# Asymptotic efficiency of ranking and selection procedures for independent Gaussian populations (joint work with Or Zuk)

Royi Jacobovic

The Hebrew University of Jerusalem

*royi.jacobovic@mail.huji.ac.il*

January 22, 2018

# General model

Consider a model of $k$ populations and a statistician who wants to pinpoint the $1 \leq s \leq k - s$ populations associated with specific relative stochastic properties, e.g. highest means, smallest variances, etc.

**Definition**

By selection procedure we refer to a sampling policy and selection rule to pinpoint the target populations (with satisfactory confidence level and low sampling cost).

# Guiding questions

1. How to define a confidence criterion in this context?
2. What should be assumed over the joint distribution of the populations in order to let the user perform the selection with predefined confidence level?
3. How many samples are needed in order to perform selection with satisfactory confidence level?

# History of selection procedures

**1950-1990:** Statisticians dealt with the question of how to select stochastic populations for $k << \infty$.

**1990-present:** Motivated by the field of discrete-event simulation, industrial engineers developed selection methods for $k \approx \infty$. Recently, more applications are in the field of gene-expression data analysis.

# Contributions of this work

1. Analytical results with general selection regime, namely $s = s_k$ as $k \to \infty$.
2. Mathematical technique to derive the analytical expressions for the asymptotic efficiency of selection procedures as $k \to \infty$.
3. Generalized Siegmund-Robbins (1968) result.
4. Asymptotic comparison between the procedures of Dudewicz *et al.* (1975) and Rinott (1978).

# Homoscedastic Gaussian model with known variance

Model: $X_{ij} \sim N(\theta_i, \sigma^2); i = 1, ..., k, j = 1, ..., N$ be independent univariate Gaussian r.vs with known variance $\sigma^2 > 0$ and unknown means $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$.

Problem: How to pinpoint the $1 \leq s \leq \lfloor \frac{k}{2} \rfloor$ populations with the largest means.

Solution: Pick the $s$ populations with the highest empirical means.

# Probability of correct selection (PCS)

## Definition

If $CS^s_{k,N}$ is the event of selecting the s populations with the highest means, then we shall require that

$$\inf_{\tilde{\theta}\in\mathbb{R}^k} \mathbb{P}\{CS^s_{k,N};\tilde{\theta}\} \geq p\,.$$

where $p \in (0,1)$ is an exogenous probability reflecting the confidence level required by the statistician. In addition, any $\theta^*$ which solves the above-mentioned infimum is called a *least-favorable configuration*.

# Bechhofer's indifference-zone approach

Problem:
$\{\gamma \cdot 1_k; \gamma \in \mathbb{R}\}$ is the set of LFC's, i.e. the probability of correct selection doesnt depend on N.

Solution (Bechhofer-1954):
Let $\Delta > 0$ be known (indifference) parameter and restrict the parameter space to

$$\Theta(\Delta, k) = \{\tilde{\theta} \in \mathbb{R}^k; \tilde{\theta}_{[k-s+1]} - \tilde{\theta}_{[k-s]} \geq \Delta\}$$

where $\tilde{\theta}_{[1]} \leq \ldots \leq \tilde{\theta}_{[k]}$ are the ordered components of $\tilde{\theta}$.

# Optimal sample-size

**Definition**

The optimal sample-size $N^*_{k,s}(p)$ with respect to $p \in \left(\frac{s!(k-s)!}{k!}, 1\right)$ is the minimal N which makes the probability of correct selection to be bigger than p. Practically, ignoring a rounding error, it is determined as the solution of the following equation in $N$:

$$\inf_{\tilde{\theta} \in \Theta(\Delta, k)} \mathbb{P}\{CS^s_{k,N}; \tilde{\theta}\} = p\,.$$

# Siegmund and Robbins (1968)

**Theorem**

*For any* $p \in (0,1)$,

$$N^*_{k,s=1}(p) \sim \frac{2\sigma^2}{\Delta^2} \ln(k-1)$$

*as* $k \to \infty$.

# Generalized Siegmund-Robbins result

## Theorem

*For any $p \in (0,1)$, let $N_k^*(p) = N_{k,s_k}^*(p)$ where $(s_k)_{k \geq 1}$ is a sequence such that*

1. $1 \leq s_k \leq k - s_k$, *for every $k$ up to a finite prefix.*
2. *There exists $\bar{s} \in \mathbb{N} \cup \{\infty\}$ such that $s_k \to \bar{s}$ as $k \to \infty$.*
3. $\exists \lim_{k\to\infty} \frac{\ln(s_k)}{\ln(k-s_k)} =: C$.

*Then,*

1. $(N_k^*(p))_{k\geq 1}$ *exists up to a finite prefix.*
2. $N_k^*(p) \sim \frac{2\sigma^2(1+\sqrt{C})^2}{\Delta^2} \ln(k - s_k)$ *as $k \to \infty$.*

**!$p$ has no impact on the first order of the optimal sample-size!**

# Hetroscedastic Gaussian model with unknown variances

Consider the same model with the following adjustments:

1. There are $k+1$ populations
2. $s_k \equiv 1$, i.e. the statistician looks for the population with the highest mean.
3. The variances of the populations are unkown and might be different.

# Two-stage procedures

1. $P_E$ - the procedure of Dudewicz and Dalal (1975).
2. $P_R$ - the procedure of Rinott (1978).

Both of these procedures share the same guideline:

Stage 1:Draw $N_0$ samples from each population and compute the empirical variance of each population.

Stage 2: Draw more samplings from each population. In particular, more samplings are taken from the noisier populations. Pick the population whose weighted average is the greater ($P_R$ uses regular average while $P_E$ works with other weights).

# Asymptotic relative efficiency

Denote the sample size taken from each population in the first stage by $N_0 \geq 1$. Let $G(\cdot)$ and $g(\cdot)$ be the c.d.f. and p.d.f. of student's T distribution with $\nu = N_0 - 1$ d.f's. Dudewicz *et al.* defined a sequence $h_k^1$ which tends to infinity as $k \to \infty$ and solves the equation:

$$\int_{-\infty}^{\infty} G^k(t+h)g(t)dt = p\,.$$

Similarly, Rinott defined another sequence $h_k^2 \geq h_k^1$ which solves the equation:

$$\left[\int_{-\infty}^{\infty} G(t+h)g(t)dt\right]^k = p\,.$$

# Asymptotic relative efficiency

It can be shown that the asymptotic expected sample sizes of the abovementioned procedures are given respectively by

$$h_k^m \sum_{i=1}^{k+1} \frac{\sigma_i^2}{\Delta^2} \quad , \quad m = 1, 2\,.$$

Thus, it is plausible to determine the asymptotic relative efficiency of these procedures by the asymptotic behavior of the ratio $h_k^2/h_k^1$ as $k \to \infty$.

Remark:
On basis of numerical calculations, Rinott (1978) claimed that if $p \geq 0.75$, then the difference $h_k^2 - h_k^1$ is not big.

# Asymptotic results

## Theorem

*Let $q_p$ be the pth quantile of $\nu$-Frchet distribution and let $\gamma_\nu$ be defined as follows:*

$$\gamma_\nu = \left[ \frac{\gamma\left(\frac{\nu+1}{2}\right)}{\nu\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)} \right]^{\frac{1}{\nu}} .$$

*Then,*

1. $h_k^1 \sim \gamma_\nu q_p k^{\frac{1}{\nu}}$ *as* $k \to \infty$.
2. $h_k^2 \sim \gamma_\nu q_p (2k)^{\frac{1}{\nu}}$ *as* $k \to \infty$.

Thus, $h_k^2 - h_k^1 \to \infty$ as $k \to \infty$ and hence Rinott's numerical insight is not valid for $k >> 1$ and regardless to the value of $p$.

# Asymptotic results

A consequence of the previous results is that

$$\lim_{\nu\to\infty}\lim_{k\to\infty}\frac{h_k^2}{h_k^1} = \lim_{\nu\to\infty} 2^{\frac{1}{\nu}} = 1\,.$$

The following theorem shows that the order of limits matters, i.e. the above-mentioned convergence is not uniform.

**Theorem**

$$\lim_{k\to\infty}\lim_{\nu\to\infty}\frac{h_k^2}{h_k^1} = \sqrt{2}\,.$$

# More things we did and don't have time to talk about :)

1. Regarding the asymptotic comparison between $P_E$ and $P_R$: is there a sequence $(\nu_k)_{k\geq 1}$ for which both procedures are asymptotically equivalent up to the first order? We have shown that under two relaxions the answer is positive.

2. Numerical validation of our analytic approximations.

3. Analytical proofs which are based on extreme-value theory.

# Possible directions for further research

1. Prove/disprove our conjecture about existence of a sequence $\nu_k$ for which $P_E$ and $P_R$ are asymptotically equivalent.

2. Generalizing more selection procedures by taking $s = s_k$ as $k \to \infty$.

3. Use our mathematical technique to derive analytical asymptotic results regarding more selection procedures.

Jacobovic, R. and O. Zuk. (2017). On the asymptotic efficiency of selection procedures for independent Gaussian populations. *Electronic Journal of Statistics.* Volume **11**, Number **2**, 5375-5405.

Thank you!