*Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.*
*http://www.simfit.org.uk*

It is obvious that outliers in a sample lead to biased parameters estimates. In some instances an experimenter is able to examine the data and make a decision to eliminate certain observations, usually extremely low or high values, that indicate a systematic source of variation beyond the usual spread of observational errors. Alternatively, to avoid subjective doctoring of data, a robust method can be used which generally involves discarding extreme values and using more appropriate numerical methods that do not assume that the sample is normally distributed.

As an example, choose statistics from the main SIMFIT menu, navigate to [Data exploration] and open the option for [Robust analysis of one sample]. The results from examining the test file `robust.tf1` after trimming 10% off the extreme values are shown below, followed by the results from handling the full data set without any trimming in the exhaustive analysis procedure.

| Robust analysis | |
|---|---|
| Data: 50 N(0,1) random numbers with 5 outliers | |
| Total sample size | 50 |
| Median value | 0.2019 |
| Median absolute deviation | 1.0311 |
| Robust standard deviation | 1.5288 |
| Trimmed mean (TM) | 0.2227 |
| Variance estimate for TM | 0.0192 |
| Winsorized mean (WM) | 0.2326 |
| Variance estimate for WM | 0.0192 |
| Number of discarded values | 10 |
| Number of included values | 40 |
| Percentage of sample used | 80% (for TM and WM) |
| Hodges-Lehmann estimate (HL) | 0.2586 |

| Exhaustive analysis | |
|---|---|
| Minimum, Maximum values | -2.208, 7.000 |
| Lower and Upper Hinges | -0.829, 1.307 |
| Coefficient of skewness | 1.690 |
| Coefficient of kurtosis | 3.566 |
| Median value | 0.202 |
| Sample mean | 0.512 |
| Sample standard deviation | 1.853: CV% = 361.736% |
| Standard error of the mean | 0.262 |
| Upper 2.5% t-value | 2.010 |
| Lower 95% confidence limit for mean | -0.014 |
| Upper 95% confidence limit for mean | 1.039 |
| Variance of the sample | 3.435 |
| Lower 95% confidence limit for variance | 2.397 |
| Upper 95% confidence limit for variance | 5.335 |
| Shapiro-Wilks W statistic | 0.851 |
| Significance level for W | 0.000 Reject normality at 1% sig.level |

Clearly the exhaustive analysis indicates that the presence of outliers has created a sample that is not normally distributed and the results from robust analysis yield better estimates for the population mean and variance

which, before adding outliers, were $\mu = 0, \sigma^2 = 1$. An outline of the theory and definitions used in this robust analysis follows.

## Theory

If the sample vector is $x_1, x_2, \ldots, x_n$ the following calculations are done.

1. Using the whole sample and the inverse normal function $\Phi^{-1}(.)$, the median $M$, median absolute deviation $D$ and a robust estimate of the standard deviation $S$ are calculated as

$$\begin{aligned} M &= \text{median}(x_i) \\ D &= \text{median}(|x_i - M|) \\ S &= D/\Phi^{-1}(0.75). \end{aligned}$$

2. The percentage of the sample chosen by users to be eliminated from each of the tails is $100\alpha\%$, then the trimmed mean $TM$, and Winsorized mean $WM$, together with variance estimates $VT$ and $VW$, are calculated as follows, using $k = [\alpha n]$ as the integer part of $\alpha n$.

$$\begin{aligned} TM &= \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_i \\ WM &= \frac{1}{n} \left\{ \sum_{i=k+1}^{n-k} x_i + k x_{k+1} + k x_{n-k} \right\} \\ VT &= \frac{1}{n^2} \left\{ \sum_{i=k+1}^{n-k} (x_i - TM)^2 + k(x_{k+1} - TM)^2 + k(x_{n-k} - TM)^2 \right\} \\ VW &= \frac{1}{n^2} \left\{ \sum_{i=k+1}^{n-k} (x_i - WM)^2 + k(x_{k+1} - WM)^2 + k(x_{n-k} - WM)^2 \right\}. \end{aligned}$$

3. If the assumed sample density is symmetrical, the Hodges-Lehman location estimator $HL$ can be used to estimate the center of symmetry. This is

$$HL = \text{median} \left\{ \frac{x_i + x_j}{2}, 1 \le i \le j \le n \right\},$$

and it is calculated along with 95% confidence limit. This would be useful if the sample was a vector of differences between two samples $X$ and $Y$ for a Wilcoxon signed rank test that $X$ is distributed $F(x)$ and $Y$ is distributed $F(x - \theta)$.