kPop: An $R$ package for the interval estimation of the mean of the selected populations

Vik Gopal$^{1,*}$, Claudio Fuentes$^2$

1. IBM Research Collaboratory, Singapore
2. Department of Statistics, Oregon State University
*Contact author: viknesh.gopal@gmail.com

Keywords: Confidence intervals, selected mean, selected populations, order statistics.

Consider the oneway balanced ANOVA experiment, where there are $p$ populations, each with mean $\theta_i$. Suppose that we can obtain a sample of size $n$ from each of these populations, and denote each sample by $X_{ij}$, where $1 \leq i \leq p$ and $1 \leq j \leq n$. The standard ANOVA assumption is that $X_{ij} \sim N(\theta_i, \sigma^2)$. Under this distributional assumption, the task of prime interest in this paper is to estimate the confidence interval of the mean of the $k$ populations that returned the largest sample averages. To be more precise, if $k = 1$, then we are interested in frequentist confidence intervals for the population mean $\theta_\ell$, chosen upon the criteria that the corresponding sample mean satisfies that $\bar{X}_\ell \geq \bar{X}_j$ for every population $j \neq \ell$.

Traditional intervals fail to maintain the nominal coverage probability as these methods are based on biased estimates and ignore the selection process. This bias becomes more pronounced as the means of the individual populations differ less. Consequently, it is important to take the selection mechanism into account in order to make correct inference.

Nowadays, this type of problems has become more relevant as researchers focus their attention on a subset of the populations under investigation. For instance, in genomic studies, researchers might be interested in identifying genes that are associated with a particular phenotype. In this context, typically thousands of genes are screened, but only a few are selected, using mechanisms such as false discovery rate (among others). Hence the inference on these selected genes has to be formally correct; otherwise the confidence coefficient would be much lower than the nominal level. In fact, the large number of ‘populations’ in the genetics setting accelerates the rate at which this confidence coefficient approaches 0.

In the $R$ package kPop, we provide a collection of functions for experimenters to obtain confidence intervals, using different methodologies, for the selected population. Apart from traditional and Bonferroni intervals, kPop provides the following intervals:

1. When $k = 1$, it is straightforward to derive symmetric confidence intervals with the desired coverage probability. The package extends this approach by including a procedure for reducing the length of the interval when the experimenter has some knowledge pertaining to the difference between the means. This approach is explained in Fuentes and Casella (2012). In addition, kPop provides bootstrap intervals for the selected population when $k = 1$, using $w$-estimators, which correct for the bias of selection while optimizing the mean square error. This general class of estimators are covered in Venter and Steel (1991) and basically correspond to weighted combinations of the order statistics of the sample means.

2. The main contribution of the package, however, is when $k > 1$. In this case, the approach in Fuentes and Casella (2012) for $k = 1$ can be generalised, thus providing a practical yet formal tool for estimating (simultaneously) the mean of several populations.

In addition, the package provides plotting functions for visualizing the different intervals. In the talk, we shall give a summary of the above techniques and their application to a dataset, which shall be included in the package kPop.
References
