Distribution theory for selection from logistic populations

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

COSOR-Memorandum 91-23

Distribution theory for selection from logistic populations

by

P. van der Laan

ISSN: 0926-4493

September, 1991

The Netherlands

DISTRIBUTION THEORY FOR SELECTION FROM LOGISTIC POPULATIONS[1]

Paul van der Laan

ABSTRACT

Assume k (integer $k \geq 2$) independent populations $\pi_1, \pi_2, \ldots, \pi_k$ are given. The associated independent random variables $X_1, X_2, \ldots, X_k$ are Logistically distributed with unknown means $\mu_1, \mu_2, \ldots, \mu_k$, respectively, and common known variance. The goal is to select the best population, this is the population with the largest mean. Some distributional results are derived for subset selection as well as for the indifference zone approach. The probability of correct selection is determined. Exact and numerical results concerning the expected subset size are presented for the subset selection approach.
Finally, some remarks are made for a generalized selection goal using subset selection. This goal is to select a non-empty subset of populations that contains at least one ε-best (almost best) treatment with confidence level P^*. For a set of populations an ε-best treatment is defined as a treatment with location parameter on a distance less than or equal to $\varepsilon(\varepsilon \geq 0)$ from the best population.

AMS Subject classification: Primary 26F07; secondary 62E15.

Key words and phrases: distribution theory, selection, subset selection, indifference zone approach, almost best population, best population, relative efficiency, Logistic distribution

[1]Paper presented at the Symposium on Biostatistics and Statistics in Honour of Charles W. Dunnett, 29 May – 1 June 1991, Hamilton, Ontario, Canada.

1. Introduction

This paper is mainly dealing with statistical selection from Logistic populations. The research can be characterized by three quotations. The first quotation is from Betrand Russell. The last two are from John W. Tukey. The quotations are:

Although this may seem a paradox all exact science is dominated by the idea of approximation.

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

A good asymptotic theory is one that works well for $n = 1$.

2. The Logistic distribution

Assume a collection V of k (integer $k \geq 2$) populations is given. The populations are indicated by $\pi_i (i = 1, 2, \ldots, k)$. So

$$V = (\pi_1, \pi_2, \ldots, \pi_k) .$$

Suppose that with population π_i there is associated a random variable $X_i (i = 1, 2, \ldots, k)$ which has a Logistic distribution with cumulative distribution function $G(\mu_i, \sigma^2)$, where the expectation

μ_i is unknown $(i = 1, 2, \ldots, k)$ and

the variance is assumed to be known. The ranked random variables (statistics) are denoted by $X_{[1]} \leq X_{[2]} \leq \ldots \leq X_{[k]}$. Ties occur only with probability zero.
The standard Logistic distribution has density

$$g(z) = \frac{\pi}{\sqrt{3}} e^{-\frac{\pi}{\sqrt{3}} z} \{1 + e^{-\frac{\pi}{\sqrt{3}} z}\}^{-2} \quad .$$

In the next chapter we shall consider statistical selection for the best population from a collection V of k Logistic populations. For the Logistic distribution it is possible to solve analytically certain distributional problems which are characteristic for statistical selection following the Indifference Zone approach as well as the Subset Selection approach. This makes it of interest to investigate the selection problem for Logistic populations. Moreover, there are some other arguments to consider the Logistic distribution, namely:

- Logistic population are interesting in their own right.
- The Logistic distribution has heavier tails than the Normal distribution. So it is possible to use the Logistic distribution in cases where probably the Normal distribution is appropriate but where we expect somewhat heavier tails due to outliers.
- An interesting point is the striking resemblance between the Logistic and the Normal curve (for a suitable choice of location and scale parameters). Therefore it is perhaps possible to use Logistic results as approximations for the Normal model.

For the last point we consider as an illustration the following problem. The Subset selection rule of Gupta is defined as follows:

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

where the selection constant d has to be determined under Normality assumption such that the probability requirement $P(CS) \geq P^*$ is satisfied. CS mean Correct selection and is defined as a selection for which the best population (that is the population with largest expectation [2]) is selected into the subset. The probability P^* is a known value, for instance

[2] If there is more than one best population due to ties, then one of them is appropriately tagged as the best.

0.90. Now, one can determine the actual lower bound of $P(CS)$ if in fact the distribution sampled from is Logistic (also with variance 1). The following results (a subset of the results obtained) illustrate in a certain sense the difference between Normal and Logistic population concerning subset selection:

P^*		
k	.90	.95
2	.906	.951
5	.900	.944
10	.890	.937

Table 1. Minimal probabilities of CS for the Normal subset selection procedure where in reality the random variables are Logistically distributed.

These minimal probabilities of CS are computed exactly for Logistically distributed observations.

If $X_i (i = 1, 2, \ldots, k)$ are sample means based on $n \geq 2$ individual and Logistically distributed observations, then with the Central Limit Theorem in mind one should expect smaller deviations. Simulation show this phenomenon.

3. Selection from Logistic populations

In this chapter we shall consider the selection problem in more detail. The expectations $\mu_1, \mu_2, \ldots, \mu_k$ associated with the k Logistic populations are ranked in increasing order of magnitude. The ordered parameters are denoted by

$$\mu_{[1]} \leq \mu_{[2]} \leq \ldots \leq \mu_{[k]}$$

and are associated with

$$\pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(k)} \ .$$

We suppose that the relation between ranked expectations and the populations is not known. The goal of the selection problem is to select $\pi_{(k)}$, the so-called best population, i.e. the population with the largest expectation. If there are more than one candidates for the best, then one of them is supposed to be appropriately tagged as the best. There are two main approaches in the literature in order to select the best population. The first one is the Indifference Zone approach. The second one is the Subset Selection approach. The Indifference Zone approach is especially of practical importance in designing an experiment, whereas the Subset selection approach has been developed for application after the experiment has already been executed. In this last case the numbers of observations are already fixed. Also in this last case the Indifference Zone approach can be applied, but a certain adaptation of the selection parameters is necessary.

Indifference Zone approach

The Indifference Zone approach of Bechhofer (1954) runs as follows.
Select the population with $X_{[k]}$, thus the population for which the associated random variable has the largest outcome.
The probability requirement is as follows:

$$P(CS) \geq P^*$$

for $\mu \in \Omega(\delta^*) = \{\mu = (\mu_1, \mu_2, \ldots, \mu_k) : \ \mu_{[k]} - \mu_{[k-1]} \geq \delta^*\}$, where

$CS =$ correct selection, i.e. selection of the best population,

$$\delta^* \geq 0 \ ,$$

and

$\Omega(\delta^*)$ is a subspace of the parameterspace $\Omega = \ \{\mu = (\mu_1, \mu_2, \ldots, \mu_k)\}$.

For Logistic populations one can prove that $P(CS)$ is minimal for the so-called Least Favourable Configuration LFC: $\mu_{[1]} = \mu_{[k-1]} = \mu_{[k]} - \delta^*$. Then the following holds:

$$P_{LFC}(CS) = \int_{-\infty}^{\infty} G^{k-1}(x+\delta^*)g(x)dx$$

$$= a^{k-1}\int_{0}^{\infty} \frac{ds}{(s+1)^2(s+a)^{k-1}}$$

with

$$a = e^{\frac{\pi}{\sqrt{3}}\delta^*} \quad .$$

From results proved in Van der Laan (1989) the next theorem can easily be proved.

Theorem 1. The probability requirement can be written as follows

$$P^* = 1 - \frac{k-1}{a}C_{k-1}(a)$$

where $C_m(a)$ for all integer m is defined as follows

$$C_m(a) = \left(\frac{a}{a-1}\right)^{m+1}\left[\ln a - \sum_{i=1}^{m}\frac{1}{a}(1-\frac{1}{a})^i\right]$$

with

$$\sum_{i=1}^{m}\frac{1}{a}(1-\frac{1}{a})^i = 0 \quad \text{for} \quad m \leq 0 \quad .$$

Using Theorem 1 some results for δ^* are given in Table 2.
(Between brackets Normal results, $\sigma^2 = 1$).

P^* / k	.90	.95
2	1.76	2.31
	(1.81)	(2.33)
5	2.60	2.13
	(2.60)	(3.06)
10	3.06	3.59
	(2.98)	(3.42)

Table 2. The value of δ^* for some values of k and P^*. Between brackets the value of δ^* is given for Normal populations.

Subset selection

The subset selection procedure of Gupta (1965) runs as follows. Select π_i in the subset if and only if $X_i \geq X_{[k]} - d$.
The selection constant d must be determined in such a way that the probability requirement

$$\inf_{\mu} P(CS) = P^*$$

is met. Using the approach of Gupta CS is defined as the best population is selected in the subset. The Least Favourable Configuration LFC for Gupta's procedure is attained for all μ's equal to each other, thus

$$\mu_{[1]} = \mu_{[k]} \ .$$

Thus the probability requirement can be written as follows:

$$\begin{aligned} P^* &= P_{LFC}(CS) \\ &= P(CS|\mu_{[1]} = \mu_{[k]}) \ . \end{aligned}$$

Using Theorem 1 one finds for all $k \geq 2$

$$P^* = 1 - \frac{k-1}{a} C_{k-1}(a)$$

with $a = e^{\lambda d}$ and λ the scale parameter of the Logistic distribution.

Expected size of the subset
It can easily be proved (cf. Gupta (1965)) that

$$\max_{\mu \in \Omega} E(S) = kP^* \ ,$$

where S is the size of the subset, i.e. the number of selected populations in the subset.
It is of interest to investigate the maximum value of $E(S)$ in the subspace $\Omega(\delta)$ of Ω. Using results from Van der Laan (1990; 1991) Theorem 2 can be verified.

Theorem 2. With

$$M = \max_{\Omega(\delta)} E(S)$$

the following holds for all $k \geq 2, d \neq \delta$ and $b = e^{\lambda \delta}$:

$$\begin{aligned} M = \ & 1 - \frac{k-1}{ab} C_{k-1}(ab) + (k-1)\frac{a}{a-b}[1- \\ & - \frac{k-2}{a} C_{k-2}(a) - \frac{b}{a-b}\{C_{k-3}(a) - C_{k-3}(b)\}] \ . \end{aligned}$$

For $d = \delta$ a somewhat more simple formula can be derived (see Van der Laan (1990)). Some results for $\max_{\mu \in \Omega(\delta)} E(S)$ are given in Table 3.

k \ d	1	2	4	
2	1.4	1.8	2.0	$\leftarrow \delta = 1$
	1.0	1.0	1.2	$\leftarrow \delta = 5$
5	2.3	3.7	4.9	
	1.0	1.1	1.9	
10	3.2	6.4	9.6	
	1.0	1.2	3.0	
100	5.5	23.0	83.5	
	1.3	2.5	22.4	

Table 3. Some values of $\max_{\mu \in \Omega(\delta)} E(S)$ for some values of k, the selection constant d, and for Logistic populations. The first number is for $\delta = 1$. Under this value one finds the result for $\delta = 5$.

4. Subset selection of an almost (ε-best) population

In practice one can expect that in a number of cases where one is interested in indicating the best population, selecting a population which is very near to the best one, is also acceptable. Such a population is called an almost best population or an ε-best population, where an ε-best population is defined as follows.

Definition 1. A population π_i is called an ϵ-best population or an almost best population if and only if

$$\mu_i \geq \mu_{[k]} - \varepsilon \quad (\varepsilon \geq 0;\ i = 1, 2, \ldots, k) \ .$$

Subset selection of an ε-best population is defined as follows.
It means selection of a subset in which at least one ε-best population is selected with minimal probability P^*. A CS is now defined as selecting a subset in which an ε-best population has been selected.
It is of practical and theoretical interest to compare subset selection of the best population with subset selection of an ε-best population. This has been done using two different criteria. The first criterion is based on the minimal probability of correct selection. The second criterium is based on the expected subset size.
The selection rule is as follows:

$$\pi_i \text{ in subset if and only if } X_i \geq X_{[k]} - c$$

where $X_{[k]} = \max_{1 \leq j \leq k} X_j$ and with c (≥ 0) such that

$$P(CS) \geq P^* \ .$$

The selection constant c has been determined for some values of P^* and ε. Using this selection constant c the minimal probability of selecting the best population in the subset has been determined. One may expect that this probability is smaller. An interesting point is the difference between P^* and this smaller value. In Table 4 some results are given for Normal populations with $\sigma^2 = 1$.

k \ ε	.5	1
2	.83	.71
4	.82	.69
10	.81	.67
50	.80	.65

Table 4. The minimal probability of correct selection of the best Normal population ($\sigma^2 = 1$) using the selection rule for an ε-best with $P^* = 0.90$.

In the second case we can define a relative efficiency r.e. as follows:

$$r.e. = \frac{\sup_{\Omega} E(S_B)}{\sup_{\Omega(\epsilon)} E(S_E)} \ ,$$

where

S_B = size subset for the best population

and

S_E = size subset for an ε−best population.

In Table 5 some results are given for Normal and Logistic populations, both with variance 1.

r.e.	$P^* = .90$		
k \ ε	1	2	
2	1.06	1.25	← Normal
	1.07	1.27	← Logistic
5	1.05	1.26	
	1.04	1.25	
10	1.03	1.20	
	1.03	1.17	

Table 5. The relative efficiency r.e. for some values of k and ε. The first value of each pair is for Normal populations, the second value at the bottom of the pair is for Logistic populations (both with variance 1).

5. Concluding remarks

Firstly, from the computed values it can be concluded that there is a good agreement between Normal and Logistic results. So it seems reasonable to consider the possibility of using Logistic results as approximations for Normal problems. Further research is needed. Secondly, it seems worthwile to consider the generalized selection goal to select an almost (ε-best) population instead of the best one. For sufficiently small ε, selection of an ε-best population will in practice be more or less of equal value. The gain in probability of correct selection and the decrease in expected subset size may be in a number of cases practically important.

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.

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

Van der Laan, P. (1989). Selection from Logistic Population. *Statistica Neerlandica* **43**, 169-174.

Van der Laan, P. (1990). On subset selection from Logistic populations. *Memorandum COSOR 90-12*, Eindhoven University of Technology, Dept. of Mathematics and Computing Science.

Van der Laan, P. (1991). On subset selection from logistic populations. *Statistica Neerlandica* **46**, nr. 2 (to appear).

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science
PROBABILITY THEORY, STATISTICS, OPERATIONS RESEARCH
AND SYSTEMS THEORY
P.O. Box 513
5600 MB Eindhoven, The Netherlands

Secretariate: Dommelbuilding 0.03
Telephone : 040-473130

-List of COSOR-memoranda - 1991

Number	Month	Author	Title
91-01	January	M.W.I. van Kraaij W.Z. Venema J. Wessels	The construction of a strategy for manpower planning problems.
91-02	January	M.W.I. van Kraaij W.Z. Venema J. Wessels	Support for problem formulation and evaluation in manpower planning problems.
91-03	January	M.W.P. Savelsbergh	The vehicle routing problem with time windows: minimizing route duration.
91-04	January	M.W.I. van Kraaij	Some considerations concerning the problem interpreter of the new manpower planning system formasy.
91-05	February	G.L. Nemhauser M.W.P. Savelsbergh	A cutting plane algorithm for the single machine scheduling problem with release times.
91-06	March	R.J.G. Wilms	Properties of Fourier-Stieltjes sequences of distribution with support in [0,1).
91-07	March	F. Coolen R. Dekker A. Smit	Analysis of a two-phase inspection model with competing risks.
91-08	April	P.J. Zwietering E.H.L. Aarts J. Wessels	The Design and Complexity of Exact Multi-Layered Perceptrons.
91-09	May	P.J. Zwietering E.H.L. Aarts J. Wessels	The Classification Capabilities of Exact Two-Layered Peceptrons.
91-10	May	P.J. Zwietering E.H.L. Aarts J. Wessels	Sorting With A Neural Net.
91-11	May	F. Coolen	On some misconceptions about subjective probability and Bayesian inference.

COSOR-MEMORANDA (2)

91-12	May	P. van der Laan	Two-stage selection procedures with attention to screening.
91-13	May	I.J.B.F. Adan G.J. van Houtum J. Wessels W.H.M. Zijm	A compensation procedure for multiprogramming queues.
91-14	June	J. Korst E. Aarts J.K. Lenstra J. Wessels	Periodic assignment and graph colouring.
91-15	July	P.J. Zwietering M.J.A.L. van Kraaij E.H.L. Aarts J. Wessels	Neural Networks and Production Planning.
91-16	July	P. Deheuvels J.H.J. Einmahl	Approximations and Two-Sample Tests Based on P - P and Q - Q Plots of the Kaplan-Meier Estimators of Lifetime Distributions.
91-17	August	M.W.P. Savelsbergh G.C. Sigismondi G.L. Nemhauser	Functional description of MINTO, a Mixed INTeger Optimizer.
91-18	August	M.W.P. Savelsbergh G.C. Sigismondi G.L. Nemhauser	MINTO, a Mixed INTeger Optimizer.
91-19	August	P. van der Laan	The efficiency of subset selection of an almost best treatment.
91-20	September	P. van der Laan	Subset selection for an ε-best population: efficiency results.
91-21	September	E. Levner A.S. Nemirovsky	A network flow algorithm for just-in-time project scheduling.
91-22	September	R.J.M. Vaessens E.H.L. Aarts J.H. van Lint	Genetic Algorithms in Coding Theory - A Table for $A_3(n,d)$.
91-23	September	P. van der Laan	Distribution theory for selection from logistic populations.