Distributions with a Specified Distnace between the Mean and Median
August 20, 2007

Mixture families can be found which will have a prespecified distance between the mean (μ) and the median ($\tilde{\mu}$). This is how Tom, Simon and I (2003) showed the poor behavior of the MM test; see display (5.8) of that paper. Here is a simpler example: a family of skewed contaminated normals. It too shows the nonrobustness of the MM procedure.

Let $0 < \epsilon < 1$ and $-\infty < \mu_c < \infty$. Let Z be $N(0,1)$, $I_{1-\epsilon}$ be 1 or 0 with probability $1-\epsilon$ and ϵ, respectively, and let Z and $I_{1-\epsilon}$ be independent. Take the response to be Y where

$$Y = I_{1-\epsilon}Z + (1 - I_{1-\epsilon})(Z + \mu_c).$$

Then the mean of Y is

$$\mu_Y = \epsilon\mu_c. \tag{1}$$

Also the cdf of Y is

$$F_Y(y) = (1-\epsilon)\Phi(y) + \epsilon\Phi(y - \mu_c),$$

where Φ is the cdf of a standard normal random variable. Let $\tilde{\mu}_Y$ be the median of Y. Then $\tilde{\mu}_Y$ solves the equation

$$\frac{1}{2} = (1-\epsilon)\Phi(\tilde{\mu}_Y) + \epsilon\Phi(\tilde{\mu}_Y - \mu_c). \tag{2}$$

Problem: Given $\Delta = \tilde{\mu}_Y - \mu_Y$ and ϵ, determine the median and mean of Y. Note that $\Delta = \tilde{\mu}_Y - \mu_c\epsilon$; hence, $\tilde{\mu}_Y = \Delta + \mu_c\epsilon$. Now rewrite (2) as

$$(1-\epsilon)\Phi(\Delta + \mu_c\epsilon) + \epsilon\Phi[\Delta - \mu_c(1-\epsilon)] - \frac{1}{2} = 0. \tag{3}$$

Therefore, the solution is to solve (3) for μ_c.

I have written a few simple R functions which solve this equation by Newton's method. Caution, though, on the needed initial guess. Anyways, here is an example.

Example: Set $\epsilon = 0.15$ and $\Delta = -4$. Then the solution is $\mu_c \doteq 28.15$. Hence, $\mu_Y = 4.222$ and $\tilde{\mu}_Y = 0.222$.

This example also shows the nonrobustness of the MM method. If we sample from this contaminated distribution, 85% of our data will be between ± 2 and 15% of our data will be between 26 to 30. These outliers will kill the standardization of the MM procedure. Further, if we consider the contaminated model in the article, we can kill the numerator of the MM procedure, by driving the mean with the appropriate choices of μ_L and μ_R.