Estimates for Very Small Samples

S. Morgenthaler[1]

[1] Ecole polytechnique fédérale de Lausanne (EPFL), FSB IMA, Station 8, 1015 Lausanne, Switzerland

Keywords: Robustness, H-distributions, minimax, optimality

1 Abstract

In many statistical problems, replication of measurements is too expensive and as a result only very small samples are obtained. Many analytical chemistry measurements are of this kind. In typical microarray experiments it is, for example, unusual to make more than about $k = 4$ measurements per treatment group. Even though only a few data points are available, one would like to find a good summary in the form of a central value. This is the topic of this presentation. Let $Y_1, \ldots, Y_k$ be a sample of independent observations, with $3 \leq k \leq 5$. What function $m^*(Y_1, \ldots, Y_k)$ is a good choice for summarizing the data?

If we demand location and scale equivariance of our function $m()$ as well as symmetry with regard to its arguments, the problem can be simplified by considering first the ordered sample $Y_{(1)} \leq \cdots \leq Y_{(n)}$ and the scaling it to $-1 \leq C_2 \leq \ldots \leq C_{k-1} \leq 1$, where

$$C_i = \frac{Y_{(i)} - Y_{(1)}}{Y_{(n)} - Y_{(1)}} \,.$$

In this case, we have

$$m^*(Y_1, \ldots, Y_k) = Y_{(1)} + (Y_{(n)} - Y_{(1)}) m(C_2, \ldots, C_{k-1}) \,,$$

for some appropriate function $m()$.

For $k = 3$, there is a single variable, C_2, to consider. Well-known choices of $m()$ include the average $C_2/3$, the median C_2, symmetrically trimmed or winsorized means $C_2/(3 - 6\alpha)$ for $0 \leq \alpha \leq 1/3$, the midrange 0, and so on. They all agree in the assessment that $m(0) = 0$, but disagree on what to assign to $m(\pm 1)$. Clearly, the median stakes out the most extreme position by choosing $m(\pm 1) = \pm 1$. Is there a better choice? The cases $k = 4$ and $k = 5$ can be described in an analogous manner.

In this talk, we will describe the form of optimal estimators and compare them to known forms. Our definition of optimal estimation is linked to the choice of a suitable set of distributions $F \in \mathcal{F}$ of the observations Y_i. Given such a set, one can then compute the optimal estimator, either in the minimax sense or using some other criterion. Following Morgenthaler and Tukey (2000), the set $\mathcal{F}$ is defined by simple transformations of normal variates.

Morgenthaler and Tukey (2000) study the estimation problem for $k = 5$ using as an (inverse) information measure $i(F)$ the length of a confidence intervals for the median. If there is time, we will also present their results and compare them to the more direct formulation chosen here.

References

S. Morgenthaler and J.W. Tukey (2000). Fitting Quantiles: Doubling, HR, HQ, and HHH Distributions. *J. Computational and Graphical Statistics*, 9, 180–195.