Austral. J. Statist. 36(1), 1994, 59–66

SHOULD WE ESTIMATE THE PROBABILITY OF CORRECT SELECTION?

EVE BOFINGER[1]

University of New England

Summary

Various authors, given k location parameters, have considered lower confidence bounds on (standardized) differences between the largest and each of the other $k-1$ parameters. They have then used these bounds to put lower confidence bounds on the probability of correct selection (PCS) in the same experiment (as was used for finding the lower bounds on differences). It is pointed out that this is an inappropriate inference procedure. Moreover, if the PCS refers to some later experiment it is shown that if a non-trivial confidence bound is possible then it is already possible to conclude, with greater confidence, that correct selection has occurred in the first experiment. The short answer to the question in the title is therefore 'No', but this should be qualified in the case of a Bayesian analysis.

Key words: Probability of correct selection; subset selection; confidence bounds; posterior probability.

1. Introduction

Let Y_i $(i = 1, \ldots, k)$ be independently distributed with absolutely continuous cumulative distribution functions (cdfs) given by

$$F\left(\frac{y-\theta_i}{\beta}\right)$$

the location parameters θ_i being unknown. When the common scale parameter β is unknown we assume an estimator $\hat{\beta}$ is available, and that $\hat{\beta}/\beta$ has a known distribution.

The Y_i may be thought of as estimators of the θ_i, based on samples from populations Π_i, and the goal of the experiment may be the selection of the 'best' population, $\Pi_{\rho(k)}$, where the θ_i are ordered by $\rho(\cdot)$ so that

$$\theta_{\rho(1)} \leq \cdots \leq \theta_{\rho(k)}.$$

(If it happens that $\theta_{\rho(k-1)} = \theta_{\rho(k)}$ for some k then we suppose that some tagging system is used so that $\rho(k)$ is well defined.)

Received March 1992; revised April 1993; accepted November 1993.

[1]Dept of Mathematics, Statistics and Computing Science, University of New England, Armidale, NSW 2351, Australia.

Acknowledgements. The author thanks the editor for his helpful comments.

Now suppose that the Y_i are ordered by $R(\cdot)$ so that

$$Y_{R(1)} < Y_{R(2)} < \cdots < Y_{R(k)}$$

and that population $\Pi_{R(k)}$ is selected as the 'apparently best'. Correct selection is defined as the event

$$\{R(k) = \rho(k)\}$$

and the probability of correct selection (PCS) is given, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, by

$$PCS(\boldsymbol{\theta}) = \int \prod_{i \neq \rho(k)} F\left(\frac{y - \theta_i}{\beta}\right) dF\left(\frac{y - \theta_{\rho(k)}}{\beta}\right).$$

With $\delta_i = \theta_{\rho(k)} - \theta_{\rho(i)}$ for all $i = 1, \ldots, k-1$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{k-1})$ this may be written as

$$P\left(\frac{\boldsymbol{\delta}}{\beta}\right) = \int \prod_{i=1}^{k-1} F\left(x + \frac{\delta_i}{\beta}\right) dF(x), \tag{1}$$

and a lower bound is given by

$$P_k\left(\frac{\delta}{\beta}\right) = \int \left\{F\left(x + \frac{\delta}{\beta}\right)\right\}^{k-1} dF(x), \tag{2}$$

with $\delta = \delta_{k-1}$ (for simplicity in later exposition).

Many authors, including Olkin, Sobel & Tong (1982), Anderson, Bishop & Dudewicz (1977), Kim (1986), Lam (1989), Gupta, Leu & Liang (1990), Gupta & Liang (1991), Driessen (1991) and Sohn & Kang (1992), have used for the δ_i estimators or lower confidence bounds based on the experiment. Hence they obtain estimators or lower confidence bounds for $P(\boldsymbol{\delta}/\beta)$ or $P_k(\delta/\beta)$ (in both normal and non-normal situations).

These appear to be inappropriate inference procedures for the following reasons:

(a) the PCS is only of interest *prior* to the experiment and should be considered when the experiment is designed, as in Bechhofer (1954);
(b) after the experiment is completed a confidence statement on the selected population is of more interest;
(c) although $P_k(\delta/\beta)$ and $P(\boldsymbol{\delta}/\beta)$ may be thought of as parameters, they actually depend on the design of the experiment and hence are not necessarily parameters of interest when the experiment is completed;
(d) on the other hand, if another experiment of *exactly the same design* is contemplated, then estimators posterior to the first experiment may be used in estimating $P(\boldsymbol{\delta}/\beta)$ (or $P_k(\delta/\beta)$) which is now the probability (or lower bound to the probability) of correct selection in the second experiment.

Such a design seems a rather unlikely allocation of resources and it is only when the second experiment is of some different design — possibly involving larger sample sizes — that the estimation of PCS seems useful.

The question of a second experiment is explored in Sections 2 and 3. We show in Section 2 that if a confidence statement about the PCS (in the second experiment) is based on a lower bound for δ/β and if a non-trivial statement is possible, then, with greater confidence, it is possible to conclude that correct selection has already been made (in the first experiment). Similar arguments apply, and are given in Section 3, when the lower bound for PCS is based on lower bounds for δ_i/β for all $i = 1, \ldots, k-1$.

An approach which may be confused with the 'estimation' of PCS is that of a Bayesian analysis, as in Berger & Deely (1988) or Bofinger (1990). When shrinkage estimators for the δ_i are used, as in McCulloch & Dechter (1985) and Ventner (1989), the 'estimated PCS' may be regarded as the posterior probability of the event that the selected population is better than the others, or

$$\Pr\{\theta_{R(k)} \geq \theta_i \quad \text{for all} \quad i \neq R(k) \mid Y_1, \ldots, Y_k\}.$$

To avoid confusion between confidence statements and probability statements we introduce notation for observed values. Let $y_1, \ldots, y_k$ be the observed values of $Y_1, \ldots, Y_k$ and let the function $r(\cdot)$ give the ordering for the y_i so that $y_{r(1)} < \cdots < y_{r(k)}$. (We ignore the possibility of ties but in practice some method for dealing with them should be contemplated.)

Hsu (1984, 1985) and Bofinger (1983) give a confidence interval approach which results in statements on $\delta_{r(k)}$. However, this paper is restricted to inference on the parameter $\rho(k)$.

2. Demonstrably Best

We make the inference $r(k) = \rho(k)$ provided that $y_{r(k)} - y_{r(k-1)}$ is sufficiently large, and otherwise we make no inference. Alternatively we may think of this procedure as that of establishing a confidence set for $\rho(k)$ as shown below.

2.1. Known β

Consider a confidence set for the parameter $\rho(k)$ given by S_0, the observed value of S, where

$$S = \begin{cases} \{R(k)\} & \text{if } Y_{R(k)} - Y_{R(k-1)} > d_\alpha \beta, \\ \{1, \ldots, k\} & \text{otherwise,} \end{cases} \tag{3}$$

where, for $0 < \alpha < 1$, d_α is chosen to satisfy

$$\sum_{i \neq \rho(k)} \Pr\{Y_i - Y_{R(k-1)} > d_\alpha \beta\} \leq \alpha. \tag{4}$$

From (4) we have $\Pr\{\rho(k) \in S\} \geq 1-\alpha$ and hence the confidence for the set S_0 is at least $1-\alpha$. In other words, we infer, with confidence $1-\alpha$, that correct selection has occurred provided $y_{r(k)} - y_{r(k-1)} > d_\alpha\beta$, and make no inference otherwise. Bofinger (1988) has referred to $d_\alpha\beta$ as the 'least significant spacing'.

For distributions with the monotone likelihood ratio (MLR) property we can define d_α by

$$P_2(-d_\alpha) = 2^{-1}\frac{k\alpha}{k-1} \tag{5}$$

as the following theorem shows.

Theorem 2.1. *Inequality (4) holds for MLR distributions when d_α is given by (5).*

Proof. The LHS of (4) may be seen to be decreasing in $\theta_{\rho(k)}$ by writing it as

$$\Pr\{\max_{i\neq\rho(k)} Y_i - Y_{R(k-1)} > d_\alpha\beta\}$$

and noticing that $\max_{i\neq\rho(k)} Y_i$ does not depend on $\theta_{\rho(k)}$ while $Y_{R(k-1)}$ is stochastically increasing in $\theta_{\rho(k)}$. Hence its maximum value occurs when $\theta_{\rho(k)} = \theta_{\rho(k-1)}$ which we now assume to be the case.

From the lemma in Kim (1986) or Lemma 2.1 in Gupta *et al.* (1990) we have

$$\begin{aligned}\Pr\{Y_{R(k)} - Y_{R(k-1)} > d_\alpha\beta\} &\leq \Pr\{|Y_{\rho(k)} - Y_{\rho(k-1)}| > d_\alpha\beta\} \\ &= 2P_2(-d_\alpha)\end{aligned} \tag{6}$$

by our assumption. Now

$$\Pr\{Y_{\rho(k)} - Y_{R(k-1)} > d_\alpha\beta\} \geq \Pr\{Y_i - Y_{R(k-1)} > d_\alpha\beta\} \quad \text{for all } i \neq \rho(k), \tag{7}$$

and hence the LHS of (7) is greater than or equal to k^{-1}(LHS of (6)). Hence

$$\sum_{i\neq\rho(k)} \Pr\{Y_i - Y_{R(k-1)} > d_\alpha\beta\} \leq (1-k^{-1})\Pr\{Y_{R(k)} - Y_{R(k-1)} > d_\alpha\beta\},$$

from which the result follows.

For the case of normal distributions considerable improvement is possible: it amounts to replacing the RHS of (5) by α. This was done in Bofinger (1988) as follows. Let $F(\cdot) = \Phi(\cdot)$, the standard normal cdf, and set $d_\alpha = 2^{1/2}z_\alpha$ where z_α is the upper α point of $\Phi(\cdot)$. Then provided $d_\alpha > \ln 2$, which implies $\alpha < 0.2733$, inequality (4) holds, with equality when $\theta_{\rho(k-1)} = \theta_{\rho(k)}$ and $\theta_{\rho(k-2)} \to -\infty$. Hence the confidence is exactly $1-\alpha$.

We note that this procedure leads more often to the inference that $\Pi_{r(k)}$ is 'demonstrably best' than the conditional procedure of Gutmann & Maymin (1987) where $2^{1/2}z_{\alpha/2}$ is used in place of $2^{1/2}z_\alpha$. For non-normal distributions the corresponding improvement given by Theorem 2.1 is slight for moderately large k since the Gutmann & Maymin procedure uses a d_α given by $P_2(-d_\alpha) = \alpha/2$ instead of (5).

2.2. Unknown β

For the case where β is estimated we replace β by $\hat{\beta}$ in the definition at (3) of the set S. Then, for MLR distributions, we define d_α to satisfy

$$\mathrm{E}_{\hat{\beta}}\left\{P_2\left(\frac{-d_\alpha\hat{\beta}}{\beta}\right)\right\} = 2^{-1}\frac{k\alpha}{k-1}. \tag{8}$$

As in Theorem 2.1, by working first conditionally on $\hat{\beta}$, we see that the confidence for the set S_0 is at least $1-\alpha$.

For normal distributions an improvement is possible and we conclude, with *nominal* confidence $1-\alpha$, that correct selection has occurred provided

$$y_{r(k)} - y_{r(k-1)} > (2/n)^{1/2} t_\alpha^{(\nu)} s, \tag{9}$$

where s is the usual estimate of the standard deviation σ based on $\nu = k(n-1)$ degrees of freedom and $t_\alpha^{(\nu)}$ is the upper α point of Student's t-distribution with ν degrees of freedom, because the Y_i are sample means of n independent observations.

The actual confidence level may be very slightly less than $1-\alpha$ but Bofinger (1988) indicates that the change in α is less than $\alpha/20$ for $\alpha \leq 0.10$ and $\nu \geq 10$. (Bofinger also indicates a method of iterating towards a conservative replacement for $t_\alpha^{(\nu)}$.)

2.3 Estimation of δ/β

Kim (1986) and Gupta *et al.* (1990), refining and extending the work of Anderson *et al.* (1977) to the case of distributions with the MLR property, derive a $1-\alpha$ lower confidence bound on δ/β which they use to obtain a lower confidence bound on $P_k(\delta/\beta)$ as given by (2). As discussed in Section 1, this can be regarded as (a lower bound on) the probability of correct selection in a second experiment. Notice, however, that if the lower bound on δ/β is zero then the lower bound on $P_k(\delta/\beta)$ is just $1/k$. On the other hand, if the lower bound is positive we have a non-trivial lower bound on $P_k(\delta/\beta)$ but we also find that $S_0 = \{r(k)\}$ and hence we may infer with confidence $1-\alpha$ that correct selection has already occurred. Also, even when the lower bound on δ/β is zero it is still possible that $S_0 = \{r(k)\}$. In other words, non-trivial inferences are more often made using the method above than by putting lower bounds on δ/β.

For an example on selecting a most profitable production plan Kim (1986) finds a 90% lower confidence bound on δ/β and hence concludes, "We can state with 90% confidence that PCS$\geq$ .856." Also, using the same example and finding lower bounds for δ_i/β for all $i = 1, \ldots, k-1$ Lam (1989) finds a 95% lower confidence bound on PCS of 0.685. However, for the same example, using the method above, we can state that the correct selection *has* been made with 99.5% confidence.

For an example on insulating fluids Gupta *et al.* (1990) find, with confidence 95%, a lower bound for PCS of just 0.5356 . However, using the method above, we may state, with 95% confidence, that the correct selection has been made.

3. Subset Selection

For Gupta's subset selection approach, as described in Gupta & Panchapakesan (1979, 203–209), we consider a confidence set for $\rho(k)$ of GS_0, the observed value of GS, where

$$GS = \{i \mid Y_i > Y_{R(k)} - \lambda_{k,\alpha}^{(\nu)}\hat{\beta}\}$$

and $\lambda_{k,\alpha}^{(\nu)}$ is given by

$$E_{\hat{\beta}}\left\{P_k\left(\frac{\lambda_{k,\alpha}^{(\nu)}\hat{\beta}}{\beta}\right)\right\} = 1 - \alpha. \tag{10}$$

In other words, we select populations Π_i for all $i \in GS_0$, and claim, with confidence at least $1-\alpha$, that the best population $\Pi_{\rho(k)}$ is one of those selected. (Of course, this method is of doubtful interest to an experimenter who wishes to select just *one* population but it is certainly of great interest in a 'screening' experiment.)

For the case of the normal distribution Gupta & Liang (1991) consider $1-\alpha$ lower confidence bounds on δ_i/β which are then used to put a lower bound on $P(\boldsymbol{\delta}/\beta)$. As in Section 1 this may be considered to refer to a second experiment.

However, if the lower bounds on all the δ_i are positive, then we find that the number of populations selected, denoted by $\#GS_0$, is equal to 1. Hence, instead of inferring (with confidence $1-\alpha$) that the probability of correct selection in the next experiment is at least the lower bound obtained, we may infer (with confidence $1-\alpha$) that correct selection has *already* occurred in the first experiment.

On the other hand, if $\#GS_0 = t > 1$, so that a screening experiment would result in t populations being selected, then we notice that the lower bounds for at least $t-1$ of the δ_i/β are equal to zero so that the lower bound on $P(\boldsymbol{\delta}/\beta)$ is at *most* $1/t$. The same inequality applies (to an estimated lower bound for the PCS) even when the second experiment is not identical, no matter how sensitive this experiment may be.

Now if, instead of a second experiment on the k populations, we consider a second experiment involving just the t populations left after screening, we may infer with confidence $1-\alpha$ that $\Pi_{\rho(k)}$ is one of these t and hence that the probability of correct selection is at *least* $1/t$. This holds even when the second experiment is quite insensitive and in particular when one of the t is selected at random.

The estimation procedure in Lam (1989) may be seen to have the same kind of problems as discussed above for the Gupta & Liang procedure. Also, Lam considers estimating the probability of correct *subset* selection, with a generalisation being given in Driessen (1991). Defining, for some $d > 0$, the subset

$$T = \{i \mid Y_i > Y_{R(k)} - d\hat{\beta}\},$$

he finds, by using lower confidence bounds on the δ_i/β, a lower bound for the probability of correct subset selection, defined as the event $\{\rho(k) \in T\}$. This lower bound, being based on the first experiment, is to be taken as referring to a second experiment.

Now, if $\#GS_0 = t$, consider a second experiment which uses just these t populations and which selects that subset (of the t) which have Y values within $d\hat{\beta}$ of the largest Y. Since we infer (with confidence $1-\alpha$) that $\Pi_{\rho(k)}$ is one of the t populations we may also infer that the probability is at *least* $E_{\hat{\beta}}\{P_t(d\hat{\beta}/\beta)\}$ that $\Pi_{\rho(k)}$ belongs to the subset selected by the second experiment.

On the other hand, using Lam's lower bound, we see that at least $t-1$ of the lower bounds for the δ_i/β are equal to zero and hence that his lower bound on the probability of correct subset selection (in an identical second experiment) is at most $E_{\hat{\beta}}\{P_t(d\hat{\beta}/\beta)\}$.

References

ANDERSON, P.O., BISHOP, T.A. & DUDEWICZ, E.J. (1977). Indifference zone ranking and selection: confidence intervals for true achieved P(CD). *Comm. Statist. A—Theory Methods* **6**, 1121–1132.

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.

BERGER, J.O. & DEELY, J.J. (1988). A Bayesian approach to ranking and selection of related means with alternatives to AOV methodology. *J. Amer. Statist. Assoc.* **83**, 364–373.

BOFINGER, EVE (1983). Multiple comparisons and selection. *Austral. J. Statist.* **25**, 198–207.

— (1988). Least significant spacing for 'one versus the rest' normal populations. *Comm. Statist. A—Theory Methods* **17**, 1697–1716.

— (1990). Posterior probability of correct selection. *Comm. Statist. A—Theory Methods* **19**, 599–616.

DRIESSEN, S. (1991). Multiple comparisons with and selection of the best treatment in (incomplete) block designs. *Comm. Statist. A—Theory Methods* **20**, 179–217.

GUPTA, S.S., LEU, L.Y. & LIANG, T. (1990). On lower confidence bounds for PCS in truncated location parameter models. *Comm. Statist. A—Theory Methods* **19**, 527–546.

— & LIANG, T. (1991). On a lower confidence bound for the probability of a correct selection: analytical and simulation studies. In *Proc. ICOSCO — I*, Vol.II 26, eds. A. Ozturk and E.C. van der Meulen, pp.77–95. Columbus, Ohio: American Sciences Press.

— & PANCHAPAKESAN, S. (1979). *Multiple Decision Procedures.* New York: Wiley.

GUTMANN, S. & MAYMIN, Z. (1987). Is the selected population the best? *Ann. Statist.* **15**, 456–461.

HSU, J.C. (1984). Constrained two-sided simultaneous confidence intervals for multiple comparisons with the best. *Ann. Statist.* **12**, 1136–1144.

— (1985). A method of unconstrained multiple comparisons with the best. *Comm. Statist. A—Theory Methods* **14**, 2009–2028.

KIM, W-C. (1986). A lower confidence bound on the probability of a correct selection. *J. Amer. Statist. Assoc.* **81**, 1012–1017.

LAM, K. (1989). The multiple comparison of ranked parameters. *Comm. Statist. A—Theory Methods* **18**, 1217–1237.

McCULLOCH, C.E. & DECHTER, A. (1985). An empirical Bayes approach to estimating the probability of correct selection. *Comm. Statist. B—Simulation Comput.* **14**, 173–186.

OLKIN, I., SOBEL, M. & TONG, Y.L. (1982). Bounds for a k-fold integral for location and scale parameter models with applications to statistical ranking and selection problems. In *Statistical Decision Theory and Related Topics III*, Vol.2, eds. S.S. Gupta and J.O. Berger, pp.193–212. New York: Academic Press.

SOHN, J.K. & KANG, S.G. (1992). On the estimation of the true probability of a correct selection. *Comm. Statist. B—Simulation Comput.* **21**, 445–462.

VENTNER, J.H. (1989). Estimating the Probability of Correct Selection. *J. Statist. Comput. Simulation* **32**, 165–176.