JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Minimax Error Selection of a Discrete Univariate Distribution With Prescribed Componentwise Ranking

A. J. Goldman

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(September 27, 1968)

The topic treated is that of finding a reproducible, plausible and computationally simple method of selecting a discrete frequency distribution with a prescribed ranking of its components. The problem is shown to be tractable when a minimax error selection criterion is employed, and "error" is measured by maximum absolute deviation between components. The vertices of the polyhedron of optimal solutions can also be found explicitly, and so their centroid can be calculated if unique specification is required.

Key Words: Mathematical models; minimax estimation; probability distribution.

1. Introduction

In the mathematical modeling efforts associated with an operations research study, one may well have only incomplete information on which to base a representation of the probabilities of the various outcomes of some pertinent chance event. Under these circumstances, one should of course examine the consequences of several alternative probability distributions, each consistent with the information at hand. It still seems desirable, however, to have a systematic and reproducible method for arriving at a *single* "nominal" distribution, to serve as a base-point for such sensitivity analyses.

This note works out the mathematics of one approach, based on a "minimax error" criterion, to the selection of a nominal distribution. The "incomplete information" is assumed to consist of a *ranking* of the individual terms of the probability distribution.[1] A previous paper[2] considered the case in which the given information consists of prescribed upper and lower bounds on these terms.

Let P be the polyhedron of real n-vectors $\mathbf{x}$ (where $n > 1$) whose components x_i satisfy

$$0 \leq x_1 \leq x_2 \leq \ldots \leq x_n, \tag{1.1}$$

$$\Sigma_i x_i = 1. \tag{1.2}$$

Our objective is to choose $\mathbf{x} \epsilon P$ to minimize

$$F(\mathbf{x}) = \max \{d(\mathbf{x}, \mathbf{y}) : \mathbf{y} \epsilon P\}, \tag{1.3}$$

where d is the metric on n-space given by

$$d(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|. \tag{1.4}$$

[1] Study of this situation was suggested by J. McLynn of Davidson, Talbird and McLynn, Inc.

[2] A. J. Goldman and P. R. Meyers, Minimax error selection of a discrete univariate distribution with prescribed componentwise bounds, J. Res. NBS **72B** (Math. Sci.) No. 4 263-271 (1968).

Other metrics may be investigated subsequently.

The following section accomplishes the explicit evaluation of $F(\mathbf{x})$ in highly tractable form. The minimization of $F(\mathbf{x})$ over P, which begins section 3, is then trivial. It turns out, however, that for $n > 2$ there will be a convex polyhedron of optimal $\mathbf{x}$'s rather than a single one; section 3 goes on to propose the centroid of this polyhedron's vertices as a plausible "representative" choice, and includes a determination of both vertices and centroid.

2. Evaluation of *F*

The particular metric (1.4) has the pleasant property that $F(\mathbf{x})$—for *any* compact set P, not only the one defined by (1.1) and (1.2)—can readily be evaluated in terms of the quantities

$$M_i = \max \{y_i : \mathbf{y} \epsilon P\}, \tag{2.1}$$

$$m_i = \min \{y_i : \mathbf{y} \epsilon P\}. \tag{2.2}$$

Indeed, we have

$$F(x) = \max_y \max_i \max (y_i - x_i, x_i - y_i)$$

$$= \max_i \max \{\max_y (y_i - x_i), \max_y (x_i - y_i)\},$$

or finally

$$F(x) = \max_i \max \{M_i - x_i, x_i - m_i\}. \tag{2.3}$$

For the particular polyhedron P which figures here, it is easily verified that

$$M_i = 1/(n+1-i) \qquad (1 \leqslant i \leqslant n), \tag{2.4}$$

$$m_i = 0 \qquad (1 \leqslant i < n), \tag{2.5}$$

$$m_n = 1/n. \tag{2.6}$$

Thus we have

$$F(\mathbf{x}) = \max \{\max_i \{1/(n+1-i) - x_i\}, x_{n-1}, x_n - 1/n\}.$$

The term corresponding to $i = n$ is

$$1 - x_n = \Sigma_1^{n-1} x_i \geqslant x_{n-1},$$

so we can replace the last equation by

$$F(\mathbf{x}) = \max \{\max_i \{1/(n+1-i) - x_i\}, x_n - 1/n\}. \tag{2.7}$$

Next, suppose the term with $i = n-1$ yields the maximum in (2.7). Then we should have

$$1/2 - x_{n-1} \geqslant x_n - 1/n, \tag{2.8}$$

$$1/2 - x_{n-1} \geqslant 1 - x_n. \tag{2.9}$$

Addition yields

$$1-2x_{n-1} \geq 1-1/n,$$

which is equivalent to $x_{n-1} \leq 1/2n$. Thus $x_i \leq 1/2n$ for $1 \leq i < n-1$, and so

$$1-x_{n-1}-x_n = \Sigma_1^{n-2} x_i \leq (n-2)/2n.$$

This and (2.8) imply

$$(n+2)/2n - x_{n-1} \leq x_n = 1/n + (x_n - 1/n) \leq 1/n + (1/2 - x_{n-1}) = (n+2)/2n - x_{n-1}. \tag{2.10}$$

Since equality holds in (2.10), it must hold throughout the preceding sequence of equalities. In particular it holds in (2.8) and (2.9), so that the maximum in (2.7) does not occur *uniquely* for the term with $i = n-1$, and hence this term can be deleted from (2.7):

$$F(\mathbf{x}) = \max \{\max_{i<n-1}\{1/(n+1-i) - x_i\},\ 1-x_n,\ x_n - 1/n\}. \tag{2.11}$$

For $n=2$, (2.11) gives

$$F(\mathbf{x}) = \max \{1-x_2,\ x_2 - 1/2\}. \tag{2.12}$$

We next assume $n \geq 3$.

Now suppose the maximum in (2.11) holds for the term corresponding to some $i < n-1$. Then we should have

$$1/(n+1-i) - x_i \geq 1 - x_n, \tag{2.13}$$

$$1/(n+1-i) - x_i \geq x_n - 1/n. \tag{2.14}$$

Addition yields

$$2/(n+1-i) - 2x_i \geq 1 - 1/n,$$

or equivalently

$$\begin{aligned} 2x_i &\leq 2/(n+1-i) + 1/n - 1 \\ &= [2n + (n+1-i) - n(n+1+i)]/n(n+1-i) \\ &= [1 + 2n - n^2 + (n-1)i]/n(n+1-i) \\ &= [2 - (n-1)^2 + (n-1)i]/n(n+1-i) \\ &= [2 - (n-1)(n-1-i)]/n(n+1-i). \end{aligned}$$

Since $x_i \geq 0$, we must have

$$2 \geq (n-1)(n-1-i). \tag{2.15}$$

But $n-1 \geq 2$ and $n-i > 1$. Thus this situation is impossible for $n > 3$, i.e.,

$$F(\mathbf{x}) = \max \{1-x_n,\ x_n - 1/n\} \qquad (n > 3). \tag{2.16}$$

And it is possible for $n=3$ only when $i=n-2=1$, in which case equality holds in (2.15), hence throughout the preceding inequalities, so that the maximum in (2.11) does not hold *uniquely* for the term corresponding to $i=1$. We have now shown that, for *all* $n>1$,

$$F(\mathbf{x})=\max\{1-x_n,\ x_n-1/n\}. \tag{2.17}$$

3. Determination of Optimal Distributions

The minimization of $F(\mathbf{x})$ over P is trivial in view of (2.17), which can be rewritten

$$\begin{aligned} &F(\mathbf{x})=1-x_n \text{ for } 0\leq x_n\leq (n+1)/2n,\\ &F(\mathbf{x})=x_n-1/n \text{ for } (n+1)/2n\leq x_n\leq 1. \end{aligned} \tag{3.1}$$

We see at once that

$$F_{\min}=(n-1)/2n, \tag{3.2}$$

and that the minimizing $\mathbf{x}\epsilon P$ are precisely those with

$$x_n=(n+1)/2n. \tag{3.3}$$

One such $\mathbf{x}$ is given by

$$x_i=1/2n \qquad \text{for } 1\leq i\leq n-1. \tag{3.4}$$

But this choice is somewhat arbitrary, and moreover honors the original ranking (for $x_1, \ldots, x_{n-1}$) only in a technical sense. In view of the original intention to fasten on a single $\mathbf{x}$, it seems (to the writer) less arbitrary to select the *centroid* of the vertices of the polyhedron of $F-$ minimizing $\mathbf{x}\epsilon P$. This centroid is obtained by adjoining an nth component $(n+1)/2n$, as in (3.3), to the centroid of the polyhedron Q in $(n-1)$-space consisting of those $\mathbf{z}$ with

$$0\leq z_1\leq \ \ldots\ \leq z_{n-1}\leq (n+1)/2n, \tag{3.5}$$

$$\Sigma_1^{n-1}z_i=1-(n+1)/2n=(n-1)/2n. \tag{3.6}$$

Note that in view of (3.6), the last inequality in (3.5) can be dropped.

To determine the vertices of Q, note that a nonsingular linear transformation preserves extreme points of polyhedra. This applies in particular to the transformation associated with the substitutions

$$\begin{aligned} &u_1=z_1,\\ &u_i=z_i-z_{i-1} \qquad (2\leq i\leq n-1), \end{aligned}$$

which convert Q into the polyhedron Q' defined in $\mathbf{u}$-space by

$$u_i\geq 0 \qquad (1\leq i\leq n-1), \tag{3.7}$$

$$\Sigma_{i=1}^{n-1}(n-i)u_i=(n-1)/2n. \tag{3.8}$$

Q' consists of the intersection of the hyperplane H, defined by (3.8), with the nonnegative orthant.

Thus its extreme points are the intersections of H with the $(n-1)$ positive coordinate axes, i.e., the $(n-1)$ points $\mathbf{u}^{(j)}$ given by

$$u_i^{(j)} = \delta_{ij}(n-1)/2n(n-j). \tag{3.9}$$

The corresponding extreme points $\mathbf{z}^{(j)}$ of Q are given by

$$z_i^{(j)} = 0 \qquad (1 \leq i \leq j),$$

$$z_i^{(j)} = (n-1)/2n(n-j) \qquad (j \leq i \leq n-1).$$

Thus the centroid $\mathbf{z}^*$ of Q has coordinates

$$z_i^* = (n-1)^{-1}\Sigma_{j \leq i}(n-1)/2n(n-j) = (1/2n)\Sigma_{j \leq i}(n-j)^{-1}, \tag{3.10}$$

which unlike (3.4) are strictly increasing in i.

We conclude by pointing out two somewhat disquieting features of the solution obtained above. The first [see (3.3)] is that the chance-event outcome ranked most probable is assigned a probability greater than 1/2, though only slightly so for large n. The second [see (3.2)] is the size of $F_{\min}$ . . . nearly 1/2 for large n. But perhaps this should not be surprising in view of the extremely conservative nature of minimax decision criteria.

(Paper 72B4–277)