To trim or not to trim, that is the question

Pere Puig[1,2]

[1]ppuig@mat.uab.cat, Department of Mathematics, Universitat Autònoma de Barcelona
[2] Centre de Recerca Matemàtica, Campus de la UAB

In clinical settings, the trimmed mean can be used to analyze measurements of vital signs, such as heart rate, respiratory rate, and blood pressure. By removing extreme values caused by stress or other factors, the trimmed mean can provide a more accurate estimate of the patient's condition. Trimmed mean was first documented in an anonymous work in 1821:

...to determine the mean yield of a property of land, there is a custom to observe this yield during twenty consecutive years, to remove the strongest and the weakest yield and then to take one eighteenth the sum of the others. Annales de Mathématiques pures et appliquées, tome 12 (1821-1822), p. 181-204, translated by Huber in 1972.

The α-trimmed mean (TM_α) of the observations is calculated by sorting all the values, discarding $\alpha 100\%$ of the smallest and $\alpha 100\%$ of the largest values, and computing the average of the remaining values. Note that TM_0 is the sample mean and $\text{TM}_{0.5}$ is the sample median, $\text{TM}_{0.25}$ is sometimes called the sample *midmean*. Trimmed means are commonly used in many disciplines and are very useful in machine learning algorithms.

Suppose that we have a data set coming for an arbitrary unimodal symmetric distribution and we want to use an α-trimmed mean to estimate the location parameter μ. How to choose the trimming proportion α? To answer this question we characterize all symmetric distributions with smooth densities such that the α-trimmed mean is an asymptotically efficient location parameter estimator. These are composed of two families of distributions, one of which is unimodal and is referred to as the "H distribution", with density function,

$$h(x-\mu;a,b)=\frac{be^{2b}}{ac(b)}\begin{cases}\exp\left(-\frac{((x-\mu)^2+a^2)b}{a^2}\right) & |x-\mu|\le a\\ \exp\left(-\frac{2b|x-\mu|}{a}\right) & |x-\mu|>a\end{cases}, \tag{1}$$

where $c(b)=(\sqrt{b})\sqrt{\pi b}e^b+1$ and $(.)$ is the error function. The location parameter (population mean) is indicated as μ, and b is a shape parameter that directly determines the truncation proportion α of the trimmed mean from the equation $c(b)=1/(2\alpha)$. This result suggests that an α-trimmed mean will be a good choice for estimating the location parameter μ when the underlying distribution of the data will be similar to the H distribution. Therefore, we propose to fit the data with the H distribution and to estimate the trimming proportion as $\hat{\alpha}=1/(2c(\hat{b}))$ where $\hat{b}$ is the MLE of parameter b. This is an automatic procedure that can be incorporated in a machine learning algorithm. Several exemples of application will be discussed.

Keywords: Asymptotically efficient estimator; Characterization of distributions; Symmetric location models.