Analysis of Censored Data
IMS Lecture Notes – Monograph Series (1995) Volume 27

Ranking and Subset Selection Procedures for Populations With Censored Data: A Review

Harshinder Singh and Amar Nath Gill
Department of Statistics
Panjab University
Chandigarh-160014 (INDIA)

In this paper, we review the work carried out on the problems of ranking and subset selection procedures for populations with censored data. Such problems arise in various situations such as industrial life testing, medical studies and biological experiments. Some of the specific situations are described as under:

1. Components of k types are available and interest is to select the best component (e.g. the component with largest mean life time) or we want to select the t best components $1 < t < k - 1$. The data is censored and could be Type-I, Type-II or random.

2. Possible k therapies of a disease (e.g. cancer, AIDS) are available. The problem is to select the best therapy when the data on survival times of patients are available. Similarly, the selection of best drug out of k drugs available for dissolving kidny stone, when the data of time to dissolve the stone in patients are censored, is a problem of selection under censoring.

3. There are k different locations in a river with different levels of potential heavy metal contamination at different locations. The interest is to select a site having least contamination based on the data of certain heavy metal concentrations in the muscle of fish caught from these locations. It is not possible to measure the concentration below certain level. In this case the data is Type-I left censored.

In Section 1, we discuss selection procedures for selecting exponential population with largest location parameter when the data is either Type-I or Type II censored. Section 2 contains procedure for selecting exponential population with largest scale parameter in the presence of Type-II censoring and in Section 3, in the presence of random censoring. In Section 4, we briefly discuss other parametric selection procedures in the presence of censoring. Selection Designs for Pilot Studies based on survival, assuming Cox Proportional Hazard Model, are discussed in Section 5. Section 6 contains a discussion on nonparametric subset selection procedure for selecting a subset containing the distribution associated with largest location parameter in the presence of Type-I censoring.

1. Subset selection Procedure for Selecting Exponential Population With Largest Location Parameter Based on Type-I and Type-II Censoring

Following the subset selection approach of Gupta (1956, 1965), Berger and Kim (1985) proposed subset selection procedure for selecting a subset containing the population associated with the largest location parameter based on Type-I censored data from k exponential populations. Let $\Pi_1, \ldots, \Pi_k$ denote $k \geq 2$ independent exponential populations with density functions

$$f(x, \lambda_i) = \frac{1}{\theta} e^{\left[\frac{-(x-\lambda_i)}{\theta}\right]}, x \geq \lambda_i, (1)$$

where θ is assumed to be known; i=1, ...,k. Let $\lambda_{[1]} \leq \cdots \leq \lambda_{[k]}$ denote the ordered values of $\lambda_1, \ldots, \lambda_k$ and let $\Pi_{(i)}$ denote the unknown population associated with $\lambda_{[i]}$. Goal is to select a nonempty subset of the k populations containing $\Pi_{(k)}$. A correct selection (CS), is the selection of any subset which contains $\Pi_{(k)}$. Let X_{ij}, j=1, ...,n be a random sample of size n from Π_i i=1, ...,k which is subject to Type-I censoring at time T. Observable data are $X_{ij}^* = min(X_{ij}, T)$. Let
$Y_i = min(X_{i1}^*, \ldots, X_{in}^*) = min(min(X_{i1}, \ldots, X_{in}), T)$.
It is seen that the set $Y_1, \ldots, Y_k$ is sufficient for the problem. The subset selection procedure say, R_1, proposed by them is:

R_1 : Select Π_i in the subset if
$Y_i \geq max_{1 \leq j \leq k} Y_j - \frac{d\theta}{n}$,
where $d = d(k, P^*, \theta, T, n) \geq 0$ is chosen so that

$$inf_{\underline{\lambda}} P_{\underline{\lambda}}(CS|R_1) \geq P^*$$

and $P^* \in (\frac{1}{k}, 1)$ is specified by the experimenter in advance.

It is seen that $inf_{\underline{\lambda}} P_{\underline{\lambda}}$ (CS $|R_1$) is attained when $\lambda_{[1]} = \cdots = \lambda_{[k]}$ and the expression for inf equated to P^* gives the solution of d. The authors have tabulated the values of d for various choices of k, P^* and $\frac{nT}{\theta}$. The procedure has been shown to be monotonic, i.e., if $\lambda_i \leq \lambda_j$ then $P[Select\ \Pi_i|R_1] \leq P[Select\ \Pi_j|R_1]$. Moreover, if T $\leq \lambda_{[k]}$ or $\lambda_{[k-1]} < \lambda_k < T$,

$$\lim_{n\to\infty} P[CS|R_1] = 1$$

and for $\lambda_{[k-1]} < min\left[\lambda_{[k]}, T\right]$,

$$\lim_{n\to\infty} E[S|R_1] = 1,$$

where S is the size of the selected subset.

Ofusu (1974) discussed the above problem under Type-II censored data. In sample of size n from each population, observable from population Π_i are the times of first r_i failures with $X_{i[1]} \leq, \ldots, \leq X_{i[r_i]}$ as the corresponding ordered times. The set $X_{1[1]}, \ldots, X_{k[1]}$ is sufficient for the problem and the procedure say, R_2, suggested is

R_2 : select Π_i in the subset if
$X_{i[1]} \geq max_{1\leq j\leq k}\ X_{j[1]} - \frac{d\theta}{n}$,
where $d = d(k, P^*) \geq 0$ is a solution of the equation

$$\frac{1}{k}\left[e^d\left(1-e^{-d}\right)^k\right] = P^*.$$

The procedure R_2 satisfies the following properties:
(i) it is monotonic (ii) Sup $E_{\underline{\lambda}}(S|R_2) = kP^*$
(iii) If $\lambda_{[k-1]} < \lambda_{[k]}$ then

$$lim_{n\to\infty} P_{\underline{\lambda}}[CS|R_2] = lim_{n\to\infty} E_{\underline{\lambda}}(S|R_2) = 1.$$

2. Selection Procedures for selecting Largest Scale Parameter of Exponential Population With Type-II Censored Data

For this problem the indifference and subset selection formulations respectively are due to

(I) Berger and Kim (1985): The selection procedure for selecting exponential populations in terms of scale parameters based on Type-II censored data under indifference zone approach is as follows:

Let $\Pi_1, \ldots, \Pi_k$ be k ($k \geq 2$) exponential populations such that the pdf of population Π_i is

$$f(x, \theta_i) = \frac{1}{\theta_i} exp(\frac{-x}{\theta_i}), (2.1)$$

where $x > 0$ and $i = 1, \ldots, k$.

Let $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$ denote the ordered values of $\theta_1, \ldots, \theta_k$. The unknown population associated with $\theta_{[i]}$ is denoted by $\Pi_{(i)}$. Under Type-II censoring, the first r failures from each population are observed. Let these ordered failures from the population Π_i be denoted by $X_{i[1]} \leq \ldots \leq X_{i[r]}$, i=1,...,k. The maximum likelihood estimator of θ_i is given by $\hat{\theta}_i = \frac{Z_i}{r}$, where $Z_i = \sum_{j=1}^{r} X_{i[j]} + (n-r)X_{i[r]}$, i.e, the total time spent on observing the n items from population Π_i undergoing life testing until rth failure. Note that $\hat{\theta}_i$ is complete and sufficent for θ_i and $\frac{2r\hat{\theta}_i}{\theta_i}$ is distributed as $\chi^2_{(2r)}$. The authors cosidered the goal to select t, $1 \leq t \leq$ k-1, populations associated with t largest scale parameters $\theta_{[j]}$, j= k-t+1, ..., k. The selection of the subset containing exactly the t largest scale parameters will be called the correct selection (CS). The preference zone $\underline{\theta}(\delta)$ is defined by

$$\underline{\theta}(\delta) = \left\{ \underline{\theta} : \frac{\theta_{[k-t]}}{\theta_{[k-t+1]}} < \delta \right\}$$

where $\delta (0 < \delta < 1)$ is fixed by the experimenter. Let the ordered values of Z_i be denoted by $Z_{[1]} \leq \cdots \leq Z_{[k]}$. The procedure, say R_3 , defined by them is

R_3 : Select the t populations associated with $Z_{[k-t+1]}, \ldots, Z_{[k]}$.

Having specified the values of δ and $P^* \left(\frac{1}{\binom{k}{t}} < P^* < 1 \right)$ the experimenter is interested in determining the smallest r to satisfy the requriment

$$inf_{\underline{\theta} \in \underline{\theta}(\delta)} P_{\underline{\theta}} [CS | R_3] = P^*$$

which reduces to

$$(k-1) \int_0^\infty F_r^{k-t-1}(u)[1 - F_r(\delta u)]^t dF_r(u) = P^*, \quad (1)$$

where

$$F_r(u) = \frac{1}{\Gamma(r)} \int_0^u x^{r-1} e^{-x} dx.$$

Tables of Gibbons, Olkin and Sobel (1977) can be used to determine the minimum r satisfying (1). Moreover, it is seen that:

(i) for fixed r, $1 \leq r \leq n$, $\lim_{\delta \to 1} \inf_{\underline{\theta} \in \underline{\theta}(\delta)} P_{\underline{\theta}} [CS | R_3] = 1$.

(ii) let n $\rightarrow \infty$ in such a way that $\frac{r}{n} \rightarrow q \in (0,1)$ then for any $\underline{\theta}\varepsilon\underline{\theta}(\delta)$

$$\lim_{n\rightarrow\infty} P_{\underline{\theta}}(CS|R_3) = 1$$

(II) Huang and Huang (1980): Using the same notations as in Section 2(I) the subset selection procedure proposed, by the authors to select a subset of k exponential populations containing the one associated with $\theta_{[k]}$, is as follows

R_4 : SelectΠ_i in the subset if
$Z_i \geq$ c $(max_{1\leq j\leq k} Z_j)$,
where c $(0 < c < 1)$ is a positive constant satisfying the probability requirment (also called the P^* condition)

$$inf P(CS|R_4) = P^*. \,(2.2)$$

Here CS means that the selected subset contains the best population. It is seen that (2.2) reduces to

$$\int_0^\infty F_r^{k-1}(\frac{u}{c})dF_r(u) = P^*.$$

The Tables of Gupta (1963) can be used to find c for various values of k, r and P^*.

3. Subset Selection Procedure for Exponential Populations Under Random Censoring : Scale Parameters Case

Let the Population Π_i has the pdf (2.1), i=1, ...,k. The life times of n items from population Π_i represented by the random variables X_{ij}, are subject to random censoring by iid random variables Y_{ij} with exponential distribution G(y) $= 1 - e^{\frac{-y}{\alpha}}, y > 0$, j=1, ...,n; i=1, ...,k. X_{ij} and Y_{ij} are assumed to be independent . The observable random variables are $Z_{ij} = min(X_{ij}, Y_{ij})$.Let

$$\delta_{ij} = \begin{cases} 1 & \text{if } X_{ij} \leq Yij \\ 0 & \text{otherwise.} \end{cases}$$

The data at hand are the independent pairs of random variables (Z_{ij} , δ_{ij}). Let

$$T_i = \begin{cases} \frac{1}{D_i}\sum_{j=1}^n Z_{ij} & \text{if } D_i > 0 \\ \infty & \text{if } D_i = 0, \end{cases}$$

where $D_i = \sum_{j=1}^{n} \delta_{ij}$.Note that $\sum_{j=1}^{n} Z_{ij}$ and D_i are jointly sufficient for θ_i and T_i is MLE of θ_i.

Kim (1988) proposed the following subset selection procedure, say R_5, to select a subset containing the population associated with the largest scale parameter $\theta_{[k]}$

R_5 : Select Π_i in the subset if

$T_i \geq c\ (max_{1 \leq j \leq k} T_j)$,

where c = $c(k, n, P^*) \in (0, 1)$ and $P^* \in (\frac{1}{k}, 1)$ are prespecified constants such that

$inf_{\underline{\theta}}\ P_{\underline{\theta}}(CS|R_5) \geq P^*$.

It is interesting to note that the above infimum of probability of correct selection is independent of mean life length α of the censoring mechanism.

Unlike usual procedures, here P^* can not be arbitrarily prespecified independent of k. There is dependency between k and P^*. If P^* is chosen too high, the rule R_5 may not be well defined in the sense that the constant c satisfying the P^*-condition may not exist. Rule R_5 is well defiend if P^* is specified such that $P^* < 1 - k^{-\frac{1}{k-1}} + k^{-\frac{k}{k-1}}$. Upper bound of P^* diminishes as k $\rightarrow \infty$. So if k is large P^* has to be fixed quite small. This can be prevented, of course, by increasing the sample size n. This procedure may not be most desirable when k is large.

The following Theorem gives the limiting behaviour of T_i when $\frac{\alpha}{\theta_i} \rightarrow \infty$.

Theorem : $T_i \rightarrow X_i$ in distribution as $\frac{\alpha}{\theta_i} \rightarrow \infty$, where X_i is a gamma variate with shape parameter n and scale parameter $\frac{\theta_i}{n}$.

The Tables of Gupta (1963), for selecting gamma populations with largest scale parameter, can be used to find c if $\frac{\alpha}{\theta_i}$ is large for all i=1, ...,k. Results are reliable if $\frac{\alpha}{\theta_i}$ is at least 50 for i=1, ...,k. The monotonicity property of procedure R_5 was also established.

4. Other Parametric selection Procedures in the Presence of Censoring

Kingston and Patel (1980) have discussed procedures for selecting the best two parameter Weibull population in the presence of type-II censored data. The ranking of the populations is done by comparing their reliabilities at certain fixed time or by comparing their α-th quantiles. The procedures are based on MLEs and simplified linear estimators of the unknown parameters. In selected cases the selection constants, needed for the implimentation of the procedures, are tabulated using Monte Carlo methods.

Gupta and Liang (1993) have discussed a Bayes selection procedure for selecting the exponential population associated with the largest mean life among several exponential populations based on Type-I censored data. They proved the monotonicity of the Bayes selection procedure and an early selection procedure was also proposed.

Abughalous and Bansal (1994) have studied the selection rules of the population with the smallest scale (hazard rate) parameter using minimaxity criterion and Bayes risk principle under the '0 - 1' loss function and a regret loss function from any one of the gamma, Weibull and exponential populations with Type-I censored data.

Tseng and Wu (1990) have proposed locally optimal selection rule, under Type-II censoring, to select Weibull populations that are more reliable than a standard when the shape parameters are known. A modification of the said selection rule has also been proposed when the unknown shape parameters have some prior distributions. Chang et. al. (1992) developed a selection rule for selecting most reliable design for an acclerated life test using the Arhenius Model for exponential population when the data is Type II censored.

5. Selection Designs for Pilot Studies Based on Survival

In cancer clinical trials new regimens are typically tested for antitumour activities in patients with advanced disease. The promising ones are then compared to the standard treatment in a randomized study, sometimes performed on patients with earlier stage disease. When there are multiple promising regiments, it may not be possible to compare all of them to the control group because of the prohibitive sample size and study length requirements. Liu et. al. (1993) proposed a design, assuming Cox regression model, to select a best treatment based on survival before the randomised comparison. They presented sample sizes required to attain the probability of correct selection of 0.90 using asymptotic results for Weibull survival distributions with parameters in a range of practical importance. Through simulation results they concluded that: (i) the asymptotic approximations to the correct selection probabilities are quite satisfactory and (ii) the procedure is reasonably robust to the proportional hazard assumption.

6. A Nonparametric Subset Selection Procedure for Selecting Largest Location Parameter in the Presence of Type-I Censoring

Singh and Gill (1994) have developed a subset selection procedure under Type-I censoring as explained below :

Let $\Pi_1, \ldots, \Pi_k$ be k ($k \geq 2$) independent populations and $F_i(x) = F(x - \theta_i)$ be the cdf of Π_i, where F is an absolutely continuous cdf. Let $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$ be the ordered values of $\theta_1, \ldots, \theta_k$. Suppose we have random samples of sizes $n_1, \ldots, n_k$ from the k populations which are subject to Type-I censoring on the right at time T. Let $\Pi_{(i)}$ be the population associated with parameter $\theta_{[i]}$, i=1, ...,k. Let $n_{[1]} \leq \cdots \leq n_{[k]}$ be the ordered values of $n_1, \ldots, n_k$. Uncensored observations from all the k samples are combined and ranked. Let

X_i = Number of uncensored observations (i.e.$\leq T$) in a sample of size n_i from population Π_i. Note that X_i has binomial distribution with parameters n_i and $F_i(T)$

S_i = Sum of the ranks of uncensored observations from population Π_i in the combined arrangement,

$M = \sum_{i=1}^{k} X_i$

$N = \sum_{i=1}^{k} n_i$,

$$T_i = \begin{cases} \frac{S_i}{X_i} & \text{if } X_i \geq 1 \\ \infty & \text{if } X_i = 0. \end{cases}$$

Intuitivly higher values of T_i are indicative of higher location parameter θ_i if T is reasonably large and sample sizes $n_1, \ldots, n_k$ are also large. Let $T_{(i)}$ be the value among T_js associated with sample coming from population $\Pi_{(i)}$ with location parameter $\theta_{[i]}$. Let $T_{[1]} \leq \ldots \leq T_{[k]}$ be the ordered values of $T_1, \ldots, T_k$. Their subset selection procedure is

R : Select Π_i in the subset if

$T_i \geq T_{[k]} - d\sqrt{\frac{MN}{12}\left(\frac{1}{n_i} + \frac{1}{n_{[1]}}\right)}$

Let

$Z_j = \frac{T_{(j)} - T_{(k)}}{\sqrt{\frac{NM}{12}\left(\frac{1}{n_{(j)}} + \frac{1}{n_{(k)}}\right)}}\; j = 1, \ldots, k-1.$

The following theorem follows from Atkinson and Mount (1994).

Theorem 6.1: Under $\theta_1 = \cdots = \theta_k$, as $n_i \to \infty$ such that $\frac{n_i}{N} \to \lambda_i, 0 < \lambda_i < 1$,the joint limiting distribution of $Z_1, \ldots, Z_{k-1}$ is (k-1) variate normal distribution with mean $\underline{0}$ and correlations

$$\rho_{ij} = \left[1 + \frac{\lambda_{(k)}}{\lambda_{(i)}}\right]^{\frac{-1}{2}} \left[1 + \frac{\lambda_{(k)}}{\lambda_{(j)}}\right]^{\frac{-1}{2}}.$$

$i, j = 1, \ldots, k-1$; $i \neq j$.

The probability of correct selection is given by

$$P\left[T_{(k)} \geq T_{[k]} - d\sqrt{\frac{MN}{12\left(\frac{1}{n_{(k)}} + \frac{1}{n_{[1]}}\right)}}\right] =$$

$$P\left[Z_j \leq d\sqrt{\frac{\frac{1}{n_{(k)}} + \frac{1}{n_{[1]}}}{\frac{1}{n_{(j)}} + \frac{1}{n_{(k)}}}}, j = 1, \ldots, k-1\right].$$

Let

$$\rho = min_{ij}\ \rho_{ij} = \left[1 + \frac{\lambda_{[k]}}{\lambda_{[1]}}\right]^{\frac{-1}{2}} \left[1 + \frac{\lambda_{[k]}}{\lambda_{[2]}}\right]^{\frac{-1}{2}}.$$

Also

$$min_j\sqrt{\frac{\frac{1}{n_{(k)}} + \frac{1}{n_{[1]}}}{\frac{1}{n_{(j)}} + \frac{1}{n_{(k)}}}} = 1$$

Thus by using Theorem 6.1 and Slepian Theorem (refer Slepian 1962), we have, under $\theta_1 = \cdots = \theta_k$ and for the limiting case (i.e $n_i \to \infty$) probability of correct selection, $P_0[CS]$, satisfies

$$P_0[CS] \geq P[V_i \leq d, i = 1, \ldots, k-1],$$

where $\underline{V}$ has multivariate normal distribution with mean vector $\underline{0}$ and equal correlations ρ.

For a given value of P^*, Tables of Gupta, Nagel and Panchapakesan (1973) can be used to read the value of d, so that $P_0[CS] \geq P^*$.

The procedure suffers from the drawback that in general $P_0[CS]$ may not be the infimum of probability of correct selection, $P[CS]$, (refer Rizvi and Woodworth 1970).

For k=3, the authors carried out simulations from Exponential, Weibull and Gamma populations with different sample sizes and different possible location parameter configurations. In simulation the selection constants were taken from the Tables of Gupta et. al. (1973) for some specific values of P^*. The simulated values of $P[CS]$ indicated that inf $P[CS]$ is attained at $\theta_1 = \theta_2 = \theta_3$ and that the asymptotic normal approximations to the correct selection probabilities are quite satisfactory. It is conjectured that for these life distributions, commonly arrising in Life Testing Models, inf $P[CS]$ is attained at the equality of location parameters for general k.

As pointed earlier, in general $P_0[CS|R]$ may not be the infimum value of the probability of correct selection. In order to get rid of this limitation the authors are attempting to develop a subset selection procedure based on pairwise ranking of the censored observations. There is lot of scope for further research in the area of nonparametric selection procedures in the presence of censoring. No doubt the problems are challenging as the methodology becomes complicated.

References

[1] Abughalous, M.M. and Bansal, N.K. (1994). On the problem of selecting the best population in life testing models. Commun. Statist. Theory and Methods, 23(5), 1471-1481.

[2] Atkinson, G.F. and Mount, K. (1994). A nonparametric test for Type-I censored data: r samples. The Canadian Jl. of Statistics, 22(1), 149-162.

[3] Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist., 25, 16-39.

[4] Berger, R.E.,and Kim, J.S. (1985). Ranking and subset selection procedures for exponential populations with Type-I and Type-II censored data. The Frontiers of Medern Statistical Inference Procedures. American Science Press, Inc.

[5] Chang, D.S., Huang, D.Y. and Tseng, S.T. (1992) Selecting the most reliable design under Type II censored acclerated testing. IEEE Transactions on Reliablity, 41(4), 588-592.

[6] Gibbons. J.D., Olkin, I. and Sobel, M. (1977). Selecting and ordering populations: A new statistical methodology. John Wiley, New York.

[7] Gupta, S.S. (1956). On a decision rule for a problem of ranking means. Memograph Series No. 150, Institute of Statistics, University of North carolina, Chapel Hill, North Carolina.

[8] Gupta, S.S. (1963). On selection and ranking procedure for gamma populations. Ann. Inst. Statist. Math., 14, 199-126.

[9] Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. Technometrics, 7, 225-245.

[10] Gupta, S.S. and Liang, T. (1993). Selecting the best exponential population based on Type-I censored data: A Bayesian approach. Technical Report No. 87-4, Dept. of Statistics Purdue University.

[11] Gupta, S.S., Nagel, K. and Panchapakesan, S. (1973). On the order statistics from equally correlated normal random variables. Biometrika, 60, 403-413.

[12] Huang, W.T. and Huang, K.C. (1980). Subset selections of exponential populations based on censored data. Proceedings of conference on Recent Development in Statistical Methods and Applications. Academia Sinica, Taipei, Taiwan.

[13] Kim, J.S. (1988). A subset selection procedure for exponential populations under random censoring. Commun. Statist. Theory & Methods, 17(1), 183-206.

[14] Kingston, J.V. and Patel, J.K. (1980). Selecting the best one of several Weibull populations. Commun. Statist. Theory & Methods, 9(4), 383-398.

[15] Liu, P.Y.,Duhlberg, S. and Crowly, J.(1993). Selection designs for pilot studies based on survival. Biometrics, 49, 391-398.

[16] Ofusu. J.B. (1974). On selection procedures for exponential distributions. Bull. Math. Statist., 16, 1-9.

[17] Rizvi, M.H., and Woodworth, G.G. (1970). On selection procedures based on ranks: counterexamples concerning the least favourable configurations. Ann. Math. Statist., 41, 1942-1952.

[18] Singh, H. and Gill, A.N. (1994). Some selection procedures for Type-I censored data based on ranks. Communicated.

[19] Slepian, D. (1962). The one sided barrier problem for Gaussian noise. Bell system Tech. Jl., 41, 463-501.

[20] Tseng, S.T. and Wu, H.J. (1990). Selecting, under Type-II censoring, Weibull populations that are more reliable. IEEE Transactions on Reliability, 39(2), 193-198.