Working Paper

Robust interference

Author(s):
Hampel, Frank R.

Publication Date:
2000

Permanent Link:
https://doi.org/10.3929/ethz-a-004065953 →

ROBUST INFERENCE

by

Frank Hampel

Research Report No. 93
December 2000

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

ROBUST INFERENCE

Frank Hampel

Seminar für Statistik, ETH-Zentrum, CH-8092 Zürich, Switzerland

Introduction

Many scientists who are told for the first time to use a t-test, are wondering how they are supposed to know that the data are distributed according to the normal or Gaussian distribution, which according to mathematical statistics is one of the assumptions for the validity of the t-test. Statistical goodness-of-fit tests can never (statistically) prove, only disprove normality, and when there are enough data, they usually do disprove it. However, such tests can often prove "approximate" normality (to be defined in a suitable sense), in accordance with frequent informal practical experience. Now, the scientists are often told (or hope) that small deviations from normality do not matter, that the t-test is "robust" against small deviations from the mathematical assumption of normality. But this is true, roughly speaking, only for the level of the t-test; the power (and the corresponding length of confidence intervals, as well as the efficiency of the arithmetic mean) is very sensitive even to small deviations from normality. For not too small samples, there are other tests, such as the Wilcoxon- (Mann-Whitney U-) test, with a much better behavior.

If the problem considered by the scientist is a two-sample problem, there is also the assumption of equality of variances for the validity of the t-test. It can be shown that this assumption is of limited importance unless the sample sizes differ considerably. But even in recent years there have been textbooks written which suggest or require first an F-test for the equality of the variances. Only when this test comes out nonsignificant, is the application of the two-sample t-test permitted. This pseudologic ignores not only the fact that exact equality of variances can never be statistically proven, it ignores the much more important fact, known since around 1930, that already the level of chisquare- and F-tests for variances is so sensitive to tiny deviations from normality that J.W. Tukey later suggested these tests might better be used as tests of normality.

There is another assumption for the t-test, which is hardly ever discussed in the literature, but for example according to the highly experienced practical statistician Cuthbert Daniel it is the most important one: the assumption of independence. Great data analysts, such as K. Pearson, Gosset ("Student") and Jeffreys, have for a long time been aware of the nonvalidity of this assumption and the dangers arising from this fact, but only fairly recently has there been some systematic work trying to bring more light into these issues.

Overview

Basic aims and some results

Robust inference is inference which is insensitive to (smaller or larger) deviations from the assumptions under which it is derived. Some very commonly used assumptions in statistics are normality, independence, identical distributions, linearity (e.g., in regression), and stationarity of stochastic processes. As a rule, such assumptions are only approximations to reality, and the questions arise what deviations tend to occur in practice, what effects they have on known statistical procedures, and how to develop better, "more robust" procedures when deemed desirable. For answering these questions, new statistical concepts have proven useful.

The need for robust procedures was always clear to eminent practical statisticians such as Newcomb [94] [95], K. Pearson [99], "Student" [120], and Jeffreys [65] Ch. 4.4, 5.6 and 5.7). It was convincingly demonstrated by E.S. Pearson for tests and by J.W. Tukey for estimators. E.S. Pearson [96] [97] showed the nonrobustness even of the level of chisquare- and F-tests for variances; in this context, Box [11] and Box & Andersen [12] introduced the term "robust". Tukey [122], (summarizing earlier work) showed the nonrobustness of the arithmetic mean even under slight deviations from normality.

In order to assess robustness properties quantitatively, and to develop better and even in some senses optimal robust procedures, an extensive theory of robustness has been developed (and is still being worked out further), starting with Huber [55] and described largely in the books by Huber [63] and Hampel et al. [52]. The theory treats not only deviations from the model of normally or Gaussian distributed data, but in fact from any reasonable parametric model (such as two- and three-parameter Gamma distributions, binomial and Poisson distributions), except that often the necessary computer programs still need to be worked out. The concepts and methods developed (such as influence function, breakdown point, and M-estimators, see below) are not only useful for parametric models, but also in semi-parametric and nonparametric statistics. The robust methods derived under this theory also take largely care of non-identically distributed data and in particular both of the accommodation and identification of outliers (Hampel [45]), which occur almost everywhere and which are one of the most obvious reasons for the necessity of at least informally robust procedures. The avoidable efficiency losses of least squares (and other "classically" optimal procedures) compared with good robust procedures are more typically in the range of 10% - 100% than in the range of 1% - 10% (as some confusing papers and a misinterpretation of the Gauss-Markov theorem would seem to suggest, cf. below); even informal robust procedures (such as setting aside outliers "by eye"), though highly to be recommended over purely formal computer calculations without follow-up interpretation, typically lose about 10%-20% efficiency unnecessarily.

Robustness and good statistical practice

However, it should be kept in mind that good scientific questions, good observations, a good choice of models and transformations, and a thorough interpretation of the data and the results of the purely statistical part of the inference in the light of the scientific knowledge and experience available, are much more important for scientific inference than a 20% efficiency gain in an intermediate step. Thus, if no good robust computer program is available; if one is not in a routine situation, which would often make it worthwhile to develop one's own program; if the data set is not too complicated or too large to be looked at informally; and if one does not want to squeeze out the last 5-20% information in the data, it may be most expedient to use the available tools, such as least squares combined with graphical and interactive methods, in a clever, informally robust way. It could even be argued that, at present, often a good qualitative understanding of the dangers of nonrobustness, especially of outliers in somewhat complex data structures, is more important than the use of formal robust methods (cf., e.g., the informally robust but thorough data analyses in Daniel and Wood [19], Daniel [18], and Hampel [47]). On the other hand, there are situations where formal robust methods are either necessary or at least a vast improvement over any other method (cf., e.g., Mallows [76], Rocke et al. [103], Stahel et al. [115]; cf. also Kleiner et al. [69]). And sometimes the data are so expensive and precious that it may be worthwhile to develop a custom-tailored efficient robust method, including a computer program; an example is the robust randomization test for regression developed for the Swiss weather modification (hail prevention) project "Grossversuch IV" (Federer et al. [25]) (the usual randomization test, using the mean for ordering the permutations, is robust with respect to the level, but not with respect to the power).

For a discussion of the place of robustness (and statistics in general) in the field of tension between pure mathematics and scientific applications, including the "human factor", see Hampel [51].

Violation of the independence assumption

An assumption which is even deeper than the assumption of normality (or of any other parametric model distribution) and which is also commonly made in nonparametric statistics, but which is often ignored in data analysis, is the assumption of independence. The situation here is much more difficult than for deviations from some marginal model distribution. There is no usable theory for "all kinds" of deviations of a certain size; in particular, there is no minimax solution which is still nearly optimal under the ideal model. In fact, the least favorable case of serial correlations up to a certain size is the case of a constant systematic error, which is known to be hopeless once the data are in. Hence for robustness theory about deviations from independence, one has to make stronger assumptions, such as that the serial correlations tend to zero with increasing lag (but keeping the dangers of systematic errors in practice in mind). The question of typical true correlation structures of supposedly i.i.d. (independent identically distributed) data is much more important than that of typical true marginal distributions (instead of, e.g., normality), because

the techniques used as countermeasures depend much more on the former.

There seem to be very few long series of supposedly i.i.d. measurement data (say, of more than 100 observations each); but almost all of them that could be found exhibit not only highly significant correlations, but even long-range correlations, with the long-range correlation intensity parameter (Hurst parameter) being almost uniformly distributed between independence and non-stationarity (explaining the few non-significant cases); see below for more details and references. This is a surprising extension of the results by Mandelbrot and Wallis [82] that practically all geophysical time series they looked at (river flows, tree ring indices, temperatures, geophysical layers, precipitation, earthquakes, etc.) turned out to be not only weakly or short-range correlated (and thus describable by ARMA models), but strongly or long-range correlated, necessitating new probabilistic tools (such as self-similar processes, or fractional ARIMA processes) for their proper treatment.

The main consequences of long-range correlations in supposedly i.i.d. data are a bit complicated to describe. The effects are mild for point estimation, but drastic for standard errors, confidence intervals and tests for not very small samples, and they increase exponentially with the size of the data set. To cite a typical example, the true variance of the arithmetic mean of 130 observations can easily be 20 times the variance derived under the independence assumption. However, if we do not consider absolute constants, but contrasts (such as effects and interactions in ANOVA, or slopes in regression) in a well-mixed (e.g., randomized) experimental design, then the first-order effects of the correlations on the level of tests and confidence intervals cancel out asymptotically (Künsch et al. [73]). This is in full accordance with the folklore in applied statistics that ANOVA and regression (if well handled) are successful and reliable statistical techniques, while the usual standard error of the arithmetic mean estimating an absolute constant is quite rightly often shunned by physicists and others as having no meaning in reality. Thus it appears to be a remarkably wise custom to leave out the grand mean and intercept in the ANOVA table. It should be noted that even for contrasts there are still losses of efficiency and power due to the correlations, but they in turn can be greatly reduced by blocking and randomization.

Thus, the main problems due to long-range correlations arise for long data series estimating absolute constants. To resolve them, one has to estimate the long-range correlation parameter; but this cannot be done very informatively for fewer than 50-100 observations. Fortunately, for shorter series the effect of the correlations tends to be still small.

For a more detailed introduction into the problems of violation of the independence assumption, cf. Hampel et al. [52], Ch. 8.1.

Long-range correlations and time series

Long-range correlations are also important in time series which are, or have been, frequently treated as short-range correlated, such as the mean earth temperatures or the ozone concentrations in the higher atmosphere for every year. The more realistic models with long-range correlations generally lead to larger confidence intervals and less power than traditional time-series methods, making it harder to "prove"

changes due to human interference, for example; but this greater caution is only realistic, and it might help avoiding part of the ill-founded controversies about the future development of the earth. On the other hand, one should not forget the difference between statistical significance and practical relevance, and the "gray zone" between a well-fitting model and a statistically rejected model. A potential effect may be too small to be significant, but could be of great practical importance; and scientists often go for "hints" (say, P-values somewhat above 5% or 10%, or log likelihood ratios between -1 and -2) which is ok as long as they remain conscious that such effects could well be merely random phenomena, and which is often a good ("exploratory") scientific strategy although classical statistical theory, as it stands, has nothing to say about this. We should strive for good, realistic models, but we also should not forget the dangers of errors of the second kind and the limitations of statistical results.

Other problems

There has been relatively little work on small deviations from assumed linearity in regression (cf. Huber [61] and subsequent work).

Additive and innovative outliers and other robustness problems in time series are discussed extensively in the time series literature (for some early references, cf., e.g., Hampel et al. [52], Ch. 8.3 for a start).

Both topics will not be discussed anymore below.

Some clarifications and further references

Robust statistics is often mixed up with nonparametric statistics, but conceptually they are entirely different. Robust statistics belongs to parametric statistics and can be viewed as a natural further development both of classical Fisherian parametric statistics and of the parametric part of the Neyman-Pearson-Wald theory. It is the "statistics of approximate parametric models". In the background is the belief that parametric models, although usually inaccurate, are still often very useful for concise data description (data condensation) and data interpretation.

Nonparametric methods can also be considered and used in a parametric framework; they are typically robust with respect to the level, but not always with respect to power or length of confidence intervals. (Cf. also above and below.)

It is somewhat surprising that Neobayesians have so little to say about deviations from, e.g., normality (except for the use of some ad hoc and debatable "supermodels"). What is called "Bayesian robustness", is usually just robustness of inference under changes of the apriori distribution, an important new robustness problem Neobayesians have, but its solutions solve only half of the overall robustness problem. (Bayesian robustness will not be considered here. Cf., e.g., Kadane [68].)

Besides the books by Huber [63] and Hampel et al. [52], some other books on robust statistics are: Rousseeuw and Leroy [105], Staudte and Sheather [118], Stahel and Weisberg [116], Morgenthaler et al. [90]; more on the mathematical side: Rieder [102], Jurečková and Sen [67]; applied books with relevance for robustness: Mosteller and Tukey [92], Box et al. [13], Hoaglin et al. [53], Gnanadesikan [34], Box et al. [14];

on special related topics: Müller [93], Morgenthaler and Tukey [91]; on computer programs for robust statistics: Marazzi [83].

Local Robustness; The Influence Function

Small deviations from a parametric model; historical background

We shall consider mainly the simplest and best-developed case of robustness theory as an introduction to its concepts, namely that of location and scale estimation for supposedly i.i.d. (independent identically distributed) normal or Gaussian data. This is also the simplest case to which the method of least squares can be applied (resulting in the arithmetic mean and the empirical variance as estimates).

The normal law had been introduced as error distribution by Gauss [33] (cf. Huber [59]) in order to make the arithmetic mean the optimal estimator of some quantity based on several observations of equal accuracy. By the end of the 19th century, almost everybody believed in the dogma of the normal law of errors: the users of statistics because they believed it to be a mathematical theorem, and the mathematicians because they believed it to be an empirical fact. But the central limit theorem suggests at best approximate normality in reality (short of the limit at infinity); and already Bessel [10] and later Newcomb [94] noticed clear deviations from normality towards "longer tails" (higher kurtosis or standardized 4th moment) in their data.

While the mean was still uncontested, astronomers, such as Eddington, used the mean deviation (now obsolete) instead of the standard deviation (or variance) because they claimed it to be a more accurate measure of the variability of their data, according to their practical experience. But then Fisher [30] wrote a mathematical paper proving the mean deviation to lose 12% asymptotic efficiency ("wasting 12% of the data") compared with the optimal (and "sufficient") standard deviation if the data were strictly normal. He also conceded that the mean deviation would be asymptotically optimal (as maximum likelihood estimator) if the errors had a double-exponential distribution; but clearly such a distribution would be rather unrealistic. What Fisher did not know (and Eddington could not prove) was that the mean deviation is better not only for an approximate double-exponential distribution, but even for an approximate normal distribution (outside a tiny neighborhood of the normal).

It was Tukey [122] (cf. also Huber [63]) who showed that in his mixture model

$$F(x) = (1-\epsilon)\Phi(x) + \epsilon \cdot \Phi\left(\frac{x}{3}\right) \tag{1}$$

(Φ being the cumulative standard normal) which (putting $x = (y-\mu)/\sigma$) generates a location and scale model with slightly longer tails than the normal, it suffices to take $\epsilon = 0.0018$ (!) to make the mean deviation better than the standard deviation, and for $\epsilon = 0.05$ (a rather common frequency of gross errors, cf. below) it is even twice as efficient. (This example shows drastically the dangers of mathematical optimality theorems without any suitable robustness or stability considerations.)

Tukey also considered location estimators and showed that, as ϵ went from 0 to 0.1 in his mixture model, the arithmetic mean quickly loses asymptotic efficiency (always compared with an asymptotically best, e.g. maximum likelihood estimator under the corresponding mixture model), while the asymptotic efficiency of the median, which is only $2/\pi$ under strict normality ($\epsilon = 0$), slowly improves to meet that of the mean (near about 70%). On the other hand, the asymptotic efficiency of the 6%-trimmed (or truncated) mean stays above 97% over the whole range of ϵ from 0 to 0.1, so there was some first hope of finding estimators which are known to be nearly optimal even though the exact true distribution is not known in practice. (The α-trimmed mean, for $0 < \alpha < 1/2$, deletes the $\alpha \cdot n$ smallest and $\alpha \cdot n$ largest observations and takes the arithmetic mean of the remaining ones. It is not a rejection rule, but locally equivalent to a corresponding Huber-estimator, cf. below.) – Cf. also Tukey [123].

Huber's minimax approach

Huber [55] (cf. also Hampel [49]) gave the first theory of robustness. He considered the more general gross-error model or ϵ-contamination model

$$F(x) = (1-\epsilon)\Phi(x) + \epsilon \cdot H(x) \qquad (0 \le \epsilon < 1), \tag{2}$$

where H is symmetric about 0, but otherwise completely arbitrary (for asymmetry, see below); and he considered the class of M-estimators of location (also called estimating equations; generalized maximum likelihood estimators, and often, though not always, maximum likelihood estimators under a different location model) described by some suitable ρ or $\psi = \rho'$, with the estimate t being the solution of the minimization problem, or of the implicit equation (easy for the computer, and with excellent and simple explicit one-step approximations)

$$\sum \rho(x_i - t) = \min \text{ or } \sum \psi(x_i - t) = 0. \tag{3}$$

(If f is a density, then $\rho = -\log f$ and $\psi = -f'/f$ yield the MLE.)

In a fictional 2-person 0-sum game of the statistician (choosing ψ) against Nature (choosing H and hence F), with gain for Nature and loss for the statistician the asymptotic variance, which turns out to be $V(\psi, F) = \int \psi^2 dF / (\int \psi' dF)^2$, Huber asks for a saddlepoint (ψ_0, F_0) such that $V(\psi_0, F) \le V(\psi_0, F_0) \le V(\psi, F_0)$ for all ψ and F allowed. The solution (ψ_0, F_0) exists; ψ_0 is the Huber-estimator (for location) with parameter k (depending on ϵ) given by

$$\psi_0(x) = x \text{ for } |x| \le k, \text{ and } = k \cdot \text{ sign } (x) \text{ for } |x| > k; \tag{4}$$

and F_0 (Huber's least favorable distribution) has a density proportional to $\exp(-\rho_0(x))$, with $\psi_0 = \rho_0'$, that is, normal in the middle and (double-)exponential in the tails.

The Huber-estimator is a minimax solution: it minimizes the maximum asymptotic variance over all F in the gross-error model. While sometimes minimax solutions are "too pessimistic," this is not the case here: the loss of efficiency of the Huber-estimator against the arithmetic mean under strict normality is only 11% for $k = 1.0$, 4% for $k = 1.5$, and 1% for $k = 2.0$ (corresponding to $\epsilon =$

14%, 4%, and 0.8%, resp.), and a wrong choice of k in the range 1 to 2 also hardly matters. But for any $\epsilon > 0$ in the gross-error model, all Huber-estimators can be infinitely better than the arithmetic mean.

The gross-error model can be interpreted as yielding exactly normal data with probability 1 - ϵ, and gross errors (or some other, "contaminating" distribution) with the small probability ϵ (usually between 0% and 10%). The asymptotic variance is often, and also here, a very good approximation for n times the (practically intractable) actual variance of the estimator down to $n = 20$ or 10, as has been shown by Monte Carlo studies (cf. Andrews et al. [1]). The Huber-estimator can be intuitively interpreted as the mean of a modified sample (transformed by ψ) in which outliers are "brought in" towards the bulk of the data in a smooth way. The computation is done iteratively, as for most "ordinary" maximum likelihood estimators. Limiting cases for Huber-estimators (as for α-trimmed means) are arithmetic mean (for $k \to \infty$, and $\alpha \to 0$, resp.) and median (for $k \to 0$, and $\alpha \to 1/2$, resp.). Commonly used values of k are $k = 1.5$ (as default value, and for data of known medium quality), $k = 1$ (for data with relatively many gross errors), and $k = 2$ (for data of very high quality).

In practice, one usually needs a scale estimator simultaneously. By taking logarithms, Huber [55] also solved the minimax problem for scale (with an asymmetric model distribution!), which for $k > 1.1$ yields Huber's scale estimator s as solution of

$$\sum \psi^2 \left(\frac{x_i}{s} \right) / n = E_\Phi \psi^2(X) =: \beta, \tag{5}$$

where ψ is the Huber location ψ (called ψ_0 above). A locally equivalent scale estimator is the α-trimmed variance, cf. Table 6.4 in Johnson and Leone, I [66]. For simultaneous estimation of location and scale, one commonly used method (besides that using the MAD, see below) is Huber's "proposal 2": solve for t and s the system of equations

$$\sum \psi \left(\frac{x_i - t}{s} \right) = 0 \quad \text{and} \quad \sum \psi^2 \left(\frac{x_i - t}{s} \right) = n \cdot \beta \tag{6}$$

for a Huber location ψ with parameter k. (Another scale estimator very suitable in conjunction with the α-trimmed mean is the α-Winsorized variance, cf. Huber [58].)

Huber ([56] [57]; Huber & Strassen [64]) developed a second theory, or sub-branch of robustness theory, for censored likelihood ratio tests and exact finite-sample confidence intervals, using more general neighborhoods of the normal model described by Choquet-capacities of order 2 (a special kind of upper and lower probabilities). This approach may be mathematically deepest, but seems very hard to generalize and therefore plays hardly any role in applications.

Hampel's infinitesimal approach; the influence function

Hampel [39] [40] [42] (cf. Hampel et al. [52]) developed a third, also very closely related robustness theory which is more generally applicable than Huber's first (and second) theory and also contains an explicit treatment, with practical consequences, of asymmetry. Three main concepts are introduced: qualitative robustness, which is

essentially continuity of the estimator viewed as functional in the weak topology; the influence curve (IC) or influence function (IF), which describes the first derivative of the estimator, as far as existing; and the breakdown point (BP), a global robustness measure which describes how many percent gross errors are still tolerated. This corresponds to the stability aspects of, say, a bridge: small perturbations should have small effects; a first order Taylor expansion describes the small effects quantitatively (often in very good approximation); and the breakdown point tells under which load the bridge breaks down (or the estimator is totally unreliable). Thus, we can call robustness theory the stability theory of statistical procedures. The three concepts complement each other; in particular, the breakdown point often gives an indication in practice up to what contamination (say, BP/4 or even BP/2) the linear extrapolation given by the IF is still reasonably accurate.

To arrive at the influence curve, we first note that many estimators, more precisely sequences of estimators (one for every n), depend exactly (e.g., MLEs and M-estimators), approximately (e.g., α-trimmed means with $\alpha \cdot n$ rounded up to the nearest integer) or asymptotically (e.g., Bayes-estimators) only on the empirical cumulative distribution function (e.c.d.f.) F_n (e.g., not on the time sequence of the observations, which however may give valuable information on unsuspected serial correlations, see below). As F_n can be viewed as a (random) probability measure, we can then often consider our estimator as (or replace it by) a functional (a real-valued mapping) $T(F)$ defined in a natural way also for other probability distributions, for example the ideal model distributions $\{F_\theta, \theta \in \Theta\}$ of a parametric model. An example (which however is not defined for all F) is the arithmetic mean generalized to the expectation. If $T(F_\theta) \equiv \theta, \ \forall \ \theta$, that is, if at an exact model distribution T gives always the correct parameter, T is called Fisher-consistent. (This, as opposed to the asymptotic concept of consistency, is Fisher's original idea of consistency.)

Consider now a functional (estimator) T and a fixed sample $x_1, ..., x_n$, yielding an e.c.d.f. F_n. We can investigate the local changes of $T(F_n)$ by throwing in an additional observation x_{n+1} anywhere, by taking out an observation (as in the jackknife), or by moving one given observation from $-\infty$ to $+\infty$. $T(F_{n+1}) - T(F_n)$ as a function of x_{n+1} (with $F_{n+1} = (nF_n + \delta(x_{n+1}))/(n+1)$, where $\delta(x)$ is the distribution function of the point mass 1 in x) is also called a finite-sample sensitivity function. The change caused by a new observation in a fixed place x is often approximately proportional to $\frac{1}{n}$ (we can double, triple, ... the other observations, obtaining the same F_n). When T is well-behaved, we can consider the standardized change (i.e., multiplying by n) of T due to some new pointmass $\delta(x)$ of size $\epsilon = 1/(n+1)$ in a large sample from some distribution F. Passage to the limit yields the definition of the influence function

$$\mathrm{IF}(x; T, F) := \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon \cdot \delta(x)) - T(F)}{\epsilon}, \tag{7}$$

if the limit is defined for all x.

The influence function is a heuristic tool which provides a wealth of quantitative information about the local behavior of T (cf. Hampel [42]). For i.i.d. Xs with $EX = \mu$ and Var $(X) = \sigma^2$, it is $x - \mu$ for the arithmetic mean, and $(x-\mu)^2 - \sigma^2$

for the empirical variance. For M-estimators of location, given by some ψ, we get

$$\mathrm{IF}(x;\psi,F) = \psi(x) \Big/ \int \psi'(x)dF(x) \propto \psi(x), \tag{8}$$

and boundedness of ψ is almost equivalent with qualitative robustness. A monotone ψ gives a unique (or at least convex) solution for the implicit equation, but gives outliers still maximum influence; a ψ redescending to 0 for large $|x|$ rejects distant outliers completely. Examples for smooth redescenders are the 25A to 12A group, including the "2-4-8 estimator" (three-part redescenders), and Tukey's biweight estimator, but also (not quite reaching 0 with ψ) the location MLE for the Cauchy distribution (cf. Andrews et al. [1], Hampel [42]). The ψ-function of the "2-4-8 estimator" (scaled by the MAD) is

$$\psi(x) = \begin{cases} x & \text{for} \quad |x| \le 2, \\ 2 \cdot \text{ sign } (x) & \text{for} \quad 2 \le |x| \le 4, \\ (8 \cdot \text{ sign } (x) - x)/2 & \text{for} \quad 4 \le |x| \le 8, \\ 0 & \text{for} \quad |x| \ge 8. \end{cases} \tag{9}$$

Its behavior lies between that of 21A and 17A, it is thus a simple "general purpose redescender". Since IF $\propto \psi$, we can, like here, prescribe the properties of the IF except for a constant and can define custom-tailored estimators (this cannot be done for so-called R- and L-estimators, cf. Hampel [44]).

Mathematically speaking, the influence function is the set of all partial derivatives of the functional T in the direction of the point masses. For functionals, there exist several concepts of differentiation; Gâteaux, Hadamard or compact, and Fréchet derivative have been used in statistics, the Fréchet derivative being the strongest concept and formerly considered to be very rarely applicable; but the main reason for this belief seems to be the nonrobustness of most classical estimators, while at least some (if not most) smooth M-estimators are indeed Fréchet-differentiable (Clarke [15][16], Bednarski [3]; cf. also Fernholz [27]). The IF describes the derivative of a functional in whatever sense it exists.

Two different generalizations of the IF to time series are by Künsch [71] and by Martin and Yohai [86].

Gross-error sensitivity and related concepts derived from the IF; optimal robust compromises

The worst possible local effect of a gross error is described by the gross-error sensitivity $\gamma^*(T,F) = \sup_x |\mathrm{IF}(x;T,F)|$. The asymptotic variance of an estimator T at some F is usually given by $V(T,F) = \int(\mathrm{IF}(x;T,F))^2 dF(x)$. This can be seen from the approximations $T(G) - T(F) \approx \int \mathrm{IF}(x;T,F)d(G-F)(x) = \int \mathrm{IF} dG$ (because of $\int \mathrm{IF}(x;T,F)dF(x) = 0$) for G "near" F, and specifically $T(F_n) - T(F) \approx \sum \mathrm{IF}(x_i;T,F)$. There are still other useful concepts connected with the IF (local-shift sensitivity, rejection point, change-of-variance sensitivity, change-of-bias sensitivity, cf. Hampel [41] [42] and below).

The most important robustness lemma (Lemma 5 in Hampel [39]) shows how to minimize V under a bound on γ^* among all locally linear, Fisher-consistent estimators (and thus "robustify" the asymptotic Cramèr-Rao or information inequality);

it does so by putting symmetric bounds from above and below on the ψ-function (score function, $\propto$ IF) of the MLE and shifting the score function up or down so as to retain Fisher-consistency. The solutions in the normal case are again the Huber-estimators of location and scale. But contrary to Huber's minimax approach, which needs the log density of the model distribution to be concave, this approach is perfectly general and works also, for example, for location of the Cauchy distribution. The "most robust" location estimator, with minimum (positive) γ^*, is again the median med; the "most robust" scale estimator (and counterpart of the median) is the median deviation (in close analogy with mean deviation and standard deviation) or median absolute deviation

$$\text{MAD } \{x_i\} := \text{ med}_i\{|x_i - \text{med}_j\{x_j\}|\}. \tag{10}$$

The MAD has been very successfully used for the scaling of robust M-estimators of location, and as a basis for the safe rejection of outliers (see below).

As the asymptotic variance $V(T, F)$ of an estimator T under a distribution F (or its logarithm) is also a functional, we can also compute the effect of an "infinitesimal contamination" of F in any x on $V(T, F)$ (instead of on T as for the IF) and thus obtain the change-of-variance curve CVC or change-of-variance function CVF, and from it in turn its supremum (in some sense), the change-of-variance sensitivity χ^* or κ^*. (There are several slightly different definitions of both concepts around, cf. Hampel et al. [52] and the references therein.) We shall here use the approximation

$$\log V(T, G) - \log V(T, F) \approx \int CVF \; d(G - F) = \int CVF \; dG \leq \epsilon \cdot \kappa^*(T, F) \tag{11}$$

if the total variation norm $||G - F|| \leq \epsilon$, analogously to the approximation

$$|T(G) - T(F)| \approx \left| \int \text{IF} \; dG \right| \leq \epsilon \cdot \gamma^*(T, F) \tag{12}$$

which puts a bound on the estimation bias due to asymmetric contamination of size ϵ. Both approximations (linear extrapolations using 1-step Taylor expansions, a common tool everywhere in mathematics and its applications) are numerically extremely accurate for smooth robust M-estimators (in case of κ^* in some examples up to 3 decimal places for $\epsilon = 1\%$, about 2 decimal places for $\epsilon = 10\%$, and still quantitatively useful for $\epsilon = 50\%$, cf. Hampel et al. [52]). In case of the normal location model, the M-estimators minimizing $V(T, \Phi)$ under a bound for κ^* are again the class of Huber-estimators. This is the "infinitesimal imitation" (with extrapolation) of Huber's minimax approach.

Asymmetric contamination

We now can define a new paradigm which also takes asymmetry into account (Hampel et al. [52] p. 49-52). Given an amount ϵ of arbitrary gross errors, some bias due to asymmetry is unavoidable and cannot be fully eliminated from the data. About all we can do is to choose a small γ^* to keep the unknown bias small. We also want to estimate the correct parameter if the parametric model were exactly true,

that is, we want a Fisher-consistent functional (this solves the problem of "what to estimate" in the case of asymmetry); and given this, we want a small "asymptotic mean squared error" (MSE) as the simplest measure of quality. Now, using the asymptotic concepts as approximations for finite n,

$$\begin{aligned} \mathrm{MSE}(T, G; n) &= \left(\int \mathrm{IF}\ dG \right)^2 + V(T, F) \cdot \exp\left(\int \mathrm{CVF}\ dG \right) / n \\ &\leq \epsilon^2 \gamma^{*2}(T, F) + V(T, F) \exp\left(\epsilon \kappa^*(T, F)\right)/n \end{aligned} \tag{13}$$

for $||G - F|| \leq \epsilon$. The solution in the simplest case appears to be again the Huber-estimator; in general, we get at least a useful and simple upper bound for the MSE. The most remarkable feature is that the bound (and hence the solution) depends explicitly on n. As every good applied statistician (but not every statistician) knows, the bias becomes relatively more and more important with increasing n, while the variance goes to zero and eventually can be forgotten. Thus, for very large samples the median is the unique best estimator, in accordance with some folklore in applications (apart from the possibilities of sometimes eliminating part - but not all - of the bias, such as by redescending M-estimators or by estimators of the "shorth" type, cf. Andrews et al. [1]). For small samples, however, the variance is more important. To give some crude numerical indications: the optimal Huber-estimators for asymmetric contamination and for $n = 5$ have $k \approx 2.0$ for $\epsilon = 1\%$ and $k \approx 1.2$ for $\epsilon = 10\%$ which are about the same values as for symmetric contamination; but for $n = 40$, the values are $k \approx 1.8$ for $\epsilon = 1\%$ and $k \approx 0.7$ for $\epsilon = 10\%$, so especially for the larger ϵ, the (quadratic) effect of the bias is already quite noticeable. Apart from the (partly practical) question whether we always want (or have) to assume the worst possible effect of bias, we thus can find the Huber-estimator which minimizes the maximum MSE in the approximated gross-error model for every ϵ and every n. (Since the asymptotic variance is also only an - albeit good - approximation for the actual variance, the additional approximation by linear extrapolation is even more legitimate; but the main point is that the new paradigm, centering fully at the parametric model instead of partly at the least favorable distribution, and extrapolating everything from there into the full "neighborhood" considered, allows for a simple and elegant treatment of arbitrary asymmetric contamination.)

Confidence intervals and tests

Asymptotic confidence intervals and tests using robust estimators, which should be usable except for rather small samples, can be derived from the asymptotic normal distribution and the asymptotic variance, either "parametrically" by plugging the estimated parameters into the formula for the asymptotic variance under the ideal model, or "nonparametrically" by putting the observed e.c.d.f. into the formula. There are generalizations of the influence function ("test IF", "level IF", and "power IF", all three $\propto$ IF) and change-of-variance function ("change-of-efficacy function") to tests. In more complex situations, such as regression, there are a number of different (asymptotically equivalent) asymptotic formulas, and it is not always clear which are to be preferred in a given situation. For rather small samples (say, from $n = 3$ to 10), there are sometimes excellent "small sample asymptotic"

approximations (cf. Field and Hampel [28], Field and Ronchetti [29]) related to the saddlepoint method (cf. Daniels [20]).

But the problem of bias due to asymmetric gross errors or contamination arises also for confidence intervals and tests. Due to unfortunate circumstances, it has been neglected until quite recently, although first numerical investigations of asymmetry go at least back to 1971 (cf. Hampel [50]). Given an ideal model distribution mixed with an asymmetric contamination, any (robust) confidence interval devised for symmetric contamination will contain a fixed bias, and its length will shrink to zero as $n \to \infty$, hence it will soon not contain the true, ideal parameter anymore. Hence, again, we should decrease γ^* with increasing n and perhaps try to keep the bias a fraction (such as 1/4) of the standard deviation. This implies that it does not make sense to collect too many bad data for the same information. We might partly "symmetrize" the data by smoothly rejecting distant contamination (often easy to do), but the main problem is with unidentifiable (for every given n) asymmetric contamination hidden in the "flanks" of the model distribution (say, 1-2 standard deviations from the center). - For some recent work on asymmetry in regression, cf. Fraiman et al. [32]. Cf. also Samarov [108].

Higher dimensions; leverage points

The concepts of one-dimensional robustness theory can often be directly generalized to higher dimensions, such as more-parameter and higher-dimensional distributions, including multivariate statistics, and structured (non-i.i.d.) data, such as in regression (cf. Hampel et al. [52]). For example, the generalizations of "Lemma 5" minimize the trace of the asymptotic covariance matrix under a bound on a suitably defined generalized gross-error sensitivity. But there are also new aspects in more general data structures.

Consider briefly regression, both because of its importance and because it has been so extensively explored. Given a regression data set, let r_i be the residuals from any fitted model, then the least squares (LS) fit is the one that minimizes $\sum r_i^2$. This can immediately be generalized to the requirement of minimizing $\sum \rho(r_i)$ for any symmetric, nondecreasing (for $r \geq 0$), (convex or nonconvex) ρ ("Huber-type regression"). If $\rho(x)$ increases more than linearly for $|x| \to \infty$, the estimator is not robust (e.g., the estimators minimizing the L_p-norm for $p > 1$ are not robust). If ρ is convex, the set of solutions is convex; if ρ is bounded, distant outliers are (practically) rejected. If ρ is the Huber location ρ (see above), the estimator is the Huber regression estimator. The limiting case for $k \to 0$ is called L_1-regression, as a certain counterpart of the median. Cf. Huber ([59][60][62][63], Hampel et al. [52]). But already in simple regression (of y on x) a new phenomenon arises. If one x is very far away from all the others, it will dominate the fit of Huber-type regression with monotone ψ (and LS) and will have a very small residual, no matter whether its y fits the line through the other points or is an outlier. The L_1-solution will even go exactly through this (x, y)-point as soon as the mean of all xs is outside the remaining xs. More generally (and vaguely defined), an outlying point in design space (together with its y), which tends to dominate the fit (unless special precautions are taken), is called a leverage point. (For outliers, cf. also below.)

Leverage points can be good or bad; if they are proper observations, they contain a lot of information, much more than the other points, and can be extremely valuable (e.g., an isolated report on a solar eclipse in antiquity, together with many modern data). But if they are gross errors, they can completely spoil the fit to the "good" data. It is therefore of utmost importance in practice to identify and study all leverage points and try to find out whether they are trustworthy or dangerously misleading (this is mostly not a statistical problem!).

One way of checking potentially dangerous points is of course to leave them out tentatively and to compare the different solutions. Another, more elegant way is to "downweight" all leverage points automatically, and then to compare with the solution without downweighting. Two classes of such methods are popular. "Mallows-type regression" downweights all outlying xs, no matter what their ys, with a suitably decreasing function of the "robustified Mahalanobis distance" of each x from the point cloud of xs; it then computes a "weighted Huber-type regression" with these weights so that the influence of any (downweighted) leverage point stays bounded. "Schweppe-type regression" downweights leverage points only if they do not lie very close to the fit determined by the majority of points. In case of "good" leverage points, it is thus more efficient than Mallows-type regression, because it fully uses their rich information, but it is more intricate and locally less stable and may also sometimes be numerically more difficult. (By the way, Merrill and Schweppe "reinvented" robust regression for the purposes of power system control from the practical viewpoint, because they were highly dissatisfied with the bad behavior of least squares in the presence of gross errors, cf. Merrill and Schweppe [87].)

The methods just described are also called "bounded influence regression" because they put a bound on the total influence of each data point, which can be written as the product of the (ordinary) "influence of residual" and the (new) "influence of position in factor space". As a rule, their results should be compared with those of a Huber-type regression, and the differences should be interpreted. Unfortunately, these methods are well usable only for regression with rather few independent variables (x-variables, or "carriers") because of global robustness problems (with the breakdown point) otherwise (see below).

For robustness work in some other areas of statistics, cf. Künsch et al. [74]; Fellner [26], Stahel and Welsh [117].

Outliers and Global Robustness; The Breakdown Point

Outliers and gross errors

The most obvious reason for the necessity of some kind of (informal or formal) robust methods is the widespread occurrence of gross errors in real data. Gross errors, bad data or blunders are data where "something went wrong": reading, copying or transmission errors, intermittent phenomena such as thunderstorms or earthquakes, mix-up of two different experimental conditions, or inadvertent measurement of a member of a different population are some of their possible reasons; and even with

functioning fully automatic data recording equipment, there can be transient effects causing huge and undesired outliers (Hampel et al. [52] p. 26). It is clear that a single distant gross error, if left untreated, causes havoc for least squares (and other nonrobust methods). The frequency of gross errors depends of course on the reliability and care with which the data are obtained, and there are large high-quality data sets with less than 1% (including zero) gross errors; but it turns out that routine data in the sciences ("everyday" data not taken with special care) typically contain 1-10% gross errors (Hampel et al. [52]). We have to face this fact; we often cannot afford to treat every observation with very special care; and even if we could, nobody is completely perfect (there are examples with 0.01% rather hidden outliers and gross errors). Outside the exact sciences, for example in medicine, there can easily be more than 10% gross errors.

Outliers are data which "do not fit the pattern suggested by the majority of the data". This concept is an ill-defined concept, with no exact limits, but rather with a transitional zone of increasing doubt of its properness as a suspect observation moves farther and farther out. It is still a useful concept.

Outliers are not the same as gross errors. Most gross errors show themselves as outliers, but some are hidden among the "good" observations. Many outliers are gross errors, but some outliers are proper observations (e.g., from a longer-tailed distribution, cf. also the "Noah-effect" in Mandelbrot and Wallis [80]), and sometimes the outliers are the most valuable observations of the whole sample, indicating a new, unsuspected effect or discovery (so that sometimes statisticians get patents on outliers). An example is the discovery of the ozone hole over Antarctica (the data were clear outliers compared with what was previously known). It is therefore not advisable to discard ("reject") outliers without at least a second look (and thought; even gross errors can sometimes give useful information).

By the way, the widespread rumor that the Americans did not discover the ozone hole first despite their superior equipment because their "sophisticated" computers had automatically rejected the measurements as outliers, has been challenged (cf. Pukelsheim [100]); according to this article, the data had been automatically "flagged" by the computer, as needing special attention (as it should be); they were then compared with another, related series which however ran totally discordant; only then were the first data seriously doubted. But later it turned out that the second(!) series contained a serious systematic(!) error, measuring something very different from what it should, and therefore misleading the scientists.

There are two main aims in data analysis concerning outliers: still to get reliable information about the pattern of the majority of the data ("accommodation of outliers"), and identification of all outliers and suspect outliers for special treatment, in case they have something interesting to tell. The first aim is precisely one aim of robust statistics; and the residuals from a robust(!) fit are a very good basis for identification of special points (if care is also taken concerning leverage points). There is a more primitive technology of "rejection of outliers" which can (and should) be incorporated into robust statistics, but which by itself gives simple-looking solutions to rather debatable and misleadingly incomplete questions (see below); however, the simplicity of its answers is often considered a virtue.

The breakdown point

The main question in robustness theory concerning outliers is: how many percent arbitrary gross errors (or outliers) are tolerated by a statistical procedure before it can give arbitrarily misleading results? This fraction is called the "breakdown point" BP of the procedure. For example, the arithmetic mean (and LS) can be carried to $\pm\infty$ by a single arbitrary outlier, hence its BP = 0 (for the mean combined with rejection of outliers, see below); the median tolerates slightly less than 50% outliers on one side before it has nothing to do with the "good" data (its asymptotic BP = 1/2); the α-trimmed mean (see above) has (approximate and asymptotic) BP = α ($0 < \alpha < 1/2$). There are several slightly different definitions of the BP (Hampel [39][40], Hampel et al. [52]; Donoho and Huber [23]), mainly the asymptotic BP and the "lower finite-sample gross-error BP" described above; the corresponding "upper" BP gives the "smallest amount of free contamination that can carry the estimator over all bounds" and is larger by $1/n$, hence asymptotically equivalent.

The idea of considering the whole "bias curve" which connects the gross-error sensitivity (its slope at zero) with the BP (place of its nearest pole)(cf. Hampel [39]), has also been investigated more recently (cf.,e.g., Maronna et al. [85] and Berrendero and Zamar [9]).

In view of the frequency of gross errors, the BP should ordinarily be above 10%, preferably (for safety) even above 20%, while 50% is obviously the maximum for equivariant estimators and "mean" (nasty) contamination (imagine a mock sample, consisting of $n/2$ data, looking "good", anywhere else: without prior knowledge, we could not distinguish between the "good" and the "bad" half-sample). But if some (informal, not necessarily Bayesian) prior information about the "good" data is available, the BP can be higher, even = 1. Since local robust efficiency does not help us in face of global unreliability, the breakdown point is the first and most important single number robustness property, even before the pair "asymptotic variance and gross-error sensitivity γ^*", and the other properties of the IF.

For location and scale estimators, it is easy to obtain BP = 1/2. For scale, the median deviation ("MAD", see above) has BP = 1/2, while the better known interquartile range (3rd quartile minus 1st quartile) has only BP = 1/4. This difference proved to be surprisingly important in practice even when both were only auxiliary scale estimators for robust M-estimators of location under symmetric distributions, where both are locally (asymptotically) equivalent (Andrews et al. [1], cf., e.g., Ch. 7E3). M-estimators of location, scaled by the MAD (and starting with the median in case of redescenders, see above), have BP = 1/2. By contrast, Huber's proposal 2 with $k = 1.5$ has "only" BP = 26% (still high enough for most purposes).

There is much literature on so-called L-estimators (linear functions of the order statistics) and R-estimators (estimators derived from rank tests) of location, although few of them are commonly used in practice. The L-estimators have BP = 0 even if $\gamma^* < \infty$, unless the weight function is zero in two regions at either end (as is the case for α-trimmed and α-Winsorized means). The R-estimators have BP > 0 even if $\gamma^* = \infty$, as for the normal scores estimator (derived from the Fisher-Yates-Terry-Hoeffding-van der Waerden or normal scores test) which has BP = 24% and is asymptotically fully efficient under the normal model; but because of $\gamma^* = \infty$,

it becomes worse than the Hodges-Lehmann estimator (H/L) already very close to the normal (Hampel [44]). H/L (the median of all pairwise means, derived from the Wilcoxon or Mann-Whitney U-test) has BP = 29% and a bounded monotone IF and behaves rather similar to Huber-estimators; apart from the median (which is an M-, L- and R-estimator), it is the only frequently used R-estimator of location.

Rejection of outliers and robust estimation

There is a rich literature on rejection of outliers (cf., e.g., Grubbs [38]; Barnett and Lewis [2]), and a large number of rejection rules has been proposed. However, the main paradigm appears to be a test of, for example, H_o: the data are exactly normal, against H_A: two outliers on the right. The data are of course never exactly normal (but the level of the tests easily fails even with small deviations), and there are often problems of multiplicity of tests due to the frequently data-dependent choice of the alternative. (Many people will take a different test if there is only one suspect observation, or two on different sides, thereby destroying the meaning of the level and therefore of the test.) Hence the nominal level of such tests has to be taken at least with a grain of salt. But there are other problems.

Great data analysts have stressed that identification of outliers should mainly be done for reasons from the subject matter science, not for purely statistical reasons. The interpretation of the suspected outliers is important. There are even cases where each somewhat large "non-outlier" (or "almost-outlier") is entirely compatible with the "good" data, but together (and with very few more distant data) they form a clear outlying pattern demanding and obtaining a different interpretation (Hampel [47], Table 3.2). So much to identification of outliers.

The other, more clear-cut aim with outliers is bounding their (global and local) influence and rendering them harmless in order to get safe information on the majority of the data ("accommodation of outliers"). For example, how reliable (BP) and efficient (as. variance) is the combined statistical procedure: "First reject all outliers according to some rule, then use the arithmetic mean of the rest"? (This procedure, not the strict mathematical mean, is the "arithmetic mean" probably most commonly used.) The naive idea that after rejection the remaining data are perfectly normal, ignores both errors of the first and second kind of the test and is highly misleading. But for a long time, not even the asymptotic or Monte Carlo variances of combined rejection-estimation procedures under strict normality have been studied.

It turned out (Hampel [45]) that the most important information on the behavior of "rejection with subsequent estimation" is given by the breakdown point of the combined procedure, followed by the variance (or efficiency loss) under normality. These two numbers describe already to a high degree the results of a large Monte Carlo study with 10 different underlying distributions. A new, simple and very successful class of rejection rules emerged, the "Huber-type skipped means": "Reject everything outside med $\pm k \cdot$ MAD, and take the mean of the rest", with k around 5. Its BP is always 1/2, and its Monte Carlo variance (here: inverse efficiency) under normality for $n = 20$ is 1.04 for $k \approx 5.5$ ($\hat{=}$ 3.50 standard deviations under the normal) and 1.09 for $k \approx 4.5$. If the advantage of having a clear-cut (though

artificial) separation into "good" and "bad" data by a hard rejection rule is thought to outweigh the efficiency losses of all rejection rules (see below), this class (or extensions and refinements of it) appears to be the best solution.

On the other hand, there is a "rejection rule" which cannot even reject a single very distant outlier out of 20 observations (Hampel [45]), but is cited in the literature without any warning: the "studentized range" (range divided by the standard deviation of the same sample, or $(x_{(n)} - x_{(1)})/s$) for levels up to about (beyond) 5%! (The simple reason is that the ratio of range and s tends to a finite limit below the critical value for the assumed level when an outlier moves to infinity.)

Perhaps even more dangerous for statistical practice is that the "largest (or maximum) studentized residual", $\max\{|x_i - \bar{x}|\}/s$, also called "Grubbs's rule", which is probably the most widely used rejection rule (and informally shines through in residual analysis from least squares fits), has only a BP around 10% (depending on the exact level of the formal test), that means, it can barely (or not even) safely reject a single outlier out of 10 (or 2 outliers out of 20) and is thus a borderline case for practical usability.

Obviously, Dixon's rule in its most common two-sided forms has only lower finite-sample BP $= 1/n$ or $2/n$.

By the way, the often-cited "masking effect" of a second outlier on a first one can happen just when $1/n < \text{BP} < 2/n$. (And the level of the outlier test is meaningless for accommodation: some very good rejection rules have level = 50%.) See Hampel [45] for more details.

For a more recent discussion of rejection rules, see Davies and Gather [21].

Local properties of rejection rules, including subjective rejection

The influence curve of any rejection rule combined with the mean has a huge jump at the "rejection point", where it becomes zero, causing local instability and relatively large (in bad cases arbitrarily large) efficiency losses. All reliable rejection rules suffer most from contamination near the rejection point; they are best in the presence of distant contamination. On the other hand, Huber-estimators (and others with monotone IF) are best for "high quality data" distributions with somewhat elongated tails, such as a t_3-distribution. They all lose about 10-20% efficiency unnecessarily for usual amounts (5-10%) of distant contamination (because the latter is not rejected, although it could be easily identified), while all rejection rules lose at least 10-20% efficiency (often much more) unnecessarily under "high-quality data" without gross errors and clear outliers, but (as is typically the case) with somewhat "fattened flanks" and slightly elongated tails of the distribution.

A class of estimators without either efficiency losses are the "smoothly redescending" (to zero, with their ψ-function) M-estimators, such as the 25A - 12A group and Tukey's biweight (Andrews et al. [1]). They were designed to reject all clear outliers, but avoid the hard jump of rejection rules and replace it, more realistically, by a continuous transition of treating the data from "fully good" to "fully bad" (Hampel [50]). (By the way, neither R- nor L-estimators, mentioned above, can achieve this behavior, cf. Hampel [44].)

What about subjective rejection (i.e., rejection without any fixed rule or mathematical formula)? What are its efficiency losses in various situations? There has been indeed a remarkable Monte Carlo study putting 5 statisticians in front of a computer screen and letting them visually reject outliers in the samples shown, and also give a "seat-of-the-pants estimate" each time (Relles and Rogers [101]; their preceding research report also contains some nice graphs). The results, very uniformly, show again an avoidable efficiency loss of about 10-20% in the most important situations. Thus, for modest demands on efficiency, subjective rejection is fine, as long as the data can be looked at visually at all (and it has the advantage of being closer to thinking about and interpreting the outliers); but for higher demands on efficiency, it should be replaced by a smoothly redescending M-estimator.

The actual efficiency of least squares

There are still claims that least squares usually lose only a few percent (at most about 5-10%) efficiency with real data, compared with good robust estimators, although already Jeffreys [65], Ch. 5.7) had demonstrated that for 9 large high-quality data sets, presumably without gross errors, the efficiency losses of the mean are between about 10% and 50% (and for the empirical variance between 20% and 100%(!)). The findings by Jeffreys were discussed by Cox and Hinkley [17] in such a way as to suggest (without ever claiming so!) efficiency losses of at most (instead of at least) 10%. Another confusing paper was Stigler's [119] comparison of estimates derived from some old high quality sets of measurements of physical constants with their modern values. It is well-known among physicists that all such measurements contain systematic (and "semi-systematic", cf. below) errors which change over time (cf., e.g., Youden [129]), hence typically the modern value was on one side of most good robust estimates; so was often the mean because of its high variability, hence it was sometimes very good and sometimes very bad and on the average, because of the linear comparison made, hardly worse than the (accidentally) "best" robust estimator. (Stigler's theorem in the rejoinder was wrongly applied; cf. also Hampel et al. [52], p. 31f.) - A common "justification" of LS is the Gauss-Markov theorem; but this tells only that LS is optimal among all **linear** unbiased estimators; and as was already demonstrated by Fisher [31], **all linear estimators** are very bad outside a very small region of Pearson curves around the normal distribution. (Rejection of outliers is nonlinear; so are most maximum likelihood estimators.)

Linearity still has the great virtue of simplicity; and sometimes one is lucky with the efficiency of LS with high-quality data, sometimes not. Sometimes the same authors find examples for both situations (Spjøtvoll and Aastveit [111] [112]). With less reliable data, which may or do contain gross errors, no decent applied statistician will use the "arithmetic mean" (or LS) in its strict mathematical sense, but rather the mean (or LS) after (subjective or objective) "rejection of outliers"; but we have seen that with any hard rejection rule (without transition zone), including subjective rejection, typical avoidable efficiency losses are at least 10-20%. Good rejection rules (with high BP) prevent the worst and may often be sufficiently good (cf. also the remarks on 20% efficiency loss in the greater context of data analysis near the beginning); but to act as if after rejection of outliers one has an exactly

normal sample is untenable.

The breakdown point in higher dimensions and structured designs

M-estimators can be immediately generalized to multivariate statistics and multiple regression, and they retain their good local properties. However, it came as a bad surprise (related to the "curse of dimension") when Maronna ([84]; cf. also Schönholzer [109], Stahel [114]) proved that for all "nice", smooth affine equivariant estimators of multivariate location the breakdown point is $\leq 1/d$, where d is the dimension of the space in multivariate statistics, with corresponding consequences for the design space in multiple regression. Thus, for higher than 10-dimensional data, no "nice" equivariant methods exist which are absolutely safe against up to 10% gross errors.

As a reaction, a lot of research was done on "pathological" affine equivariant estimators with BP = 1/2 for any dimension. Such estimators exist; the first such regression estimator (stimulated by Tukey's "shorth", cf. Andrews et al. [1]) was the fit minimizing the MAD of the residuals (Hampel [43], investigated further and popularized under the name "least median of squares" by Rousseeuw ([104] and subsequent work); cf. also Rousseeuw and Leroy [105] which contains a number of examples). Somewhat similarly, Stahel [113] [114]) and Donoho [22], for robust covariance matrix estimation (which can then also be used in multiple regression), used a measure of outlyingness based on a search over "all directions" on a high-dimensional sphere. Such searches can at best be approximated on the computer; even so they are very computer-intensive. A third regression method with BP = 1/2, which however is not affine equivariant, Siegel's [110] "repeated median", needs only finitely many steps in principle, but suffers soon from the "combinatorial explosion". All three methods are presently only usable for somewhat less than 10 independent variables (depending of course somewhat on computer power and program). In addition, they are very inefficient; but this can be remedied by follow-up (1-step) redescending M-estimators (cf., e.g., Yohai et al. [128]).

Meanwhile there has been a lot of research, and there are many other "high breakdown point estimators", with variants and combinations to make them more efficient (such as S-estimators, τ-estimators, MM-estimators, minimum volume ellipsoids..., cf. Rousseeuw and Leroy [105] and the papers in Stahel and Weisberg [116] for a start). All this research concentrates on BP = 1/2 and on strict affine equivariance. But it can be argued that a good model for the distribution of gross errors is often not affine equivariant (gross errors tend to occur partly in single coordinates), and that a flexible robust tool box for interactive data analysis would be more useful for statistical practice, although in some situations high breakdown point estimators might be the method of choice.

Another aspect is that the concept of breakdown point has been generalized to linear models mostly in only one way, namely for random "carriers" or x-variables, allowing direct application of the theory for i.i.d. vectors. But one can also consider the conditional BP, given the design (which is known after the data are in, even if it contains gross errors); this concept provides much richer information complementing

and augmenting the one given by the unconditional BP (cf. Hampel [43], where the conditional BP is implicitly used for some examples as the most natural one before the other one had even been defined). In special situations, for example in the ANOVA, it can also be helpful to define a "local BP" (see Mili et al. [89]) and a "partial BP" (Ruckstuhl [106], [107]) and to differentiate between a BP against "wild" and against "mean" (nasty) outliers (Hampel [48]; see also Terbeck and Davies [121]).

Long-Range dependence; The Violation of the Independence Assumption

More aspects of statistical "random errors"

Compare also the sections on long-range dependence at the beginning.

Experienced data analysts know that "every observation is influenced by the date on which it is made" (Cuthbert Daniel in his lectures 1968 at UC Berkeley; W.S. Gosset ("Student") [120] and cited by Jeffreys [65], 3rd. ed., p. 298). Newcomb [95] gave a penetrating analysis of statistical errors in astronomy, and so did

"Student" [120], with many examples, for chemical data. The paper by "Student", reproduced also in his "Collected Papers" (E.S. Pearson and Wishart, eds., [98]), is worthwhile reading for every applied statistician. Karl Pearson [99] tried to analyze the "personal equation" (the personal error or effect of different observers) and found in 6 long controlled series not only (long-range) correlations over time, but also (even more mysterious) cross-correlations between different observers. Rather striking is also the behavior of several hundred measurements of the US 1 kg check standard weight obtained with utmost care (with a relative error of 10^{-10}!) for the purpose of checking on the measurement process itself by the NBS in Washington, D.C.; they should be prime examples for "i.i.d." data if ever there exist any; but instead they exhibit highly significant long-range correlations (which can be as little explained as in other cases). See Graf et al. [36], Hampel [46] and Beran [7] for more details and examples (including the about 3000 measurements of the velocity of light by Michelson et al. [88]).

A consequence of these phenomena is that "replicates" (observations under "identical" conditions) should be spread as far apart as possible (in time, space, ...) in order to get a realistic estimate of the "random error"; "replicates" right next to each other ("pseudo-replicates") give a spuriously small error. On the other hand, if the size of a "contrast" (effect or interaction in ANOVA, slope in regression) is sought, the observations should be done as close together as possible (possibly in blocks) so as not to be unduly influenced by the slowly changing correlation effects ("semi-systematic errors").

Chemists know from experience that the interlaboratory error ("reproducibility") is larger than the intralaboratory error ("repeatability") and make elaborate "interlaboratory tests" (in German: "Ringversuche"), where samples of the same well-mixed quantity are analyzed by different labs, in order to assess these errors. As noted before, these aspects are mainly important for the assessment of absolute

constants; for example, for testing whether a legal limit (threshold value, standard) for some contamination in food or the environment has been surpassed (a frequent and for society highly important task for many chemists). It could happen that 2 tests of the same material, done by different laboratories, both come out significantly above the legal limit, but that there is no significance if the interlaboratory error is taken into account.

It is by now widely known from experience that count data often exhibit "overdispersion", so that the standard errors derived from the multinomial (etc.) distribution have to be corrected by an empirical factor. Already long ago, Berkson [8] noted that in sufficiently large samples of count data, the χ^2-test virtually always rejects the null hypothesis ("making statistical testing superfluous", as he, half-jokingly, half-puzzled, remarked). It may be surmised that both phenomena are (at least in part) related to long-range correlations in the count data.

Other examples which have been shown to exhibit long-range correlations range from the density of wireworms (data given in Yates and Finney [127] and in Yates [126]) and agricultural yields in uniformity trials (H. Fairfield Smith [24]) to the weights of milk cartons filled in a fully automatic production line (Graf unpublished, cited in Hampel [46]). Compare also the many examples for long-range dependence in virtually all kinds of geophysical data (cf. the beginning) collected by Mandelbrot and Wallis [82].

As Heyde (in discussion of Hampel, [46], p. 255) noted with reference to the International Geosphere-Biosphere Programme (IGBP), the investigation of environmental data under the aspect of long-range dependence is of greatest importance, in view of the consequences for standard errors, tests and confidence intervals (see below).

The modelling of long-range dependence

The true behavior of errors of series of "independent" measurements can be considered as lying somewhere in between that of independent random errors and constant systematic errors. They were called "semi-systematic errors" by Newcomb [95] and "semi-constant errors" by "Student" [120]. H. Fairfield Smith [24] showed empirically for about 40 uniformity trials that the variance of the mean yield did not go to zero like n^{-1}, where n is the number of plots, but like $n^{-\alpha}$, for α anywhere between 0 and 1. This contradicts not only the famous textbook formula that $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_1)/n$; but also for all short-range correlated (strongly mixing) processes, including all ARMA processes, $\mathrm{Var}(\bar{X}_n) \propto n^{-1}$ at least asymptotically, and the only way out seemed the assumption that one was still in the realm of transient effects (which can always be described by sufficiently many serial correlations as parameters). When the engineer H.E. Hurst thoroughly analyzed the about 800 yearly maxima and minima of the river Nile in preparation for the construction of the Aswan dam (the longest such data series known), he arrived again at empirical conclusions which seemed incompatible with a "stationary" behavior of the Nile. But then B.B. Mandelbrot (cf. Mandelbrot [77]; Mandelbrot and Wallis [80][82]; Mandelbrot and van Ness [79]) was able to describe the "Hurst phenomenon" by means of the increment processes of so-called self-similar Gaussian processes (Kolmogorov [70]). These are

stationary, but only weakly mixing and long-range correlated processes where the lag k correlation goes to zero so slowly (hyperbolically instead of exponentially) that the sum over all lags is infinite, and the variance of the arithmetic mean goes to zero exactly like n^{2H-2}, where the parameter H (named after Hurst by Mandelbrot) usually lies between 1/2 (independence) and 1 (nonstationarity; the limiting case can also be viewed as a random, nonergodic constant systematic error). Mandelbrot called the biblical phenomenon of "seven fat years followed by seven lean years" caused by the Nile the "Joseph effect".

For a continuous time self-similar process, the sample paths always "look the same" except for scaling (are "similar to themselves") in the large as in the small ("under the magnifying glass"). Such processes can be derived by means of a "fractional integral" over white noise (cf., e.g., Mandelbrot and van Ness [79]). They are special cases of Mandelbrot's [78] "fractals". The (stationary!) increment processes (which are mathematically rather "pathological") of Gaussian self-similar processes are called "fractional Gaussian noises". For the discretized versions, the correlation between X_i and X_{i+k} as function of the lag k (which fixes a stationary Gaussian process) is $((k+1)^{2H} - 2k^{2H} + |k-1|^{2H})/2$ for $k = 0, 1, 2, \ldots$. Loosely speaking, the distant past is still noticeably correlated with the distant future.

It is instructive to study the pictures of simulations of discrete fractional Gaussian noises, see, e.g., Mandelbrot and Wallis [81], Wallis and Matalas [124], Wallis and O'Connell [125] (the latter two also with simulations of lag-one Markov processes and an ARIMA(1, 0, 1) process for comparison), Mandelbrot [78], Graf [35], Graf et al. [36] (including a visual comparison with a series of 588 monthly ozone data, part of the famous Arosa ozone series, to show the strong similarity in character), Hampel et al. [52], Hampel [46], and the monograph by J. Beran [7]. One striking feature is that there seem to be all sorts of "trends" and "cycles" ("hidden periodicities") in any limited segment (often about 2-3 "cycles", no matter what the length of the segment); but with the continuation, these "trends" and "cycles" disappear (and new ones seem to appear). Thus, one has to be careful: many "climatic cycles" (or similar-looking phenomena) may be merely artefacts of a long-range correlated process. This is not to say that all such "trends" and "cycles" are spurious; but there should be a good theoretical reason (such as an astronomical cycle in the background, or another convincing explanatory variable) or very strong empirical evidence before such apparent effects can be trusted.

Self-similar processes are not the only way to model long-range correlations; asymptotically equivalent are so-called fractional ARIMA-processes (cf. Granger and Joyeux [37], Hosking [54], Li and McLeod [75]), which were developed later, but fit better into the ARIMA-technology.

First statistical methods for the increment processes of discretized self-similar processes were already developed by Hurst and by Mandelbrot. Graf ([35], cf. Graf et al. [36]) developed asymptotically optimal methods (and robust variants) for estimating and testing H, and J. Beran ([4] [5], cf. also [6] and [7])(compare also Künsch [72]) developed a one-sample t-test for long-range correlated data. For more recent results, cf. Beran [7]. In view of the results of Künsch et al. [73], the one-sample t-test, although looking simpler at first, is much more problematic in practice (except for small samples) than the two-sample t-test, in which with well-

chosen designs the first-order effects of long-range correlations cancel out, as noted in the beginning.

References

[1] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J., 1972.

[2] Barnett, V., and Lewis, T. *Outliers in Statistical Data*. Wiley, N. Y., 1978. 3rd Edition 1994.

[3] Bednarski, T. Fréchet differentiability of statistical functionals and implications to robust statistics. In *New Directions in Statistical Data Analysis and Robustness* (1993), S. Morgenthaler, E. Ronchetti, and W. A. Stahel, Eds., Birkhäuser Verlag, Basel, pp. 25–34.

[4] Beran, J. *Estimation, testing and prediction for self-similar and related processes*. PhD thesis, Swiss Federal Institute of Technology (ETH), 1986.

[5] Beran, J. A test of location for data with slowly decaying serial correlations. *Biometrika 76* (1989), 261–269.

[6] Beran, J. Statistical methods for data with long–range dependence. *Statist. Sci. 7* (1992), 404–427, (with discussion).

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

[8] Berkson, J. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Statist. Assoc. 33* (1938), 526–536.

[9] Berrendero, J. R., and Zamar, R. H. Maximum bias curves for robust regression with non-elliptical regressors. *Annals of Statistics 29*, 1 (2001). In press.

[10] Bessel, F. W. *Fundamenta Astronomiae*. Nicolovius, Königsberg, 1818.

[11] Box, G. E. P. Non-normality and tests on variances. *Biometrika 40* (1953), 318–335.

[12] Box, G. E. P., and Andersen, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. Royal Statist. Soc. B 17* (1955), 1–34.

[13] Box, G. E. P., Hunter, W. G., and Hunter, J. S. *Statistics for Experimenters*. Wiley, N. Y., 1978.

[14] Box, G. E. P., Leonard, T., and Wu, C. F., Eds. *Scientific Inference, Data Analysis, and Robustness*. Academic Press, New York, 1983.

[15] Clarke, B. R. Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *The Annals of Statistics 11* (1983), 1196–1205.

[16] Clarke, B. R. Nonsmooth analysis and Fréchet differentiability of M-functionals. *Prob. Th. Rel. Field 73* (1986), 197–209.

[17] Cox, D., and Hinkley, D. A note on the efficiency of least-squares estimates. *J. Royal Statist. Soc. B 30* (1968), 284–289.

[18] Daniel, C. *Applications of Statistics to Industrial Experimentation.* Wiley, N. Y., 1976.

[19] Daniel, C., and Wood, F. S. *Fitting Equations to Data*, 2 ed. Wiley, N. Y., 1980.

[20] Daniels, H. E. Saddlepoint approximations in statistics. *Ann. Math. Statist. 25* (1954), 631–650.

[21] Davies, L., and Gather, U. The identification of multiple outliers. *J. Am. Statist. Assoc. 88*, 423 (1993), 782–792; with discussion 793–801.

[22] Donoho, D. L. Breakdown properties of multivariate location estimators. Ph.d. qualifying paper, Department of Statistics, Harvard University, Cambridge, Mass, 1982.

[23] Donoho, D. L., and Huber, P. J. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, P. J. Bickel, K. Doksum, and J. L. J. Hodges, Eds. Wadsworth, Belmont, 1983, pp. 157–184.

[24] Fairfield Smith, H. An empirical law describing heterogeneity in the yields of agricultural crops. *J.Agric.Sci. 28* (1938), 28,1–23.

[25] Federer, B., Waldvogel, A., Schmid, W., Schiesser, H. H., Hampel, F., Schweingruber, M., Stahel, W., Bader, J., Mezeix, J. F., Doras, N., D'Aubigny, G., DerMegreditchian, G., and Vento, D. Main results of Grossversuch IV. *Journal of Climate and Applied Meteorology 25*, 7, July (1986), 917–957.

[26] Fellner, W. H. Robust estimation of variance components. *Technometrics 28*, 1 (1986), 51–60.

[27] Fernholz, L. T. *Von Mises Calculus for Statistical Functionals.* Lecture Notes in Statistics 19. Springer, New York., 1983.

[28] Field, C., and Hampel, F. Small-sample asymptotic distributions of M-estimators of location. *Biometrika 69* (1982), 29–46.

[29] Field, C., and Ronchetti, E. *Small Sample Asymptotics.* Institute of Mathematical Statistics Monograph Series, Hayward (CA), 1990.

[30] Fisher, R. A. 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* (1920), 758–770. Reprinted in Collected Papers of R. A. Fisher, ed. J. H. Bennett, Volume 1, 188–201, University of Adelaide 1971.

[31] Fisher, R. A. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London. A 222* (1922), 309–368. Reprinted in Collected Papers of R. A. Fisher, ed. J. H. Bennett, Volume 1, 275–335, University of Adelaide 1971.

[32] Fraiman, R., Yohai, V. J., and Zamar, R. H. Optimal robust M-estimates of location. *Annals of Statistics 29*, 1 (2001). In press.

[33] Gauss, C. F. Theoria combinationis observationum erroribus minimis obnoxiae (pars prior), presented 15.2.1821. Commentationes societatis regiae scientiarium Gottingensis recentiores. In *Werke*, vol. 4. Dieterichsche Universitäts-Druckerei, 1880, 1823, pp. 1–108.

[34] Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, N. Y., 1977.

[35] Graf, H. P. *Long-range correlations and estimation of the self-similarity parameter*. PhD thesis, Swiss Federal Institute of Technology (ETH), 1983.

[36] Graf, H. P., Hampel, F. R., and Tacier, J. The problem of unsuspected serial correlations. In *Robust and Nonlinear Time Series Analysis*, J. Franke, W. Härdle, and R. D. Martin, Eds., Lecture Notes in Statist. 26. Springer, 1984, pp. 127–145.

[37] Granger, C. W. J., and Joyeux, R. An introduction to long-memory time series and fractional differencing. *J. Time Series Anal. 1* (1980), 15–30.

[38] Grubbs, F. E. Procedures for detecting outlying observations in samples. *Technometrics 11* (1969), 1–21.

[39] Hampel, F. *Contributions to the theory of robust estimation*. PhD thesis, University of California, Berkeley, 1968.

[40] Hampel, F. A general qualitative definition of robustness. *Ann. Math. Statist. 42* (1971), 1887–1896.

[41] Hampel, F. Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 27* (1973), 87–104.

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

[43] Hampel, F. Beyond location parameters: Robust concepts and methods (with discussion). In *Bull. 40th Session of the ISI* (1975), vol. XLVI, Book 1, pp. 375–391.

[44] Hampel, F. The robustness of some nonparametric procedures. In *A Festschrift for Erich L. Lehmann*, P. J. Bickel, K. Doksum, and J. L. Hodges Jr., Eds. Wadsworth, Belmont, 1983, pp. 209–238.

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

[46] Hampel, F. Data analysis and self-similar processes. In *Bull. 46th Session of the ISI, Tokyo* (1987), vol. LII, Book 4, pp. 235–254 (Discussion: pp. 255–264).

[47] Hampel, F. Design, modelling, and analysis of some biological data sets. In *Design, Data, and Analysis, by some friends of Cuthbert Daniel*, C. L. Mallows, Ed. Wiley, N. Y., 1987, pp. 93–128.

[48] Hampel, F. Some problems in statistics. In *Proc. First World Congress of the Bernoulli Society, Tashkent 1986* (1987), Y. Prohorov and V. V. Sazonov, Eds., vol. 2, VNU Science Press, Utrecht, pp. 253–256.

[49] Hampel, F. Introduction to Huber (1964): Robust estimation of a location parameter. In *Breakthroughs in Statistics, vol. 2: Methodology and Distribution*, S. Kotz and N. L. Johnson, Eds. Springer-Verlag, 1992, pp. 479–491.

[50] Hampel, F. Some additional notes on the "Princeton Robustness Year,". In *The Practice of Data Analysis: Essays in Honor of John W. Tukey* (1997), D. R. Brillinger, L. T. Fernholz, and S. Morgenthaler, Eds., Princeton University Press, Princeton, pp. 133–153.

[51] Hampel, F. Is statistics too difficult? *Canad. J. Statist. 26*, 3 (1998), 497–513.

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

[53] Hoaglin, D. C., Mosteller, F., and Tukey, J. W., Eds. *Understanding Robust and Exploratory Data Analysis*. Wiley, N. Y., 1983.

[54] Hosking, J. R. M. Fractional differencing. *J. Time Series Anal. 1* (1981), 15–29.

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

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

[57] Huber, P. J. Robust confidence limits. *Z. Wahrsch. verw. Geb. 10* (1968), 269–278.

[58] Huber, P. J. Studentizing robust estimates. In *Nonparametric Techniques in Statistical Inference*, M. L. Puri, Ed. Cambridge University Press, Cambridge, England, 1970, pp. 453–463.

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

[60] Huber, P. J. Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist. 1* (1973), 799–821.

[61] Huber, P. J. Robustness and designs. In *A Survey of Statistical Design and Linear Models* (1975), J. N. Srivastava, Ed., North Holland, Amsterdam, pp. 287–301.

[62] Huber, P. J. *Robust statistical procedures*. SIAM, Philadelphia, 1977. Second edition 1996.

[63] Huber, P. J. *Robust Statistics*. Wiley, N. Y., 1981.

[64] Huber, P. J., and Strassen, V. Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist. 1* (1973), 251–263. Corr: **2**, 223–224.

[65] Jeffreys, H. *Theory of Probability*. Clarendon Press, Oxford., 1939. Later editions: 1948, 1961, 1983.

[66] Johnson, N. L., and Leone, F. C. *Statistics and Experimental Design in Engineering and the Physical Sciences*, vol. 1. Wiley, N. Y., 1964. 2nd edition 1977.

[67] Jurečková, J., and Sen, P. K. *Robust Statistical Procedures; Asymptotics and Interrelations*. Wiley Series in Probability and Statistics. Wiley, 1996.

[68] Kadane, J. *Robustness in Bayesian Statistics*. North Holland, Amsterdam, 1984.

[69] Kleiner, B., Martin, R. D., and Thomson, D. J. Robust estimation of power spectra. *J. Royal Statist. Soc. B 41* (1979), 313–351.

[70] Kolmogorov, A. N. Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen Raum. *Acad.Sci. URSS (N.S.), C.R. (Doklady) 26* (1940), 115–118.

[71] Künsch, H. R. Infinitesimal robustness for autoregressive processes. *Ann. Statist. 12*, 3 (1984), 843–863.

[72] Künsch, H. R. Statistical aspects of self-similar processes. In *Proc. First World Congress of the Bernoulli Society, Tashkent 1986* (1987), Y. Prohorov and V. V. Sazonov, Eds., vol. 1, VNU Science Press, Utrecht, pp. 67–74. Invited paper.

[73] Künsch, H. R., Beran, J., and Hampel, F. R. Contrasts under long-range correlations. *Ann. Statist. 21*, 2 (1993), 943–964.

[74] Künsch, H. R., Stefanski, L. A., and Carroll, R. J. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Am. Statist. Assoc. 84*, 406 (1989), 460–466.

[75] Li, W. K., and McLeod, A. I. Fractional time series modelling. *Biometrika 73* (1986), 217–221.

[76] Mallows, C. L. Robust methods - some examples of their use. *Am. Statist. 33* (1979), 179–184.

[77] Mandelbrot, B. B. Une classe de processus homothétiques à soi: Application à la loi climatologique de H. E. Hurst. *C.R. Acad.Sci.Paris 260* (1965), 3274–3277.

[78] Mandelbrot, B. B. *The Fractal Geometry of Nature*. Freeman, New York., 1983. First edition: 1977.

[79] Mandelbrot, B. B., and van Ness, J. W. Fractional Brownian motions, fractional noises and applications. *SIAM Review 10* (1968), 422–437.

[80] Mandelbrot, B. B., and Wallis, J. R. Noah, Joseph, and operational hydrology. *Water Resources Research 4* (1968), 909–918.

[81] Mandelbrot, B. B., and Wallis, J. R. Computer experiments with fractional Gaussian noises. *Water Resources Research 5* (1969), 228–267.

[82] Mandelbrot, B. B., and Wallis, J. R. Some long-run properties of geophysical records. *Water Resources Research 5* (1969), 321–340.

[83] Marazzi, A. *Algorithms, Routines, and S Functions for Robust Statistics*. Wadsworth, Inc., Belmont, California, 1993.

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

[85] Maronna, R. A., Yohai, V. J., and Zamar, R. J. Bias-robust regression estimation: A partial survey. In *New Directions in Statistical Data Analysis and Robustness* (1993), S. Morgenthaler, E. Ronchetti, and W. A. Stahel, Eds., Birkhäuser Verlag, Basel, pp. 157–176.

[86] Martin, R. D., and Yohai, V. J. Influence functionals for time series. *The Annals of Statistics 14*, 3 (1986), 781–818 (Discussion pp. 819–855).

[87] Merrill, H. M., and Schweppe, F. C. Bad data suppression in power system static state estimation. *IEEE Trans. Power App. Syst. PAS-90* (1971), 2718–2725.

[88] Michelson, A. A., Pease, F. G., and Pearson, F. Measurement of the velocity of light in a partial vacuum. *Contributions from the Mount Wilson Observatory, Carnegie Institution of Washington XXII*, 522 (1935), 259–294.

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

[90] Morgenthaler, S., Ronchetti, E., and Stahel, W. A., Eds. *New Directions in Statistical Data Analysis and Robustness*. Birkhäuser Verlag, Basel, 1993.

[91] Morgenthaler, S., and Tukey, J. W., Eds. *Configural Polysampling: A Route to Practical Robustness*. Wiley, N. Y., 1991.

[92] Mosteller, F., and Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, Mass., 1977.

[93] Müller, C. H. *Robust Planning and Analysis of Experiments*, vol. 124 of *Lecture Notes in Statistics*. Springer, N. Y., 1997.

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

[95] Newcomb, S. Astronomical constants (the elements of the four inner planets and the fundamental constants of astronomy). Supplement to the American Ephemeris and Nautical Almanac for 1897, U.S. Government Printing Office, Washington, D.C., 1895.

[96] Pearson, E. S. The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika 21* (1929), 259–286.

[97] Pearson, E. S. The analysis of variance in cases of non-normal variation. *Biometrika 23* (1931), 114–133.

[98] Pearson, E. S., and Wishart, J., Eds. *"Student's" collected papers*. Cambridge, published for the Biometrika trustees, University Press, 1936.

[99] Pearson, K. On the mathematical theory of errors of judgement, with special reference to the personal equation. *Philos. Trans. Roy. Soc. Ser. A 198* (1902), 235–299.

[100] Pukelsheim, F. Letter to the editor. *IMS Bulletin 19*, 4 (1990), 540–542.

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

[102] Rieder, H. *Robust asymptotic statistics*. Springer, N. Y., 1994.

[103] Rocke, D. M., Downs, G. W., and Rocke, A. J. Are robust estimators really necessary? *Technometrics 24* (1982), 95–102.

[104] Rousseeuw, P. J. Least median of squares regression. *J. Am. Statist. Assoc. 79* (1984), 871–880.

[105] Rousseeuw, P. J., and Leroy, A. M. *Robust Regression & Outlier Detection.* Wiley, N. Y., 1987.

[106] Ruckstuhl, A. F. *Analysis of the T_2 emission spectrum by robust estimation techniques.* Ph. D. thesis no. 11170, Swiss Federal Institute of Technology (ETH), 1995.

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

[108] Samarov, A. M. Bounded-influence regression via local minimax mean squared error. *J. Amer. Stat. Assoc. 80* (1985), 1032–1040.

[109] Schönholzer, H. *Robuste Kovarianz.* PhD thesis, ETH, Zurich, 1979.

[110] Siegel, A. F. Robust regression using repeated medians. *Biometrika 69* (1982), 242–244.

[111] Spjøtvoll, E., and Aastveit, A. H. Comparison of robust estimators on data from field experiments. *Scand.J.Statist. 7* (1980), 1–13.

[112] Spjøtvoll, E., and Aastveit, A. H. Robust estimators on laboratory measurements of fat and protein in milk. *Biometrical J. 25* (1983), 627–639.

[113] Stahel, W. A. Breakdown of covariance estimators. Res. rep. 31, Fachgruppe für Statistik ETH, Zurich, 1981.

[114] Stahel, W. A. *Robust estimation: Infinitesimal optimality and covariance matrix estimators (in German).* PhD thesis, Swiss Federal Institute of Technology (ETH), 1981.

[115] Stahel, W. A., Ruckstuhl, A. F., Senn, P., and Dressler, K. Robust estimation in the analysis of complex molecular spectra. *J. Am. Statist. Assoc. 89*, 427 (1994), 788–795.

[116] Stahel, W. A., and Weisberg, S., Eds. *Directions in Robust Statistics and Diagnostics*, vol. 1, 2. Springer, N. Y., 1991.

[117] Stahel, W. A., and Welsh, A. Approaches to robust estimation in the simplest variance-components model. *Journal of Statistical Planning and Inference 57* (1997), 295–319.

[118] Staudte, R. G., and Sheather, S. J. *Robust Estimation and Testing.* Wiley, N. Y., 1990.

[119] Stigler, S. M. Do robust estimators work on real data? *Ann. Statist. 6* (1977), 1055–1098.

[120] "Student" (W. S. Gosset). Errors of routine analysis. *Biometrika 19* (1927), 151–164.

[121] Terbeck, W., and Davies, P. L. Interactions and outliers in the two-way analysis of variance. *The Annals of Statistics 26*, 4 (1998), 1279–1305.

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

[123] Tukey, J. W. The future of data analysis. *Ann. Math. Statist. 33* (1962), 1–67.

[124] Wallis, J. R., and Matalas, N. C. Correlogram analysis revisited. *Water Resources Research 7* (1971), 1448–1459. 8 (1972), 1112–1117.

[125] Wallis, J. R., and O'Connell, P. E. Firm reservoir yield – how reliable are historic hydrological records? *Hydrological Sciences Bulletin XVIII* (1973), 347–365.

[126] Yates, F. *Sampling Methods for Censuses and Surveys*, 4 ed. Charles Griffin, London, 1981. First edition 1949.

[127] Yates, F., and Finney, D. J. Statistical problems in field sampling for wireworms. *Ann. Appl. Biol. 29* (1942), 156–167.

[128] Yohai, V., Stahel, W. A., and Zamar, R. A procedure for robust estimation and inference in linear regression. In *Directions in Robust Statistics and Diagnostics* (1991), W. A. Stahel and S. Weisberg, Eds., vol. 2, Springer, New York, pp. 365–374.

[129] Youden, W. J. Enduring values. *Technometrics 14* (1972), 1–11.