Introduction to robust statistics[*]

Xuming He
National University of Singapore

To statisticians, the model, data and methodology are essential. Their job is to propose statistical procedures and evaluate how good they are.

Take a sample of size n from a Normal distribution $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2. It is well known that the "best" estimator for μ is the sample mean, the average of all n observations.

Here, given the model (Normal distribution) and data, the problem seems to be completely solved – the best estimator is found. However, unlike the mathematicians who can live in their own beautiful world of mathematics, the statisticians have to dirty their hands and face the reality.

What is the reality? Firstly, no model is exact in describing the real problem. Secondly, most classical (and optimal) statistical procedures can easily break down when some deviation from the model occurs or the data set contains just a few outliers.

A classical procedure can be shown optimal only under a series of assumptions such as Normality, linearity, symmetry, independence or finite moments. Violations of these distributional assumptions often nullify the optimality seriously. Even more dangerous is the occurrence of outliers, which may be a result of key punch errors, misplaced decimal points, recording or transmission errors, exceptional phenomena such as earthquakes or strikes, or members of a different population slipping into the sample.

For example, with just one bad point, the sample mean can go everywhere, yielding no relevant information at all. Another good example is linear regression, an important statistical tool that is routinely applied in most sciences.

[*] Written version of a colloquium talk delivered to the Department of Mathematics, National University of Singapore on 20 February 1990. The talk is intended to be nontechnical.

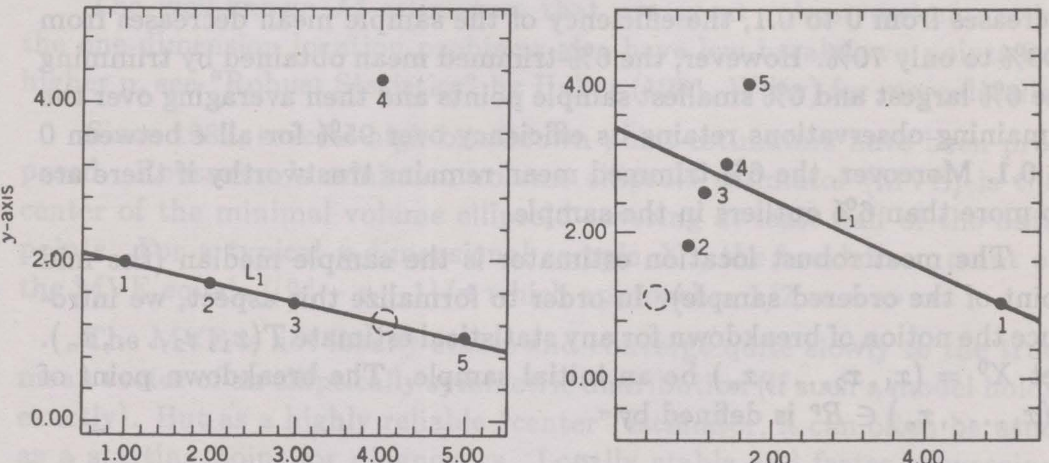
As an illustration, consider the simple model

$$y_i = a + bx_i + e_i \qquad (i = 1, 2, \ldots, n)$$

where y is the response variable (output) and x is the explanatory variable (input), $e_1, e_2, \ldots, e_n$ are independent random errors. Given observations (x_i, y_i) for $i = 1, 2, \ldots, n$, the classical fitting of the line $y = a + bx$ is the least squares method. We find the regression coefficients $(\hat{a}, \hat{b})$ by minimizing $\sum_{i=1}^{n} (y_i - a - bx_i)^2$ over every pair (a, b).

Dating back to 1809, Gauss introduced the Normal distribution as the error distribution for which LS is optimal, yielding a beautiful mathematical theory. Recently, people began to realize that real data do not completely satisfy the classical assumptions, which often have dramatic effects on the quality of the fitting. The LS fitting is fooled easily by a single outlier as shown in the following example.

Figure a contains five points $\{(x_i, y_i), i = 1, 2, \ldots, 5\}$ with a well fitting LS line. If we make an error in recording x_1, we obtain Figure b with one outlier in the x-direction, and its effect on the fit is so large that it actually tilts the LS line.



(a)

(b)

There are two possible ways that we can take for better and safer results. The first approach is to clean the data (e.g. removing outliers) before a classical treatment is applied. However, it is often impossible to detect outliers effectively without a robust method. What's more, the model is still not exact even after the data is cleaned.

A better approach is to develop robust procedures. A good robust procedure should have the following features.

(1) It performs well at the idealized model in terms of consistency, efficiency, etc.
(2) Small changes in the model or the data should have small effect on the result.
(3) A larger deviation should not nullify the analysis completely.
(4) It is computationally possible.

Robust procedures are actually in use long before the formal theory of robust statistics is developed by Huber in 1964.

For example, Turkey in 1960 considered the efficiency of trimmed means for a location model $\{F(x-\theta), \theta \in R\}$ with

$$F(x) = (1-\epsilon)\Phi(x) + \epsilon\Phi(x/3)$$

where $\Phi(x)$ is the distribution function of the standard Normal. $F(x)$ is a mixture of two Normal distributions as ϵ ranges from 0 to 10%. The resulting distribution is still symmetric with slightly flattened tails. As ϵ increases from 0 to 0.1, the efficiency of the sample mean decreases from 100% to only 70%. However, the 6%-trimmed mean obtained by trimming the 6% largest and 6% smallest sample points and then averaging over the remaining observations retains its efficiency over 95% for all ϵ between 0 to 0.1. Moreover, the 6%-trimmed mean remains trustworthy if there are no more than 6% outliers in the sample.

The most robust location estimator is the sample median (the mid point of the ordered sample). In order to formalize this aspect, we introduce the notion of breakdown for any statistical estimate $T(x_1, x_2, \ldots, x_n)$. Let $X^0 = (x_1, x_2, \ldots, x_n)$ be an initial sample. The breakdown point of $T(x_1, \ldots, x_n) \in R^p$ is defined by

$$\epsilon_n^*(T; X^0) = \min\{\frac{m}{n} : \sup_{X^m} \|T(X^m)\| = \infty\}$$

where X^m is obtained by replacing m out of n points in X^0 arbitrarily. In other words, it is the smallest fraction of contamination that can cause the estimator T to take on values beyond all bounds.

Since one outlier can drive the sample mean over all bounds, the breakdown point for the sample mean is $\frac{1}{n}$. (Typically, the breakdown point $\epsilon_n^*(T, X^0)$ does not depend on X^0.) The sample median has breakdown point close to $\frac{1}{2}$, since the median remains bounded as long as more than half sample points stay bounded.

For multivariate data, however, it is not all that easy to find a "median" with breakdown point as high as $\frac{1}{2}$, if we require affine equivariance of the estimator, i.e.

$$T(Ax_1 + b, \ldots, Ax_n + b) = AT(x, \ldots, x_n) + b$$

for any nonsingular $p \times p$ matrix A and vector $b \in R^p$.

Almost all known affine equivalent techniques before 1982 face the ubiquitous upper bound $1/(p+1)$ on their breakdown points. Those techniques include convex peeling and iterative trimming.

Convex peeling proceeds by removing the points on the boundary of the smallest convex hull containing all the data points, and repeating this several times until a sufficient number of points have been peeled away. Such a procedure may delete too many "good" points at the first few peelings, because each step removes $p + 1$ points from the sample, and there may be only one outlier among them.

The well-known M-estimators that are great robust estimators in the one-dimension location problems also have low breakdown points for higher p, see "Robust Statistics" by Huber (1981, Wiley) for more details.

Since 1982, several high breakdown point estimators have been proposed. Rousseeuw's minimum volume ellipsoid estimator (MVE) is the center of the minimal volume ellipsoid covering at least half of the data points. For a typical p-dimensional sample X^0, the breakdown point of the MVE equals $([\frac{n}{2}] - p + 1)/n$ which approaches $1/2$ as $n \to \infty$.

The MVE is not locally stable and converge quite slowly to the true mean vector of an elliptically symmetric distribution (if such a model holds exactly). But as a highly reliable "center" estimator, it can often be used as a starting point for diagnostics. Locally stable and faster converging estimator for multivariate location-scatter estimators are also available.

Let $K : R^+ \to [0,1]$ be a nondecreasing left continuous function with

(1) $K(0) = 1$ and $K(u)$ continuous at $u = 0$

(2) $K(u) > 0$ for $0 \leq u < c$,but $K(u) = 0$ for $u > c$ for some $c > 0$.

For an elliptically symmetric distribution $F((x-u)^T \sum^{-1}(x-u))$ with mean vector μ and scatter matrix $\sum$, an S-estimator $(\hat{\mu}_n, \hat{\sum}_n)$ is a solution (α^*, A^*) of the following minimization problem.

$$\text{minimize} \quad \det |A| \text{ subject to}$$

$$\frac{1}{n}\sum_{i=1}^{n} K((X_i-\alpha)^T A^{-1}(x_i-\alpha)) \geq 1-\epsilon$$

over all $\alpha \in R^p$ and $p \times p$ positive definite matrix A, where $1-\epsilon = E[K((X-\mu)^T \sum^{-1}(X-\mu))]$ under the model. Such an S-estimator has been shown to have the following properties :

(i) For $n(1-\epsilon) \geq p+1$, there is at least one solution (α^*, A^*) with $A^* > 0$.

(ii) Under the model, both $\hat{\mu}_n$ and $\hat{\sum}_n$ converge to μ and $\sum$ at the rate of $n^{-\frac{1}{2}}$ (with some regularity conditions).

(iii) The breakdown point of the S-estimator is ϵ, which can be chosen between 0 and $([\frac{n}{2}]-p+1)/n$.

(iv) The smooth S-estimator is also locally stable.

However, the S-estimator is computationally difficult. Finally, we should point out that a robust location-scatter estimator is itself a key to various robust techniques for multivariate data. For a further understanding of robust statistics, see "Robust statistics, The Approach Based on Influence Function" by Hampel, Rousseeuw, Ronchetti and Stahel (1986, John Wiley & Sons).