\theta^*) = a(\theta+;\theta,\theta^*) = \tfrac{1}{2} \quad \text{for each} \quad \theta \in \Omega .$$

Estimators satisfying $\text{Prob}\left\{\theta^* = \theta \mid \theta'\right\} > 0$ for some θ and θ' will in general not be median-unbiased. For example, for the binomial distribution

$f(x,\theta) = \binom{n}{x} \theta^x(1-\theta)^{n-x}$, $x=0,1,\ldots n$, with $0 \leq \theta \leq 1$, the range of each estimator $\theta^*(x)$ is a set of at most $n+1$ points in the closed unit interval. Then $a(\theta-,\theta,\theta^*) = 1-a(\theta+,\theta,\theta^*)$ except at points in the range of θ^*; and $a(\theta-,\theta,\theta^*)$, $a(\theta+,\theta,\theta^*)$ are continuous except at these points, where discontinuities occur. Hence $\theta^*(x)$ cannot be median-unbiased.

Discreteness of distributions $f(x,\theta)$ does not necessarily imply non-existence of median-unbiased estimators. If $S = \left\{x\right\}$ is a discrete sample space, it is possible to employ in a customary way

an augmented sample space $S' = \{(x,y(x))\}$, where $y(x)$ is an observed value of an auxiliary random variable $Y(x)$ whose continuous distribution may depend upon the observed x but not upon θ. Then estimators of the form $\theta^*(x,y(x))$ may be median-unbiased in problems for which no $\theta^*(x)$ is median-unbiased. Such augmented sample spaces S' are, in fact, formally included as possible cases of S throughout our general discussion. They are useful technically and theoretically to give a formal unity to theoretical developments. An estimator $\theta^*(x,y(x))$ whose values depend in an essential way on the value $y(x)$ of a randomization variable $Y(x)$ may be useful for some decision or prediction applications, but

it seems clear (in the writer's opinion)

that such estimators should not be used for informative inference purposes even though it is sometimes useful to consider them in the formal development of estimation theory. It follows that exact attainment of median-unbiasedness will sometimes be inconsistent with the purpose of making informative inferences.

In some problems with continuous densities $f(x,\theta)$, median-unbiased estimators exist but are all inadmissible. For example, in the estimation of the mean of a normal distribution (Example 1 of Section 4 above), if we take $\Omega = \{\theta | 0 < \theta < \infty\}$, then the classical estimator modified in the natural way,

$$\theta^*(x) = \begin{cases} \bar{y} & \text{if } \bar{y} \geq 0, \\ 0 & \text{if } \bar{y} \leq 0, \end{cases}$$

is admissible and is uniformly best among median-unbiased estimators. But if we take $\Omega = \{\theta | 0 \leq \theta < \infty\}$, the same estimator is admissible, but the condition of median-unbiasedness fails just at $\theta = 0$, where $a(0-,0,\theta^*) = 0$ while $a(0+,0,\theta^*) = \frac{1}{2}$; an estimator

$$\theta^{**}(x) = \begin{cases} \bar{y} & \text{if } \bar{y} \geq 0 , \\ -\epsilon & \text{if } \bar{y} < 0 , \end{cases}$$

where ϵ is any positive number, is median-unbiased but inadmissible.

It is useful and natural to define the median-bias of any estimator θ^* at θ as

$$B(\theta,\theta^*) = \text{Prob}\left\{\theta^*(X) > \theta \middle| \theta\right\} - \text{Prob}\left\{\theta^*(X) < \theta \middle| \theta\right\}$$

$$= a(\theta-,\theta,\theta^*) - a(\theta-,\theta,\theta^*) ,$$

so that "positive median-bias" signifies a larger probability of overestimation than of underestimation. In some problems it is useful to consider estimators which "minimize $|B(\theta,\theta^*)|$ for $\theta \in \Omega$" in some sense. For example, the modified classical estimator θ^* of the normal mean given above is admissible, and among admissible estimators none is strictly better with respect to median-bias, since any admissible estimator $\theta'(x)$ with $|B(0,\theta')| < B(0,\theta^*) = \frac{1}{2}$ gives, for some $\theta > 0$, $B(\theta,\theta') < B(\theta,\theta^*) = C$.

An estimator θ^* will be called approximately median-unbiased if $B(\theta,\theta^*)$ is close to zero for all $\theta \in \Omega$. In some problems, such as those with discrete distributions $f(x,\theta)$, it is useful to consider admissible estimators which minimize the maximum of $|B(\theta,\theta^*)|$ for $\theta \in \Omega$. Examples are given below.

Despite these necessary qualifications, the property of median-unbiasedness remains a guiding criterion for point-estimators which it is useful to consider, along with the admissibility criterion, in developing estimators which may be useful for making informative inferences.

For additional comments on various criteria of unbiasedness, see Brown [9] and Savage ([5], p. 244). For a decision-theoretic

approach to such criteria, see Lehmann [10].

5.1 Best median-unbiased estimators in simple standard problems.

The simplest standard problems of estimation are those in which the family of densities $f(x,\theta)$, $\theta \in \Omega$, satisfies the monotone likelihood ratio condition (A) of Section 4.2 above. For the problem of estimating the mean of a normal distribution (with known or unknown variance), the classical estimator $\hat{\theta} = \bar{y}$ was shown (in Section 4.4 above) to be the best median-unbiased estimator when Ω is the real line. (Other cases of Ω were discussed above in the present section.)

5.11 Normal variance or standard deviation.

In the problem of Example 2 of Section 4.4 above, it was shown that the best median-unbiased estimator of a normal variance σ^2 is

$$\tilde{\sigma}^2 = s^2 k_n^2 , \text{ where } k_n^2 = n / \chi^2_{n,.5} ,$$

and similarly for a normal standard deviation

$$\tilde{\sigma} = s\, k_n ,$$

where s^2 is the usual unbiased estimator of σ^2 based on n degrees of freedom.

Table 5.1 gives the values of k_n^2 and k_n, for various n, which can be used to compute $\tilde{\sigma}^2$ or $\tilde{\sigma}$ from values of the classical estimates s^2 or s. Since $1 < k_n < 1.02$ for $n \geq 18$ and $1 < k_n < 1.06$ for $n \geq 6$, this modification of the classical estimators is a quantitatively minor one except for small n, and for many purposes it will suffice to take the approximate values $\tilde{\sigma}^2 \doteq s^2$ and $\tilde{\sigma} \doteq s$

except for small n. Figure 5.1 gives graphs of the functions k_n^2 and k_n for $n \leq 40$.

From the standpoint of criteria for estimators, if the criteria of admissibility and median-unbiasedness are adopted for this problem, the latter relationships show that the classical estimators s^2 and s are justified as very convenient and close approximations to the optimal ones, except for small n.

These relationships provide a justification for use of the classical estimators, except for small n, which seems generally more satisfactory (despite the approximation involved) than the usual one based on mean-unbiasedness. The magnitudes of the median-bias of the usual estimators,

$$B(\sigma^2,s^2) \equiv B(\sigma,s) = \text{Prob}\left\{ s^2 > \sigma^2 | \sigma^2 \right\} - \text{Prob}\left\{ s^2 < \sigma^2 | \sigma^2 \right\},$$

are independant of σ; they are shown in Table 5.2 for various n. For example, for $n = 70$, s^2 underestimates with probability $.511 = .50 + \frac{1}{2}(.022)$.

TABLE 5.1

CONSTANTS FOR COMPUTING BEST MEDIAN-UNBIASED ESTIMATES OF THE VARIANCE OR STANDARD DEVIATION OF A NORMAL DISTRIBUTION, FROM VALUES OF THE CLASSICAL ESTIMATES

$\tilde{\sigma}^2 = k_n^2 s^2$ and $\tilde{\sigma} = k_n s$, where $s^2 = \left[\frac{\sum_{i=1}^{n+1} (x_i - \bar{x})^2}{n}\right]$ if $\mu = E(X)$ is unknown, and $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$ if μ is known.

n	k_n^2	k_n
1	2.198	1.483
2	1.443	1.201
3	1.268	1.126
4	1.192	1.092
5	1.149	1.072
6	1.122	1.059
7	1.103	1.050
8	1.089	1.044
9	1.079	1.039
10	1.070	1.035
11	1.063	1.031
12	1.058	1.029
13	1.054	1.026
14	1.050	1.025
15	1.046	1.023
16	1.043	1.021
17	1.041	1.020
18	1.038	1.019
19	1.036	1.018
20	1.034	1.017

n	k_n^2	k_n
21	1.033	1.016
22	1.031	1.015
23	1.030	1.015
24	1.028	1.014
25	1.027	1.014
26	1.026	1.013
27	1.025	1.013
28	1.024	1.012
29	1.023	1.012
30	1.023	1.011
40	1.017	1.008
50	1.013	1.007
60	1.011	1.006
70	1.010	1.005
80	1.008	1.004
90	1.007	1.004
100	1.007	1.003
1000	1.001	1.000

FIGURE 5.1

COEFFICIENTS FOR COMPUTING MINIMUM MEDIAN-UNBIASED ESTIMATES OF THE VARIANCE OR STANDARD DEVIATION OF A NORMAL DISTRIBUTION, FROM VALUES OF THE CLASSICAL ESTIMATES

$\tilde{\sigma}^2 = k_n^2 s^2$ and $\tilde{\sigma} = k_n s$, where $s^2 = \left[\sum_{i=1}^{n-1} (x_i - \bar{x})^2\right] / n$ if $\mu = E(X)$ is unknown, and $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$ if μ is known.

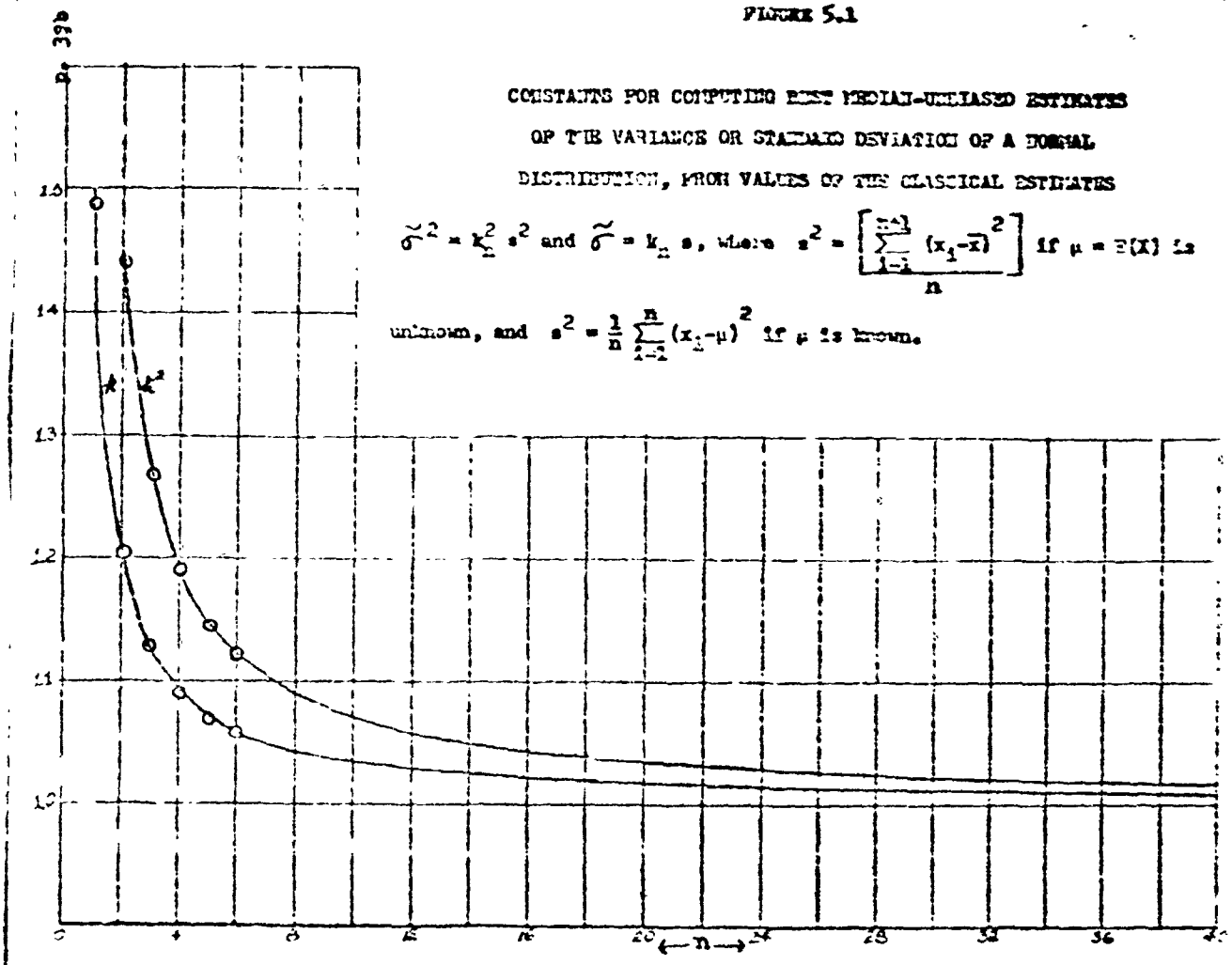


TABLE 5.2

MEDIAN-BIAS OF CLASSICAL UNBIASED ESTIMATOR s^2 OF THE VARIANCE σ^2

OF A NORMAL DISTRIBUTION

Bias $B(s^2, \sigma^2) = \text{Prob}\left\{s^2 > \sigma^2\right\} - \text{Prob}\left\{s^2 < \sigma^2\right\} = 2\left(\text{Prob}\left\{s^2 > \sigma^2\right\} - \frac{1}{2}\right)$.

n	$\frac{1}{2} B(s^2, \sigma^2)$
1	-.183
2	-.132
3	-.108
4	-.094
5	-.084
6	-.077
7	-.071
8	-.067
9	-.063
10	-.060
15	-.049
20	-.042
30	-.034
40	-.030
50	-.027
60	-.024
70	-.023

5.12 Poisson mean. Let $x = (y_1,\dots y_n)$ be a sample of n independent observations from a Poisson distribution with unknown mean θ, $0 < \theta < \infty$; then

$$f(x,\theta) = \text{Prob}\left\{(Y_1,\dots Y_n) = (y_1,\dots y_n)\middle|\theta\right\}$$

satisfies the monotone likelihood ratio condition (A) of Section 4.2, with the sufficient statistic $z = \sum_{i=1}^{n} y_i$ distributed according to

$$b(z,\theta) = \text{Prob}\left\{Z = z\middle|\theta\right\} = e^{-n\theta}(n\theta)^z/z!,\ z=0,1,2,\dots .$$

By a simple generalization of Lemma 1 (dispensing with assumption (a), which does not hold for non-randomized estimators when $f(x,\theta)$ is discrete), it can be shown that each non-decreasing function $\theta^*(z)$ of z, taking non-negative values, is an admissible estimator, and is uniformly best among estimators with the same location functions. Among such estimators, we can choose one for which $|B(\theta,\theta^*)|$ is generally relatively small, as follows: For each $\theta_o > 0$ (following the method of Section 4.1 without the restriction of condition (1)), we seek a test of H_1: $\theta < \theta_o$ and a test of H_2: $\theta \leq \theta_o$, with respective acceptance regions $\left\{z|z < z_1\right\}$, $\left\{z|z \leq z_2\right\}$ with $z_1 \leq z_2$, not necessarily integers, chosen so as to minimize the maximum of

$$\left|\frac{1}{2} - \text{Prob}\left\{Z < z_1\middle|\theta_o\right\}\right| , \left|\frac{1}{2} - \text{Prob}\left\{Z > z_2\middle|\theta_o\right\}\right| .$$

It is easily verified that (following Section 4.1) this leads to a nested sequence of acceptance regions, for all $\theta_o > 0$, and the following corresponding estimator: For each $z = 0,1,2,\dots$, $\tilde{\theta}(z)$ is the value of θ for which $\frac{z}{n}$ is the median of Z in the precise sense th

TABLE 5.3

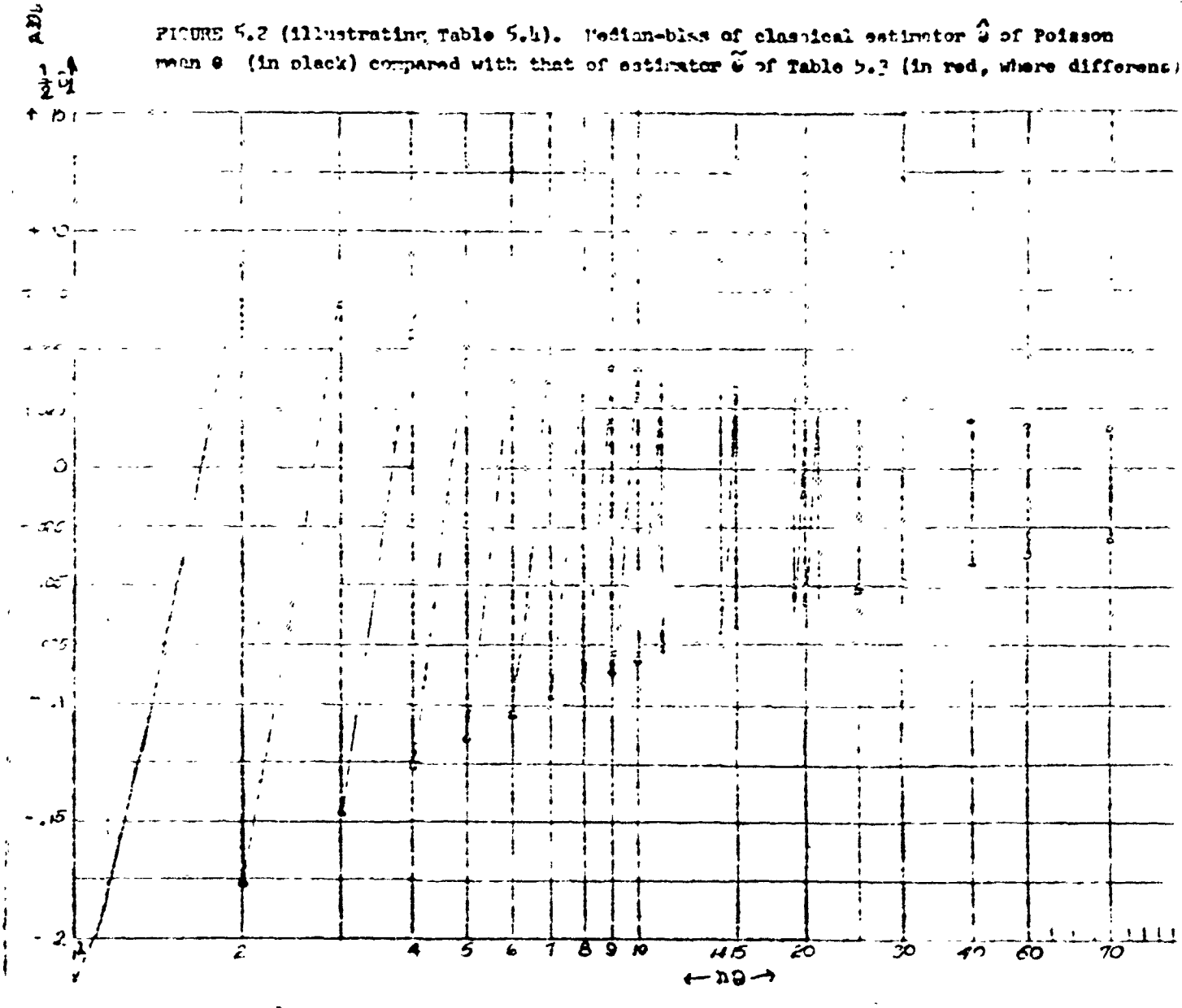
APPROXIMATELY MEDIAN-UNBIASED ESTIMATOR $\tilde{\theta}$ OF POISSON MEAN θ .

$z = \sum_{i=1}^{n} y_i$ = sample total = $n\hat{\theta}$, where $\hat{\theta}$ = classical estimate.

$\tilde{\theta}$ = approximately median-unbiased estimate.

z	$n\tilde{\theta}(z)$
0	0
1	1.146
2	2.156
3	3.159
4	4.161
5	5.162
6	
7	
8	
9	9.166
10	10.165
11	
12	
13	
14	14.165
15	
20	20.17
25	25.17

FIGURE 5.2 (illustrating Table 5.4). Median-bias of classical estimator $\hat{\theta}$ of Poisson mean θ (in black) compared with that of estimator $\tilde{\theta}$ of Table 5.2 (in red, where different).



MEDIAN-BIAS OF [illegible] $= \frac{n}{\sum_{i=1}^{n} y_i}/n$ OF THE MEAN θ OF A

POISSON DISTRIBUTION FOR THE [illegible] OF ESTIMATOR $\tilde{\theta}$ OF TABLE 5.3.
For each sample size, $\text{Prob}\{\tilde{\theta} > \theta|\theta\} = \text{Prob}\{\hat{\theta} < \theta|\theta\} = B(n\theta, n\tilde{\theta})$,
and $\text{Prob}\{\tilde{\theta} > \theta|\theta\} = \text{Prob}\{\hat{\theta} < \theta|\theta\} = B(n\theta, n\tilde{\theta})$.

$n\theta$	$\frac{1}{2}B(n\theta,n\hat{\theta})$	$\frac{1}{2}B(n\theta,n\tilde{\theta})$	$n\theta$	$\frac{1}{2}B(n\theta,n\hat{\theta})$	$\frac{1}{2}B(n\theta,n\tilde{\theta})$
0.0	[illegible]	Zero, except			
0.0+	[illegible]	where values	3.2	- .103	
.005	- .[illegible]	are given.	3.4	- .058	
.01	- .[illegible]		3.6	- .015	
			3.8	+ .027	
.05	- .[illegible]		4.0	- .129	+ .067
.10	- .[illegible]		4.2	- .090	
.15	- .[illegible]		4.4	- .051	
.20	.[illegible]		4.6	- .013	
.25	.[illegible]		4.8	+ .024	
.30	- .[illegible]		5.0	- .116	+ .060
.40	- .[illegible]		5.2	- .080	
.50	- .[illegible]		5.4	- .046	
.60	- .[illegible]		5.6	- .012	
.70	.[illegible]		5.8	+ .022	
.80	.[illegible]		6.0	- .106	+ .054
.90	.[illegible]		6.5	- .027	
1.0	- .[illegible]	+ .132	7.0	- .090	+ .050
1.1	.[illegible]		7.5	- .025	
1.2	.[illegible]		8.0	- .093	+ .047
1.3	.[illegible]		8.5	- .023	
1.4	- .[illegible]		9.0	- .087	+ .044
1.5	- .[illegible]		9.2		
1.6	- .[illegible]		9.4		
1.7	- .[illegible]		9.6		
1.8	+ .[illegible]		9.8		
1.9	+ .[illegible]		10.0	- .083	+ .042
2.0	.[illegible]	+ .094	10.5		
2.2	- .[illegible]		11.0	- .079	+ .040
2.4	- .[illegible]		14.0	- .070	+ .036
2.6	- .[illegible]		.2		
2.8	+ .[illegible]		.4		
3.0	- .[illegible]	+ .077	.6		
			.8		

Table 5.4, continued

$n\theta$	$\frac{1}{2} B(n\theta, n\hat{\theta})$	$\frac{1}{2} B(n\theta, n\tilde{\theta})$
15.0	- .068	+ .034
.2		
.4		
.6		
.8		
16.0		
19.0	- .061	+ .031
.2		
.4		
.6		
.8		
20.0	- .059	+ .030
.2		
.4		
.6		
.8		
21.0	- .058	+ .029
25.0	- .053	+ .027
30.0	- .048	+ .024
40.0	- .042	+ .021
50.0	- .038	+ .019
70.0	- .032	+ .016
100.0	- .027	+ .013

$$\text{Prob}\left\{Z < z \middle| \theta = \tilde{\theta}(z)\right\} = \text{Prob}\left\{Z > z \middle| \theta = \tilde{\theta}(z)\right\}.$$

Such values of $\tilde{\theta}(z)$ are easily determined by use of tables of the Poisson distribution, and are illustrated in Table 5.3. A single such table suffices for all sample sizes n, since the distribution of z depends on $n\theta$ but not on n and θ separately; hence the table gives, for each z, the value of $n\tilde{\theta}(z)$, which is to be divided by the sample size n occurring any particular application.

It will be seen that $\tilde{\theta}(z)$ differs only slightly from the classical estimator $\hat{\theta}(z) = z/n$: for $0 \leq z \leq$,

$$z \equiv n\hat{\theta}(z) \leq n\tilde{\theta}(z) \leq z + 0.2 \quad .$$

Figure 5.2 and Table 5.4 compare the median-bias functions of $\tilde{\theta}(z)$ and $\hat{\theta}(z)$; again the differences are slight. Thus for most purposes the classical estimator $\hat{\theta}(z)$ will serve as a convenient and close approximation to $\tilde{\theta}(z)$. From the standpoint of criteria for estimators, these comparisons provide a justification for use of the classical estimator (as being admissible and having approximately minimizes median-bias among non-randomized admissible estimators) which seems generally more satisfactory than the usual justification (based on mean-unbiasedness).

5.13 Binomial mean. Let $x = (y_1, \ldots y_n)$ be a sample of n independent Bernoulli observations, $\text{Prob}\left\{Y_i = 1|\theta\right\} = \theta$, $\text{Prob}\left\{Y_i = 0|\theta\right\} = 1-\theta$, when θ is unknown, $0 < \theta < 1$. Then, proceeding as in the Poisson case above we obtain the sufficient statistic $z = \sum_{i=1}^{n} y_i$ with the binomial distribution

$$h(z,\theta) = \text{Prob}\left\{Z = z|\theta\right\} = \binom{n}{z}\theta^z(1-\theta)^{n-z} \quad , z = 0,1,\ldots n.$$

For each sample size n, an estimator $\tilde{\theta}(z)$ can be determined which

has minimum median-bias in the same sense adopted in the Poisson case. Table 5.5 gives such estimates $\tilde{\theta}(z)$, for sample sizes $n = 3,5,10$, and 20, in comparison with the classical estimates $\hat{\theta}(z) = z/n$. For $n=2$, $\hat{\theta}(z) \equiv \tilde{\theta}(z)$. It will be seen that $|\tilde{\theta}(z) - \hat{\theta}(z)| \leq .016$ for $n \geq 5$, and $|\tilde{\theta}(z) - \hat{\theta}(z)| \leq .007$ for $n \geq 20$. For $n=10$, the median-bias of the classical estimator is compared with that of $\tilde{\theta}$ in Table 5.6 and Figure 5.3. For $n=20$, the same comparison is given in Table 5.7 and Figure 5.4. Again all of the differences are slight.

Thus if the estimator $\tilde{\theta}(z)$ is adopted on the criteria that it is admissible and has minimum median-bias among non-randomized admissible estimators, then for many purposes the classical estimator $\hat{\theta}$ will serve as a convenient and close approximation to $\tilde{\theta}$. In this way, use of the classical estimators receives a new justification which, depite the approximation involved, seems generally more satisfactory than the usual justification based on mean-unbiasedness.

6. Acknowledgment. The writer is grateful to Mr. Leslie Zurick for computing and preparing tables.

TABLE 5.5

APPROXIMATELY MEDIAN-UNBIASED ESTIMATOR $\tilde{\theta}$ OF THE BINOMIAL PARAMETER θ, COMPARED WITH THE CLASSICAL ESTIMATOR $\hat{\theta}$.

$$x = \sum_{i=1}^{n} y_i \, , \quad \hat{\theta} = x/n \, .$$

$n = 3$

x	$\hat{\theta}$	$\tilde{\theta}$
0	0	0
1	.3	.347
2	.6	.653
3	1.0	1.0

$n = 5$

x	$\hat{\theta}$	$\tilde{\theta}$
0	0	0
1	.2	.216
2	.4	.406
3	.6	.594
4	.8	.784
5	1.0	1.0

$n = 10$

x	$\hat{\theta}$	$\tilde{\theta}$
0	0	0
1	.1	.111
2	.2	.209
3	.3	.306
4	.4	.403
5	.5	.5
6	.6	.597
7	.7	.694
8	.8	.791
9	.9	.889
10	1.0	1.0

$n = 20$

x	$\hat{\theta}$	$\tilde{\theta}$
0	0	0
1	.05	.057
2	.1	.106
3	.15	.156
4	.2	.205
5	.25	.254
6	.3	.303
7	.35	.352
8	.4	.402
9	.45	.451
10	.5	.5
11	.55	.549
12	.6	.598
13	.65	.648
14	.7	.697
15	.75	.746
16	.8	.795
17	.85	.845
18	.9	.894
19	.95	.943
20	1.0	1.0

FIGURE 5.3 (illustrating Table 5.6). Median-bias of classical estimator $\hat{\theta}$ of binomial parameter θ (in black), for sample size $n = 10$, compared with that of estimator $\tilde{\theta}$ of Table 5.5 (in red, where different).

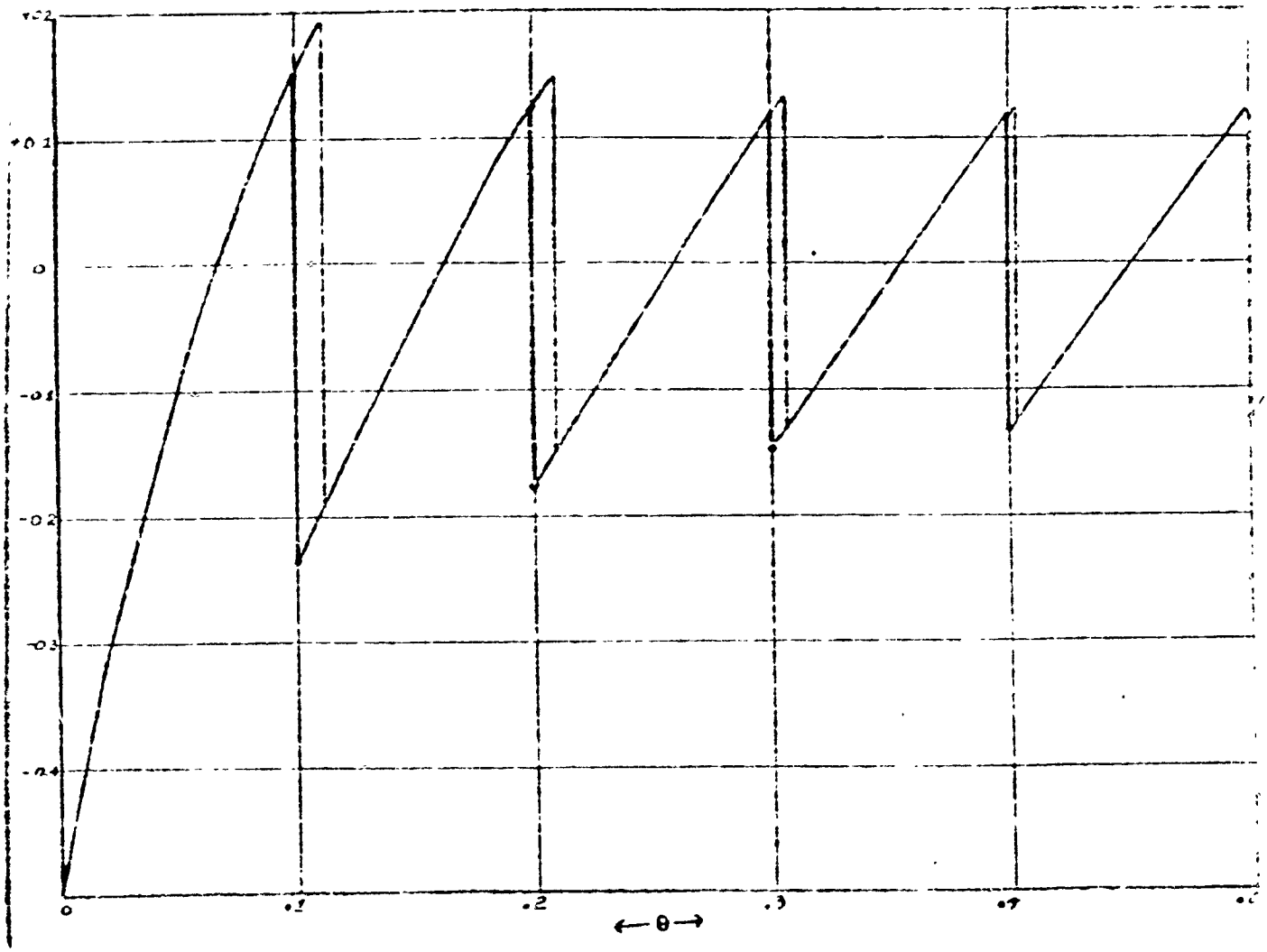


TABLE 5.6

p.41c

MEDIAN-BIAS OF CLASSICAL ESTIMATOR $\hat{\theta} = x/n$ OF A BINOMIAL PARAMETER θ, FOR SAMPLE SIZE $n = 10$, COMPARED WITH THE ESTIMATOR $\tilde{\theta}$ OF TABLE 5.5.

$$B(\theta,\hat{\theta}) = \text{Prob}\left\{\hat{\theta} > \theta|\theta\right\} - \text{Prob}\left\{\hat{\theta} < \theta|\theta\right\},$$

$$B(\theta,\tilde{\theta}) = \text{Prob}\left\{\tilde{\theta} > \theta|\theta\right\} - \text{Prob}\left\{\tilde{\theta} < \theta|\theta\right\}.$$

θ	$\frac{1}{2}B(\theta,\hat{\theta})$	$\frac{1}{2}B(\theta,\tilde{\theta})$
0.00	0.0	Same, except where value given.
0.00+	- .5	
.01	- .404	
.02	- .317	
.03	- .237	
.04	- .165	
.05	- .099	
.06	- .039	
.07	+ .016	
.08	+ .066	
.09	+ .111	
.10	- .236	+ .151
.11	- .197	+ .188
.12	- .158	
.13	- .120	
.14	- .082	
.15	- .044	
.16	- .008	
.17	+ .027	
.18	+ .061	
.19	+ .093	
.20	- .178	+ .124
.21	- .147	
.22	- .117	
.23	- .086	
.24	- .056	
.25	- .026	

θ	$\frac{1}{2}B(\theta,\hat{\theta})$	$\frac{1}{2}B(\theta,\tilde{\theta})$
.26	+ .004	
.27	+ .034	
.28	+ .062	
.29	+ .090	
.30	- .150	+ .117
.31	- .123	
.32	- .096	
.33	- .068	
.34	- .041	
.35	- .014	
.36	+ .013	
.37	+ .040	
.38	+ .066	
.39	+ .092	
.40	- .133	+ .118
.41	- .108	
.42	- .082	
.43	- .056	
.44	- .030	
.45	- .044	
.46	+ .022	
.47	+ .047	
.48	+ .073	
.49	+ .098	
.50	0.0	

FIGURE 5.4 (illustrating Table 5.7). Median-bias of classical estimator $\hat{\theta}$ of binomial parameter θ (in black), for sample size $n = 20$, compared with that of estimator $\tilde{\theta}$ of Table 5.5 (in red, where different).

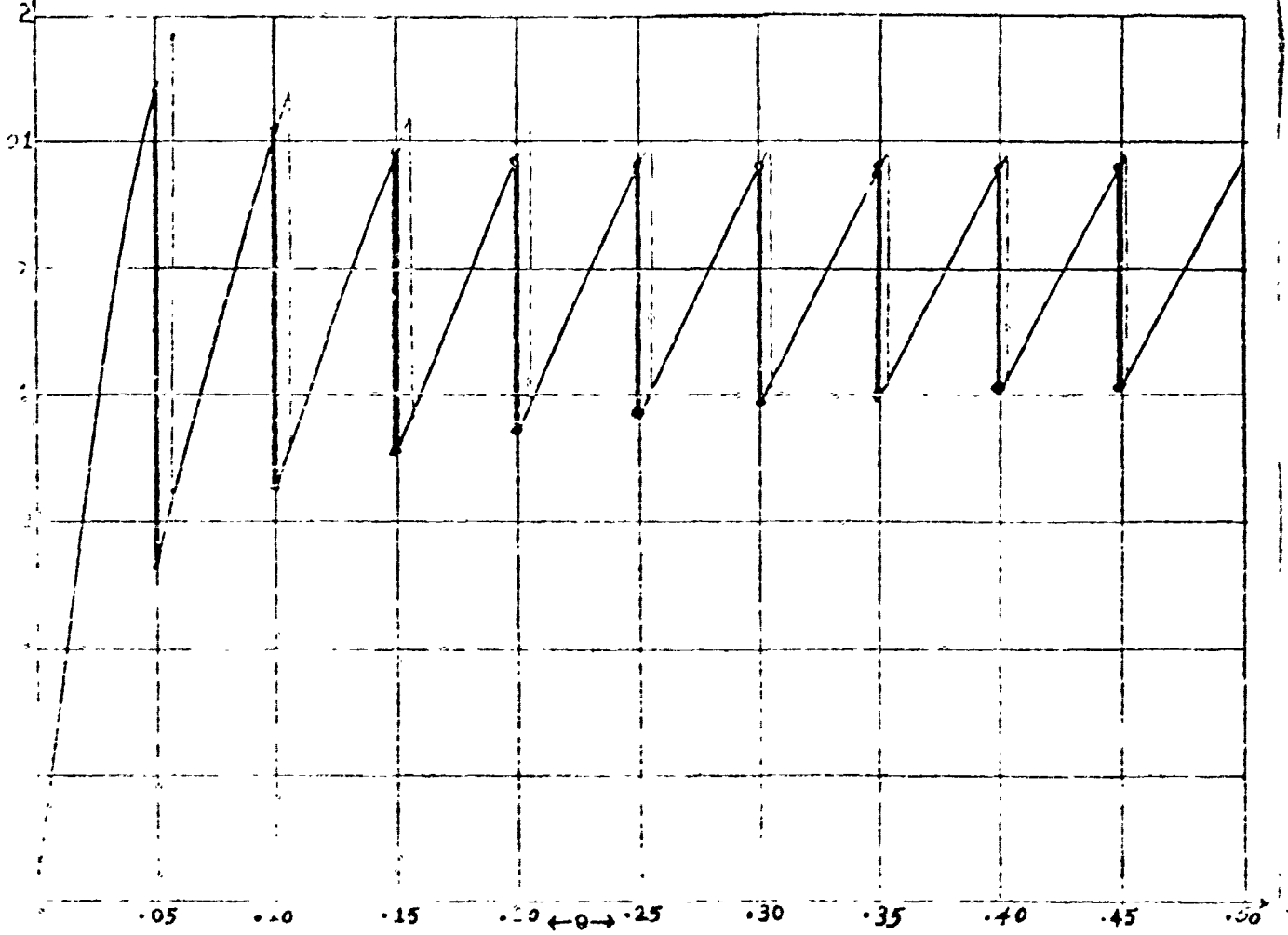


TABLE 5.7

p 11a.

MEDIAN-BIAS OF CLASSICAL ESTIMATOR $\hat{\theta} = x/n$ OF A BINOMIAL PARAMETER θ, FOR SAMPLE SIZE $n = 20$, COMPARED WITH THE ESTIMATOR $\tilde{\theta}$ OF TABLE 5.5.

$$B(\theta,\hat{\theta}) = \text{Prob}\left\{\hat{\theta} > \theta|\theta\right\} - \text{Prob}\left\{\hat{\theta} < \theta|\theta\right\},$$

$$B(\theta,\tilde{\theta}) = \text{Prob}\left\{\tilde{\theta} > \theta|\theta\right\} - \text{Prob}\left\{\tilde{\theta} < \theta|\theta\right\}.$$

θ	$\frac{1}{2}B(\theta,\hat{\theta})$	$\frac{1}{2}B(\theta,\tilde{\theta})$
.00	0.	Same, except where values given.
.00+	- .5	
.01	- .318	
.02	- .168	
.03	- .044	
.04	+ .058	
.05	- .236	+ .142
.06	- .161	+ .210
.07	- .087	
.08	- .017	
.09	+ .048	
.10	- .177	+ .108
.11	- .120	
.12	- .063	
.13	- .008	
.14	+ .045	
.15	- .148	+ .095
.16	- .099	
.17	- .050	
.18	- .003	
.19	+ .044	
.20	- .130	+ .089
.21	- .086	
.22	- .042	
.23	+ .001	
.24	+ .044	
.25	- .117	+ .085

θ	$\frac{1}{2}B(\theta,\hat{\theta})$	$\frac{1}{2}B(\theta,\tilde{\theta})$
.26	- .077	
.27	- .036	
.28	+ .005	
.29	+ .045	
.30	- .108	+ .084
.31	- .070	
.32	- .031	
.33	+ .008	
.34	+ .046	
.35	- .101	+ .083
.36	- .064	
.37	- .027	
.38	+ .011	
.39	+ .048	
.40	- .096	+ .084
.41	- .059	
.42	- .023	
.43	+ .014	
.44	+ .050	
.45	- .091	+ .086
.46	- .056	
.47	- .020	
.48	+ .017	
.49	+ .053	
.50	0.0	

REFERENCES

(1) Lindley, D.V. "Statistical Inference." Journal of the Royal Statistical Society, Series B, Vol. XV, 1953, pp.30-76.

(2) Cox, D.R. "Some Problems Connected with Statistical Inference." Annals of Math. Stat., Vo. 29, 1958, pp.357-372.

(3) Neyman, J. "Current Problems of Mathematical Statistics." Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, Vol. I, pp.1-22.

(4) Savage, L.J. Foundations of Statistics, John Wiley and Sons, Inc., New York, 1954.

(5) Stein, C. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, University of California Press, 1958, pp.197-206.

(6) Weiss, L. "A Higher Order Complete Class Theorem." Annals of Math. Stat., Vol. 24, 1953, pp.677-680.

(7) Cramer, H. Mathematical Methods of Statistics, Princeton University Press, 1946.

(8) Berkson, J. "Tables for the maximum likelihood estimate of the logistic function." Biometrika, Vol. 13 (1957), pp.28-34.

(9) Brown, G.W. "On small-sample estimation." Annals of Math. Stat., Vol. 18, 1947, pp.582-585.

(10) Lehmann, E. "A general concept of unbiasedness." Annals of Math. Stat., Vol. 22, pp.587-592.