SOME MINIMAX INVARIANT PROCEDURES FOR ESTIMATING A CUMULATIVE DISTRIBUTION FUNCTION[1]

By Om P. Aggarwal

Purdue University and University of Washington

1. Summary. Some invariant procedures, which are essentially step-functions, are considered as estimators of the cumulative distribution function of a one-dimensional random variable on which a finite fixed number of observations are given, for various loss functions. Two principal classes of loss functions are considered and it is shown that for a special loss function in one class the optimum procedure is the usual sample cumulative function.

2. Introduction. Suppose that a sample $X_1, X_2, \cdots, X_n$ of a one-dimensional chance variable X is given. In a recent paper, Birnbaum [1] has discussed various techniques for deciding whether X has a completely specified continuous cumulative distribution function (c.d.f.), $H(x) = P(X \leqq x)$. In this paper is discussed an allied problem, viz., that if $F(x) = P(X \leqq x)$ is the unknown continuous c.d.f. of X and if $\hat{F}(x)$ be an estimate of $F(x)$ based on the sample $X_1, X_2, \cdots, X_n$, what would be the best estimate $\hat{F}$ when certain forms of the loss function are given.

Consider the loss function

$$L(F, \hat{F}) = \int_{-\infty}^{\infty} |F(x) - \hat{F}(x)|^r \, dx, \tag{1}$$

where r is an integer $\geqq 1$. It is almost obvious that the only invariant procedures for estimating F under the group of all one-to-one monotone transformations of the real numbers onto themselves which leave the sample values X_i $(i = 1, 2, \cdots, n)$ invariant are those which estimate $F(x)$ by a step-function

$$\hat{F}(x) = \text{constant}, \quad \text{say } c_j \text{ for } \quad X^{(j)} \leqq x < X^{(j+1)}, \tag{2}$$

where $X^{(1)} < X^{(2)} < \cdots < X^{(n)}$ are the ordered observations and $X^{(0)}$ and $X^{(n+1)}$ denote $-\infty$ and $+\infty$ respectively.

Using this estimate $\hat{F}$, we get

$$\begin{aligned} L(F, \hat{F}) &= \sum_{j=0}^{n} \int_{X^{(j)}}^{X^{(j+1)}} |F(x) - c_j|^r \, dF(x) \\ &= \frac{1}{r+1} \sum_{j=0}^{n} [(F(X^{(j+1)}) - c_j)|F(X^{(j+1)}) - c_j|^r \\ &\qquad - (F(X^{(j)}) - c_j)|F(X^{(j)}) - c_j|^r] \end{aligned} \tag{3}$$

Received August 20, 1954; revised February 28, 1955.

[1] Work done under the sponsorship of the Office of Naval Research.

and the right-hand side of this equation is a symmetric function of $F(X_1), F(X_2), \cdots, F(X_n)$ where $X_1, X_2, \cdots, X_n$ is the unordered sample. Using the probability integral transformation, it is clear that the distribution of $L(F, \hat{F})$ does not depend on F for F continuous. Hence the risk R, being the expectation of L with respect to the distribution F, is constant and independent of F itself. We can thus take F to be a rectangular distribution over (0, 1) and write

$$(4) \qquad R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} |x - c_j|^r \, dx,$$

where $X_1 < X_2 < \cdots < X_n$ is an ordered sample of size n from this rectangular distribution over (0, 1), X_0 and X_{n+1} denote 0 and 1 respectively, and the symbol E denotes that the expectation is taken with respect to the rectangular distribution over (0, 1). In the rest of this paper, we shall use consistently the letter E to denote the fact that the expectation is to be taken with respect to the rectangular distribution over (0, 1).

The same argument applies when the loss function is of the form

$$(5) \qquad L(F, \hat{F}) = \int_{-\infty}^{\infty} \frac{|F(x) - \hat{F}(x)|^r}{F(x)[1 - F(x)]} \, dF(x)$$

and in this case by taking $\hat{F}$ as in (2) we obtain

$$(6) \qquad R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} \frac{|x - c_j|^r}{x(1 - x)} \, dx$$

where X_j, $j = 0, 1, \cdots n + 1$, are the same as in (4).

It is obvious that since risk R is constant, a minimax procedure among the class of invariant procedures being considered will be to choose c_j, $j = 0, 1, \cdots, n$, such that R is minimum. We consider in this paper the values of c_j when the loss function is of the form (1) for all integers $r \geqq 1$ and when the loss function is of the form (5) for $r = 1$ and when r is an even integer $\geqq 2$. The case when r is odd in (5) seems to be rather complicated.

3. The loss function $L(F, \hat{F}) = \int_{-\infty}^{\infty} [F(x) - \hat{F}(x)]^r \, dF(x)$ where r is any positive even integer. Let $r = 2s$, then

$$(7) \qquad R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} (x - c_j)^{2s} \, dx = \sum_{j=0}^{n} Q_j,$$

where

$$(8) \qquad Q_j = \frac{1}{2s + 1} E \sum_{k=0}^{2s+1} \binom{2s + 1}{k} (-c_j)^{2s+1-k} (X_{j+1}^k - X_j^k)\cdot$$

for $j = 0, 1, 2, \cdots, n$.

Since the distribution of the jth order statistic X_j in a sample of size n from the rectangular distribution over (0, 1) is a beta distribution with probability density

$$(9) \qquad p(y) = \frac{1}{B(j, n - j + 1)} y^{j-1}(1 - y)^{n-j}, \qquad 0 \leqq y \leqq 1,$$

it is easily seen that for any positive integer r,

$$E(X_j^r) = \frac{j(j+1)\cdots(j+r-1)}{(n+1)(n+2)\cdots(n+r)}, \tag{10}$$

$$E(X_{j+1}^r - X_j^r) = \begin{cases} \dfrac{r(j+1)(j+2)\cdots(j+r-1)}{(n+1)(n+2)\cdots(n+r)} & \text{for } r \neq 1, \\ \dfrac{1}{n+1} \quad \text{for } r = 1. & j = 0, 1, \cdots, n. \end{cases} \tag{11}$$

Substituting from (11) in (8) we obtain

$$\begin{aligned} Q_j &= \frac{1}{n+1} c_j^{2s} + \frac{1}{2s+1} \sum_{k=2}^{2s+1} \binom{2s+1}{k} \\ &\qquad \cdot (-c_j)^{2s+1-k} \frac{k(j+1)\cdots(j+k-1)}{(n+1)\cdots(n+k)} \\ &= \frac{1}{n+1}\left[c_j^{2s} + \sum_{k=2}^{2s+1} \binom{2s}{k-1} (-c_j)^{2s+1-k} \frac{(j+1)\cdots(j+k-1)}{(n+2)\cdots(n+k)}\right]. \end{aligned} \tag{12}$$

For conciseness we introduce the following notation somewhat similar to the binomial and distinguished from it by an asterisk. Let

$$\left(t - \frac{a+1}{b+1}\right)^{q*} = t^q + \sum_{k=1}^{q} (-1)^k \binom{q}{k} t^{q-k} \prod_{i=1}^{k} \frac{a+i}{b+i}, \tag{13}$$

for fixed real a and b and a positive integer q. For $q = 0$, let (13) be equal to 1. It is easily verified that for any positive integer r,

$$\frac{d^r}{dt^r}\left(t - \frac{a+1}{b+1}\right)^{q*} = \begin{cases} q(q-1)\cdots(q-r+1) \\ \qquad \left(t - \dfrac{a+1}{b+1}\right)^{(q-r)*} & \text{when} \quad r \leqq q, \\ 0 \quad \text{when} \quad r > q. \end{cases} \tag{14}$$

Using this notation we can write

$$Q_j = \frac{1}{n+1}\left(c_j - \frac{j+1}{n+2}\right)^{2s*} \tag{15}$$

We have to choose c_j so as to minimize R. Since $R = \sum_0^n Q_j$, and from (7) we see that for each j, Q_j is positive and depends only on j, it is obvious that minimizing R is equivalent to minimizing Q_j separately for each j. We obtain

$$\frac{\partial Q_j}{\partial c_j} = \frac{2s}{n+1}\left(c_j - \frac{j+1}{n+2}\right)^{(2s-1)*}, \tag{16}$$

$$\frac{\partial^2 Q_j}{\partial c_j^2} = \frac{2s(2s-1)}{n+1}\left(c_j - \frac{j+1}{n+2}\right)^{(2s-2)*} \tag{17}$$

Since $Q_j = E \int_{X_j}^{X_{j+1}} (x - c_j)^{2s} \, dx > 0$, it is clear that

$$\frac{\partial^2 Q_j}{\partial c_j^2} = 2s(2s - 1)E \int_{X_j}^{X_{j+1}} (x - c_j)^{2s-2} \, dx > 0. \tag{18}$$

Let $f(c_j) = \partial Q_j / \partial c_j$. It is easily seen that $f(0)$ is negative and $f(1)$ is positive, and since $f'(c_j) > 0$ for all real c_j, $f(c_j)$ is a strictly increasing function of c_j. Hence $f(c_j) = 0$ for one and only one real value of c_j, and this c_j necessarily lies between 0 and 1. Thus we find that Q_j, and hence R, is minimized by setting $\partial Q_j / \partial c_j = 0$ and solving for c_j the resulting equation

$$\left(c_j - \frac{j+1}{n+2}\right)^{(r-1)*} = 0. \tag{19}$$

This equation has one and only one real root which lies between 0 and 1. The minimax invariant procedure for the loss function of this section is thus to estimate $F(x)$ by

$$\hat{F}(x) = c_j; \qquad X_j \leqq x < X_{j+1}, \qquad j = 0, 1, \cdots, n, \tag{20}$$

where X_j, $j = 0, 1, \cdots, n + 1$, have been defined earlier and c_j is the real root of (19). It can further be seen from (19) that the equation remains unchanged if we replace j by $n - j$ and c_j by $1 - c_j$. Hence $c_{n-j} = 1 - c_j$, and we see that in practice the number of equations to be solved is about half the sample size.

Special case for $r = 2$. When $r = 2$, the equation (19) reduces to a linear equation

$$\left(c_j - \frac{j+1}{n+2}\right)^{1*} = 0, \tag{21}$$

which has the unique solution $c_j = (j + 1) / (n + 2)$. This result can, however, be obtained directly by writing the risk R from (7) and (12) for $s = 1$ in the form

$$R = \frac{1}{6(n+2)} + \frac{1}{n+1} \sum_{j=0}^{n} \left(c_j - \frac{j+1}{n+2}\right)^2 \tag{22}$$

We see thus that R is minimized by choosing

$$c_j = \frac{j+1}{n+2}, \qquad j = 0, 1, \cdots, n, \tag{23}$$

and hence the minimax invariant procedure is to estimate $F(x)$ by

$$\hat{F}(x) = \frac{j+1}{n+2}, \qquad X_j \leqq x < X_{j+1}, \qquad j = 0, 1, \cdots, n, \tag{24}$$

where $(X_1, X_2, \cdots, X_n)$ is the ordered sample and X_0 and X_{n+1} stand for $-\infty$ and $+\infty$ respectively.

The minimum risk corresponding to this procedure is seen to be $1/6(n + 2)$. It is of some interest to note that the risk corresponding to the usual procedure of taking $c_j = j/n$ is given by $1/6n$.

4. The loss function $L(F, \hat{F}) = \int_{-\infty}^{\infty} |F(x) - \hat{F}(x)|^r \, dF(x)$, where r is any positive integer. In this case

$$R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} |x - c_j|^r \, dx = \sum_{j=0}^{n} Q_j, \tag{25}$$

where

$$Q_j = \frac{1}{r+1} E[(X_{j+1} - c_j)\,|X_{j+1} - c_j|^r - (X_j - c_j)\,|X_j - c_j|^r]. \tag{26}$$

Using (9) we obtain

$$\begin{aligned} E[(X_j - c_j)\,|X_j - c_j|^r] = j\binom{n}{j}\Bigg[\int_{c_j}^{1} (y - c_j)^{r+1} y^{j-1} (1 - y)^{n-j} \, dy \\ - \int_0^{c_j} (c_j - y)^{r+1} y^{j-1} (1 - y)^{n-j} \, dy\Bigg], \end{aligned} \tag{27}$$

and similarly,

$$\begin{aligned} E[(X_{j+1} - c_j)\,|X_{j+1} - c_j|^r] = (n - j)\binom{n}{j} \\ \Bigg[\int_{c_j}^{1} (y - c_j)^{r+1} y^{j} (1 - y)^{n-j-1} \, dy - \int_0^{c_j} (c_j - y)^{r+1} y^{j} (1 - y)^{n-j-1} \, dy\Bigg]. \end{aligned} \tag{28}$$

From (27) and (28) we obtain

$$\begin{aligned} Q_j = \frac{1}{r+1}\binom{n}{j}\Bigg[\int_{c_j}^{1} (y - c_j)^{r+1} y^{j-1} (1 - y)^{n-j-1} (ny - j) \, dy \\ + (-1)^r \int_0^{c_j} (y - c_j)^{r+1} y^{j-1} (1 - y)^{n-j-1} (ny - j) \, dy\Bigg]. \end{aligned} \tag{29}$$

Again it is obvious that to minimize R is equivalent to minimizing Q_j for each j. Further we see that the conditions for differentiation with respect to c_j under the integral sign in (29) are satisfied, and we obtain

$$\begin{aligned} \frac{\partial Q_j}{\partial c_j} = -\binom{n}{j}\Bigg[\int_{c_j}^{1} (y - c_j)^{r} y^{j-1} (1 - y)^{n-j-1} (ny - j) \, dy \\ + (-1)^r \int_0^{c_j} (y - c_j)^{r} y^{j-1} (1 - y)^{n-j-1} (ny - j) \, dy\Bigg], \end{aligned} \tag{30}$$

$$\begin{aligned} \frac{\partial^2 Q_j}{\partial c_j^2} = r\binom{n}{j}\Bigg[\int_{c_j}^{1} (y - c_j)^{r-1} y^{j-1} (1 - y)^{n-j-1} (ny - j) \, dy \\ + (-1)^r \int_0^{c_j} (y - c_j)^{r-1} y^{j-1} (1 - y)^{n-j-1} (ny - j) \, dy\Bigg] \\ = \begin{cases} r(r-1)E \displaystyle\int_{X_j}^{X_{j+1}} |x - c_j|^{r-2} \, dx, & \text{for } r \geqq 2 \\ 2\displaystyle\binom{n}{j} c_j^j (1 - c_j)^{n-j}, & \text{for } r = 1. \end{cases} \end{aligned} \tag{31}$$

Define a function f by $f(c_j) = \partial Q_j / \partial c_j$. We see by straightforward computations that

$$f(0) = -r\,\frac{n!(r + j - 1)!}{j!(r + n)!} < 0,$$

$$f(1) = r\,\frac{n!(r + n - j - 1)!}{(n - j)!(r + n)!} > 0.$$

Since from (31) it is seen that, for all $r \geqq 2$, $f'(c_j) = \partial^2 Q_j / \partial c_j^2 > 0$ for all real c_j (the special case for $r = 1$ is given at the end of this section), f is a strictly increasing function of c_j and assumes the value zero for one and only one real value of c_j, and this value of c_j necessarily lies between zero and one. Thus we find that Q_j and hence R is minimized by setting $\partial Q_j / \partial c_j = 0$ and solving for c_j the resulting equation

$$\begin{aligned}(32)\qquad &\int_{c_j}^{1} (y - c_j)^r y^{j-1}(1 - y)^{n-j-1}(ny - j)\,dy \\ &\qquad + (-1)^r \int_0^{c_j} (y - c_j)^r y^{j-1}(1 - y)^{n-j-1}(ny - j)\,dy = 0.\end{aligned}$$

Thus the problem reduces to that of solving the above equation for $j = 0, 1, \cdots, n$. The general solution of (32) giving c_j explicitly in terms of j, n, and r does not seem to be possible. We shall, however, simplify the equation so that it should not be too difficult to obtain the solution in any given case. It can, however, be proved from (32) that $c_{n-j} = 1 - c_j$, so that the number of equations to be solved in practice will be about half the sample size.

We can write (32) as

$$\begin{aligned}(33)\qquad &\int_0^{1} (y - c_j)^r y^{j-1}(1 - y)^{n-j-1}(ny - j)\,dy \\ &\qquad = [1 - (-1)^r] \int_0^{c_j} (y - c_j)^r y^{j-1}(1 - y)^{n-j-1}(ny - j)\,dy.\end{aligned}$$

The left-hand side of equation (33) can be expressed as

$$(34)\qquad \sum_{k=0}^{r} k \binom{r}{k} (-c_j)^{r-k} B(j + k, n - j + 1),$$

which indicates that the coefficient of c_j^r is zero. For $k \neq 0$, we can utilize the fact that $\binom{r}{k}\,k = r\,\binom{r-1}{k-1}$ and reduce it further to the form

$$(35)\quad rB(j, n - j + 1) \sum_{k=1}^{r} \binom{r - 1}{k - 1} (-c_j)^{r-k} \frac{j(j + 1) \cdots (j + k - 1)}{(n + 1)(n + 2) \cdots (n + k)},$$

which by making use of the notation introduced in (13) can be written as

$$(36)\qquad (-1)^{r-1} rB(j + 1, n - j + 1) \left(c_j - \frac{j + 1}{n + 2}\right)^{(r-1)*}$$

When r is even, the right-hand side of the equation (33) reduces to zero and cancelling out the nonzero coefficient $(-1)^{r-1}rB(j+1, n-j+1)$ from the left-hand side as expressed by (36) we obtain c_j as a root of the same equation as (19) obtained earlier by a different method.

The right-hand side of the equation (33), except for the factor $[1-(-1)^r]$, can be written as

$$\sum_{k=0}^{r}\binom{r}{k}(-c_j)^{r-k}\int_0^{c_j}\sum_{s=0}^{n-j}(-1)^{s-1}(j+s)\binom{n-j}{s}y^{k+j+s-1}\,dy, \tag{37}$$

and by making use of the relation

$$\sum_{k=0}^{r}(-1)^k\binom{r}{k}\frac{1}{k+t}=B(t,r+1), \tag{38}$$

it can be reduced to

$$(-1)^{r-1}r\sum_{s=0}^{n-j}(-1)^s\binom{n-j}{s}B(r,j+s+1)c_j^{r+j+s} \tag{39}$$

Using (36) and (37) we can, thus, write the equation (33) as

$$\begin{aligned}&B(j+1,n-j+1)\left(c_j-\frac{j+1}{n+2}\right)^{(r-1)*}\\&\qquad=[1-(-1)^r]\sum_{s=0}^{n-j}(-1)^s\binom{n-j}{s}B(r,j+s+1)c_j^{r+j+s}.\end{aligned} \tag{40}$$

This equation is to be solved for c_j to get a minimax invariant procedure for estimating F when the loss function is given by (1). When r is even, the factor $1-(-1)^r=0$ and we get an equation of degree $(r-1)$. When r is odd, the factor $1-(-1)^r=2$ and the equation reduces to

$$\begin{aligned}&\sum_{s=0}^{n-j}(-1)^s\binom{n-j}{s}B(r,j+s+1)c_j^{r+j+s}\\&\qquad-\tfrac{1}{2}B(j+1,n-j+1)\left(c_j-\frac{j+1}{n+2}\right)^{(r-1)*}=0\end{aligned} \tag{41}$$

which is an equation of degree $n+r$. In either case there is one and only one real root which lies between 0 and 1 and the set of such roots for $j=0, 1, \cdots, n$ minimizes R.

An alternative way of expressing the right-hand side of (33) is to rewrite (39) in the form:

$$(-1)^{r-1}r!\sum_{s=0}^{n-j}(-1)^s\binom{n-j}{s}\frac{c_j^{r+j+s}}{(j+s+1)(j+s+2)\cdots(j+s+r)}. \tag{42}$$

It is easily verified that (42) is equal to

$$\begin{aligned}&(-1)^{r-1}r!\sum_{s=0}^{n-j}(-1)^s\binom{n-j}{s}\int_0^{c_j}\int_0^{z_r}\cdots\int_0^{z_2}z_1^{j+s}\,dz_1\cdots dz_r\\&\qquad=(-1)^{r-1}r!\int_0^{c_j}\int_0^{z_r}\cdots\int_0^{z_2}z_1^j(1-z_1)^{n-j}\,dz_1\cdots dz_r.\end{aligned} \tag{43}$$

The equation (40) can, therefore, also be expressed as

$$B(j + 1, n - j + 1)\left(c_j - \frac{j+1}{n+2}\right)^{(r-1)*} = [1 - (-1)^r](r - 1)! \int_0^{c_j} \int_0^{z_r} \cdots \int_0^{z_2} z_1^j(1 - z_1)^{n-j}\, dz_1 \cdots dz_r. \tag{44}$$

Special case for $r = 1$. When $r = 1$, (30) is easily seen to reduce to

$$\frac{\partial Q_j}{\partial c_j} = 2\binom{n}{j}\left[\int_0^{c_j} z^j(1 - z)^{n-j}\, dz - \tfrac{1}{2}B(j + 1, n - j + 1)\right], \tag{45}$$

from which follows easily the result given in (31), viz.,

$$\frac{\partial^2 Q_j}{\partial c_j^2} = 2\binom{n}{j} c_j^j(1 - c_j)^{n-j}. \tag{46}$$

Setting $\partial Q_j / \partial c_j = 0$ and solving we obtain c_j as the median of the beta distribution with density

$$g(z) = \frac{1}{B(j + 1, n - j + 1)} z^j(1 - z)^{n-j}, \qquad 0 \leqq z \leqq 1, \tag{47}$$

for $j = 0, 1, 2, \cdots, n$. Since (46) shows that $\partial^2 Q_j / \partial c_j^2 > 0$ for $0 < c_j < 1$, it follows that this solution for c_j in fact minimizes Q_j for $j = 0, 1, \cdots, n$, and hence minimizes R. The equation (44) for c_j obtained for $r \geqq 2$ thus holds good for $r = 1$ as well and the minimax invariant procedure is seen to estimate $F(x)$ by

$$\hat{F}(x) = c_j\,; \qquad X_j \leqq x < X_{j+1}\,, \qquad j = 0, 1, \cdots, n, \tag{48}$$

where $(X_1, X_2, \cdots, X_n)$ is the ordered sample, X_0 and X_{n+1} stand for $-\infty$ and $+\infty$ respectively, and c_j $(j = 0, 1, \cdots, n)$ is the median of the beta distribution with density (47). It is rather interesting to note that the value $(j + 1)/(n + 2)$ for c_j obtained in the last section for $r = 2$ is the mean of the same beta distribution.

The actual computation of the values of c_j $(j = 0, 1, \cdots, n)$ can be easily carried out, for a given n, with the help of the tables of the incomplete beta function [2]. In the notation of those tables

$$I_x(p, q) = \frac{\int_0^x x^{p-1}(1 - x)^{q-1}\, dx}{\int_0^1 x^{p-1}(1 - x)^{q-1}\, dx} \tag{49}$$

Thus we have to find the value of x for each j such that

$$I_x(j + 1, n - j + 1) = \tfrac{1}{2}. \tag{50}$$

Using the relation

$$I_x(p, q) = 1 - I_{1-x}(q, p), \tag{51}$$

TABLE I

Values of c_j $(j = 0, 1, \cdots, n)$ for $n = 1, 2, \cdots, 12$

n	c_j												
	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
1	.29	.71											
2	.21	.50	.79										
3	.16	.39	.61	.84									
4	.13	.31	.50	.69	.87								
5	.11	.26	.42	.58	.74	.89							
6	.09	.23	.36	.50	.64	.77	.91						
7	.08	.20	.32	.44	.56	.68	.80	.92					
8	.07	.18	.28	.39	.50	.61	.72	.82	.93				
9	.07	.16	.26	.35	.45	.55	.65	.74	.84	.93			
10	.06	.15	.23	.32	.41	.50	.59	.68	.77	.85	.94		
11	.06	.14	.22	.30	.38	.46	.54	.62	.70	.78	.86	.94	
12	.05	.13	.20	.27	.35	.42	.50	.58	.65	.73	.80	.87	.95

it is seen that as in the general case,

$$c_{n-j} = 1 - c_j \,. \tag{52}$$

The values of c_j $(j = 0, 1, \cdots, n)$ for $n = 1, 2, \cdots, 12$ correct to two decimal places are computed and tabulated as shown in Table I.

5. The loss function $L(F, \hat{F}) = \int_{-\infty}^{\infty} [F(x) - \hat{F}(x)]^r / F(x)[1 - F(x)]\, dF(x)$ where r is any positive even integer.

Let $r = 2s$; then

$$R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} \frac{(x - c_j)^{2s}}{x(1 - x)}\, dx = \sum_{j=0}^{n} Q_j, \tag{53}$$

where

$$Q_j = E \int_{X_j}^{X_{j+1}} \frac{(x - c_j)^{2s}}{x(1 - x)}\, dx. \tag{54}$$

Since $X_0 = 0$ and $X_{n+1} = 1$, it is clear that in order to obtain finite risk it is necessary and sufficient that $c_0 = 0$ and $c_n = 1$. For $j \neq 0, n$, we can write

$$\begin{aligned} Q_j = E \Bigg[\sum_{h=0}^{2s-2} \frac{1}{h+1}\, a_h (X_{j+1}^{h+1} - X_j^{h+1}) + c_j^{2s} (\log X_{j+1} - \log X_j) \\ - (1 - c_j)^{2s} \{\log (1 - X_{j+1}) - \log (1 - X_j)\} \Bigg], \end{aligned} \tag{55}$$

where

$$a_h = - \sum_{i=0}^{2s-2-h} \binom{2s}{i} (-c_j)^i; \qquad h = 0, 1, 2, \cdots, 2s - 2. \tag{56}$$

The probability density of X_j is given by (9), from which we obtain

$$E\,(\log X_j) = j\binom{n}{j}\int_0^1 y^{j-1}(1-y)^{n-j}\log y\,dy. \tag{57}$$

In order to evaluate (57) we use the following lemma.

LEMMA 5.1.

$$\int_0^1 y^{j-1}(1-y)^{n-j}\log y\,dy = \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(n+1)}\,[\psi(j) - \psi(n+1)], \tag{58}$$

where $\psi(k) = \Gamma'(k)\,/\,\Gamma(k)$.

PROOF. Let $f(\alpha) = \int_0^1 y^{\alpha-1}(1-y)^{n-j}\,dy$. The left-hand side of (58) is $f'(\alpha)$ evaluated at $\alpha = j$ as can be seen by differentiating under the integral sign. But $f(\alpha) = \Gamma(\alpha)\Gamma(n-j+1)\,/\,\Gamma(\alpha+n-j+1)$, and the desired result is obtained by evaluating the logarithmic derivative of $f(\alpha)$ at $\alpha = j$.

From the lemma 5.1 and (57) we get

$$E(\log X_j) = \psi(j) - \psi(n+1). \tag{59}$$

In the same way, we obtain

$$E\log(1 - X_j) = \psi(n-j+1) - \psi(n+1). \tag{60}$$

Further, since $\Gamma(k+1) = k\Gamma(k)$, $\Gamma'(k+1) = \Gamma(k) + k\Gamma'(k)$, we see that $\psi(k+1) = \Gamma'(k+1)\,/\,\Gamma(k+1) = 1/k + \psi(k)$, and hence the function ψ satisfies the difference equation

$$\psi(k+1) - \psi(k) = 1/k. \tag{61}$$

From (59), (60), and (61) we get

$$E(\log X_{j+1} - \log X_j) = 1/j, \qquad \text{for } j = 1, 2, \cdots, n, \tag{62}$$

and

$$E[\log(1 - X_{j+1}) - \log(1 - X_j)] = -1\,/\,(n-j), \qquad \text{for } j = 0, 1, \cdots, n-1. \tag{63}$$

Substituting from (11), (62), and (63) in (55) we get

$$Q_j = \sum_{h=0}^{2s-2}\frac{(j+h)!\,n!}{(n+h+1)!\,j!}\,a_h + \frac{1}{j}\,c_j^{2s} + \frac{1}{n-j}\,(1-c_j)^{2s}, \tag{64}$$

and substituting from (56), we can write

$$Q_j = \frac{n!}{j!}\sum_{h=0}^{2s-2}\frac{(j+h)!}{(n+h+1)!}\sum_{i=0}^{2s-2-h}(-1)^{i+1}\binom{2s}{i}c_j^i + \frac{1}{j}\,c_j^{2s} + \frac{1}{n-j}\,(1-c_j)^{2s}. \tag{65}$$

This is a $2s$th degree polynomial in c_j. Collecting the coefficients of like powers of c_j we obtain, for $k = 0, 1, 2, \cdots, 2s-2$,

$$Q_j = \frac{n}{j(n-j)}\,c_j^{2s} - \frac{2s}{n-j}\,c_j^{2s-1} + \sum_{k=0}^{2s-2} g_k c_j^k, \tag{66}$$

where

$$g_k = (-1)^{k+1}\binom{2s}{k}\left[\frac{n!}{j!}\sum_{h=0}^{2s-2-k}\frac{(j+h)!}{(n+h+1)!} - \frac{1}{n-j}\right]. \tag{67}$$

To simplify (66) further, we state and prove the following lemma.

Lemma 5.2. *If j and n are positive integers and $j < n$, then*

$$\frac{n!}{j!}\sum_{h=0}^{q}\frac{(j+h)!}{(n+h+1)!} = \frac{1}{n-j}\left[1 - \prod_{\alpha=1}^{q-1}\frac{j+\alpha}{n+\alpha}\right]. \tag{68}$$

Proof. The left-hand side is equal to

$$\begin{aligned}\binom{n}{j}\sum_{h=0}^{q}\frac{(j+h)!\,(n-j)!}{(n+h+1)!} &= \binom{n}{j}\sum_{h=0}^{q}\int_0^1 x^{j+h}(1-x)^{n-j}\,dx\\ &= \binom{n}{j}\int_0^1 (x^j - x^{j+q+1})(1-x)^{n-j-1}\,dx\\ &= \text{the right-hand side, after simplification.}\end{aligned}$$

Substituting in (67) from (68) when $q = 2s - 2 - k$, we obtain

$$g_k = (-1)^k\frac{1}{n-j}\binom{2s}{k}\prod_{\alpha=1}^{2s-1-k}\frac{j+\alpha}{n+\alpha} \qquad \text{for } k = 0, 1, 2, \cdots, 2s-2, \tag{69}$$

and substituting now in (66) we obtain

$$\begin{aligned}Q_j &= \frac{n}{j(n-j)}\left[c_j^{2s} + \sum_{k=0}^{2s-1}\binom{2s}{k}(-c_j)^k\prod_{\alpha=0}^{2s-1-k}\frac{j+\alpha}{n+\alpha}\right]\\ &= \frac{n}{j(n-j)}\left(c_j - \frac{j}{n}\right)^{2s*},\end{aligned} \tag{70}$$

using the notation introduced in (13).

Now with the same reasoning as in Section 3 it will be seen that Q_j and hence R is minimized by setting $\partial Q_j / \partial c_j = 0$ and solving for c_j the resulting equation

$$(c_j - j/n)^{(r-1)*} = 0. \tag{71}$$

This equation, by the same argument as in Section 3, has one and only one real root which lies between 0 and 1. Since for $j = 0$, (71) reduces to $c_0^{r-1} = 0$ giving $c_0 = 0$ as the only real root, and for $j = n$, it reduces to $(c_n - 1)^{r-1} = 0$, giving $c_n = 1$ as the only real root, it follows that we can say that the minimax invariant procedure for the loss function of this section is to estimate $F(x)$ by

$$\hat{F}(x) = c_j\,; \qquad X_j \leqq x < X_{j+1}\,, \qquad j = 0, 1, \cdots, n,$$

where X_j, $j = 0, 1, \cdots, n+1$, have been defined earlier and c_j is the real root of (71). Again the number of equations to be solved in practice will be about half the sample size since it can be easily seen that (71) remains unchanged by replacing j by $n - j$ and c_j by $1 - c_j$, so that $c_{n-j} = 1 - c_j$.

Special case for $r = 2$. When $r = 2$, the equation (71) reduces, for each j, to a linear equation

$$(c_j - j/n)^{1*} = 0,$$

which has the unique solution $c_j = j/n$. This can also be seen by using (35), (70), (62), and (63) for $r = 2$ and writing the risk R in the form

$$R = \frac{1}{n} + \sum_{j=1}^{n-1} \frac{n}{j(n-j)} \left(c_j - \frac{j}{n}\right)^2. \tag{72}$$

Thus the minimax invariant estimate $\hat{F}$ for the loss function in the special case for $r = 1$ in this section turns out to be the usual sample cumulative function

$$\hat{F}(x) = c_j = j/n, \quad \text{when} \quad X_j \leqq x < X_{j+1}, \qquad j = 0, 1, \cdots, n, \tag{73}$$

where $X_1 < X_2 < \cdots < X_n$ is an ordered sample from the c.d.f. F, X_0 and X_{n+1} standing for $-\infty$ and $+\infty$ respectively. The actual value of the risk corresponding to this estimate is $1/n$.

6. The loss function $L(F, \hat{F}) = \int_{-\infty}^{\infty} |F(x) - \hat{F}(x)| \,/\, F(x)[1 - F(x)]\, dF(x)$. In this case we obtain

$$R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} |x - c_j| \,/\, x(1-x)\, dx = \sum_{j=0}^{n} Q_j, \tag{74}$$

where

$$Q_j = E \int_{X_j}^{X_{j+1}} |x - c_j| \,/\, x(1-x)\, dx. \tag{75}$$

As in the last section, it will be seen that for finite risk the necessary and sufficient condition is that $c_0 = 0$ and $c_n = 1$. For $j \neq 0, n$, we obtain

$$\begin{aligned} Q_j = E[c_j\, &|\log c_j - \log X_j| - c_j\, |\log c_j - \log X_{j+1}| \\ &+ (1 - c_j)|\log(1 - c_j) - \log(1 - X_{j+1})| \\ &\qquad - (1 - c_j)|\log(1 - c_j) - \log(1 - X_j)|]. \end{aligned} \tag{76}$$

The distribution of X_j has probability density $p(y)$ given by (9) and the distribution of X_{j+1} has the probability density

$$q(y) = \frac{1}{B(j+1, n-j)}\, y^j (1 - y)^{n-j-1}, \qquad 0 \leqq y \leqq 1. \tag{77}$$

Using (9) and (77) we can express Q_j in the form

$$Q_j = \binom{n}{j} \left[\int_0^{c_j} g(c_j, y)\, dy - \int_{c_j}^{1} g(c_j, y)\, dy \right], \tag{78}$$

where

$$g(c_j, y) = [c_j \log c_j + (1 - c_j) \log (1 - c_j) - c_j \log y - (1 - c_j) \log (1 - y)]y^{j-1}(1 - y)^{n-j-1}(j - ny). \tag{79}$$

Straightforward integration leads to

$$\int g(c_j, y)\, dy = y^j(1 - y)^{n-j}[c_j(\log c_j - \log y) + (1 - c_j)(\log (1 - c_j) - \log (1 - y))] + \int (c_j - y)y^{j-1}(1 - y)^{n-j-1}\, dy + \text{constant}, \tag{80}$$

which enables us to obtain Q_j as

$$Q_j = \binom{n}{j}\left[\int_0^{c_j} (c_j - y)y^{j-1}(1 - y)^{n-j-1}\, dy - \int_{c_j}^1 (c_j - y)y^{j-1}(1 - y)^{n-j-1}\, dy\right], \tag{81}$$

for $j = 1, 2, \cdots, n - 1$. Since Q_0 and Q_n are fixed, and each Q_j is positive and depends only on j, minimizing R is equivalent to minimizing Q_j for each j. We see that

$$\frac{\partial Q_j}{\partial c_j} = \binom{n}{j}\int_0^{c_j} y^{j-1}(1 - y)^{n-j-1}\, dy - \binom{n}{j}\int_{c_j}^1 y^{j-1}(1 - y)^{n-j-1}\, dy, \tag{82}$$

$$\frac{\partial^2 Q_j}{\partial c_j^2} = 2\binom{n}{j}c_j^{j-1}(1 - c_j)^{n-j-1} \tag{83}$$

Setting $\partial Q_j / \partial c_j = 0$ and solving we obtain c_j as the median of the beta distribution with density

$$h(z) = \frac{1}{B(j, n - j)} z^{j-1}(1 - z)^{n-j-1}, \qquad 0 \leqq z \leqq 1, \tag{84}$$

for $j = 1, 2, \cdots, n - 1$. Since (83) shows that $\partial^2 Q_j / \partial c_j^2 > 0$ for $0 < c_j < 1$, it follows that this solution for c_j in fact minimizes Q_j and hence minimizes R. To summarize, the minimax invariant procedure for the loss function considered in this section is to estimate $F(x)$ by

$$\hat{F}(x) = c_j\, ; \qquad X_j \leqq x < X_{j+1}, \qquad j = 0, 1, \cdots, n, \tag{85}$$

where X_j, $j = 0, 1, \cdots, n + 1$, have been defined earlier, $c_0 = 0$, $c_n = 1$ and for $j = 1, 2, \cdots, n - 1$, c_j is the median of the beta distribution with density (84). Again it is interesting to note that the value j/n for c_j obtained in the last section for $r = 2$ is the mean of the same beta distribution.

Further it is obvious that $c_{n-j} = 1 - c_j$ and only about half the total number of c values are to be actually computed. These can be obtained with the help of the tables of the incomplete beta-function [2] as indicated in Section 4. However, if a table for c values like Table I has been constructed, no fresh computations are needed, since the value of c_j $(j = 1, 2, \cdots, n - 1)$ for any n in this case is equal to the value of c_{j-1} for $n - 2$ in Table I. For example, when $n = 10$, the values of c_j $(j = 0, 1, \cdots, 10)$ correct to two decimal places are

$$0, .07, .18, .28, .39, .50, .61, .72, .82, .93, 1. \tag{86}$$

I am thankful to Professors Z. W. Birnbaum and H. Rubin for some helpful discussions during the preparation of this paper.

REFERENCES

[1] Z. W. Birnbaum, "Distribution-free tests of fit for continuous distribution functions," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 1-8.

[2] Karl Pearson, *Tables of the Incomplete Beta-Function*, The Biometrika Office, University College, London.