The Annals of Mathematical Statistics
1971, Vol. 42, No. 1, 157–168

ON DISTRIBUTION-FREE STATISTICAL INFERENCE WITH UPPER AND LOWER PROBABILITIES[1]

BY R. J. BERAN

University of California, Berkeley

1. Introduction. This paper sets forth a new theory of distribution-free inference for the general statistical model

$$\mathbf{x} = T_\theta^{-1}\mathbf{e}. \tag{1.1}$$

By assumption, $\{T_\theta : \theta \in \Omega\}$ is a known family of nonsingular transformations mapping R^N into R^N, Ω is a Borel subset of a real Euclidean space, $\mathbf{x} = (x_1, \cdots, x_N)$ is a sample of N observations, and the components of $\mathbf{e} = (e_1, \cdots, e_N)$ are realized values of N independent, identically distributed random variables with common continuous distribution function F on the real line.

The scientific background for model (1.1) is as follows. An experiment is performed, resulting in N measurements $\mathbf{x}$. The observed $\mathbf{x}$ are generated from underlying realized errors $\mathbf{e}$ by the transformation (1.1). While the vector $\mathbf{x}$ is an observed constant, the values of $\mathbf{e}$ and θ giving rise to $\mathbf{x}$ through (1.1) are unknown. Our main goal in this paper is to draw inferences about the unknown constant θ from the observed $\mathbf{x}$ and the model, first with no knowledge of the distribution function F, apart from continuity, and secondly under the additional assumption that F is symmetric about the origin.

The theory described in this paper both extends and applies ideas introduced by Fraser [3] and by Dempster [1]. While technically the results fall within the general framework of Dempster's upper and lower probabilities, the statistical rationale differs. Upper and lower probabilities are introduced in Section 2 to measure the reliability of certain simple decision procedures, with reliability being assessed by relative frequency of success. This approach gives a well-defined statistical interpretation to upper and lower probabilities and has the advantage of leading naturally to a decision theory.

Model (1.1) encompasses many of the common models treated in classical nonparametric theory. As examples, one and two-sample versions of location models, scale models, regression models, and auto-regressive models are analyzed in Section 3 and Section 4. In particular, an optimality property is established for the one and two-sample Wilcoxon tests, for the Hodges and Lehmann [4] location parameter estimates (based on the Wilcoxon tests), for a test of scale parameter studied by Sukhatme [6], and for a regression parameter estimate proposed by Theil [5] and Sen [7].

Received November 13, 1969; revised July 1970.

[1] This research was partially supported by a University of California Summer Faculty Fellowship and National Science Foundation Grant GP-15283.

2. Statistical analysis.

2.1. *No assumptions on F.* Consider first the analysis of model (1.1) under no assumptions on the distribution function F, apart from continuity. Let F^{-1} be the inverse of F, let $U^N = \{\mathbf{u} \in R^N : 0 \leqq u_i \leqq 1\}$ be the N-dimensional unit cube, and let $\mathbf{F}^{-1}$ be the function that maps $\mathbf{u} \in U^N$ into $(F^{-1}(u_1), \cdots, F^{-1}(u_N))$. If $\mathbf{u}$ is a realization of a random variable distributed uniformly over U^N, the components of $\mathbf{F}^{-1}(\mathbf{u})$ are realizations of independent random variables, each distributed according to F. Model (1.1) can then be rewritten as

$$\mathbf{x} = T_\theta^{-1}\{\mathbf{F}^{-1}(\mathbf{u})\}. \tag{2.1}$$

Thus, performing the experiment mentioned in the Introduction amounts to realizing, through physical operations, a specific triple $(\mathbf{x}, (\theta, F), \mathbf{u})$, where $\mathbf{x} \in R^N$, $\mathbf{u} \in U^N$, $\theta \in \Omega$ and $F \in \mathscr{F}$, the set of all continuous distribution functions on the real line. Before the experiment is carried out (or before the outcome $\mathbf{x}$ is noted), the following prospective assertions can be made about the triple to be realized: the chance that $\mathbf{u}$ lies in a measurable subset $B \subset U^N$ is $P(B)$, where P is the uniform probability measure on U^N; (θ, F) is an unspecified element of $\Omega \times \mathscr{F}$; the observable $\mathbf{x}$ is related to $\mathbf{u}$ and (θ, F) through (2.1).

Once the experiment has been performed and $\mathbf{x}$ has been observed, the particular triple $(\mathbf{x}, (\theta, F), \mathbf{u})$ that was realized can be described more precisely. The first component, $\mathbf{x}$, is now known. Typically, not every combination of $\mathbf{u} \in U^N$ and $(\theta, F) \in \Omega \times \mathscr{F}$ will result in the observed $\mathbf{x}$, under model (2.1). If

$$S_x(\mathbf{u}) = \{(\theta, F) \in \Omega \times \mathscr{F} : \mathbf{x} = T_\theta^{-1}\{\mathbf{F}^{-1}(\mathbf{u})\}\}, \tag{2.2}$$

it is clear that the $\mathbf{u}$ realized in the experiment must lie in $U_x = \{\mathbf{u} \in U^N : S_x(\mathbf{u}) \neq \phi\}$ and, whatever that $\mathbf{u}$ is, the realized (θ, F) must belong to $S_x(\mathbf{u})$.

To obtain a more convenient description of U_x, we introduce the following definition and notations.

DEFINITION 2.1. The Ω-region associated with a rank vector $\mathbf{r}$ by the sample $\mathbf{x}$ under model (2.1) is $\{\theta \in \Omega : \operatorname{rank}(T_\theta \mathbf{x}) = \mathbf{r}\}$.

Let $\Omega_1, \Omega_2, \cdots, \Omega_M$, where $M \leqq N!$, denote the non-void Ω-regions associated with $\mathbf{r}$ as $\mathbf{r}$ is varied over the $N!$ permutations of $\{1, 2, \cdots N\}$. Under the model, $M > 0$ with probability one. Clearly $\Omega_i \cap \Omega_j = \phi$ if $i \neq j$. Define Ω_0 to be the residual set $\Omega_0 = \Omega - \bigcup_{i=1}^{M} \Omega_i$. For any $\theta \in \Omega_0$, at least two components of $T_\theta \mathbf{x}$ will be equal.

Let $\mathbf{r}_i$ denote the (unique) rank vector associated with Ω-region Ω_i, $1 \leqq i \leqq M$, and let $I_M = \{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_M\}$. Let I_0 denote the set of all distinct improper rank vectors associated, as in Definition 2.1, with the various $\theta \in \Omega_0$.

Since F is continuous, F^{-1} is strictly monotone and the model (2.1) implies that $\operatorname{rank}(\mathbf{u}) = \operatorname{rank}(T_\theta \mathbf{x})$. Consequently,

$$U_x = \{\mathbf{u} \in U^N : \operatorname{rank}(\mathbf{u}) \in I_M \cup I_0\}. \tag{2.3}$$

Let $P[B \mid U_x]$ denote the conditional probability that $\mathbf{u} \in B$, given that $\mathbf{u} \in U_x$; for any measurable subset $B \subset U_x$,

$$P[B \mid U_x] = P[B]/P[U_x], \tag{2.4}$$

where P is the uniform probability measure on U^N.

After the experiment has been performed and $\mathbf{x}$ has been observed, the following statements can be made about the realized triple $(\mathbf{x}, (\theta, F), \mathbf{u})$: $\mathbf{x}$ is as observed; $\mathbf{u} \in U_x$ and the chance that $\mathbf{u} \in B$, where B is a measurable subset of U_x, is $P[B \mid U_x]$; whatever $\mathbf{u}$ is, (θ, F) is an unspecified element of the corresponding set $S_x(\mathbf{u})$. This collection of assertions about the triple $(\mathbf{x}, (\theta, F), \mathbf{u})$ will be called the *reduced model* for the experiment. Both Fraser [3] and Dempster [1] have previously considered reductions of this type, though without explicitly introducing experimental triples.

Since the realized experiment $(\mathbf{x}, (\theta, F), \mathbf{u})$ is described more precisely by the reduced model than by the original model, we propose to evaluate statistical procedures of interest by their average behavior over a hypothetical sequence of independent experiments, each of which is generated under the assumptions of the reduced model. We begin by examining the reliability of the simple decision that the realized $(\theta, F) \in D$, where $D \subset \Omega \times \mathscr{F}$. Let $\{(\mathbf{x}, (\theta_i, F_i), \mathbf{u}_i);\ i = 1, 2, \cdots\}$ be a sequence of independent experiments generated under the reduced model; in other words, $\mathbf{u}_1, \mathbf{u}_2, \cdots$ are independent and each is distributed according to the conditional probability $P[\cdot \mid U_x]$; (θ_i, F_i) is selected arbitrarily from $S_x(\mathbf{u}_i)$; $\mathbf{x}$ is the observed value. For each i, the equation $\mathbf{x} = T_{\theta_i}^{-1}\{F_i^{-1}(\mathbf{u}_i)\}$ will necessarily be satisfied.

Let the general notation $\text{prop}_n(\Pi_i)$ denote the proportion of true propositions among the propositions $\{\Pi_1, \Pi_2, \cdots, \Pi_n\}$. Evidently, for every $n > 0$,

$$\text{prop}_n[S_x(\mathbf{u}_i) \subset D, S_x(\mathbf{u}_i) \neq \phi] \leqq \text{prop}_n[(\theta_i, F_i) \in D] \leqq \text{prop}_n[S_x(\mathbf{u}_i) \cap D \neq \phi]. \tag{2.5}$$

As $n \to \infty$, the upper and lower bounds for $\text{prop}_n[(\theta_i, F_i) \in D]$ converge w.p. 1 to

$$P^*(D) = P[\mathbf{u}: S_x(\mathbf{u}) \cap D \neq \phi \mid U_x]$$
$$P_*(D) = P[\mathbf{u}: S_x(\mathbf{u}) \subset D, S_x(\mathbf{u}) \neq \phi \mid U_x] \tag{2.6}$$

respectively, provided the arguments on the right are measurable. All cluster points of the sequence $\{\text{prop}_n[(\theta_i F_i) \in D];\ n = 1, 2, \cdots\}$ lie between $P_*(D)$ and $P^*(D)$ w.p. 1.

Thus, $P_*(D)$ and $P^*(D)$ measure, respectively, the smallest and largest possible reliability to be attached to the assertion $(\theta, F) \in D$; reliability is assessed relative to the reduced model for the experiment. P^* and P_* are, in fact, Dempster's [1], [2] upper and lower probabilities for model (2.1). On those subsets D for which $P^*(D) = P_*(D)$ (see the remarks following Proposition 2.1 for examples), the common value amounts to what Fraser [3] terms a structural probability. Unlike the approach taken above, Dempster [1], [2] defines upper and lower probabilities through a purely formal mapping of probability from U_x to $\Omega \times \mathscr{F}$. Fraser [3] does similarly within his framework. In this writer's view, the frequency approach described here is preferable in that it gives a well-defined statistical interpretation to P^* and P_* which leads naturally (later in the section) to analogues of risk functions and to a decision theory.

In general, the computation of $P^*(D)$ and $P_*(D)$ may be quite complicated. For inferences about θ alone, however, it is often enough to consider sets D of the form $D = A \times \mathscr{F}$, where $A \subset \Omega$. In such cases, $P^*(D)$ and $P_*(D)$ are more readily calculated. For $1 \leqq i \leqq M$, let

$$\delta^*(i, A) = 1 \quad \text{if} \quad A \cap \Omega_i \neq \phi, \tag{2.7}$$

$$= 0 \quad \text{otherwise};$$

$$\delta_*(i, A) = 1 \quad \text{if} \quad A \supset \Omega_i,$$

$$= 0 \quad \text{otherwise.}$$

PROPOSITION 2.1. *Let A be an arbitrary subset of Ω. Then under model* (2.1),

$$P^*[A \times \mathscr{F}] = M^{-1} \textstyle\sum_{i=1}^{M} \delta^*(i, A) \tag{2.8}$$

$$P_*[A \times \mathscr{F}] = M^{-1} \textstyle\sum_{i=1}^{M} \delta_*(i, A).$$

PROOF. The result follows from the identities

$$\{\mathbf{u} \in U_x : S_x(\mathbf{u}) \subset A \times \mathscr{F}, S_x(\mathbf{u}) \neq \phi\} \tag{2.9}$$

$$= \{\mathbf{u} \in U_x : \operatorname{rank}(\mathbf{u}) \in \Delta_* \cup \textstyle\bigcup_{\{i:\delta_*(i,A)\neq 0\}} \{\mathbf{r}_i\}\}$$

and

$$\{\mathbf{u} \in U_x : S_x(\mathbf{u}) \cap A \times \mathscr{F} \neq \phi\} \tag{2.10}$$

$$= \{\mathbf{u} \in U_x : \operatorname{rank}(\mathbf{u}) \in \Delta^* \cup \textstyle\bigcup_{\{i:\delta^*(i,A)\neq 0\}} \{\mathbf{r}_i\}\}$$

where Δ_* and Δ^* are suitable subsets of I_0.

Proposition 2.1 shows in particular that $P_*[\Omega_i \times \mathscr{F}] = P^*[\Omega_i \times \mathscr{F}] = 1/M$ for $1 \leqq i \leqq M$ and $P_*[\Omega_0 \times \mathscr{F}] = P^*[\Omega_0 \times \mathscr{F}] = 0$. On the Boolean algebra generated by the partition $\{\Omega_0, \Omega_1, \cdots \Omega_M\}$ of Ω, P^* (or P_*) is a probability measure. For $D \subset \Omega \times \mathscr{F}$, $0 \leqq P_*(D) \leqq P^*(D) \leqq 1$, if defined. Further properties of upper and lower probabilities are described in Dempster [2]. More generally, P^* is a completely alternating set function while P_* is a completely monotone set function.

It is possible, within the statistical framework of this paper, to set up an analog to standard decision theory. Let $\mathscr{D}$ denote a space of decisions or actions, and let $l : \Omega \times \mathscr{F} \times \mathscr{D} \to R^+$ be a real-valued, nonnegative loss function. Suppose $d \in \mathscr{D}$ is a decision whose consequences we wish to evaluate relative to the reduced model.

Let $\{(\mathbf{x}, (\theta_i, F_i), \mathbf{u}_i); i = 1, 2, \cdots\}$ again denote a hypothetical sequence of independent experiments generated under the reduced model. The average loss incurred over the first n of these experiments as a result of taking action d is $n^{-1} \sum_{i=1}^{n} l(\theta_i, F_i, d)$. Since $l \geqq 0$,

$$n^{-1} \textstyle\sum_{i=1}^{n} l(\theta_i, F_i, d) = \int_0^\infty \operatorname{prop}_n [l(\theta_i, F_i, d) > t]\, dt. \tag{2.11}$$

If $A(t, d) = \{(\theta, F) \in \Omega \times \mathscr{F} : l(\theta, F, d) > t\}$, equation (2.11) and the inequalities (2.5) imply that

$$\begin{aligned} (2.12) \qquad & \int_0^\infty \text{prop}_n [S_x(\mathbf{u}_i) \subset A(t, d), S_x(\mathbf{u}_i) \neq \phi]\, dt \\ & \leqq n^{-1} \sum_{i=1}^n l(\theta_i, F_i, d) \\ & \leqq \int_0^\infty \text{prop}_n [S_x(\mathbf{u}_i) \cap A(t, d) \neq \phi]\, dt. \end{aligned}$$

Subject to measurability restrictions, the upper and lower bounds converge w.p. 1, as $n \to \infty$, to

$$\begin{aligned} (2.13) \qquad R^*(l, d) &= \int_0^\infty P^*[A(t, d)]\, dt \\ R_*(l, d) &= \int_0^\infty P_*[A(t, d)]\, dt. \end{aligned}$$

All cluster points of the sequence $\{n^{-1} \sum_{i=1}^n l(\theta_i, F_i, d); n = 1, 2, \cdots\}$ of average losses are bounded w.p. 1 from above and below by $R^*(l, d)$ and $R_*(l, d)$, respectively.

Thus, the *lower risk* $R_*(l, d)$ and the *upper risk* $R^*(l, d)$ measure, respectively, the smallest and largest possible long-run average loss incurred as a consequence of decision d. The evaluation is with respect to the reduced model. The relative desirability of various decisions d may be assessed by reference to the corresponding risks.

It is readily seen that R_* and R^* are equivalent, for nonnegative l, to the lower and upper expectations defined by Dempster [1], [2]. (The considerations of (2.12) could be extended to real-valued loss functions; in this more general case, the analog of (2.13) would still be equivalent to Dempster's upper and lower expectations). Dempster suggests in [1] that upper and lower expectations be used "as guides for betting or decision procedures whose loss functions are linear in [the parametric function]," but gives no example. From the viewpoint of this paper, linearity in the loss function seems unimportant. The frequency interpretation for R_* and R^* motivates the following definition of a minimax decision.

DEFINITION 2.2. A decision $d \in \mathscr{D}$ is minimax with respect to loss function l if $R^*(l, d) \leqq R^*(l, d')$ for every $d' \in \mathscr{D}$ and if $R_*(l, d) \leqq R_*(l, d'')$ for every $d'' \in \mathscr{D}$ which satisfies $R^*(l, d'') = R^*(l, d)$.

For decisions involving θ alone, the loss function $l(\theta, F, d)$ might be replaced by one of the form $l(\theta, d)$. The arguments of the upper and lower probabilities which appear as the integrands of $R^*(l, d)$ and $R_*(l, d)$ would then be product sets $A \times \mathscr{F}$, with $A \subset \Omega$. Proposition 2.1 handles these. Examples of minimax decisions are given in Section 3 and Section 4.

2.2. *Symmetry assumption on F.* We turn to the analysis of model (1.1) under the assumptions that the distribution function F is symmetric about the origin and continuous. Let F_+ be the distribution function of $|e_1|$ and let F_+^{-1} be its inverse. Define the function sign $(\cdot)$ through

$$\begin{aligned} (2.14) \qquad \text{sign}(x) &= 1 && \text{if} \quad x > 0 \\ &= 0 && \text{if} \quad x = 0 \\ &= -1 && \text{if} \quad x < 0. \end{aligned}$$

Let $V^N = \{\mathbf{v} \in R^N : -1 \leqq v_i \leqq 1\}$ and let $\mathbf{G}^{-1}$ be the function that maps $\mathbf{v} \in V^N$ into $(\text{sign}(v_1)F_+^{-1}(|v_1|), \cdots \text{sign}(v_N)F_+^{-1}(|v_N|))$. If $\mathbf{v}$ is a realization of a random variable distributed uniformly over V^N, the components of $\mathbf{G}^{-1}(\mathbf{v})$ are realizations of independent random variables, each distributed according to F. Model (1.1) can then be rewritten as

$$\mathbf{x} = T_\theta^{-1}\{\mathbf{G}^{-1}(\mathbf{v})\}. \tag{2.15}$$

The form (2.15) is preferred over (2.1) because it brings out explicitly the symmetry of F.

Thus, performing the experiment described in the Introduction amounts, in this case, to realizing a triple $(\mathbf{x}, (\theta, F_+), \mathbf{v})$, where $\mathbf{x} \in R^N$, $\mathbf{v} \in V^N$, $\theta \in \Omega$ and $F^+ \in \mathscr{F}^+$, the set of all continuous distribution functions on the positive real line. Prior to the performance of the experiment, the following assertions can be made about the triple to be realized: the chance that $\mathbf{v}$ lies in a measurable subset $B \subset V^N$ is $P^+(B)$, where P^+ is the uniform probability measure over V^N; (θ, F^+) is an unspecified element of $\Omega \times \mathscr{F}^+$; the observable $\mathbf{x}$ is related to $\mathbf{v}$ and (θ, F^+) through (2.15).

Once the experiment has been carried out and $\mathbf{x}$ has been observed, the particular triple $(\mathbf{x}, (\theta, F^+), \mathbf{v})$ that was realized can be described more precisely. The first component is the observed $\mathbf{x}$. If

$$S_x^+(\mathbf{v}) = \{(\theta, F^+) \in \Omega \times \mathscr{F}^+ : \mathbf{x} = T_\theta^{-1}\{\mathbf{G}^{-1}(\mathbf{v})\}\}, \tag{2.16}$$

it is evident that the $\mathbf{v}$ realized in the experiment necessarily lies in $V_x = \{\mathbf{v} \in V^N : S_x^+(\mathbf{v}) \neq \phi\}$ and, whatever that $\mathbf{v}$ is, the realized (θ, F^+) must belong to $S_x^+(\mathbf{v})$.

The argument continues as in Section 2.1. Appropriate Ω-regions are described by

DEFINITION 2.3. The Ω-region associated with a rank vector $\mathbf{r}$ and a sign vector $\mathbf{s}$ by the sample $\mathbf{x}$ under model (2.15) is $\{\theta \in \Omega : \text{rank}(|T_\theta \mathbf{x}|) = \mathbf{r}, \text{sign}(T_\theta \mathbf{x}) = \mathbf{s}\}$.

In this definition, the notation $|\mathbf{Z}|$ stands for $(|Z_1|, |Z_2|, \cdots |Z_N|)$. Let $\Omega_1^+, \Omega_2^+, \cdots \Omega_L^+$, where $L \leqq 2^N N!$, denote the non-void Ω-regions associated with $(\mathbf{r}, \mathbf{s})$ as $\mathbf{r}$ is varied over the $N!$ permutations of $\{1, 2, \cdots N\}$ and $\mathbf{s}$ is varied over the 2^N N-dimensional vectors with ± 1 components. Under the model, $L > 0$ with probability one. Clearly $\Omega_i^+ \cap \Omega_j^+ = \phi$ if $i \neq j$. Define Ω_0^+ to be the residual set $\Omega_0^+ = \Omega - \bigcup_{i=1}^L \Omega_i^+$. For any $\theta \in \Omega_0^+$, at least two components of $|T_\theta \mathbf{x}|$ will be equal, or at least one component of $T_\theta \mathbf{x}$ will be zero.

Let $(\mathbf{r}_i, \mathbf{s}_i)$ denote the (unique) rank and sign vector pair associated with Ω-region Ω_i^+, $1 \leqq i \leqq L$, and let $I_L^+ = \{(\mathbf{r}_1, \mathbf{s}_1), (\mathbf{r}_2, \mathbf{s}_2), \cdots (\mathbf{r}_L, \mathbf{s}_L)\}$. Let I_0^+ denote the improper rank and sign vector pairs associated with the various $\theta \in \Omega_0^+$.

It follows from (2.15) that

$$V_x = \{\mathbf{v} \in V^N : (\text{rank}(|\mathbf{v}|), \text{sign}(\mathbf{v})) \in I_L^+ \cup I_0^+\}. \tag{2.17}$$

Upper and lower probabilities or risks are now established as in Section 2.1, with V_x in place of U_x. For $1 \leqq i \leqq L$ and $A \subset \Omega$, let

$$\begin{aligned} \delta^*(i, A)_+ &= 1 \quad \text{if} \quad A \cap \Omega_i{}^+ \neq \phi, \\ &= 0 \quad \text{otherwise;} \\ \delta_*(i, A)_+ &= 1 \quad \text{if} \quad A \supset \Omega_i{}^+, \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{2.18}$$

The analogue to Proposition 2.1 is

PROPOSITION 2.2. *Let A be an arbitrary subset of Ω. Then, under model* (2.15),

$$\begin{aligned} P^*[A \times \mathscr{F}^+] &= L^{-1} \textstyle\sum_{i=1}^{L} \delta^*(i, A)_+ \\ P_*[A \times \mathscr{F}^+] &= L^{-1} \textstyle\sum_{i=1}^{L} \delta_*(i, A)_+ . \end{aligned} \tag{2.19}$$

3. Examples for Section 2.1. This section gives examples of upper and lower probabilities and minimax decisions for several common models which are special cases of (2.1). Ties and other anomalies arising because of round-off errors are not discussed; the theoretical solution to such difficulties is to include the rounding-off process in the model itself.

For later use, we define the median of a finite set $A = \{a_1, a_2, \cdots, a_T\}$, whose elements are real and arranged in order of increasing magnitude, as follows:

$$\begin{aligned} \text{median}(A) &= a_{s+1} \quad \text{if} \quad T = 2s+1 \\ &= \tfrac{1}{2}(a_s + a_{s+1}) \quad \text{if} \quad T = 2s. \end{aligned} \tag{3.1}$$

3.1. *Two-sample location model.* In this model $\mathbf{x} = (x_1, \cdots, x_m, y_1, \cdots, y_n)$, $m+n = N$, $\theta = \mu$, $\Omega = (-\infty, \infty)$ and $T_\mu \mathbf{x} = (x_1, \cdots, x_m, y_1 - \mu, \cdots, y_n - \mu)$. To construct the Ω-regions, define the differences $\{d_{ij} = y_j - x_i,\ 1 \leqq i \leqq m,\ 1 \leqq j \leqq n\}$ and let $a_1 < a_2 < \cdots < a_{M-1}$, where $M = mn+1$, denote the ordered $\{d_{ij}\}$. Under the model, this strict ordering can be accomplished with probability one.

LEMMA 3.1. *The Ω-regions $\{\Omega_i,\ 1 \leqq i \leqq M\}$ for the two-sample location model are the M open sets $(-\infty, a_1), (a_1, a_2), \cdots, (a_{M-1}, \infty)$. The region $\Omega_0 = \{a_1, a_2, \cdots, a_{M-1}\}$.*

PROOF. For any pair of observations x_i and y_j, the point $d_{ij} = y_j - x_i$ divides the real line into two open half-lines. For values of μ belonging to one of these half-lines, $y_j - \mu > x_i$, while for values of μ in the other half-line, the inequality is reversed.

An arbitrary set Ω_k, $1 \leqq k \leqq M$ (as defined in the lemma), may be constructed by appropriately selecting, for each pair (i, j), one of the open half-lines defined by d_{ij} and by then forming the set-theoretic intersection of the mn half-lines chosen. Therefore, rank$(T_\mu \mathbf{x})$ is constant for $\mu \in \Omega_k$.

Any other region Ω_l, $l \neq k$, $1 \leqq l \leqq M$, must be contained in a half-line (corresponding to some pair i, j) which does not contain Ω_k. Hence, rank$(T_\mu \mathbf{x})$ is different when $\mu \in \Omega_l$ than when $\mu \in \Omega_k$. The lemma follows.

Proposition 2.1, together with Lemma 3.1, gives the upper and lower probabilities relevant to μ alone. These may be applied to estimation as follows. Let the loss function be $l(\mu, d) = |\mu - d|$. Replace $\Omega = (-\infty, \infty)$ by $\Omega = [a_0, a_M]$, where $-\infty < a_0 < a_1$, and $a_{M-1} < a_M < \infty$. (This is to keep $R^*(l, d) < \infty$). Suppose $d \in [a_i, a_{i+1}]$, $0 \leqq i \leqq M-1$. Then

$$R^*(l,d) = M^{-1}\textstyle\sum_{j \neq i} |a_j - d| \qquad \text{if} \quad |d-a_i| < |d-a_{i+1}|, \tag{3.2}$$
$$= M^{-1}\textstyle\sum_{j \neq i+1} |a_j - d| \qquad \text{otherwise;}$$

where j ranges from 0 to M, and

$$R_*(l,d) = M^{-1}\textstyle\sum_{j=1}^{M-1} |a_j - d|. \tag{3.3}$$

When M is even, any $d \in [\frac{1}{2}(a_{M/2-1} + a_{M/2}), \frac{1}{2}(a_{M/2} + a_{M/2+1})]$ minimizes $R^*(l, d)$; among such d, only $d = \text{median}\{a_1, a_2, \cdots, a_{M-1}\}$ minimizes $R_*(l, d)$. When M is odd, $d = \text{median}\{a_1, a_2, \cdots a_{M-1}\}$ is the unique value minimizing $R^*(l, d)$. Thus, the minimax estimate of μ with respect to the loss function $l(\mu, d) = |\mu - d|$ is

$$\hat{\mu} = \text{median}\{a_1, a_2, \cdots a_{M-1}\}. \tag{3.4}$$

It is noteworthy that $\hat{\mu}$ is the Hodges and Lehmann [4] location shift estimate derived from the Wilcoxon test.

In testing a hypothesis $H: \theta \in A$ versus the alternative $K: \theta \in \mathscr{C}A$, $A \subset \Omega$, the decisions are d_0 and d_1—"accept H" and "reject H" respectively. The step loss function is defined by

$$l(\theta, d_0) = \omega_0 I_{\mathscr{C}A}(\theta) \tag{3.5}$$
$$l(\theta, d_1) = \omega_1 I_A(\theta)$$

where, conventionally, $\omega_1 > \omega_0 > 0$. Clearly

$$R^*(l,d) = \omega_0 P^*(\mathscr{C}A) \qquad \text{if} \quad d = d_0, \tag{3.6}$$
$$= \omega_1 P^*(A) \qquad \text{if} \quad d = d_1;$$

with a corresponding equation for $R_*(l, d)$. The minimax decision under the step loss is: reject H if $\omega_0 P^*(\mathscr{C}A) > \omega_1 P^*(A)$; accept H if the inequality is reversed; base the decision on $R_*(l, d)$ if there is equality.

In particular, for testing $H: \mu \leqq 0$ versus $K: \mu > 0$, the minimax test rejects H if and only if the number of pairs $\{(x_i, y_j), 1 \leqq i \leqq m, 1 \leqq j \leqq n\}$ for which $y_j > x_i$ is sufficiently large. This amounts to the Mann–Whitney or Wilcoxon tests.

3.2. *Two-sample scale model.* In this model, $\mathbf{x} = (x_1, \cdots, x_m, y_1, \cdots, y_n)$, $m+n = N$, $\theta = \sigma$, $\Omega = (0, \infty)$ and $T_\sigma \mathbf{x} = (x_1, \cdots, x_m, y_1/\sigma, \cdots, y_n/\sigma)$. Suppose m^+ of the x's are positive and m^- are negative. Similarly suppose n^+ of the y's are positive and n^- are negative. Define the ratios $\{c_{ij} = \max(y_j/x_i, 0), 1 \leqq i \leqq m, 1 \leqq j \leqq n\}$ and let $b_1 < b_2 < \cdots < b_{M-1}$, where $M = m^+n^+ + m^-n^- + 1$, denote the ordered non-zero and non-infinite ratios $\{c_{ij}\}$. Under the model, this strict ordering may be accomplished with probability one. Arguing as in the proof of Lemma 3.1, we establish

LEMMA 3.2. *The Ω-regions $\{\Omega_i, 1 \leqq i \leqq M\}$ for the two-sample scale model are the M open sets $(0, b_1), (b_1, b_2), \cdots, (b_{M-1}, \infty)$. The region $\Omega_0 = \{b_1, b_2, \cdots, b_{M-1}\}$.*

Estimation and hypothesis testing are handled as in Section 3.1. If $\Omega = (0, \infty)$ is replaced by $\Omega = (0, b_M)$, where $b_{M-1} < b_M < \infty$, the minimax estimate for σ under the loss $l(\sigma, d) = |\sigma - d|$ is

$$\hat{\sigma} = \text{median}\{b_1, b_2, \cdots, b_{M-1}\}. \tag{3.7}$$

The minimax test (step loss function) for $H: \sigma \leqq 1$ versus $K: \sigma > 1$ rejects H if and only if the number of pairs $\{(x_i, y_j), 1 \leqq i \leqq m, 1 \leqq j \leqq n\}$ for which $\text{sign}(x_i) = \text{sign}(y_j)$ and $|y_j| > |x_i|$ is sufficiently large. Sukhatme [6] studied this test from a classical viewpoint. Without further assumptions, the test is not nonparametric in the classical sense.

3.3. *Two-sample location-scale model.* In this model, $\mathbf{x} = (x_1, \cdots, x_m, y_1, \cdots, y_n)$, $m+n = N$, $\theta = (\mu, \sigma)$, $\Omega = (-\infty, \infty) \times (0, \infty)$, and $T_{\mu,\sigma}\mathbf{x} = (x_1, \cdots, x_m, (y_1 - \mu)/\sigma, \cdots, (y_n - \mu)/\sigma)$. The definition of the Ω-regions is more complex than in the previous two examples. M is defined implicitly in the following lemma, whose proof parallels that of Lemma 3.1, with half-planes replacing half-lines.

LEMMA 3.3. *The Ω-regions $\{\Omega_i, 1 \leqq i \leqq M\}$ for the two-sample location-scale model are the M open regions in Ω partitioned off by the mn lines $\{\mu + \sigma x_i = y_j, 1 \leqq i \leqq m, 1 \leqq j \leqq n\}$. The region Ω_0 consists of the partitioning line segments themselves.*

Upper and lower probabilities for (μ, σ) are provided by Proposition 2.1 and the lemma above. For testing $H: \mu \leqq 0$ versus $K: \mu > 0$, the minimax test (step loss function) rejects H if and only if the number of Ω-regions intersecting with $(0, \infty) \times (0, \infty)$ is large enough. Other hypotheses on (μ, σ) may be tested analogously.

3.4. *One-sample linear regression model.* In this model, $\mathbf{x} = (x_1, \cdots, x_N)$, $\theta = (\mu, \beta)$, $\Omega = (-\infty, \infty) \times (-\infty, \infty)$, a regressor $\mathbf{z} = (z_1, \cdots, z_N)$ is given and $T_{\mu,\beta}\mathbf{x} = (x_1 - \mu - \beta z_1, \cdots, x_N - \mu - \beta z_N)$. Argument similar to that for Lemma 3.3 shows that the Ω-regions are partitioned off by the lines $\{x_i - \mu - \beta z_i = x_j - \mu - \beta z_j, \text{ all } i \neq j\}$. Hence

LEMMA 3.4. *The Ω-regions $\{\Omega_i, 1 \leqq i \leqq M\}$ for the linear regression model* are the *M open regions in Ω partitioned off by the lines $\{\beta = (x_i - x_j)/(z_i - z_j), \text{ all } i \neq j\}$, excluding the possible lines $\beta = \pm\infty$. The region Ω_0 consists of the partitioning lines themselves.*

The Ω-regions for this model are strips parallel to the μ-axis. The sample $\mathbf{x}$ gives no information about μ in the following sense: if $B \subset (-\infty, \infty)$, the assertion $(\mu \in B, \beta \in (-\infty, \infty), F \in \mathscr{F})$ has upper probability 1 and lower probability 0, regardless of B, unless $B = \phi$ or $B = (-\infty, \infty)$.

Let $c_1 < c_2 < \cdots < c_{M-1}$ denote the finite values among $\{(x_i - x_j)/(z_i - z_j), \text{ all } i \neq j\}$. If β is restricted to a closed interval containing c_1 and c_{M-1}, the minimax estimate for β under the loss $l(\beta, d) = |\beta - d|$ is

$$\hat{\beta} = \text{median}\{c_1, \cdots, c_{M-1}\}. \tag{3.8}$$

Theil [7] and Sen [5] treated this estimate in a classical context. Under the step-loss function of Section 3.1, the minimax test for $H: \beta \leqq 0$ versus $K: \beta > 0$ rejects H if and only if the number of quadruples $\{(x_i, z_i, x_j, z_j),\ 1 \leqq i < j \leqq N\}$ for which $(x_i - x_j)/(z_i - z_j) > 0$ is large enough.

3.5. *First order auto-regressive model.* In this model, $\mathbf{x} = (x_1, \cdots, x_N)$, $\theta = (\mu, \rho)$, $\Omega = (-\infty, \infty) \times (-1, 1)$, a starting value x_0 is given, and $T_{\mu,\rho}\mathbf{x} = (x_1 - \mu - \rho(x_0 - \mu), \cdots, x_N - \mu - \rho(x_{N-1} - \mu))$. The Ω-regions are described in the following lemma, whose justification parallels that of Lemma 3.4.

LEMMA 3.5. *The Ω-regions $\{\Omega_i, 1 \leqq i \leqq M\}$ for the first order auto-regressive model are the M open regions in Ω partitioned off by those of the lines $\{\rho = (x_i - x_j)/(x_{i-1} - x_{j-1}),$ all $i \neq j\}$ which lie in Ω. The region Ω_0 consists of the partitioning lines themselves.*

In the sense of the preceding regression example, no information concerning μ is to be had from the upper and lower probabilities implied by Lemma 3.5. Optimal decisions concerning β in the regression model have analogs in ρ for the auto-regression model. Let $d_1 < d_2 < \cdots < d_{M-1}$ denote, in increasing order of magnitude, those values among $\{(x_i - x_j)/(x_{i-1} - x_{j-1}),\ 1 \leqq i < j \leqq N\}$ which lie in $(-1, 1)$. Under the loss function $l(\rho, d) = |\rho - d|$, the minimax estimate for ρ is

$$\hat{\rho} = \text{median}\,\{d_1, \cdots, d_{M-1}\}. \tag{3.9}$$

3.6. *A general two-sample model.* In this model, $\mathbf{x} = (x_1, \cdots, x_m, y_1, \cdots, y_n)$, $m + n = N$, $\theta = (H(y_1), \cdots, H(y_n))$, $\Omega = \{(z_1, \cdots, z_n): z_i \in (-\infty, \infty),\ 1 \leqq i \leqq n,$ and $z_i < z_j$ if $y_i < y_j$, all $i \neq j\}$, and $T_\theta \mathbf{x} = (x_i, \cdots, x_m, H(y_1), \cdots, H(y_n))$. Though this statistical model appears unfamiliar, it is a restatement of the more usual model in which $x_1, \cdots, x_m$ are realizations of independent identically distributed random variables with continuous distribution function F, while $y_1, \cdots, y_n$ are realizations of independent identically distributed random variables with continuous distribution function G, and $H = F^{-1}G$.

For notational convenience in describing the Ω-regions, relabel the observations so that $x_1 < \cdots < x_m$ and $y_1 < \cdots < y_n$; the strict orderings are possible with probability one. Then, $\Omega = \{\mathbf{z}: -\infty < z_1 < \cdots < z_n < \infty\}$. Let $Q_0 = (-\infty, x_1)$, $Q_i = (x_i, x_{i+1})$ if $1 \leqq i \leqq m-1$, and $Q_m = (x_m, \infty)$. Denote all possible n-fold product sets of the form $\{Q_{i_1} \times \cdots \times Q_{i_n}, i_1 \leqq \cdots \leqq i_n\}$ by $A_1, A_2, \cdots, A_M$.

LEMMA 3.6. *The Ω-regions $\{\Omega_i, 1 \leqq i \leqq M\}$ for the general two-sample model are $\Omega_i = \{\mathbf{z} \in \Omega: \mathbf{z} \in A_i\}$, $1 \leqq i \leqq M$. The region Ω_0 consists of the remaining points in Ω.*

PROOF. The sets $\{\Omega_i, 1 \leqq i \leqq M\}$ are associated, in the sense of Definition 2.1, with every possible ranking of the components of $T_\theta \mathbf{x}$ that is compatible with the relative order of $x_1, \cdots, x_m$ and the relative order of $y_1, \cdots, y_n$.

Proposition 2.1 gives the upper and lower probabilities for $(H(y_1), \cdots, H(y_n))$. The hypothesis that both samples do not differ significantly in distribution may

be formulated as $H:(H(y_1),\cdots, H(y_n))\in N(\mathbf{y})$ where $N(\mathbf{y})\subset\Omega$ is a specified neighborhood of $\mathbf{y}$. The minimax test for hypothesis H versus alternative $K:(H(y_1),\cdots, H(y_n))\notin N(\mathbf{y})$ under a step loss function is of the general form derived in Section 3.1.

4. Examples for Section 2.2. In this section, the unknown distribution function F is assumed to be symmetric about the origin. Several special cases of model (2.15) are treated.

4.1. *One-sample location model.* In this model $\mathbf{x}=(x_1,\cdots, x_N)$, $\theta=\mu$, $\Omega=(-\infty,\infty)$, $T_\mu\mathbf{x}=(x_1-\mu,\cdots, x_N-\mu)$, and F is symmetric about the origin. To define the Ω-regions, construct the averages $\{a_{ij}=\frac{1}{2}(x_i+x_j),\ 1\leqq i\leqq j\leqq N\}$ and let $d_1<d_2<\cdots<d_{L-1}$, where $L=N(N+1)/2+1$, denote the ordered averages. This strict ordering can be accomplished with probability one.

LEMMA 4.1. *The Ω-regions $\{\Omega_i^+,\ 1\leqq i\leqq L\}$ for the one-sample location model with symmetric error are the L open sets $(-\infty, d_1)$, $(d_1, d_2),\cdots,(d_{L-1},\infty)$. The region $\Omega_0^+=\{d_1, d_2,\cdots, d_{L-1}\}$.*

PROOF. For any pair $i\neq j$, the points x_i, x_j and $\frac{1}{2}(x_i+x_j)$ divide the real line into four open sets L_1, L_2, L_3, L_4. Without loss of generality, assume $i<j$ and $x_i<x_j$. Then $L_1=(-\infty, a_{ii})$, $L_2=(a_{ii}, a_{ij})$, $L_3=(a_{ij}, a_{jj})$ and $L_4=(a_{jj},\infty)$. For all $\mu\in L_1$, $|x_i-\mu|<|x_j-\mu|$ and $\operatorname{sign}(x_i-\mu, x_j-\mu)=(1,1)$. For all $\mu\in L_2$, $|x_i-\mu|<|x_j-\mu|$ and $\operatorname{sign}(x_i-\mu, x_j-\mu)=(-1,1)$. For all $\mu\in L_3$, $|x_i-\mu|>|x_j-\mu|$ and $\operatorname{sign}(x_i-\mu, x_j-\mu)=(-1,1)$. Finally, for all $\mu\in L_4$, $|x_i-\mu|>|x_j-\mu|$ and $\operatorname{sign}(x_i-\mu, x_j-\mu)=(-1,-1)$.

An arbitrary set Ω_k^+, $1\leqq k\leqq L$ (as defined in the lemma), may be constructed by appropriately selecting, for each pair $i\neq j$, one of the open line segments defined by a_{ii}, a_{jj} and a_{ij} (or a_{ji} if $i>j$) and by forming the set-theoretic intersection of these chosen segments. Therefore, $(\operatorname{rank}(|T_\mu\mathbf{x}|), \operatorname{sign}(T_\mu\mathbf{x}))$ is constant for $\mu\in\Omega_k^+$.

Any other region Ω_l^+, $l\neq k$, $1\leqq l\leqq L$, must be contained in a line segment (corresponding to some pair $i\neq j$) which does not contain Ω_k^+. Hence, $(\operatorname{rank}(|T_\mu\mathbf{x}|, \operatorname{sign}(T_\mu\mathbf{x}))$ is different when $\mu\in\Omega_l^+$ than when $\mu\in\Omega_k^+$. The lemma follows.

Proposition 2.2 and the lemma above give the upper and lower probabilities concerning μ alone. Optimal decisions involving μ may be found as in Section 3.1. For example, if μ is restricted to a closed interval containing d_1 and d_{L-1}, the minimax estimate for μ under the loss function $l(\mu, d)=|\mu-d|$ is

$$\hat{\mu}=\operatorname{median}\{d_1,\cdots, d_{L-1}\}. \tag{4.1}$$

This estimate was proposed on classical grounds by Hodges and Lehmann [4]. The minimax test for $H:\mu\leqq 0$ versus $K:\mu>0$ rejects H if and only if the number of pairs $\{(x_i, x_j),\ 1\leqq i\leqq j\leqq N\}$ for which $x_i+x_j>0$ is sufficiently large. This amounts to the one-sample Wilcoxon test.

4.2. *One-sample linear regression model.* This is the model of Section 3.4, with the additional assumption that F is symmetric about the origin.

LEMMA 4.2. *The Ω-regions $\{\Omega_i^+, 1 \leqq i \leqq L\}$ for the linear regression model with symmetric error are the L open regions in Ω partitioned off by the lines $\{\mu+\beta z_i = x_i$, $\mu+\beta z_j = x_j$, $\beta = (x_i-x_j)/(z_i-z_j)$, $2\mu+\beta(z_i+z_j) = x_i+x_j$, all $i \neq j\}$, excluding the possible lines $\beta = \pm\infty$. The region Ω_0^+ consists of the partitioning lines themselves.*

PROOF. For any pair $i \neq j$, the four lines $\{\mu+\beta z_i = x_i, \mu+\beta z_j = x_j, \mu+\beta z_i-x_i = \mu+\beta z_j-x_j, \mu+\beta z_i-x_i = -\mu-\beta z_j+x_j\}$, which amount to the lines in the statement of the lemma, intersect in a point, thereby dividing Ω into eight regions (except when the point is at infinity). Each of these regions is characterized by whether $|x_i-\mu-\beta z_i| > |x_j-\mu-\beta z_j|$ or vice versa, whether $\operatorname{sign}(x_i-\mu-\beta z_i)$ is 1 or -1, and by whether $\operatorname{sign}(x_j-\mu-\beta z_j)$ is 1 or -1. The rest of the proof parallels that of Lemma 4.1.

Unlike the upper and lower probabilities of Section 3.4, those implied by Lemma 4.2 and Proposition 2.2 bear information about both μ and β. Minimax tests (step loss function) for hypotheses concerning (μ, β) are easily found theoretically, but require considerable computation to apply.

4.3. *First order auto-regressive model.* This is the model of Section 3.5, with the additional assumption that F is symmetric about the origin. Essentially the same argument as for Lemma 4.2 establishes

LEMMA 4.3. *The Ω-regions $\{\Omega_i^+, 1 \leqq i \leqq L\}$ for the first order auto-regressive model with symmetric error are the L open regions in Ω partitioned off by the lines $\{x_i = \mu+\rho(x_{i-1}-\mu)$, $x_j = \mu+\rho(x_{j-1}-\mu)$, $\rho = (x_i-x_j)/(x_{i-1}-x_{j-1})$, $x_i+x_j = 2\mu+\rho(x_{i-1}+x_{j-1}-2\mu)$, all $i \neq j\}$, excluding the possible lines $\rho = \pm\infty$. The region Ω_0^+ consists of those points in Ω that lie on the partitioning lines themselves.*

The remarks following the previous regression example carry over, with the obvious changes, to this model.

REFERENCES

[1] DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355–374.

[2] DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38** 325–339.

[3] FRASER, D. A. S. (1968). *The Structure of Inference.* Wiley, New York.

[4] HODGES, J. L., JR., and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

[5] SEN, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.* **63** 1379–1389.

[6] SUKHATME, B. V. (1957). On certain two-sample non-parametric tests for variances. *Ann. Math. Statist.* **28** 188–194.

[7] THEIL, H. (1950). A rank-invariant method of linear and polynomial regression analysis, I, II, and III. *Nederl. Akad. Wetensch. Proc.*, Ser. A **53** 368–392, 521–525, 1397–1412.