popKorn User Manual

Vik Gopal and Claudio Fuentes

July 11, 2014

Contents

1 Introduction

Given a set of p available technologies (treatments, machines, etc.), researchers must often determine which is the best, or simply rank them according to a certain pre-specified criteria. For instance, researchers may be interested in determining which treatment is most efficient in fighting a certain disease, or in ranking vehicles according to a safety standard. These types of problems are known as ranking and selection problems. Solutions and procedures have been proposed since the second half of the 20th century, starting from [1] and [2].

There are typically two formulations of the problem; the application is usually what determines which formulation is more appropriate. Suppose that we have p populations, with unknown means θ_i for $1 \leq i \leq p$. Assuming that we can obtain a sample from each population, we might have to either

1. Attempt to select the population that has the largest parameter, $\max\{\theta_1, \ldots, \theta_p\}$, and estimate its value, or

2. Select the population which gives the largest sample mean, and then estimate its population mean.

This article presents an R package for the latter (number 2). In sections **??** and **??**, we highlight the problems that arise from not taking into account the selection procedure. Section **??** outlines how poorly traditional intervals perform. Section **??**, we demonstrate another situation when a correction is carried out, but the selection is not handled correctly.

1.1 Outline of Paper

To solve the shortcomings of traditional intervals, the author in [3] explores the minimisation (over the population parameters) of an expression for the coverage probability of an interval in the case of selecting one population out of the p. More recently, in [4], the authors propose an empirical Bayes approach for constructing simultaneous confidence intervals for k selected means. The subject of this article, the popKorn package, implements the methodology developed in [5] - a frequentist approach to interval estimation when k populations are selected, $1 \leq k \leq p$.

The rest of this paper is laid out as follows. Section 2 introduces the theory needed to understand the intervals. For further details and proofs of theorems that are touched upon here, the reader is referred to the full paper [5]. Via pedagogical examples, Section 3 outlines how an experimenter will be able to use the package to carry out an analyses on his or her dataset.

2 Theory

The formal problem set-up is as follows. For $1 \leq i \leq p$, let $X_{i1}, \ldots, X_{in}$ be a random sample from a population π_i with mean θ_i and variance σ^2. Assume that the populations π_i are independent and normally distributed, so that for any i, the sample mean $X_i = n^{-1} \sum_{j=1}^{n} X_{ij}$ follows a $N(\theta_i, \sigma^2/n)$ distribution. In addition, define the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$ to be the sample X_i's placed in descending order. In other words, the order statistics satisfy $X_{(1)} \geq \ldots \geq X_{(p)}$. In this context, we wish to construct confidence intervals for the mean of the population that gives the largest sample mean in the experiment. Formally, for any $0 < \alpha < 1$ specified prior to the experiment, we aim to construct confidence intervals for $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$, based on $X_{(1)}$, such that the confidence coefficient is at least $1 - \alpha$.

In Section 1, we saw how how dramatic the failure of traditional intervals can be when applied in this context. Now, let us investigate why traditional intervals fail so. Suppose that all the populations π_i are normally distributed with the same mean θ and variance $\sigma^2 = 1$. Then for samples of size $n = 1$ we have $X_{(1)}, \ldots, X_{(p)} \sim iid\, N(\theta, 1)$. It follows that $\Pr(X_{(1)} \leq x) = \Phi^p(x - \theta)$, where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. Since the mean of the selected population $\theta_{(1)}$ is in fact θ, we obtain (in the case when we know the true variance σ^2)

$$\Pr(\theta_{(1)} \in X_{(1)} \pm z_{\alpha/2}) = \Phi^p(z_{\alpha/2}) - \Phi^p(-z_{\alpha/2})$$

In particular, when $p = 3$, we can derive that

$$\begin{aligned}
\Pr(\theta_{(1)} \in X_{(1)} \pm z_{\alpha/2}) &= (2\Phi(z_{\alpha/2}) - 1)(1 - \Phi(z_{\alpha/2}) + \Phi^2(z_\alpha/2)) \\
&= (1 - \alpha)(1 - \Phi(z_{\alpha/2}) + \Phi^2(z_\alpha/2)) \\
&< 1 - \alpha
\end{aligned}$$

because $1 - \Phi(z_{\alpha/2}) + \Phi^2(z_\alpha/2) < 1$ for any $\alpha \in (0, 1)$. It follows that the true coverage probability of the standard confidence interval is smaller than the nominal level. In general, the coverage probability maintains the nominal level for $p = 1, 2$ and then decreases as p increases.

The authors in [5] have devised a frequentist interval estimation procedure that *does* do this. Here we state two theorems from that paper, that describe their estimation procedure.

Theorem 1 (Known Variance) *Let $0 < \alpha < 1$ and for $i = 1, \dots, p$, suppose that $X_{i1}, \dots, X_{in}$ is a random sample from a $N(\theta_i, \sigma^2)$, where θ_i is unknown but σ^2 is known. Then a confidence interval for $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$ with a confidence coefficient of at least $1 - \alpha$ is given by*

$$\left(X_{(1)} - \frac{\sigma}{\sqrt{n}} c, \quad X_{(1)} + \frac{\sigma}{\sqrt{n}} \lambda c \right) \tag{1}$$

where $\lambda_0 < \lambda < 1$ for some λ_0 and the value of c satisfies

$$\Phi^p(c) - \Phi^p(-\lambda c) = 1 - \alpha$$

Theorem 2 (Unknown Variance) *Let $0 < \alpha < 1$ and for $i = 1, \dots, p$, suppose that $X_{i1}, \dots, X_{in}$ is a random sample from a $N(\theta_i, \sigma^2)$, where θ_i and σ^2 are both unknown. Consider the estimate of σ^2 given by $s^2 = [p(n-1)]^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n} (X_{ij} - X_i)^2$. Then, a confidence interval for $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$ with a confidence coefficient of at least $1 - \alpha$ is given by*

$$\left(X_{(1)} - \frac{s}{\sqrt{n}} c, \quad X_{(1)} + \frac{s}{\sqrt{n}} \lambda c \right) \tag{2}$$

where $\lambda_0 < \lambda < 1$ for some λ_0 and the value of c satisfies

$$\int_0^\infty (\Phi^p(ct) - \Phi^p(-\lambda ct)) f(t) \, \mathrm{d}t = 1 - \alpha$$

Note that f is the pdf of s/σ.

Notice that the intervals in both theorems are based on a value λ, which is between zero and 1, and a constant c. These give rise to an asymmetric interval around $X_{(1)}$, thus reducing the upward bias that is present when traditional symmetric intervals are used following a selection process. As λ controls the degree of asymmetry, these intervals will be referred to as shrinkage intervals in the remainder of the paper. The optimal values of λ and c to be used will change depending on n, p, k, and whether we know the true variance or not. The next section summarises how the optimal λ and c values are numerically obtained in `popKorn`.

2.1 Shrinkage Intervals

Confidence intervals based on a shrinkage parameter λ were derived in [5] for two classes of subproblems:

1. When the true σ^2 is known, and
2. When the true σ^2 is unknown.

Depending on which of the above cases we happen to be in, the intervals are of the form in equations (1) and (2). Before proceeding with a summary of how the intervals are computed, we introduce a little more notation. For all i, j, let

$$\Delta_{ij} = \sqrt{n}(\theta_i - \theta_j)/\sigma \tag{3}$$

and let $\mathbf{\Delta}$ represent the vector given by

$$(\Delta_{12}, \Delta_{23}, \Delta_{34}, \ldots, \Delta_{p-1,p}) \tag{4}$$

Only the vector $\mathbf{\Delta}$ (of length $p-1$) is needed in order to retrieve all individual Δ_{ij}'s, of which there are $p(p-1)/2$.

Now suppose that the variance σ^2 is known and that we are selecting $k = 1$ population from $p = 3$. As pointed out in [5], the probability of the random shrinkage interval given in equation (1) containing $\theta_{(1)}$ is given by

$$\begin{aligned}\Pr(X_{(1)} - c\sigma/\sqrt{n} \leq \theta_{(1)} \leq X_{(1)} + \lambda c\sigma/\sqrt{n}) \\ = \sum_{i=1}^{p} \Pr(X_i - c\sigma/\sqrt{n} \leq \theta_i \leq X_i + \lambda c\sigma/\sqrt{n}, X_i = X_{(1)})\end{aligned} \tag{5}$$

We shall refer to equation (5) as the true coverage probability. Consider just the first term in this equation. If we re-parametrise the θ's in terms of the Δ_{ij}'s, then the explicit expression for this first term is

$$\begin{aligned}\Pr(X_1 - c\sigma/\sqrt{n} \leq \theta_1 \leq X_1 + \lambda c\sigma/\sqrt{n}, X_1 = X_{(1)}) \\ = \int_{-\lambda c}^{c} \Phi(z - \Delta_{12})\Phi(z - \Delta_{23} - \Delta_{12})\phi(z)\,\mathrm{d}z\end{aligned}$$

Thus computing the true coverage probability is contingent on knowing the true $\mathbf{\Delta}$ values, which we do not have in practice. In order to obtain a confidence interval that achieves the nominal coverage probability, the authors minimise over this vector of $\mathbf{\Delta}$'s, set the minimal value to the $1-\alpha$ and find the values of c and λ to be used. As suggested in the two theorems above, the minimisation differs according to the value of λ. For λ closer to 1, the coverage probability is minimised at $\Delta_{ij} = 0$. For smaller values of λ, it is minimised at $\Delta_{ij} = \infty$. Hence for a fixed $0 < \lambda < 1$, the optimal c to use can be found from one of the following two equations

$$g_0(c, \lambda, n, p, k) = 1 - \alpha, \text{ when the minimum occurs at } \Delta_{ij} = 0 \tag{6}$$

$$g_\infty(c, \lambda, n, p, k) = 1 - \alpha, \text{ when the minimum occurs at } \Delta_{ij} = \infty \tag{7}$$

For instance, when $k = 1$, the explicit form of the two equations is given by

$$\begin{aligned}g_0(c, \lambda, n, p, k) &= \Phi^p(c) - \Phi^p(-\lambda c) \\ g_\infty(c, \lambda, n, p, k) &= \Phi(c) - \Phi(-\lambda c)\end{aligned}$$

To go from the equations above to the interval in equation (1), we shall also follow the approach outlined in [5]. In any particular problem, n, p and k are fixed. We then begin by evaluating the above equations over a grid of (λ, c) values. From the intersection of the values that give rise to function evaluations that are greater than $1-\alpha$, we can isolate the optimal (λ, c) for the shortest interval.

In the case of the unknown variance, the form of the g_0 and g_∞ described above changes. In this new situation, we essentially have to marginalise over the estimate of the variance when deriving the optimal values of λ and c. Once again depending on whether the minimisation is at 0 or ∞, the interval (2) is based on solving equations of the form

$$\int_0^\infty g_0(c, \lambda, n, p, k, t) f(t)\,\mathrm{d}t = 1 - \alpha, \text{ when the minimum occurs at } \Delta_{ij} = 0 \tag{8}$$

$$\int_0^\infty g_\infty(c, \lambda, n, p, k, t) f(t)\,\mathrm{d}t = 1 - \alpha, \text{ when the minimum occurs at } \Delta_{ij} = \infty \tag{9}$$

The density f corresponds to the density of s/σ, where s^2 is an estimate of the unknown σ^2. The functions in the package apply to the case when the usual ANOVA estimate of the variance is used. The density f for this case is derived in the appendix. In the package the `integrate` function within `R` is used to carry out the integration numerically.

2.2 True Coverage Probabilities

The shrinkage intervals derived in Section 2.1 are in fact conservative; the intervals are wider than they might actually need to be. With the `popKorn` package, we can investigate the use of shorter intervals.

The conservative slant of the intervals arises from the fact that we do not know the true differences between the population means. In practice, we only have estimates of this vector, from the difference between the sample means of each population. In many cases though, we have some prior knowledge regarding these $\mathbf{\Delta}$ values, or we are willing to trust the sample estimates of this vector to a certain extent. With the functions provided in the `popKorn` package, it is possible to (informally) ascertain whether it is possible to shorten the confidence intervals without losing the nominal coverage probability. Section 3 demonstrates how this can be done.

3 Usage

In this section, we demonstrate the use of the key function in the package using simple datasets. Suppose that we have $p = 10$ populations, each with $n = 10$ replicates drawn from the following set-up:

$$X_{ij} \sim N(0, 1) \text{ for } i = 1, \ldots, p, \quad j = 1, \ldots, n$$

Observe that we are in the situation where all the true population means are in fact equal. In some sense, this is the case when ignoring the selection procedure is most costly. A boxplot of the generated sample can be seen in Figure 1.

```
> p <- 10; n <- 10
> set.seed(18)
> Xmat <- matrix(rnorm(p*n), nrow=n, ncol=p)
> colnames(Xmat) <- paste("p.", 1:p, sep="")
```

3.1 Data Input Format

In order to use `popKorn`, the data should be in the form of a matrix, with n rows and p columns. Each column of the matrix contains the data for a separate population, and each row corresponds to a replicate of data for that population. Note that this implies that the design has to be balanced.

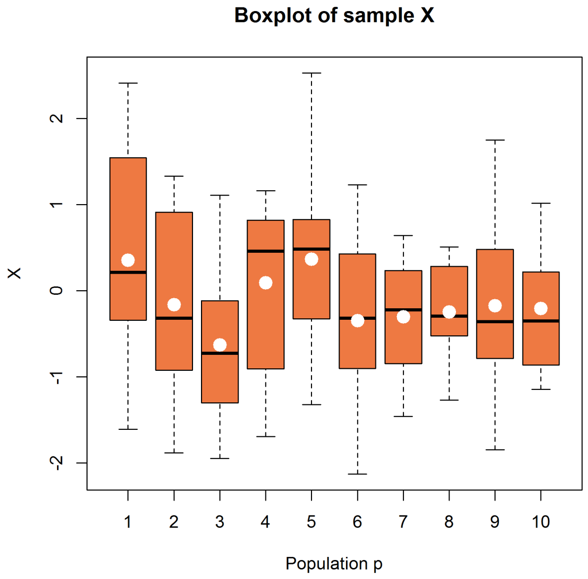

Figure 1: A boxplot of the data generated, with $p = n = 10$. The white circles indicate the mean of that sample population. In this case, the maximum sample mean was 0.368, for population number 5.

Order	1	2	3	4	5	6	7	8	9	10
Value	0.368	0.357	0.094	-0.162	-0.172	-0.204	-0.247	-0.302	-0.347	-0.631
Population	5	1	4	2	9	10	8	7	6	3

Table 1: The order index denotes the largest to the smallest sample means. Thus it was population 5 that returned the largest sample mean, and that value was 0.368.

3.2 Confidence Intervals

Table 1 contains the ordered sample means, along with the populations from which they were drawn. Thus if we selected the population with the largest sample mean, we would be forming a confidence interval for population 5.

In order to do so using the Bonferroni correction, we can call upon the `bonferroniIntervals` function within the package. The inputs to this function are the X matrix, the α-level and the number of populations to selected. Suppose we only wish to select one population. The interval for population 5 is then given to be $(-0.518, 1.253)$, which has a width of 1.770.

```
> bonferroniIntervals(Xmat, k=1)

          fit        lwr     upr
p.5 0.367693 -0.5175339 1.25292
```

Now let us derive the same interval using the methodology from [5]. The command to be used is called `asymmetricIntervals`. When compared to `bonferroniIntervals`, there is one extra argument here - the `var` argument allows an experimenter to supply a known variance to the function. The interval is reduced to $(-0.447, 0.898)$ which has a width of 1.345. Moreover, the new interval is shifted to the left to correct for the upward bias resulting from selection.

```
> asymmetricIntervals(Xmat, k=1, eps=0.05)

          fit        lwr       upr
p.5 0.367693 -0.4474376 0.8975278
```

If we wished to select more populations, the modifications to the function calls are straightforward. The following commands carry out the selection of 3 populations and create a plot of the confidence intervals. The resulting plot in Figure 2 again highlights how the shrinkage intervals are shorter and shifted downwards.

```
> asym.out <- asymmetricIntervals(Xmat, k=3)
> bonf.out <- bonferroniIntervals(Xmat, k=3)
> library(plotrix)
> plotCI(1:3, bonf.out[,1], ui=bonf.out[,"upr"], li=bonf.out[,"lwr"],
+   col="red", scol="darkgreen", slty=2, lwd=2, pch=20, cex=1.8, xaxt='n',
+   xlab="", ylab="")
> plotCI(1:3, add=TRUE, asym.out[,1], ui=asym.out[,"upr"], li=asym.out[,"lwr"],
+   scol="darkorange1", lwd=2, pch=NA, sfrac=0.02, gap=0.02)
> title("Bonferroni vs. Asymmetric Intervals")
> axis(side=1, at=1:3, labels=rownames(asym.out))
```

3.3 Workhorse Functions

The functions that do the work behind the scenes are `optimalLambda` and `optimalC`. These are the functions that carry out the numerical optimisation described in 2.1.

For instance, for the configuration in the previous section, $n = 10$ and $p = 10$ and $k = 3$ populations were selected. The optimal λ and c to use, if the true variance is unknown, can be returned with the following command.

```
> optimalLambdaC(alpha=0.05, n=10, p=10, k=3, var.known=FALSE)

$lambda
[1] 0.8

$c.val
[1] 2.9
```

Together, these values can be used in the formula in equation (2) to compute the intervals seen in Figure 2 by hand.

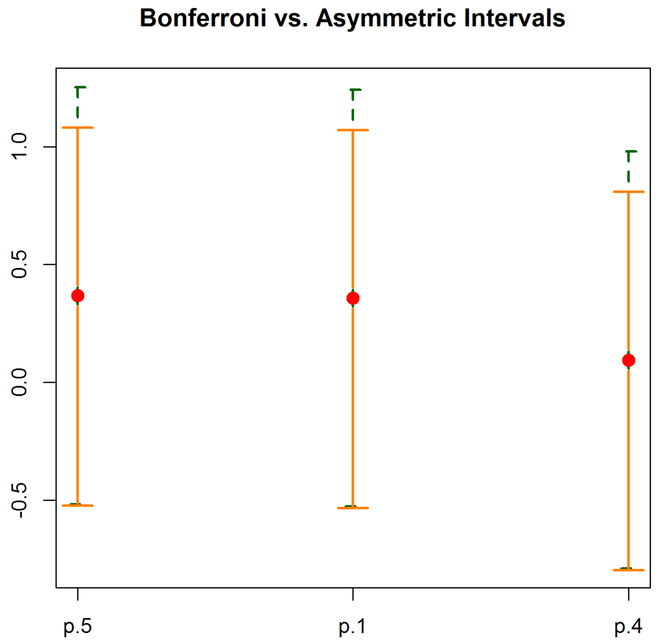

Figure 2: The error bars of the asymmetric intervals can be seen to be smaller than the Bonferroni ones. The asymmetric intervals are in orange, and the Bonferroni ones are in dark green.

3.4 A Heuristic Approach to Shortening Intervals

In this section, we outline an approach that an experimenter can take in order to shorten intervals, using information from his sample. It should be highlighted that this is not a rigorous statistical approach; it is meant as a demonstration of what is possible, and to motivate a continuing research in this area.

Suppose we have a slighly different dataset from the one in Section 3. Now, there truly is a population with a larger mean. Let us fix that population to be population 1, and for it to have a mean of 3.

```
> p <- 10; n <- 10
> set.seed(18)
> Ymat <- matrix(c(rnorm(n, 3), rnorm((p-1)*n)), nrow=n, ncol=p)
> colnames(Ymat) <- paste("p.", 1:p, sep="")
```

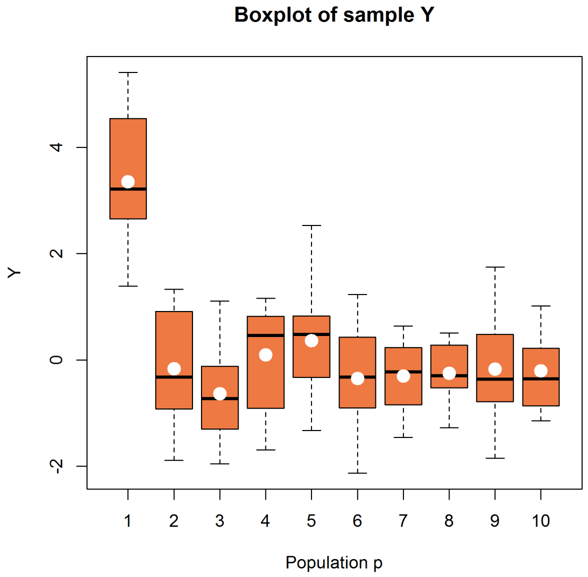

Figure 3: A boxplot of the new data generated, with $p = n = 10$. $\theta_1 = 3$, while the remaining population means are equal to 0. The white circles indicate the mean of that sample population. In this case, the maximum sample mean was 3.357, for population number 1.

The shrinkage interval upon selection of the maximum sample is given by $(2.54, 3.89)$.

```
> asymmetricIntervals(Ymat, k=1, eps=0.05)

         fit      lwr      upr
p.1 3.357259 2.542128 3.887094
```

The width is equivalent to before, but now let us try to assess how efficient this interval is. The vector of sample mean differences is obtained by

```
> sample.means <- colMeans(Ymat)
> sorted.sample.means <- sort(sample.means, dec=TRUE)
> theta.diff <- sorted.sample.means[1:(p-1)] - sorted.sample.means[2:p]
```

Suppose for a minute that this was indeed the true difference between the population means. In other words, that we had hit upon the true values of θ_i for each i, with our sample means X_i. Then we can use the function `exactCoverageProb` to compute the true coverage probability of the interval that we are using. The function is an implementation of equation (5)

```
> opt.lam.c <- optimalLambdaC(0.05, 10, 10, 1, FALSE, eps=0.05)
> exactCoverageProb(theta.diff=theta.diff, c.val=opt.lam.c$c.val,
+ lambda=opt.lam.c$lambda, sigma.2=1, n=10)
```

```
[1] 0.9647987
```

It shows that there could be a slight amount of overkill in the shrinkage interval, in that it's coverage probability could be higher than necessary for our nominal level. Let us reduce the interval width, still assuming that we have fortuitously found the true differences between the population means. The following code sets up a small grid of c values close to the optimal value found earlier, and searches through them in order to identify the right c value, that corresponds to the nominal coverage of 95%.

```
> grid.c <- seq(0.8*opt.lam.c$c.val, opt.lam.c$c.val, by=0.001)
> exactCovProbVec <- Vectorize(exactCoverageProb, vectorize.args="c.val")
> epsilon <- abs(exactCovProbVec(theta.diff=theta.diff, c.val=grid.c,
+ lambda=opt.lam.c$lambda, sigma.2=1, n=10)-0.95)
> (new.c <- grid.c[which.min(epsilon)])
```

```
[1] 2.525
```

The new interval is then equal to

$$\left(X_{(1)} - \frac{s}{\sqrt{n}} c_{new},\ X_{(1)} + \lambda c_{new} \frac{s}{\sqrt{n}} \right) = (2.581, 3.862) \tag{10}$$

The width of this interval is 1.281. Compared to the earlier interval, this one is 5% shorter.

A Computing Integrals

This section derives the density of s/σ in the case when the usual ANOVA estimate s^2 is used for σ^2. First, let us explicitly derive the expression for f, the density of s/σ in the unknown variance case. Let

$$\begin{aligned} Y &= \frac{p(n-1)s^2}{\sigma^2} \sim \chi^2_{p(n-1)} \\ X &= \frac{s}{\sigma} = \sqrt{\frac{Y}{p(n-1)}} = g(Y) \end{aligned}$$

Then we can work out that

$$y = g^{-1}(x) = p(n-1)x^2$$

and that

$$\frac{d}{dx}g^{-1}(x) = 2p(n-1)x$$

Hence the density of X is given by

$$\begin{aligned} f(x) &= 2p(n-1)x\frac{1}{\Gamma\left(\frac{p(n-1)}{2}\right)2^{0.5p(n-1)}}\left[p(n-1)x^2\right]^{0.5p(n-1)-1}e^{-0.5p(n-1)x^2} \quad (11)\\ &= \frac{2}{\Gamma\left(\frac{p(n-1)}{2}\right)2^{0.5p(n-1)}}\left[p(n-1)\right]^{0.5p(n-1)}x^{p(n-1)-1}e^{-0.5p(n-1)x^2} \quad (12)\\ &= 2 \times Gamma\left(x, \frac{p(n-1)}{2}, \frac{2}{p(n-1)}\right)x^{0.5p(n-1)}e^{0.5p(n-1)(x-x^2)} \quad (13)\end{aligned}$$

where $Gamma(x, a, b)$ denotes the gamma pdf with shape a and scale b evaluated at x. Within R, we shall evaluate the gamma density using the in-built function, and compute the rest on the log scale before exponentiating.

References

[1] Bechhofer, R., "A single-sample multiple decision procedure for ranking means of normal populations with known variances", *The Annals of Mathematical Statistics"*, 25(1), 16-39.

[2] Gupta, S. and Sobel, M., "On a statistics which arises in selection and ranking problems",*The Annals of Mathematical Statistics"*, 28(4), 957-967.

[3] Venter, J., "Estimation of the mean of the selected population", *South African Journal of Science"*, 84, 340-342.

[4] Qiu, J. and Hwang, J., "Sharp Simultaneous Intervals for the Means of Selected Populations with Application to Microarray Data Analysis", *Biometrics*, 63, 767-776.

[5] Fuentes, C., Casella, G., and Wells, M. T., "Interval estimation for the mean of the selected populations", *In progress*, 2013.