Robust statistics: A brief introduction and overview

by

Frank Hampel
E-Mail: hampel@stat.math.ethz.ch

Research Report No. 94
March 2001

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Invited talk in the Symposium "Robust Statistics and Fuzzy Techniques in Geodesy and GIS"
held in ETH Zurich, March 12-16, 2001.

Robust statistics: a brief introduction and overview

Frank Hampel

Seminar for Statistics, ETH Zurich, Switzerland
E-Mail: hampel@stat.math.ethz.ch

Abstract *The paper gives a highly condensed first introduction into robust statistics and some guidance for the interpretation of the literature, with some consideration for the uses in geodetics in the background.*

Robust statistics is the stability theory of statistical procedures. It systematically investigates the effects of deviations from modelling assumptions on known procedures and, if necessary, develops new, better procedures.

Common modelling assumptions are those of normality and of independence of the random errors. For both exist sophisticated theories with important practical consequences, and specifically normality can be replaced by any other reasonable parametric model (cf. Huber 1981, Hampel et al. 1986, and also Beran 1994).

As a simple example, let us consider n "independent" measurements of the same quantity, for example a distance. In general they will differ, namely by what we now call the "random error", and the question arises which value we should take as the "best estimate" of the unknown "true value". This question was already considered by Gauss (1821; cf. also Huber 1972), and he, noticing that he needed the unknown error distribution to answer this question, turned the problem upside down and asked for that error distribution which made a rule "generally accepted as a good one", namely the arithmetic mean, optimal (in location or shift models). This led to the normal distribution as assumed error distribution, and to a deeper justification (other than by simplicity) of the method of least squares.

Less than a hundred years later, almost everybody believed in the "dogma of normality", the mathematicians because they believed it to be an empirical fact, and the users of statistics because they believed it to be a mathematical theorem. But the central limit theorem, being a limit theorem, only suggests approximate normality under well-specified conditions in real situations; and empirical investigations, already by Bessel (1818) and later by Newcomb (1886), Jeffreys (1939) and others, show that typical error distributions of high-quality data are slightly but clearly longertailed (i.e., with higher kurtosis or standardized 4th moment) than the normal. Gauss (1821) had been careful to talk about "observations of equal accuracy", but obviously real data have different accuracy, as modeled by Newcomb (1886).

The implicit or explicit hope that under approximate (instead of exact) normality least squares would still be approximately optimal - a belief still held by many statisticians today - was thwarted by Tukey (1960; cf. also Huber 1981 or Hampel et al. 1986) who showed among other things that already under tiny deviations from normality the mean deviation - formerly much used as scale measure until the dispute between Fisher (1920) and Eddington - was better than the standard deviation, despite its efficiency loss of 12 percent under strict normality. In fact, the (avoidable) efficiency losses of least squares under high-quality data are more typically between 10 percent and 50 percent or even

100 percent than between 0 percent and 10 percent (and are even bigger for the scale estimate)(cf. Hampel et al. 1986).

So far we have talked about high-quality data obtained with greatest care and without any noticeable gross errors or blunders. But real data normally contain gross errors; for scientific routine data, not taken with utmost care, their fraction is typically between 1 percent and 10 percent (!). (Nobody is perfect. In some areas, like medicine, the fraction of gross errors can easily be above 20 percent. But there are – rarely – data sets of thousands of observations where nothing could be found wrong (C. Daniel, orally). On the other hand, even fully automatic data recordings can contain transient effects or occasional equipment failures. Cf. Hampel et al. 1986 or Hampel 1985.)

Gross errors often show themselves as outliers, but not all outliers are gross errors. Some outliers are genuine and may be the most important observations of the sample. For example, if a geodetic point seems suddenly to be in a different position, it may mean a gross error of some sort, or it may mean a shift of the underground, and some redundancy (or experience) is needed to distinguish these possibilities.

Outliers themselves – "data that don't fit the pattern set by the majority of the data" – are an illdefined concept, without clear boundaries; nevertheless they are a useful concept as long as one does not forget that there is a continuous transition to "ordinary" observations.

It is clear that a single outlier – if located sufficiently far away – can completely spoil a least squares analysis. A common reaction to this danger is (subjective or "objective") "rejection of outliers", although in principle outliers should be set aside for separate treatment.

There is a considerable literature on "rules for rejection of outliers" (Barnett & Lewis 1994), but apart from the facts that some of these rules cannot even reject one distant outlier out of 20 observations, and that the most commonly used rule (maximum studentized residual, or Grubbs's rule) can barely detect one distant outlier out of 10 (Hampel 1985), the basic philosophy given for the rules appears to be faulty, as has been stressed repeatedly by eminent statisticians. Nevertheless, one might legitimately consider the combined procedure "first reject all outliers according to some rule, then use least squares for the remaining data" as a new estimation procedure designed to prevent disasters due to distant gross errors. With "good" rejection rules which are able to find a sufficiently high fraction of distant gross errors (which have a sufficiently high "breakdown point", cf. below), this is a viable possibility; but it typically loses at least 10-20 percent efficiency compared with better robust methods for high-quality data (Hampel 1985). It is interesting to note that also subjective rejection has been investigated empirically by means of a small Monte Carlo study with 5 subjectively rejecting statisticians (Relles & Rogers 1977); the avoidable efficiency losses are again about 10-20 percent (Hampel 1985). This seems ok for fairly high, but not for highest standards.

Soon after Tukey's (1960) inspiring paper, the foundations for 3 closely related robustness theories were laid by Huber (1964), Huber (1965) and Hampel (1968). There is no space here to go into details; cf. Huber (1981) and Hampel et al. (1986). Some key concepts are: gross-error model, a rather full (semi-)"neighborhood" of a parametric model (robust statistics in principle has nothing to do with nonparametric statistics, although there are some confusing historical connections); M-estimators (or estimating equations; a slight generalization of maximum likelihood estimators, often considered under a different model); influence curve or influence function (Hampel 1974), describing the first derivative of a statistic considered as functional on the space of (empirical) probability distributions and thus allowing Taylor approximations for the local behavior of a statistic (under slight

changes of the data), with several derived numbers such as gross-error sensitivity (measuring local robustness) and the wellknown asymptotic variance (measuring local efficiency or "goodness" under the ideal model); and breakdown point, which gives the largest fraction of arbitrary gross errors tolerated before the statistic "breaks down" and becomes totally unreliable.

The breakdown point BP is thus a global robustness measure (of reliability); it is often the first and most important number to be looked at before going into the details of local robustness properties. It is also often quite simple: for the arithmetic mean it is 0, for the median it is 1/2 (slightly less than 1/2 of the data can move to infinity while the median still stays in the range of the "good" data). Among scale estimators, standard deviation, mean deviation and range all have BP = 0, while the interquartile range (difference between 3rd and 1st quartile, perhaps with a factor) has BP = 1/4. But the counterpart of the median among scale estimators is the median (absolute) deviation or "MAD" (Hampel 1968, 1974), which is the median of the absolute differences of the data from their median, and which has BP = 1/2. It is a very useful basis for scaling of M-estimators and for reliable rejection of outliers.

M-estimators can almost equivalently be described by a ρ-function (posing a minimization problem) or by its derivative, a ψ-function (yielding a (set of) implicit equation(s)), which is proportional to the influence function. In the location case, boundedness of ψ (almost) yields robustness, as in the case of the famous Huber-estimator (whose ψ is constant – linearly increasing – constant); this estimator solves a minimax problem for the gross-error model, thus being an optimal compromise for a whole neighborhood of the normal model, while still being numerically almost optimal under (fictitious) strict normality. Its ψ is monotone, so that the set of solutions of the implicit equation is unique (or at least convex), but so that distant outliers still have maximum (though bounded) influence and lead to avoidable efficiency losses of about 10-20 percent in typical cases with outliers.

To avoid these losses, one can use smoothly redescending M-estimators, such as 25A, the 2-4-8 estimator, or Tukey's biweight, with ρ being bounded and ψ continuously becoming zero for large $|x|$. They reject distant outliers completely, but not suddenly, allowing a transitional zone of increasing doubt, and are therefore much more efficient than "hard" rejection rules; they are usually about as good to clearly better than Huber-estimators, being only marginally worse in the case where Huber-estimators are optimal. They may lead to multiple solutions: this can become a problem especially in higher dimensions, where a reliable starting value is needed.

M-estimators can be generalized directly to multiple regression or linear (and nonlinear) models. Already in simple linear regression, we meet the important additional concept of "leverage points", highly influential outlying points in the design space, which always should be checked separately; they may be most informative or most detrimental (namely if they are gross errors). So-called Huber-type regression is not robust against a leverage point (with its y) moving to ∞; but so-called Mallows-type and Schweppe-type regression can tolerate a small positive fraction of gross errors if the number p of parameters (or, roughly, independent variables) is still small. Unfortunately, Maronna (1976) and others showed that for all "nice" procedures, BP $\leq \frac{1}{p}$. There are now various "high breakdown point procedures" (cf. Rousseeuw & Leroy 1987) which have nominal BP = 1/2 for all dimensions, but which typically can only be approximated and need a lot of computing power, so they, too, can only be used for rather low-dimensional parameter spaces.

The breakdown aspects for structured designs (regression, analysis of variance, electric power networks, nonlinear models...) warrant more than a single number to capture the relevant phenomena (which outliers can cause what damage?), and creative solutions for

specific models are still being asked for (cf. Hampel 2000, Mili et al. 1990, and Ruckstuhl 1997).

References

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, Wiley, N. Y.

Beran, J. (1994). *Statistics for Long-Memory Processes*, Monographs on Statistics and Applied Probability 61, Chapman & Hall, N. Y.

Bessel, F. W. (1818). *Fundamenta Astronomiae*, Nicolovius, Königsberg.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error, *Monthly Not. Roy. Astr. Soc.* **80**: 758–770. *Reprinted in Collected Papers of R. A. Fisher, ed. J. H. Bennett, Volume 1, 188–201, University of Adelaide 1971.*

Gauss, C. F. (1821). Theoria combinationis observationum erroribus minimis obnoxiae (pars prior), presented 15.2.1821. Commentationes societatis regiae scientiarium Gottingensis recentiores, *Werke*, Vol. 4, Dieterichsche Universitäts-Druckerei, 1880, pp. 1–108.

Hampel, F. (1968). *Contributions to the theory of robust estimation*, PhD thesis, University of California, Berkeley.

Hampel, F. (1974). The influence curve and its role in robust estimation, *J. Am. Statist. Assoc.* **69**: 383–393.

Hampel, F. (1985). The breakdown points of the mean combined with some rejection rules, *Technometrics* **27**: 95–107.

Hampel, F. (2000). Robust inference, *Research Report 93*, Seminar für Statistik, ETH Zürich. *To appear: Encyclopedia of Environmetrics, Wiley.*

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, N. Y.

Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Statist.* **35**: 73–101.

Huber, P. J. (1965). A robust version of the probability ratio test, *Ann. Math. Statist.* **36**: 1753–1758.

Huber, P. J. (1972). Robust statistics: A review, *Ann. Math. Statist.* **43**: 1041–1067.

Huber, P. J. (1981). *Robust Statistics*, Wiley, N. Y.

Jeffreys, H. (1939). *Theory of Probability*, Clarendon Press, Oxford. *Later editions: 1948, 1961, 1983.*

Maronna, R. A. (1976). Robust M-estimators of location and scatter, *Ann. Statist.* **4**: 51–67.

Mili, L., Phaniraj, V. and Rousseeuw, P. J. (1990). High breakdown point estimation in electric power systems, *Proceedings of the 1990 IEEE International Symposium on Electric Power Systems*, pp. 1843–1846. *New Orleans, May 1-3.*

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result, *Am. J. Math.* **8**: 343–366.

Relles, D. A. and Rogers, W. H. (1977). Statisticians are fairly robust estimators of location, *J. Am. Statist. Assoc.* **72**: 107–111.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N. Y.

Ruckstuhl, A. F. (1997). Partial breakdown in two-factor models, *Journal of Statistical Planning and Inference* **57**: 257–271. *Special Issue on Robust Statistics and Data Analysis, Part II.*

Tukey, J. W. (1960). A survey of sampling from contaminated distributions, *in* I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann (eds), *Contributions to Probability and Statistics.*, Stanford University Press, Stanford, Calif., pp. 448–485.