

Subset Selection Based on Order Statistics from Logistic Populations

Mark J. van der Laan¹, Paul van der Laan²

Division of Biostatistics¹

University of California, Berkeley

Department of Mathematics and Computing Science²

Eindhoven University of Technology

The Netherlands

March 20, 2001

Abstract

Consider k equal size treatment groups and let the outcome of interest be a survival time. Suppose that a known monotone transformation of the survival times is logistically distributed and that the treatment only affects the location parameter. We obtain exact results for the problem of selecting a subset of treatments, based on the i -th ordered survival times, which contains with specified probability P^* the optimal treatment.

Since the distribution of the median of a small logistic i.i.d. sample is shown to approximate very well the normal distribution our exact results for the median can also be used for subset selection based on estimates of location parameters which are normally or approximately normally distributed.

AMS classification: Primary 62F07; secondary 62E15.

Key Words and Phrases: Survival Analysis, subset Selection, order statistic, probability of correct selection, generalized logistic distributions, log-logistic distribution.

1 Introduction and motivation.

Subset selection has been introduced by Gupta (1965). Many contributions to this field of statistical interest have been given. Given are k (≥ 2) random variables X_1, \dots, X_k , which

may be sample means, sample quantiles or other estimates of location parameters, associated with k populations π_1, \dots, π_k , respectively. We assume that the distributions of these random variables differ only in their location parameter. If the distribution of the estimates of the location parameters have different scale parameters, then it makes sense to standardize the estimates with their scale parameter (if known) or some upper bound of their scale parameter. We are interested in choosing the best population, that is the population with the largest value of the location parameter. If there are more than one contenders for the highest rank, we suppose that one of these is appropriately tagged. Subset selection has as its goal to indicate a subset of the collection of k populations which includes the best population with a given confidence and with the requirement that the size of the subset is as small as possible. Gupta's subset selection rule is defined by:

$$\text{Select } \pi_i \text{ if and only if } x_i \geq \max_{1 \leq j \leq k} x_j - d ,$$

where x_i is the observed value of X_i ($i = 1, \dots, k$). The selection constant d (≥ 0) has to be chosen such that the probability is at least P^* ($k^{-1} < P^* < 1$) that the subset contains the best population. A correct selection CS is defined as a selection of any subset which includes the best population.

The selection constant and the probability of CS depend on the form of the underlying distribution. The probability of CS is equal to

$$P(CS) = P\left(X_{(k)} \geq \max_{1 \leq j \leq k} X_j - d\right),$$

where $X_{(k)}$ is the unknown random variable associated with $\theta_{[k]}$ and where the ranked location parameters $\theta_1, \dots, \theta_k$ are denoted by $\theta_{[1]} \leq \dots \leq \theta_{[k]}$. Let Ω be the space of all parameters configurations for $\theta_{[1]} \leq \dots \leq \theta_k$. From Gupta (1965) we have

$$\inf P(CS) = \int_{-\infty}^{\infty} F^{k-1}(x+d)dF(x), \tag{1}$$

which is attained for the least favourable configuration (LFC): $\theta_{[1]} = \theta_{[k]} = \theta$, and where $F(\cdot)$ is the cumulative distribution function (cdf) of the populations under the LFC: since (1) does not depend on θ we will standardize F to have location parameter zero. To be sure that $P(CS) \geq P^*$ for all configurations of $\theta_1, \dots, \theta_k$ the smallest value of the selection constant d has to be chosen for which

$$\int_{-\infty}^{\infty} F^{k-1}(x+d)dF(x) = P^*. \tag{2}$$

Instead of taking the infimum in (1) over the whole parameter space one could also compute the infimum over $\Omega(\delta)$, where $\Omega(\delta)$ consists of all possible configurations for which $\theta_{[k-1]}$ and $\theta_{[k]}$ are at least δ apart. In this case, we have

$$\inf_{\Omega(\delta)} P(CS) = \int_{-\infty}^{\infty} F^{k-1}(x + d - \delta) dF(x), \quad (3)$$

which is attained for the least favourable configuration (LFC): $\theta_{[1]} = \theta_{[k-1]} = \theta_{[k]} - \delta$. To be sure that $P(CS) \geq P^*$ for all configurations in $\Omega(\delta)$ the smallest value of the selection constant d has to be chosen for which

$$\int_{-\infty}^{\infty} F^{k-1}(x + d - \delta) dF(x) = P^*. \quad (4)$$

It is of interest to note that if one chooses d by solving (4), then, even under the least favourable configuration $\theta_{[1]} = \dots = \theta_{[k]}$, the probability that the subset contains the population with a location parameter not more than δ apart from $\theta_{[k]}$ is at least P^* . Therefore, in situations where one is satisfied with selection of a δ -almost best population, one should use the subset selection rule based on (4), even when it is not known that the best location parameter is at least δ apart from the other location parameters. The left-hand side of the equation (2) also equals the minimal probability of correct selection if we only select $\max_i X_i$ and where the minimum is over all configurations for which $\theta_{[k]} - \theta_{[k-1]} \geq d$.

The left-hand side of the equation (2) have been analytically determined for the logistic distribution by Van der Laan (1989, 1992) for Bechhofer's indifference zone approach (1954) and Gupta's subset selection approach, respectively. Lorentzen and McDonald (1981) considered the subset selection problem based on sample medians from logistic populations: i.e. now F represents the distribution of the sample median of n logistically distributed observations. In the case that the sampling distribution is logistic, the median is indeed a better candidate for X_i , $i = 1, \dots, k$, than the sample mean since it has a smaller variance than the sample mean, in contrary with the normal distribution where the mean is known to be an efficient estimator of its median (which equals its mean).

In the next section we consider subset selection, and thus the principal equation (2), based on the i -th order statistic of equal size i.i.d. samples of logistic populations which only differ in their location parameter. This problem is of particular practical relevance in survival analysis. Suppose that one has k equal size treatment groups and that the survival time is the outcome of interest. Assume that one has succeeded in selecting a monotone transformation

of the survival times (e.g. log) which is such that the transformed survival times have logistic distributions which only differ in their location parameter: the log-logistic model is a widely used model in survival analysis (see Collett, 1994). One can also estimate the monotone transformation by fitting the semiparametric proportional odds model (see Murphy, Rossini, van der Vaart, 1998, and Shen, 1998). If the life of a person is at risk it is important to have methods available which do not need all the data so that certain treatments can be excluded at an early stage. At the moment that one has observed the i -th failure in each group one can construct a subset which contains with probability (at least) P^* the best treatment, using our exact results for correct selection.

In van der Laan et al. (1998) it is shown numerically the distribution of the median of a sample of size 9 from a standard logistic distribution approximates the standard normal cumulative distribution within a distance 0.001 and that for most practical purposes the approximation of the standard normal distribution is in general already rather good for a sample of size 5. This approximation result shows that our subset selection result in section 2 can be used to carry out subset selection based on approximately normally distributed random variables X_j , $j = 1, \dots, k$ (e.g. sample means or maximum likelihood estimators of location parameters) whose distributions only differ in their location parameter.

2 Subset selection based on order statistics from logistic populations.

The cumulative distribution of the l -th order statistic $X_{(l)}$ of a sample of n i.i.d. random variables with cdf the standard logistic $G(x) = 1/(1 + \exp(-x))$ is given by:

$$F_{l,n}(x) = \frac{1}{(1 + \exp(-x))^n} \sum_{j=0}^{n-l} \binom{n}{j} \exp(-jx) \quad (5)$$

and the survival function is given by:

$$S_{l,n}(x) = \frac{1}{(1 + \exp(-x))^n} \sum_{j=0}^{l-1} \binom{n}{j} \exp(-(n-j)x). \quad (6)$$

For $l < n/2$ one can use $F_{l,n}(x) = 1 - S_{l,n}(x)$ so that $F_{l,n}$ does only involve a summation over l terms. Thus the density of $X_{(l)}$ for the standard logistic distribution is given by:

$$f_{l,n}(x) = \frac{1}{B(l, n+1-l)} \frac{\exp(-(n+1-l)x)}{(1 + \exp(-x))^{n+1}}, \quad (7)$$

where $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$ is the complete Beta function.

For Gupta's subset selection rule given in section 1 the minimal probability of correct selection for $F = F_{l,n}$ (see (5)) is given by

$$A_{l,n}(k) = \int_{-\infty}^{\infty} F_{l,n}^{k-1}(x+d)f_{l,n}(x)dx, \quad (8)$$

where $d > 0$ is the selection constant.

The theorem below provides us with an explicit formula for $A_{l,n}(k)$ in terms of binomial and multinomial coefficients. For notational convenience, we will denote the multinomial coefficients with $C_{i_0, \dots, i_{n-l}}^{k-1}$:

$$C_{i_0, \dots, i_{n-l}}^{k-1} = \binom{k-1}{i_0 \ i_1 \ \dots \ i_{n-l}}, \text{ with } 0 \leq i_0, \dots, i_{n-l} \leq k-1 \text{ and } \sum_{j=0}^{n-l} i_j = k-1. \quad (9)$$

We also define $s = s(i_0, \dots, i_{n-l}) \equiv \sum_{j=0}^{n-l} j i_j$, $\alpha = \exp(-d)$ and

$$C_q(\alpha) \equiv \left(\frac{\alpha}{\alpha-1} \right)^{q+1} \left\{ \log(\alpha) - \sum_{i=1}^q \frac{1}{i} \left(1 - \frac{1}{\alpha} \right)^i \right\} \text{ for integer } q \geq 1.$$

We have the following theorem.

Theorem 2.1 *We have*

$$A_{l,n}(k) = \exp(-ld)n \binom{n-1}{l-1} \times \sum_{\sum_{j=0}^{n-l} i_j = k-1, i_0, \dots, i_{n-l} \geq 0} C_{i_0, \dots, i_{n-l}}^{k-1} \prod_{j=0}^{n-l} \binom{n}{j}^{i_j} \sum_{i=0}^{n-l+s} \binom{n-l+s}{i} (-1)^i A((k-2)n - s + l + i, n+1, \alpha) \quad (10)$$

where $(k-2)n - s + l \geq 1$ and $n+1 \geq 2$. Here $A((k-2)n - s + l + i, n+1, \alpha)$ is given by the following exact formulas for $A(p, q, \alpha) = \int_0^\infty dx / \{(\alpha+x)^q(1+x)^p\}$ for positive integers p, q with $p \geq 1, q \geq 2$: We have

$$A(1, q, \alpha) = \frac{\alpha^{-q+1}}{q-1} - \frac{1}{q-1} A(2, q-1, \alpha), \text{ if } q \geq 3 \quad (11)$$

$$A(1, 2, \alpha) = \frac{1 - \alpha + \alpha \log(\alpha)}{(\alpha-1)^2 \alpha}. \quad (12)$$

For $q \geq 2$:

$$A(2, q, \alpha) = \frac{\alpha - q C_q(\alpha)}{\alpha^{q+1}}. \quad (13)$$

If $p \geq 3$ and $q \geq 2$, then we have

$$A(p, q, \alpha) = \frac{\alpha^{-q}}{p-1} + \sum_{i=1}^{p-3} (-1)^i \frac{q(q+1) \dots (q+i-1)}{(p-1)(p-2) \dots (p-1-i)} \alpha^{-(q+i)} \\ + (-1)^{p-2} \frac{q(q+1) \dots (q+p-3)}{(p-1)(p-2) \dots 2} A(2, q+p-2, \alpha). \quad (14)$$

Here the summation equals zero if $p = 3$.

For the special case $l = n$ this expression simplifies because the first summation is now only over the indice $i_0 = k - 1$.

Corollary 2.1 *We have for integer n*

$$A_{n,n}(k) = 1 + \binom{n(k-1)+n-2}{n-1} \times \\ \left[\sum_{i=1}^{n-1} (-1)^i \binom{n(k-1)+n-2}{n-i-1}^{-1} \alpha^i + (-1)^n \{n(k-1)+n-1\} \alpha^n C_{n(k-1)+n-1}(1/\alpha) \right].$$

In general, the expression (10) for $A_{l,n}(k)$ becomes more involved when l becomes smaller since the first summation in (10) is over $n-l-1$ indices. Therefore, if $l < n/2$ it is useful to obtain the result for the l -th order statistic from the result for the $n-l$ -th order statistic as follows.

Lemma 2.1 *We have*

$$A_{l,n}(k) = \sum_{j=0}^{k-1} \binom{k-1}{j} (-1)^j A_{n+1-l,n}(j+1), \quad (15)$$

where d on the right-hand side is replaced by $-d$.

Appendix: Proof of theorem 2.1.

Substitution of the representation (5) for $F_{l,n}$ into (8) provides us with:

$$A_{l,n}(k) = \int_{-\infty}^{\infty} \left\{ \sum_{i=0}^{n-l} \binom{n-l}{i} \frac{\exp(-i(x+d))}{(\exp(-x-d)+1)^n} \right\}^{k-1} n \binom{n-1}{l-1} \frac{\exp(-(n+1-l)x)}{(1+\exp(-x))^{n+1}} dx. \quad (16)$$

We define $a_i(x) \equiv \binom{n-l}{i} \frac{\exp(-i(x+d))}{(\exp(-x-d)+1)^n}$, $i = 0, \dots, n-l$. Then we can represent the term between accolades as $\sum_{i=0}^{n-l} a_i(x)$. We have $\left(\sum_{i=0}^{n-l} a_i(x) \right)^{k-1} =$

$\sum_{i_0, \dots, i_{n-l} \geq 0, \sum_{j=0}^{n-l} i_j = k-1} C_{i_0, \dots, i_{n-l}}^{k-1} a_0(x)^{i_0} \dots a_{n-l}(x)^{i_{n-l}}$. Substituting this expression into (16) yields

$$A_{l,n}(k) = n \binom{n-1}{l-1} \sum_{\sum_{j=0}^{n-l} i_j = k-1} C_{i_0, \dots, i_{n-l}}^{k-1} \int_{-\infty}^{\infty} \prod_{j=0}^{n-l} a_j^{i_j}(x) \frac{\exp(-(n+1-l)x)}{(1+\exp(-x))^{n+1}} dx.$$

Hence, it remains to determine $M(i_0, \dots, i_{n-l}) \equiv \int_{-\infty}^{\infty} \prod_{j=0}^{n-l} a_j(x)^{i_j} \frac{\exp(-(n+1-l)x)}{(1+\exp(-x))^{n+1}} dx$. We have that $\prod_{j=0}^{n-l} a_j(x)^{i_j} = \prod_{j=0}^{n-l} \binom{n}{j}^{i_j} \frac{\exp(-s(x+d))}{(1+\exp(-(x+d))^{n(k-1)}}$, where $s = s(i_0, \dots, i_{n-l}) = \sum_{j=0}^{n-l} j i_j$. So we have that $A_{l,n}(k) = n \binom{n-1}{l-1} \sum_{\sum_{j=0}^{n-l} i_j = k-1} C_{i_0, \dots, i_{n-l}}^{k-1} \prod_{j=0}^{n-l} \binom{n}{j}^{i_j} M_1$ with $M_1(i_0, \dots, i_{n-l}) \equiv \int_{-\infty}^{\infty} \frac{\exp(-s(x+d))}{(1+\exp(-(x+d))^{n(k-1)}} \frac{\exp(-(n+1-l)x)}{(1+\exp(-x))^{n+1}} dx$. Now, do the substitutions $\exp(-x) = z$, $\alpha z = x$, respectively, and write $(1+x/\alpha)^{n+1} = (\alpha+x)^{n+1}/\alpha^{n+1}$ to obtain $M_1(i_0, \dots, i_{n-l}) = \alpha^l \int_0^{\infty} \frac{x^{n-l+s}}{(\alpha+x)^{n+1}(1+x)^{n(k-1)}} dx$. Since $n-l+s \leq n-1+s$ and $s \leq (n-l)(k-1) \leq (n-1)(k-1)$ it follows that $n-l+s \leq n(k-2)$ which shows that the integral M_1 is convergent.

By writing $x^{n-l+s} = ((x+1)-1)^{n-l+s}$ it follows that $M_1 = \alpha^l \sum_{i=0}^{n-l+s} \binom{n-l+s}{i} (-1)^i A((k-2)n-s+l+i, n+1, \alpha)$, where $(k-2)n-s+l \geq n \geq 1$ and $A(p, q, \alpha) = \int_0^{\infty} \frac{dx}{(1+x)^p (\alpha+x)^q}$, $p \geq 1, q \geq 2$. This proves the expression (10) in terms of the integrals $A(p, q, \alpha)$, $p \geq 1, q \geq 2$. It remains to prove the closed form solution of $A(p, q, \alpha)$, $p \geq 1, q \geq 2$. Theorem 1 in van der Laan (1992) proves expression (13) for $A(2, q, \alpha)$ for positive integer $q \geq 2$. Now, let $p \geq 3$. Then by integration by parts we have: $\int (z+1)^{-p} (z+\alpha)^{-q} dz = \frac{\alpha^{-q}}{p-1} - \frac{q}{p-1} \int (z+1)^{-p+1} (z+\alpha)^{-q-1} dz$. By repeating this one proves expression (14) for $A(p, q, \alpha)$, for positive integers $p, q, q \geq 2$. It remains to solve $A(1, q, \alpha)$ for $q \geq 2$. Expression (11) for $A(1, q, \alpha)$ for $q \geq 3$ follows by integration by parts and expression (12) for $A(1, 2, \alpha)$ is trivial. This completes the proof.

References

- 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.
- Collett, D. (1994), *Modelling Survival Data in Medical Research*, Chapman & Hall, London.
- Dubey, S.D. (1969). A new derivation of the logistic distribution. *Naval Res. Logist. Quart.* **16**, 37-40.
- Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* **7**, 225-245.
- Laan, M. van der and Laan, P. van der (1998). Subset selection based on order statistics from logistic populations, Technical Report #70, Group in Biostatistics, University of

California, Berkeley.

Laan, P. van der (1989). Selection from logistic populations. *Statistica Neerlandica* **43**, 169-174.

Laan, P. van der (1992). On subset selection from logistic populations. *Statistica Neerlandica* **46**, 153-163.

Lorentzen, T.J. and McDonald, G.C. (1981). Selecting Logistic Populations using the sample medians. *Commun. Statist. - Theor. Meth.* **A10**(2), 101-124.

Murphy, S., Rossini, A. J. van der Vaart, A.W. (1997), MLE in the proportional odds model, *JASA* **92**, 968-976.

Shen, X. (1998), Proportional odds regression and universal sieve maximum likelihood estimation, technical report, Department of Statistics, The Ohio State University, Ohio.