

DIETER SCHOTT

Some Remarks on a Statistical Selection Procedure of Bechhofer for Expectations

ABSTRACT. Following a Bechhofer statistical selection procedure we discuss from an analytical and from a probabilistic point of view why the real function

$$F(x) = \int_{-\infty}^{+\infty} \Phi^{a-t}(z + r\sqrt{x}) \cdot t(1 - \Phi(z))^{t-1} \varphi(z) dz$$

is for $x \geq 0$ and fixed integer parameters $a > 0$, $t \in]0, a[$ as well as real parameters $r > 0$ and $\beta \in]0, 1[$ strictly monotone increasing and bounded by 1. Here φ and Φ denote the p.d.f. and c.d.f. of the standard normal distribution. Numerical procedures are described to determine the minimal natural n satisfying the inequality $F(n) \geq K$ where $0 < K < 1$. The dependence of n on the parameters a , t and r is investigated, too. Some simulation results are given and discussed for $t > 1$.

KEY WORDS. Monotone Functions, Inequalities, Selection Procedures for Expectations, Bechhofer Selection Problem, Indifference Zone Selection

Acknowledgement. I thank the referee of the journal supplying valuable suggestions for improvement.

1 Introduction

In the textbooks [2], p. 489f., and [3], p. 513f., statistical selection procedures are described to separate the t in some sense best populations from a collection of a normally distributed populations with unknown expectations and the same known variance σ^2 for a given risk $\beta \in]0, 1[$ of wrong decision. More precisely, the populations G_i ($i = 1, \dots, a$) in the collection

$$L = \{G_1, G_2, \dots, G_a\}$$

with characteristics \mathbf{u}_i are assumed to be ranked according to increasing (not decreasing) expectations (means) μ as follows:

$$(G_{(1)}, \mu_{(1)}), (G_{(2)}, \mu_{(2)}), \dots, (G_{(a)}, \mu_{(a)}) .$$

Then L is to partition into two sets

$$M = \{G_{(a)}, \dots, G_{(a-t+1)}\}, \quad N = \{G_{(a-t)}, \dots, G_{(1)}\}$$

where M is the top set with the t best populations. As a prerequisite we assume that there is a gap

$$d(M, N) = |\mu_{(a-t+1)} - \mu_{(a-t)}| \geq \delta > 0 \quad (1)$$

between the top populations and the remaining ones. The populations are selected taking stochastically independent random samples

$$(\mathbf{u}_{i1}, \mathbf{u}_{i2}, \dots, \mathbf{u}_{in})$$

of G_i ($i = 1, \dots, a$) with constant size n and ranking G_i according to their sample means $\bar{\mathbf{u}}_i$ obtaining a set \mathbf{M}_s of t populations G_i . The question is if

$$P(\mathbf{M}_s = M | d(M, N) \geq \delta) \geq 1 - \beta.$$

The number n has to be great enough such that \mathbf{M}_s satisfies this condition, but not too great to reduce the effort of ranking. Hence, we look for a minimal or at least nearly minimal such n .

Following [1] this problem is solved under the gap assumption (1) and the natural condition

$$\frac{1}{\binom{a}{t}} < 1 - \beta \quad (2)$$

if n fulfils the inequality

$$\int_{-\infty}^{+\infty} \Phi^{a-t} \left(z + \frac{\delta}{\sigma} \sqrt{n} \right) \cdot t (1 - \Phi(z))^{t-1} \varphi(z) dz \geq 1 - \beta \quad (3)$$

(see also [2], p. 494 and [3], p. 517, respectively). Here φ is as usual the probability density function (or p.d.f) and Φ the cumulative distribution function (or c.d.f.) of the standard normal distribution $N(0, 1)$. For an analytical investigation it is useful to introduce the real function

$$F(x) := \int_{-\infty}^{+\infty} \Phi^{a-t} (z + r\sqrt{x}) \cdot t (1 - \Phi(z))^{t-1} \varphi(z) dz, \quad x \geq 0 \quad (4)$$

and the constants

$$r := \frac{\delta}{\sigma}, \quad K := 1 - \beta.$$

It is easy to see that the improper integral in (4) exists and $F(x)$ is defined. Besides $F(x)$ is a real extension of the left-hand side of the inequality in (3). Hence, a reformulation of condition (3) reads

$$F(n) \geq K. \quad (5)$$

The function F in (4) is called *Bechhofer function* in the following. Observe that in the paper [5] the function $f(x) = \frac{1}{t}F(x)$ is used instead of $F(x)$.

2 Problem Analysis and Probabilistic Interpretation

The condition (1) is chosen from a practical point of view to get a reasonable separation of the two sets M and N . On the other hand: in the case $\delta = 0$, i.e. $r = 0$, we would get in (4) simply a constant $F(x) = F(0)$. Thus (5) would be fulfilled for all natural n or for no natural n .

The condition (2) has also an important practical meaning. If it is violated, then no selection process is necessary. Without sampling one could denote any of the $\binom{a}{t}$ subsets of size t by \mathbf{M}_s satisfying the above probability condition.

The condition (3) can be given a simple probabilistic interpretation. Taking the extreme cases with

$$\mu_{(a)} = \dots = \mu_{(a-t+1)} = m + \delta, \quad \mu_{(a-t)} = \dots = \mu_{(1)} = m$$

for some $m \in R$ into consideration then the probability of a correct decision is just

$$P(\mathbf{V}_M + r\sqrt{n} \geq \mathbf{V}_N) = F(n), \quad (6)$$

where

$$\mathbf{V}_M = \min_{i=a-t+1, \dots, a} \mathbf{v}_{(i)}, \quad \mathbf{V}_N = \max_{i=1, \dots, a-t} \mathbf{v}_{(i)}$$

and the $\mathbf{v}_{(i)}$ are the independently and identically distributed (i.i.d.) standard normal random variables obtained from the corresponding sample means $\bar{\mathbf{u}}_{(i)}$. Replacing $r\sqrt{n}$ by the real variable y and the statistic $\mathbf{V}_N - \mathbf{V}_M$ by \mathbf{D} this can be rewritten in the generalized form

$$G(y) = P(\mathbf{D} \leq y) = \int_{-\infty}^{+\infty} F_N(z + y) f_M(z) dz \quad (7)$$

with the functions

$$F_N(z) := \Phi^{a-t}(z), \quad f_M(z) := t(1 - \Phi(z))^{t-1} \varphi(z).$$

Here F_N is the c.d.f. of \mathbf{V}_N and f_M is the p.d.f. of \mathbf{V}_M . Besides, we have the relation

$$F(x) = G(r\sqrt{x}), \quad x \geq 0. \quad (8)$$

In [1], [2] and [3] tables for $r\sqrt{n}$ can be found for some special parameters though only for small values of a . We look for a general solution of the problem to support an extensive simulation study [4] and further experiments.

The Bechhofer problem simplifies for $t = 1$. Then an effective formula for a minimal n exists using β -quantiles of an $(a-1)$ -dimensional normal distribution (see again [2] and [3]). Hence, we treat especially the cases $t > 1$.

3 Global Properties of the Bechhofer Function

The behavior of the Bechhofer function is crucial for the solution of the Bechhofer selection problem. It is interesting to compare different approaches for proving statements about this function. We use here on the one hand basic facts of analysis and on the other hand basic facts of probability calculus given as supplements.

We start with the integrand in (4) denoted by

$$I(x, z) = I(x, z; a, t, r) := \Phi^{a-t}(z + r\sqrt{x}) \cdot t(1 - \Phi(z))^{t-1} \varphi(z), \quad (9)$$

where $a \in \mathbb{N}$, $t \in \mathbb{N}$, $t < a$, $r \in \mathbb{R}$, $r > 0$ are parameters and $z \in \mathbb{R}$, $x \in \mathbb{R}_+$ are variables. By the way, another representation is

$$I(x, z) = -\Phi^{a-t}(z + r\sqrt{x}) \cdot \frac{d}{dz} (1 - \Phi(z))^t.$$

In the paper [5] modified integrands $i(x, z) := \frac{1}{t} \cdot I(x, z)$ are plotted for $a = 10$, $t = 3$, $r = 1$ and $x = n = 0, 1, \dots, 5$ using MATLAB. These integrands turn out to be unimodal for fixed n . The global maxima $i(n, z_{max})$, and also $I(n, z_{max})$, increase and their positions z_{max} walk to the left with increasing n .

We consider the two functions

$$F(x) = F(x; a, t, r) = \int_{-\infty}^{+\infty} I(x, z) dz,$$

$$F_N(x) = F_N(x; a, t, r) = \int_{-N}^{+N} I(x, z) dz$$

where N is an appropriate positive number. The cut function F_N of the Bechhofer function F comes into play if numerical integration is used to compute the improper integral. In the paper [5] a modified cut function $f_N(x) := \frac{1}{t} \cdot F_N(x)$ is plotted for $a = 10$, $t = 3$, $r = 1$ and $N = 5$ using MATLAB. We state some general properties of these functions.

Proposition 3.1 *The Bechhofer function $F(x)$ and its cut version $F_N(x)$ are continuous, strictly monotone increasing and bounded for all $x \geq 0$ as well as smooth for all $x > 0$. The difference of both functions satisfying the relation $0 < F_N(x) < F(x)$ for all $x \geq 0$ can be made (uniformly in x) arbitrary small for sufficiently large N .*

Proof: The integrand $I(x, z)$ is composed of continuous functions and is itself continuous with respect to z and x . Therefore both $F(x)$ and $F_N(x)$ are continuous.

The integrand $I(x, z)$ is strictly monotone increasing with respect to x (for fixed z) taking the strict monotony of \sqrt{x} and the powers of $\Phi(z)$ into account. Consequently, both $F(x)$ and $F_N(x)$ are strictly monotone increasing. Further, the integrand satisfies the estimations

$$0 < I(x, z) < t [1 - \Phi(z)]^{t-1} \varphi(z) < t \varphi(z)$$

since the values of Φ are contained in the interval $]0, 1[$. Thus we have

$$0 < F_N(x) < F(x) < t \int_{-\infty}^{+\infty} \varphi(z) dz = t.$$

Hence, F and F_N are bounded (e.g. by t).

The functions in the integrand $I(x, z)$ are arbitrarily often differentiable for all arguments with the exception of \sqrt{x} where $x > 0$ is necessary. Since

$$F^{(k)}(x) = \int_{-\infty}^{+\infty} \frac{\partial^k}{\partial x^k} I(x, z) dz, \quad x > 0$$

holds for $k \in \mathbb{N}$, the smoothness follows for $x > 0$ using rules of differential calculus.

Finally we get

$$\begin{aligned} F(x) &= F_N(x) + R_N(x), \\ 0 < R_N(x) &= \int_{-\infty}^{-N} I(x, z) dz + \int_N^{+\infty} I(x, z) dz < 2t \int_N^{+\infty} \varphi(z) dz \\ &= 2t(1 - \Phi(N)) \end{aligned}$$

Consequently, the difference $R_N(x)$ of both functions can be made smaller than ε for $N > \Phi^{-1}\left(1 - \frac{\varepsilon}{2t}\right)$. ■

Supplement 3.1 The continuity and the monotony of $F(x)$ can also be derived from (7) and (8) since G is a continuous c.d.f. It follows also that $F(x)$ is even bounded by 1.

4 Local Properties of the Bechhofer Function

First we present an interesting statement for improper integration with respect to functions involving p.d.f. and c.d.f. of the normal distribution.

Lemma 4.1 *For nonnegative integers l and m it holds*

$$\begin{aligned} J_{l,m} &:= \int_{-\infty}^{+\infty} \Phi^l(z) \cdot (1 - \Phi(z))^m \varphi(z) dz = \sum_{i=0}^m \binom{m}{i} \cdot \frac{(-1)^i}{l+i+1} \\ &= \frac{m!}{(l+1) \cdot (l+2) \cdot \dots \cdot (l+m+1)} = \frac{l! \cdot m!}{(l+m+1)!} \\ &= \frac{1}{(m+1) \cdot \binom{l+m+1}{m+1}}. \end{aligned}$$

Proof: The first assertion can be shown by partial integration and the other ones by mathematical induction and simple transformations. In [6] a corresponding result is proved in detail with a more general integrand. ■

Remark 4.1 The final result in Lemma 4.1 can also be proved by applying the substitution $u = \Phi(z)$ in the integral. Then $du = \varphi(z) dz$, and the integral is reduced to the Euler beta function and gamma function, respectively. Namely, it is

$$\begin{aligned} J_{l,m} &= \int_0^1 u^l (1-u)^m du = B(l+1, m+1) = \frac{\Gamma(l+1) \cdot \Gamma(m+1)}{\Gamma(l+m+2)} \\ &= \frac{l! \cdot m!}{(l+m+1)!}. \end{aligned}$$

Proposition 4.1 *The Bechhofer function $F(x)$ starts with the value*

$$F(0) = \frac{1}{\binom{a}{t}}$$

and tends from below to

$$F(\infty) = \lim_{x \rightarrow +\infty} F(x) = 1.$$

Its range is

$$R(F) = \left[\frac{1}{\binom{a}{t}}, 1 \right].$$

Further, the ascent (gradient) of $F(x)$ starts vertical and ends horizontal, that is

$$F'(0+0) = \lim_{x \rightarrow +0} F'(x) = +\infty, \quad F'(\infty) = \lim_{x \rightarrow +\infty} F'(x) = 0.$$

Proof: We get from Lemma 4.1 putting $l = a - t$ and $m = t - 1$

$$F(0) = t \cdot J_{a-t, t-1} = t \cdot \frac{1}{t \binom{a}{t}} = \frac{1}{\binom{a}{t}}$$

and putting $l = 0$ and $m = t - 1$

$$F(\infty) = t \cdot J_{0, t-1} = t \cdot \frac{1}{t} = 1$$

since then the continuous function Φ in the integrand tends to 1. As F is strictly monotone increasing and continuous by Proposition 3.1 the limit is reached from below and the range $R(F)$ is as asserted.

Now we calculate under observation of (9)

$$\frac{\partial}{\partial x} I(x, z) = \frac{rt(a-t)}{2} \frac{\Phi^{a-t-1}(z+r\sqrt{x}) \cdot \varphi(z+r\sqrt{x})}{\sqrt{x}} \cdot (1-\Phi(z))^{t-1} \varphi(z)$$

which is defined for all $x > 0$. This implies

$$\begin{aligned} F'(0+0) &= \lim_{x \rightarrow +0} F'(x) = \lim_{x \rightarrow +0} \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} I(x, z) dz \\ &= \lim_{x \rightarrow +0} \frac{rt(a-t)}{2\sqrt{x}} \cdot \int_{-\infty}^{+\infty} \Phi^{a-t-1}(z) (1-\Phi(z))^{t-1} \cdot \varphi^2(z) dz \\ &= +\infty. \end{aligned}$$

Here we remark that the improper integral exists and is finite because

$$\begin{aligned} & \int_{-\infty}^{+\infty} \Phi^{a-t-1}(z) (1 - \Phi(z))^{t-1} \cdot \varphi^2(z) dz \\ & < \int_{-\infty}^{+\infty} \Phi^{a-t-1}(z) (1 - \Phi(z))^{t-1} \cdot \varphi(z) dz \\ & = J_{a-t-1, t-1} = \frac{1}{(a-t) \cdot \binom{a-1}{t}} \end{aligned}$$

considering $0 < \varphi(z) < 1$ for all z and Lemma 4.1. Finally, we have

$$\begin{aligned} F'(\infty) &= \lim_{x \rightarrow +\infty} \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} I(x, z) dz \\ &= \lim_{x \rightarrow +\infty} \frac{rt(a-t)}{2\sqrt{x}} \cdot \int_{-\infty}^{+\infty} 0 dz = 0. \quad \blacksquare \end{aligned}$$

Supplement 4.1 The value $F(0)$ in Proposition 4.1 is also obvious from the probabilistic point of view. Consider that F is continuous and $x = n = 0$ corresponds to the case of no sampling where each of the $\binom{a}{t}$ selections is equiprobable.

The value $F(\infty) = 1$ is intuitively clear because for sampling sizes $n \rightarrow \infty$ the means \bar{u}_i converge to the expectations μ_i for all $i \in \{1, \dots, a\}$. Another argument is based on (7) and (8). G is a c.d.f. and $F(\infty) = G(\infty) = 1$. Besides, we have also $F'(\infty) = G'(\infty) = 0$ because of

$$F'(x) = G'(r\sqrt{x}) \cdot \frac{r}{2\sqrt{x}}.$$

5 About the solution of the Bechhofer Problem

Theorem 5.1 For each constant $K \in \left] \frac{1}{\binom{a}{t}}, 1 \right[$ there is a unique argument $x = x_K > 0$ of the Bechhofer function F satisfying

$$F(x) = K.$$

This x_K can be written as $x_K = F^{-1}(K)$. The solution set of

$$F(x) \geq K$$

is given by the elements $x \geq x_K$.

Proof: The assertions are direct consequences of Proposition 3.1 and Proposition 4.1. Besides, Proposition 3.1 ensures that the inverse function of F exists. \blacksquare

Principally x_K can also be derived from the c.d.f. G given in (7) taking (8) into account. Then we get

$$x_K = \left(\frac{G^{-1}(K)}{r} \right)^2.$$

An analogue statement exists for the cut function F_N . The approximative solution $\tilde{x}_K = F_N^{-1}(K)$ of $F_N(x) = K$ satisfies $\tilde{x}_K > x_K$.

Corollary 5.1 *Under the conditions (1) and (2) there is a minimal natural $n = n_B$ such that (3) is fulfilled for all natural $n \geq n_B$.*

Proof: We consider the reformulation (5) of (3) using the Bechhofer function (4) which involves already the gap number δ from (1). Here it is $K = 1 - \beta$. If (2) holds we choose $\frac{1}{\binom{a}{t}} < K < 1$ and can apply Theorem 5.1. Then $F(x) \geq K$ is fulfilled just for all $x \geq x_K = F^{-1}(K)$ and (5) just for all natural $n \geq n_B := \lceil x_K \rceil$. ■

Remark 5.1 The Corollary shows that the number n_B indicates the minimal sampling size for the Bechhofer selection problem.

Let us now turn to practical solution methods of our problem. In statistically relevant cases condition (2) is fulfilled such that we can concentrate on condition (3), i.e. the Bechhofer function $F(x)$. Using mathematical software the p.d.f. φ and the c.d.f. Φ of the standard normal distribution are predefined or can be easily defined. We can work with natural arguments n or real arguments x to solve the Bechhofer selection problem. We start with the first possibility which is realized by a MATLAB program in [5].

Then we declare the integrand $I(n, z)$. The cut number N is chosen great enough, $K = 1 - \beta$ is determined and a numerical procedure to calculate the definite integrals $F_N(n)$ is used (trapezoidal rule, Simpson's rule or some more sophisticated method) such that a certain accuracy is realized for integration. It is not necessary to use numerical top methods since F_N has a very simple behavior. Starting with $n = 0$ the natural number n is increased step by step as long until $F_N(n)$ jumps over K (e.g. by using a while-loop). We know that $F(n)$ is than over K , too. The program should print this last $n = n_B$ as the one we looked for. Considering the numerical errors it can happen if we are not careful enough that n_B is chosen one unit too large. But this is meaningless for practical applications.

Another possibility is to replace n by real $x \geq 0$. Then F_N becomes a continuous, strongly monotone function $F_N(x)$ which takes the value K exactly once, say at $x = \tilde{x}_K \approx x_K$. Then numerical methods can be used to determine the zero of $F_N(x) - K = 0$ for sufficient large N (bisection, Newton's method or other ones). In this case we get n_B by rounding \tilde{x}_K off to the next integer. This approach will be used in this paper on the basis of MATLAB and in the simulation study [4] on the basis of the statistical software R.

6 About Simulation Results

In the paper [5] a MATLAB program is presented implementing the solution method mentioned above for natural n and using the function 'quad' for numerical quadrature (adaptive Simpson's rule). For

$$a = 30, 50, 100, 200; \quad t = 2, 3$$

$$r = 0.5, 1; \quad \beta = 0.05$$

the minimal sample sizes $n = n_B$ are calculated. The simulation shows that n_B increases for decreasing r , increasing t and increasing $a - t$ assuming the other parameters are fixed. This can also easily be seen by theoretical considerations.

Considering the expression $r\sqrt{x}$ in (4) we have

$$cr \cdot \sqrt{\frac{x}{c^2}} = r\sqrt{x}, \quad c > 0.$$

Consequently, multiplying r with the factor c means dividing $x = x_K$ by c^2 . Thus we can restrict ourselves to $r = 1$.

In [5] the relation between the calculated a and n_B is fit for $t = 2$, $r = 1$ and the mentioned values of a very well by a logarithmic term. This fact can be explained at least for large a and moderate fixed t . It is well-known that for independent standard normal random variables $\mathbf{v}_{(i)}$ ($i = 1, \dots, a - t$) the relation

$$\mathbf{V}_N = \max_{i=1, \dots, a-t} \mathbf{v}_{(i)} \approx \sqrt{2 \ln(a - t)}$$

holds as $a - t$ gets large. Using this in (6) with $r = 1$ we have

$$P(\mathbf{V}_M + \sqrt{n_B} \geq \sqrt{2 \ln a}) \approx F(n_B) \approx K$$

replacing $\ln(a - t)$ by $\ln a$ for relatively small t and taking in (5) the limit case. Since \mathbf{V}_M is fixed for fixed t we obtain

$$n_B \approx 2 \ln a + c_{t,K}.$$

with a constant $c_{t,K}$ depending only on t and K .

This asymptotic estimate can also be used with x_K instead of n_B .

By Proposition 4.1 and because of $K = 1 - \beta = 0.95$ we get for certain $a > 2$ and $t = 2$

$$F(0) = \frac{2}{a \cdot (a - 1)}, \quad F(\infty) = 1$$

and for certain $a > 3$ and $t = 3$

$$F(0) = \frac{3}{a \cdot (a - 1) \cdot (a - 2)}, \quad F(\infty) = 1.$$

Thus, the Bechhofer function starts with small or very small values $F(0)$. The critical lower bound K and the supremum $F(\infty)$ of F form a narrow strip.

As already mentioned, in this paper we use a MATLAB program based on a zero method for real arguments x to determine first $x = x_K$ and then $n = n_B$. The next tables contain the simulation results for x_K , rounded to two digits after the decimal point, and for n_B .

Table 1: case $r = 1$ and $t = 2$

a	30	50	100	200
x_K	17.71	19.36	21.50	23.59
n_B	18	20	22	24

Table 2: case $r = 1$ and $t = 3$

a	30	50	100	200
x_K	19.19	20.93	23.18	25.35
n_B	20	21	24	26

If we multiply x_K with 4 and round off to the next integer we get the corresponding results for $r = 0.5$ instead of $r = 1$ (compare results in [5]). If we consider the difference $x_K(a) - 2 \ln a$ for the given a , then it increases only slowly. This seems to manifest the above derived asymptotic estimate. But for a more accurate analysis we would need values $x_K(a)$ for still greater a .

References

- [1] **Bechhofer, R. E.** : *A Single Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances.* Ann. Math. Statist. 25, 16–39 (1954)
- [2] **Rasch, D., and Schott, D.** : *Mathematische Statistik für Mathematiker, Natur- und Ingenieurwissenschaftler.* Wiley-VCH 2016
- [3] **Rasch, D., and Schott, D.** : *Mathematical Statistics.* Wiley 2018
- [4] **Rasch, D., Takuya, Y., Schott, D., and Pilz, J.** : *Statistical Selection Procedures for Expectations – a Review and Recent Simulation Results.* Paper for the 10th International Workshop on Simulation and Statistics, Salzburg 2019
- [5] **Schott, D.** : *How to get in the Top Ten? An Analysis of the Bechhofer Selection Problem in Statistics.* In: Proceedings of the 1st Northern-Light Symposium, Hafencity University Hamburg, April 2018. Wismarer Frege-Reihe, Heft 02/2018, 7–22

[6] **Schott, D.** : *Monotone Functions Generated by Improper Integrals and Applications.*
To be submitted

received: May 22, 2018

Author:

Dieter Schott
Hochschule Wismar
Fakultät für Ingenieurwissenschaften
Bereich Elektrotechnik und Informatik
Philipp-Müller-Str. 14
D-23966 Wismar

e-mail: dieter.schott@hs-wismar.de