Metrika, Volume 28, 1981, page 257–262. ©Physica-Verlag, Vienna.

Estimation Problems for Rectangular Distributions (Or the Taxi Problem Revisited)

By *J.S. Rao*, Santa Barbara[1])

Abstract: The problem of estimating the unknown upper bound θ on the basis of a sample of size n from a uniform or rectangular distribution on $[0, \theta]$ has considerable interest. This or the analogous discrete version is variously known as the "Taxi-problem" or the "German bomb (or Tank) problem" and has a long history. The emphasis here is on estimation of θ through the lengths of the observed gaps or spacings which seem natural for this problem.

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on $[0, \theta]$. Estimation of the unknown upper bound θ is if interest, for instance, in connection with estimating the total number of taxis in a town on the basis of observed registration numbers or in estimating the number of enemy bombs (or tanks) on the basis of observed serial numbers, providing of course, some obvious assumptions hold. See, for instance, *Noether* [1971, 2–5] for an elementary discussion. A continuous uniform distribution will be assumed here, which provides a good approximation to the results in the case of a discrete uniform on the integers $\{1, 2, \ldots, \theta\}$. In fact, analogous results may be obtained for the latter case.

When $(X_1, \ldots, X_n)$ is a random sample from $R(0, \theta)$, the rectangular (or uniform) distribution on $(0, \theta)$, the following results are known and stated for completeness. Let

$$0 \leqslant X_{1n} \leqslant X_{2n} \leqslant \ldots \leqslant X_{nn} \leqslant \theta \tag{1.1}$$

denote the order statistics. The sample maximum X_{nn} is a complete sufficient statistic and has the cumulative distribution function (cdf)

$$F_{X_{nn}}(x) = (x/\theta)^n, \qquad 0 < x < \theta. \tag{1.2}$$

From (1.2), it is seen that $E_\theta(X_{nn}) = (n/n+1)\theta$ and hence

$$T_n = \frac{n+1}{n} X_{nn} \tag{1.3}$$

[1]) *J.S. Rao*, Department of Mathematics, University of California, Santa Barbara, California 93 106, USA.

is unbiased for θ. Since this estimator is a function of the complete sufficient statistic, it follows from the Rao-Blackwell and Lehmann-Scheffé theorems that T_n is the (essentially) unique uniformly minimum variance unbiased estimate (umvue) of θ [see, for instance, *David*, p. 96].

Sample spacings or observed gaps come naturally into play in this problem since X_{nn} falls short of θ by an amount equal to the last gap. Now we introduce some basic facts about spacings. Spacings are defined to be the gaps between successive observations, i.e.

$$D_{in} = X_{in} - X_{i-1,n}, \qquad i = 1, 2, \ldots, n \tag{1.4}$$

where we put $X_{0n} \equiv 0$. Since n is held fixed in all subsequent discussions, we shall drop the second subscript n in X_{in}, D_{in}, etc. to simplify the notation. If one defines

$$U_i = X_i/\theta \text{ and } T_i = D_i/\theta, \qquad i = 1, \ldots, n \tag{1.5}$$

then $(U_1, \ldots, U_n)$ has the same distribution as a random sample from a $R(0, 1)$ distribution while $(T_1, \ldots, T_n)$ correspond to the "uniform spacings." These $\{T_i\}$ form an exchangeable set of random variables with a joint Dirichlet distribution. Recall that a k-dimensional random vector $(Y_1, \ldots, Y_k)$ has a Dirichlet distribution denoted by $D(r_1, \ldots, r_k; r_{k+1})$ if it has the joint density

$$f(y_1, \ldots, y_k) = \frac{\Gamma(r_1 + \ldots + r_{k+1})}{\prod_{i=1}^{k+1} \Gamma(r_i)} \left(\prod_{i=1}^{k} y_i^{r_i - 1}\right)(1 - y_1 - \ldots - y_k)^{r_{k+1} - 1} \tag{1.6}$$

over the simplex $S_k = \{\underset{\sim}{y}: y_i \geqslant 0, \sum_1^k y_i \leqslant 1\}$ in R^k. See *Wilks* [1962, 177–182] for an excellent discussion of the basic facts about this distribution. In particular, $(T_1, \ldots, T_n)$ has an n-variate Dirichlet $D(1, \ldots, 1; 1)$ with all the parameter values unity, i.e., with density

$$f(t_1, \ldots, t_n) = n! \tag{1.7}$$

over the simplex $S_n = \{\underset{\sim}{t}: t_i \geqslant 0, \sum_1^n t_i \leqslant 1\}$ in R^n [see, for instance, *David*, 79–80]. From (1.7) it follows that any T_i has a $D(1; n)$ or Beta $(1, n)$ distribution. From this and the fact D_i and T_i/θ have the same distribution, it can be verified that

$$\begin{aligned} &E(D_i) = \theta/(n+1) \\ &V(D_i) = \theta^2 n/(n+1)^2(n+2) \\ &\operatorname{Cov}(D_i, D_j) = -\theta^2/(n+1)^2(n+2) \qquad \text{for } i \neq j. \end{aligned} \tag{1.8}$$

It may be noted in passing that Dirichlet random variables have an additive property

namely that for $l \leqslant k$, $(\sum_{i=1}^{l} Y_i)$ has a $D(r_1 + \ldots + r_l; r_{l+1} + \ldots + r_{k+1})$ which is a Beta distribution. From this, the sampling distributions of the uniform order statistics $U_{rn} = \sum_{i=1}^{r} T_i$ and the sample range $U_{nn} - U_{1n} = \sum_{i}^{n} T_i$ can be written down immediately as the Beta $(r; n+1-r)$ and Beta $(n-1; 2)$ respectively.

2. Estimation of θ

Estimation of parameters through the use of a few or all of the order statistics has several advantages, principally their simplicity. See, for instance, *Mosteller* [1946] or *David* [1970, chapters 6 and 7]. They are especially useful in situations where trimming and censoring of the observations is part of the model and yield drastic reduction in labor over the optimal methods which can be sometimes laborious. We suggest here estimation through spacings, linear combinations in which are equivalent to linear functions of order statistics. For a discussion of linear estimation through order statistics, refer to *David* [1970, p. 102]. As pointed out earlier, the sample maximum falls short of θ by an amount equal to the last gap. Since the gaps are exchangeable, adding the length of any of the gaps or the average length of any set of gaps or merely multiplying any gap by $(n+1)$ yields unbiased estimators of θ. Thus, for $r = 1, \ldots, n$

$$\begin{aligned} T_{1r} &= X_{nn} + D_r = 2D_r + \sum_{i \neq r} D_i, \\ T_{2r} &= X_{nn} + \frac{1}{r}\sum_{i=1}^{r} D_i = \sum_{i=1}^{r} D_i \left(1 + \frac{1}{r}\right) + \sum_{r+1}^{n} D_i \end{aligned} \tag{2.1}$$

and

$$T_{3r} = (n+1)D_r$$

are all unbiased estimators of θ. From (1.8), one can verify

$$\begin{aligned} \mathrm{Var}(T_{1r}) &= 4\theta^2 n/(n+1)^2(n+2) \\ \mathrm{Var}(T_{2r}) &= \theta^2 \left(1 + \frac{1}{r}\right)/(n+1)(n+2) \\ \mathrm{Var}(T_{3r}) &= \theta^2 n/(n+2). \end{aligned} \tag{2.2}$$

Because of symmetry, the variance expressions for $\{T_{1r}\}$ and $\{T_{3r}\}$ do not depend on the specific D_r that is used while the $V(T_{2r})$ decreases with r and is a minimum for $r = n$ for which

$$T_{2n} = X_{nn} + \frac{X_{nn}}{n} = \left(1 + \frac{1}{n}\right) X_{nn} = T_n \tag{2.3}$$

defined in (1.3). Also recall that if $\bar{X}_n = \sum_1^n X_i/n$ denotes the sample mean, then

$$2\bar{X}_n = 2 \sum_{i=1}^{n} \left(\frac{n-i+1}{n}\right) D_i \tag{2.4}$$

provides yet another unbiased estimator of θ. Thus one may consider a linear combination of spacings

$$W_n = \sum_{i=1}^{n} b_{in} D_i \tag{2.5}$$

to estimate θ. Unbiasedness of W_n *implies the condition*

$$\sum_1^n b_{in} = (n+1) \tag{2.6}$$

which is, of course, the case with all the estimators in (2.1) and (2.4). It is now natural to ask for the best linear unbiased estimate of θ from among the class (2.5). An elementary calculation using (1.8) shows that the variance of W_n is minimized subject to (2.6) when $b_{in} = ((n + 1/n)$ for $i = 1, \ldots, n$, with the resulting estimator (2.3). The equal weights are to be expected on all the spacings from symmetry considerations. Since (2.3) is the umvue and is also of the from (2.5), it is no surprise that it is the best linear unbiased estimate. Indeed, equation (1.8) shows that the vector $D = (D_1, \ldots, D_n)'$ follows a linear model with expectation $(n+1)^{-1}\theta 1_n^{\sim}$ where 1_n is the column vector with all ones, and covariance matrix $Q = [(n+1) I_n - 1_n 1_n'] (\theta^2/(n+1)^2(n+2))$. Using the fact that the inverse of $[(n+1) I_n - 1_n 1_n']$ is $(n+1)^{-1}[I_n + 1_n 1_n']$, the formal Gauss-Markov least squares estimator (in its slightly generalized version since the covariance matrix is not diagonal) is given by

$$\hat{\theta} = [(n+1)^{-2} 1_n' Q^{-1} 1_n]^{-1} [(n+1)^{-1} 1_n' QD] \tag{2.7}$$

$$= \left[\frac{(n+1)^2}{n}\right] \left[\frac{1}{n+1} \sum_1^n D_i\right] = \left(\frac{n+1}{n}\right) X_{nn}$$

which is again the statistic in (2.3) with equal weights $b_{in} = (n+1)/n$.

Alternately one can approach the problem of estimating θ with the goal of minimizing the mean square error (MSE) where $\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ and relax the condition (2.6) that the estimator $\hat{\theta}$ be unbiased. If we use equal weights, say b, on all $\{D_i\}$, then the problem is to find the weight b for which the estimator $\sum_{i=1}^n bD_i = bX_{nn}$ has the smallest MSE. It is easy to verify that

$$\text{MSE}(bX_{nn}) = E(bX_{nn} - \theta)^2 = \theta^2 \left[\frac{b^2 n}{(n+2)} - \frac{2bn}{(n+1)} + 1\right] \tag{2.8}$$

which is minimized when $b = (n + 2)/(n + 1)$. Thus the estimator

$$T_{4n} = \left(\frac{n+2}{n+1}\right) X_{nn} \tag{2.9}$$

has the smallest MSE. It is interesting to compare this with the other competitors namely the umvue T_{2n} in (2.3) and the maximum lieklihood estimator X_{nn}. Taking b to be $((n + 2)/(n + 1))$, $((n + 1)/n)$ and 1 respectively in (2.8), we get

$$\begin{aligned} \mathrm{MSE}(T_{4n}) &= \theta^2/(n+1)^2 \\ \mathrm{MSE}(T_{2n}) &= \theta^2/n(n+2) \\ \mathrm{MSE}(X_{nn}) &= 2\theta^2/(n+1)(n+2) \end{aligned} \tag{2.10}$$

from which it follows that with respect to the MSE criterion, T_{4n} given in equation (2.9) is uniformly better than the umvue T_{2n} given in (2.3) which in turn is uniformly better than the maximum likelihood estimator X_{nn}. This incidentally is another instance of a situation where the umvue is not admissible under the quadratic loss function.

Another interesting way to improve the estimators given in (2.1) with respect to their MSE's is given by the following procedure: Since the coefficient of variation v (i.e., $\mathrm{Var}(\hat{\theta})/\theta^2$) is independent of θ (cf. equation (2.2)), $\hat{\theta}^* = (1 + v)^{-1}\hat{\theta}$ yields another estimator of θ with

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}^*) &= \mathrm{Var}(\hat{\theta}^*) + [\mathrm{Bias}(\hat{\theta}^*)]^2 \\ &= \theta^2\left[\frac{v}{(1+v)^2} + \frac{v^2}{(1+v)^2}\right] = \theta^2\left(\frac{v}{1+v}\right) \end{aligned}$$

which is uniformly smaller than the MSE of the original estimator $\hat{\theta}$. Thus each of the unbiased estimators in (2.1) may be improved with respect to the MSE. This yields the estimators

$$\begin{aligned} T_{1r}^* &= \frac{(n+1)^2(n+2)}{(n+1)^2(n+2)+4n}(X_{nn} + D_r) \\ T_{2r}^* &= \frac{r(n+1)(n+2)}{r(n+1)(n+2)+(r+1)} T_{2r} \end{aligned} \tag{2.11}$$

and

$$T_{3r}^* = \frac{n+2}{2} D_r$$

which have smaller MSE's than the corresponding unbiased estimators given in (2.1). While the MSE of T_{1r}^* and T_{3r}^* does not depend on r, the MSE of T_{2r}^* does depend on r and is a minimum for $r = n$. It is very interesting to note that the resulting T_{2n}^* is indeed $((n + 2)/(n + 1))\, X_{nn}$, the estimator with minimum MSE that we obtained in (2.9).

But the real advantage of using spacings in estimation of θ comes in situations of censoring where some of the order statistics at either end or in the middle are missing. The best linear unbiased estimate based on the spacings would then be to put equal weights on the available or observed gaps. In particular, if the sample is censored so that one observes only the m-th largest order statistic X_{mn} (for $m \leqslant n$), then the following are all unbiased estimators of θ

$$T_{1r}^* = X_{mn} + ((n+1)-m)D_r$$

$$T_{2r}^* = X_{mn} + \frac{n+1-m}{r} \sum_{i=1}^{r} D_i \tag{2.12}$$

$$T_{3r}^* = (n+1)D_r$$

for $r = 1, \ldots, m$. By an analysis similar to that used before, it may be shown, that the best linear unbiased estimate of θ is to take

$$T_{2m}^* = \sum_{i=1}^{m} \left(\frac{n+1}{m}\right) D_i = \frac{n+1}{m} X_{mn} \tag{2.13}$$

with variance

$$V(T_{2m}^*) = (n-m+1)\theta^2/(n+2)m. \tag{2.14}$$

Thus spacings seem to be the natural quantities to consider in the estimation of θ. *Sarhan/Greenberg* [1959] discuss the problem of censoring at both ends in rectangular populations using order statistics. This alternate approach based on spacings yields the same results, more effortlessly.

The author is very grateful to the referee for his many helpful comments and suggestions.

References

David, H.A.: Order Statistics. New York 1970.
Mosteller, F.: On some useful "inefficient" statistics. Ann. Math. Statist. 17, 1946, 377–408.
Noether, G.: Introduction to statistics – a fresh approach. Boston 1971.
Sarhan, A.E., and *G.B. Greenberg*: Estimation of location and scale parameters for the rectangular populations from censored samples. J.R. Statist. Soc. **B21**, 1959, 356–363.
Wilks, S.S.: Mathematical Statistics. New York 1962.

Received April 30, 1979
(revised version December 1979)