Some Thoughts on Multistage Selection Procedures for t > 1

Jürgen Pilz

Institut für Statistik, Universität Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
juergen.pilz@aau.at

SimStat
September 6, 2019 / Salzburg, Austria

Preliminary thoughts

Selection vs. multiple comparisons vs. sequential decision problem
Start with iid case:

$$X_i \ \widetilde{iid} \ N(\mu; \sigma^2); \quad i = 1, \ldots, a$$

We need an ordering of the means:

$$\mu^*_{(1)} \leq \mu^*_{(2)} \leq \ldots \leq \mu^*_{(a)}$$

Probability of correct selection $P(CS) =: P_C$ between

$$G_1 = \{A_1, \ldots A_{a-t}\} \text{ and}$$

$$G_2 = \{A_{a-t+1}, A_{a-t+2}, \ldots A_a\}$$

Position functions

$\Rightarrow$ **Note on order statistics (position functions) for i.i.d. observations**

For a populations, denote $\zeta = (\zeta_1^{(a)}, \zeta_2^{(a)}, \ldots, \zeta_a^{(a)})$

the vector of position functions corresponding to $(X_1, \ldots, X_a) \ \widetilde{iid} \ F$

Then it is well-known that

$$\begin{aligned} F_{ka}(x) &= P(\zeta_k^{(a)} < x) \\ &= \sum_{m=k}^{a} \left(\begin{array}{c} a \\ m \end{array} \right) (F(x))^m (1 - F(x))^{a-m} \end{aligned}$$

A useful identity

Not so widely known is the identity

$$F_{ka}(x) = \frac{a!}{(k-1)!(a-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{a-k}\,dt$$

which follows from (repeated) partial integration.

From this it is easily seen that $\zeta_k^{(a)}$ has density

$$f_{k,a}(x) = \frac{a!}{(k-1)!(a-k)!}(F(x))^{k-1}(1-F(x))^{a-k}f(x).$$

The joint distribution function of the group of the $t > 1$ largest position functions

$$\left(\zeta_{a-t+1}^{(a)}, \zeta_{a-t+2}^{(a)}, \ldots, \zeta_a^{(a)}\right)$$

has density $h(y_{a-t+1}, y_{a-t+2}, \ldots, y_a) = c(a,t) \cdot f(y_{a-t+1}) \cdot \ldots \cdot f(y_a)$

whenever

$y_{a-t+1} < y_{a-t+2} < \ldots < y_a$ and zero else

where

$$
\begin{aligned}
c(a,t) &= \frac{a!}{(a-t)!} \left[\int_{-\infty}^{y_{a-t+1}} f(x)dx \right]^{a-t} \\
&= \frac{a!}{(a-t)!} \left[F(y_{a-t+1}) \right]^{a-t}
\end{aligned}
$$

Non-identical case

More compactly,

$$h(y_{a-t+1},\ldots,y_a) = \frac{a!}{(a-t)!}\left[F(y_{a-t+1})\right]^{a-t}\cdot\prod_{j=0}^{t-1} f(y_{a-j})$$

Special case: $t = 2$

$$h(y_{a-1},y_a) = a(a-1)\left[F(y_{a-1})\right]^{a-2}\cdot f(y_{a-1})\cdot f(y_a)$$

It is straight-forward to generalize the result to the case of non-identical distributions

$$X_k \ \widetilde{i.d.}\ F_k; k = 1,\ldots,a$$

Bechhofer (1954):

Indifference zone: $\delta^* = \delta\frac{\sqrt{n}}{\sigma} = c\sqrt{n}$ (standardized difference)
$c = \delta/\sigma$ relative difference

Assume that $\mu_{a-t+1} > \mu_{a-t} + \delta^*$, we have:

$$P_C = P(\max(\overline{y}_1, \ldots, \overline{y}_{a-t}) < \min(\overline{y}_{a-t+1}, \ldots, \overline{y}_a))$$

$$\geq t * \int_{-\infty}^{+\infty} [\Phi(z + \delta^*)]^{a-t}[1 - \Phi(z)]^{t-1} * \varphi(z)dz$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$

and $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}dt$

The Bechhofer integral

$$BI = BI(n, a, t, c)$$

$$= t * \int_{-\infty}^{\infty} [\Phi(z + c\sqrt{n})]^{a-t}[1 - \Phi(z)]^{t-1}\varphi(z)dz$$

Calculated in *R* using "integrate".

The Bechhofer integral is monotone increasing in n and tends to 1 as $n \to \infty$.
This makes it possible to compute the smallest sample size for which

$$P_C \geq BI \geq 1 - \beta$$

Optimal sample size

Moreover, in Rasch, Verdooren and Pilz (Wiley, 2019)

the R-function "Bech.snr ($a, t, c, beta$)"

calculates optimal sample size n depending on the requirement $P_C \geq 1 - \beta$.

Bechhofer (1954) and Guiard (1994) showed that the max. bound is attained when

- all variances are known and equal
- $n_i = n$ for all $i = 1, \ldots, a$

for fixed a, t, δ

Preliminary simulation results for selecting $t > 1$ populations from $a = 30$ populations with $\beta = 0.03$, i.e. $\min . P_{CS} = 0.97$
$N = 100.000$ simulation runs

t	δ	PC_S	$time$	P_Be	$size$	n_BE
2	1.0	0.970	286	0.952	452.4	540
2	0.5	0.963	289	0.951	1769.0	2130
3	1.0	0.972	286	0.959	528.4	600
3	0.5	0.963	274	0.951	2080.1	2310

Unknown variance

For unknown σ^2: two-stage selection with $10 \leq n_0 \leq 30$ preliminary obs. and estimated σ^2 from ANOVA
Gupta (1956): Select subset M_G of G of random size such that

$$P(G_2 \subset \mathbf{M}_G) \geq 1 - \beta$$

Gupta and Panchapakesan (1970) proved that for $t > 1$, the selection problem G is solvable with selection rule

$$M_G = \{A_i : \overline{y}_{i.} \geq \overline{y}_{(a-t+1).} - \delta\}$$

and $\delta = d \cdot \frac{s}{\sqrt{n}}$ and the Bechhofer-Integral replaced by

Subset Selection Procedure

$$EBI = t * \int_{-\infty}^{\infty} \int_{-\infty}^{+\infty} [\Phi(z + dy)]^{a-t} [1 - \Phi(z)]^{t-1} \varphi(z) f_m(y) dz dy$$

where f_m is pdf of $Y = \sqrt{S^2/m}$ and $S^2 \sim \chi^2_m$ is an estimator of σ^2.

Furthermore, for the least favourable parameter constellation

$\mu_1 = \mu_2 = \ldots = \mu_a$ we have equality

$1 - \beta = EBI$

Next simulation task

Our conjecture: as shown by Rasch and Yanagida (2019), for $t = 1$, there will be no real difference to the case of known variance σ^2 for $t > 1$ too, provided that

$$a - t \geq 30,$$

To be verified by simulation study (we are just about doing this).

Alternative formulations

Meanwhile, we will look for alternative problem formulations which might be easier to tackle and/ or allow for more rigorous results:

- Multiple comparisons
- Multiple decision problems
- Decision-theoretic approaches to optimal subset selection
- Optimal multi-stage selections using Bayes-adaptive sampling as outlined in section 9.4 of Liese and Miescke (2008): Statistical Decision Theory. Springer, New York

Robustness concepts

Finally, allowing more natural robustness considerations in case of departures from normality such as

- Median, MAD-based decisions
- Quantile-based decisions.

Bayesian Alternative

For $X_k \sim N(\mu_k, \sigma_k^2); k = 1, \ldots, a$ with known σ_k^2 we introduce priors for the location parameters μ_k as conjugate priors:

$\mu_k \sim N(\mu_{0k}, \delta_k^2), \delta_k^2 > 0, \mu_{0k} \in R^1 =$ "prior guesses"

Then the posterior distributions become

$\mu_k | x_{k1}, \ldots, x_{kn_k} \sim N(\alpha_k + \beta_k T_k, \tau_k^2)$

where $T_k = \sum_{j=1}^{n_k} x_{kj}$ (sufficient statistics)

$\alpha_k = \frac{\sigma_k^2 \mu_{0k}}{\sigma_k^2 + n_k \delta_k^2}, \beta_k = \frac{\delta_k^2}{\sigma_k^2 + n_k \delta_k^2}$ and $\tau_k^2 = \sigma_k^2 \beta_k; k = 1, \ldots, a$

Bayesian Alternative cont'd

If, for some $\tau^2 < min(\sigma_k^2/n_k : k = 1, \ldots, a)$ we choose the prior variances δ_k^2 so as to satisfy

$\frac{1}{\delta_k^2} + \frac{n_k}{\sigma_k^2} = \frac{1}{\tau^2}$ for all $k = 1, \ldots, a$

then the Bayes selection rule is based on the statistic

$$S = \{\alpha_1 + \beta_1 \sum_{j=1}^{n_1} x_{1j}, \ldots, \alpha_a + \beta_a \sum_{j=1}^{n_a} x_{aj}\}$$

For selecting the population k^* with the largest mean this implies

$$k^* = \arg\max(\alpha_k + \beta_k T_k)$$

Bayesian Alternative cont'd

Specializing further, if $n_k = n, \sigma_k^2 = \sigma^2, \mu_{0k} = \mu_0$ and $\delta_k^2 = \delta^2$ for $k = 1, \ldots, a$; then the Bayes selection rule selects the population with the largest mean

$$\frac{1}{n}\sum_{j=1}^{n} x_{kj} = \frac{1}{n} T_k$$

Note: Bayes selection rule minimizes the Bayes risk w.r.t. (classical) Neyman-Pearson-0-1-loss function.

To do

Future work: Compare (generalized) Bayes rules with Gupta's rule and our two-stage rule for "real-world" numbers of populations ($a >> 30$).

Miescke and Ryan (2006) compared Gupta's rule with generalized Bayes rules for $a = 3$.

There is a need for large-scale simulation studies!

Any comment (still better: active help) is appreciated!