Statistical Procedures and Robust Statistics *

P. L. Davies
Fachbereich Mathematik und Informatik
Universität Essen
45117 Essen, Germany

November 4, 2002

Abstract

It is argued that a main aim of statistics is to produce statistical procedures which in this article are defined as algorithms with inputs and outputs. The structure and properties of such procedures are investigated with special reference to topological and testing considerations. Procedures which work well in a large variety of situations are often based on robust statistical functionals. In the final section some aspects of robust statistics are discussed again with special reference to topology and continuity.

1 Introduction

The view expressed in this article is that one of the main tasks of statisticians is to produce statistical procedures. By this we mean algorithms with inputs, which include of course the data, and outputs which vary from numbers to images. A statistical procedure is not a statistical analysis which will typically involve the use of several statistical procedures. The idea of a procedure orientated statistics goes back to Tukey (see the Fifth Bite of Tukey (1993)). Section 2 discusses some issues concerned with statistical procedures and in particular it is argued that the topology of statistical procedures is a weak one. It is further argued that a statistical procedure makes no assumptions about the data but rather that it should be accompanied by a brief description of the sort of data sets for which it works well and of those for which it works less well. If the range of data sets for which the procedure works well is not to be too restrictive some form of robustness will be required. Section 3 contains some comments about robust statistics in the light of the discussion of statistical procedures. Again the rôle of weak topologies is emphasized.

*Research supported in part by the Deutsche Forschungsgesellschaft (SFB 475, University of Dortmund)

2 Statistical procedures

2.1 Algorithms

We take a statistical procedure to be a computer algorithm with inputs which include the data as well as parameters which have, where possible, default values. The output will typically consist of numbers, graphs, functions and pictures. The intention is to provide user friendly software which can be used sensibly by users who are not necessarily qualified statisticians. In the ideal case the user should have to input only the data and still obtain a reasonable output.

2.2 Assumptions

A statistical procedure does not impose any conditions on the data to which it is applied. The data can be real, simulated or even deterministic. A statistical procedure should however be accompanied by a brief description of the sort of data sets for which it can be expected to work well. Consider the following description for a procedure which performs a one-way analysis of variance.

> The procedure is to be applied only to *i.i.d. normally distributed samples* without the slightest form of contamination. The sample variances *must* be equal although the procedure does allow for different sample sizes. No responsibility of any form will be accepted by the author for the use on data not conforming to the above description.

This may be compared with

> The procedure is reliable for (a) different sample sizes but with a minimum size of 2, (b) at most 50 samples to be compared, (c) different scales, (d) almost symmetric data with up to 30% symmetric or extreme outliers and/or less than 10% moderate one-sided outliers.

2.3 Topologies

The topology of data analysis is a weak topology. This means that minor changes in the data should result in minor changes in the result of the procedure. Minor will refer to small changes in the values of the data as well as large changes in a small number of data points. Although procedures make no assumptions about data they are very often based on probability models and on functionals which map probability measures into Euclidean space. The simplest example is the mean or, more generally, M-functionals. Consider therefore a mapping $T : \mathcal{P} \rightarrow \mathbb{R}^k$ where $\mathcal{P}$ is some space of probability measures. In order to define continuity and differentiability of T we need a metric on $\mathcal{P}$. The metrics we require are weak ones such as the Kolmogoroff metric

$$d_{ko}(P, Q) = \sup\{|P(C) - Q(C)| : C = (-\infty, c], c \in \mathbb{R}\}$$

or more generally

$$d_{\mathcal{C}}(P,Q) = \sup\{|P(C) - Q(C)| : C \in \mathcal{C}\} \tag{1}$$

where $\mathcal{C}$ is a Vapnik-Cervonenkis class of sets. Strong metrics and discrepancies such as total variation, Hellinger and Kullback-Leibler are density based. We refer to the table on page 588 of Donoho and Liu (1988). As a partial justification of the use of weak metrics we note that random samples are generated at the level of weak metrics. Thus if $(U_i)_1^n$ is an i.i.d. sample of uniformly distributed random variables and if F and G are two distribution functions then $X_i(F) = F^{-1}(U_i)$ and $X_i(G) = G^{-1}(U_i)$ for $i = 1, 2, \ldots, n$ are i.i.d. samples with distributions F and G respectively. If F and G are close in the Kolmogoroff metric then the samples $(X_i(F))_1^n$ and $(X_i(G))_1^n$ are close in the Euclidean metric.
As a further justification of weak metrics consider the ball with centre P and radius ε by

$$B(P, \varepsilon, d) = \{Q : d(Q,P) \le \varepsilon\} \tag{2}$$

then typically for weak metrics the empirical distribution P_n of n i.i.d. random variables with common distribution P will lie in $B(P, \varepsilon_n, d)$ with $\varepsilon_n = c/\sqrt{n}$. In other words weak metrics allow direct comparisons of data with purported models. Similar arguments in favour of weak metrics together with Fréchet differentiability of statistical functionals have been put forward by Bednarski (1993), Bednarski and Clarke (1998), Clarke (2000) etc.
One functional which plays an almost dominant rôle in statistics is the differential operator D defined by

$$D(G) = g \quad \text{for all absolutely continuous } G, \quad G(x) = \int_{-\infty}^{x} g(u)\,du \tag{3}$$

The functional D is not continuous with respect to a weak metric and is indeed pathologically discontinuous. As all likelihood methods make intimate use of D this means that all likelihood based methods are pathologically discontinuous with respect to weak topologies. There is therefore no good reason for basing statistical procedures on likelihood based methods; a unified approach to statistics is only possible if it is based on a weak topology.

2.4 Constructing statistical procedures

There is no prescribed methodology for constructing statistical procedures. Even if some stochastic model is agreed to be an adequate approximation of a data set there is not the slightest reason for basing the procedures on likelihood either in the form of maximum likelihood or Bayes (see Section 2.3). Instead we use an idea of Tukey's and write the data in the form

$$\text{DATA} \quad = \quad \text{SIGNAL} \quad + \quad \text{NOISE}. \tag{4}$$

We separate SIGNAL and NOISE by assuming that the signal is simple and the noise is complex. In particular we define what we mean by noise and then choose the simplest signal such that (4) holds. We demonstrate this in the context of nonparametric regression. The stochastic model is

$$Y(t) = f(t) + \sigma\varepsilon(t), \quad 0 \le t \le 1, \tag{5}$$

where $\varepsilon(t)$ is standard Gaussian white noise. We identify the signal with the function f. The concept of simplicity we use for f is the number of local extreme values on the interval (0, 1). We identify the noise with the errors or disturbances $\varepsilon(t)$. Given design points $0 \le t_1 < t_2 < \ldots < t_n \le 1$ we form the local standardized means

$$M_n(I) = \sum_{t_i \in I} \varepsilon(t_i)/\sqrt{n(I)}$$

where I denotes an interval and $n(I)$ is the number of design points in I. The maximum size of the standardized local means is a measure of the deviation of the Y from the function f at the design points. Under the assumptions of the model (5) we have approximately

$$\max_I |M_n(I)| \le \sigma\sqrt{\tau \log(n)} \tag{6}$$

for some constant $2 \le \tau \le 4$. We indicate how the above considerations can be used to construct a procedure. Consider data $y(t_i), i = 1, \ldots n$. For any function f we write

$$r(t_i) = y(t_i) - f(t_i), \quad i = 1, \ldots, n \tag{7}$$

and

$$\sigma_n = 1.483\text{Median}(|f(t_2) - f(t_1)|, \ldots, |f(t_n) - f(t_{n-1})|)/\sqrt{2}. \tag{8}$$

We agree that the residuals $r(t_i)$ approximate Gaussian white noise if

$$\max_I |m_n(I)| \le \sigma_n\sqrt{\tau \log(n)} \tag{9}$$

where

$$m_n(I) = \sum_{t_i \in I} r(t_i)/\sqrt{n(I)}.$$

This leads to the following problem. Determine a function f with the smallest number k_n of local extreme values such that the residuals $r(t_i)$ approximate white noise in the sense of (9). This is not, as yet, an algorithm in the sense described in Section 2.1 but it can, with some effort, be turned into one. We refer to Davies and Kovac (2001) for the details.

The condition (6) is not the only property of white noise which can be used as a definition of approximation. In (5) we can assume that the median of the ε is zero and take the definition of approximation to white noise to be based on the local means of the signs of the residuals $r(t_i)$. This gives rise to a robustified procedure which works well even on test beds with Cauchy noise.

2.5 Test beds

Before being offered for general use statistical procedures should be tested where the word "test" is now to be understood in an engineering and not in a statistical sense. A procedure may be based on considerations which derive from a stochastic model but its range of applicability extends beyond this. It is therefore of importance to assess the performance of the procedure under a variety of conditions. This may be done by testing it under the well controlled conditions of a test bed defined by a stochastic model. This means that we generate samples $(X_i(P))_1^n$ using the probability model P and then apply the procedure to the sample. The advantage of this is that we are often able to identify the output of the procedure with properties of P. As an example we again use nonparametric regression. A model P is now defined by the function f and the errors ε of (5). The result of the procedure will be a function f_n which can be directly compared with the function f. By using different f and different ε one can try to determine the behaviour of the procedure, that is, the sort of data for which it works well and the sort of data for which it works less well.
Real data sets should also be used as test beds. They have the disadvantage that they cannot be controlled and that the result of the procedure cannot be compared directly with some real f. Nevertheless such testing is important as it often reveals unsuspected properties of real data sets which were not taken into account when the procedure was constructed.

2.6 Mathematical probes

Given a parametric test bed the performance and properties of a procedure are often accessible to a mathematical analysis. If we again consider the problem of nonparametric regression described in Section 2.4 then properties such as consistency and rates of convergence can be established mathematically (Davies and Kovac (2001)) on appropriate test beds. As a further example consider the behaviour of the location M-functional T_L defined by

$$\int \psi\left(\frac{x - T_L(F)}{T_S(F)}\right) dF(x) = 0 \tag{10}$$

$$\int \chi\left(\frac{x - T_L(F)}{T_S(F)}\right) dF(x) = 0 \tag{11}$$

where

$$\psi(x) = (\exp(x/5) - 1)/(\exp(x/5) + 1) \tag{12}$$

$$\chi(x) = (x^4 - 1)/(x^4 + 1). \tag{13}$$

T_L can be analysed mathematically by calculating breakdown points as well proving asymptotic normality in a locally uniform sense (see Davies (1998) for the latter). Beran has called such a mathematical analysis a "mathematical probe" (see his discussion of Davies and Kovac (2001)).

2.7 Simulations

In many situations the behaviour of a procedure for finite sample sizes is not amenable to a mathematical analysis. In such cases valuable information about a procedure can be obtained by simulations. Again if we return to the example of nonparametric regression then the function f and the distribution of the errors ε can be varied and information obtained on the limits of applicability of the procedure. In nonparametric regression standard test beds such as those developed by Donoho, Johnstone, Kerkyacharian and Picard (1995) play an important rôle.

2.8 Efficiency and blandness

To judge by the literature one main concern when evaluating the performance of a procedure seems to be its efficiency on certain test beds. Indeed the fact that a procedure is asymptotically efficient is often taken to be a general seal of approval. To illustrate the problems of using the concept efficiency as a measure of performance we consider the procedure based on the location functional T_L of (10) and (11). To calculate the asymptotic efficiency on the test bed specified by the distribution F one calculates an asymptotically optimal (in the sense of efficiency) location functional T_L^{opt} for this F. To simplify matters we assume that F is symmetric about 0 so that there is no ambiguity about what is to be estimated. The asymptotic efficiency of T_L at F for a sample of size n is defined as the ratio

$$\mathrm{Eff}(T_L, n) = \mathbb{V}(T_L^{\mathrm{opt}}(\mathbb{P}_n(F)))/\mathbb{V}(T_L(\mathbb{P}_n(F))) \qquad (14)$$

where $\mathbb{P}_n(F)$ denotes the empirical measure defined by n i.i.d. random variables with common distribution F and $\mathbb{V}$ denotes the variance. The important question and one that is not often discussed in the literature is the choice of F. Traditionally the Gaussian distribution is chosen so that T_L^{opt} is the mean. The choice is a sensible one for the reason that it is very difficult to estimate the mean of a Gaussian distribution. More precisely the Gaussian distribution maximizes the asymptotic variance $\lim_{n\to\infty} n\mathbb{V}(T_L^{\mathrm{opt}}(\mathbb{P}_n(F)))$ amongst all distributions F with finite variance. In Tukey's words the Gaussian distribution is "bland" or "hornless". If the distribution F is chosen without due care it may offer, again in Tukey's words, a free lunch. This means that quantiles based on this distribution will lead to smaller confidence intervals than are justified by the data. The model is allowing an increase in precision at no cost. It is not always obvious that this is happening. Table 1 shows the relative efficiency of the M-functional T_L defined above by (10) - (13) for samples that follow the Cauchy and the slash (Z/U with $Z = N(0,1)$ and U uniform on [0, 1]) distributions. The relative efficiencies are with respect to the maximum likelihood estimators. The results of Table 1 show that the relative efficiency of T_L on the Cauchy test bed is about 9% lower than on the slash test bed. The reason for this is that the Cauchy distribution is not bland. It has horns which are exploited by the optimal method to give an increase in efficiency. The horns of the Cauchy distri-

n	10	20	50	100
Cauchy	47.9	50.3	51.8	50.9
Slash	53.0	58.1	59.8	59.9

Table 1: Relative efficiencies of T_L on the Cauchy and slash test beds.

bution are its peakedness at the origin (see Cohen (1991)). Figure 1 shows the two densities and the somewhat greater peakedness of the Cauchy distribution. Although this is not excessive it is sufficient to give the increase in precision. The Cauchy distribution is usually taken as an example of a distribution with a heavy tail and hence outlier prone. The greater peakedness at the origin is rarely mentioned although it is this which causes the increase in precision. For this reason Tukey favours the slash distribution.

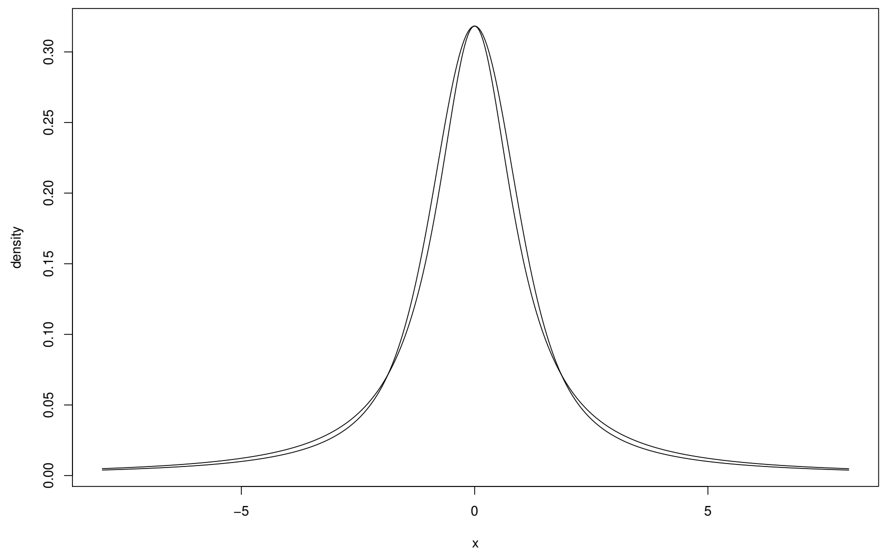

Figure 1: The figure shows the Cauchy and slash densities standardized to have the same value at 0. The Cauchy density is more peaked.

One can of course take this game even further. If Φ denotes the distribution function of the standard normal distribution then for any $\varepsilon > 0$, for example $\varepsilon = 10^{-100}$ there exist distributions F with

$$\sup_x |\Phi(x) - F(x)| < \varepsilon \tag{15}$$

and for which $\mathbb{V}(T_L^{\text{opt}}(\mathbb{P}_n(F)))$ can be made arbitrarily small. We indicate how such a distribution can be constructed. For any m there exist real numbers $e_i, i = 1, \ldots, m$ such that

$$\sum_{i=1}^{m} \alpha_i e_i = 0, \alpha_i \text{ rational } \Rightarrow \alpha_1 = \ldots = \alpha_m = 0. \qquad (16)$$

Such numbers are related to the idea of a Hamel basis for $\mathbb{R}$. For statisticians it is of interest to note that the proof of the existence of a Hamel basis given in Hewitt and Stromberg (1969) makes use of Tukey's lemma. As multiplication with a rational number does not alter the condition (16) we can for any given $\varepsilon > 0$ choose m, the e_i and weights p_i such that the following holds:

$$\begin{aligned} \sum_{i=1}^{m} p_i &= 1 & (17)\\ P &= \sum_{i=1}^{m} p_i \delta_{e_i}, \delta_x = \text{ Dirac mass at } x & (18)\\ d_{ko}(N(0,1), P) &< \varepsilon & (19) \end{aligned}$$

Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from the distribution $P(\cdot - \lambda)$ for some λ. Then with probability which can be made arbitrarily close to one by choosing m sufficiently large the sample contains at least two observation which are different. We take these to be X_1 and X_2. We can write $X_1 = \lambda + e_{i_1}$ and $X_2 = \lambda + e_{i_2}$. From this

$$X_2 - X_1 = e_{i_2} - e_{i_1}. \qquad (20)$$

It follows from (16) that the difference (20) determines e_{i_1} and e_{i_2} uniquely. From this it follows that we can determine λ exactly. In other words there exist distributions arbitrarily close to the Gaussian distribution with the property that given an i.i.d. sample from $P(\cdot - \lambda)$ we can determine λ exactly with very high probability. If we replace the δ_{e_i} in (18) by $N(e_i, \sigma_i^2)$ with σ_i very small ie

$$P_s = \sum_{i=1}^{m} p_i N(e_i, \sigma_i^2) \qquad (21)$$

then we can no longer determine λ exactly but we can determine λ with an arbitrarily high precision. Further more this smoothed version P_s of P has an infinitely differentiable density. Between P_s and the normal distribution we have a continuum of possibilities so there is *no* breakpoint between P_s and $N(0, 1)$. We note that if the real data are such that a goodness-of-fit test for normality is not rejected neither will the same test reject the distribution P_s.
For the distribution P_s and similar distributions all the information about the location parameter is carried in the tail of the decimal expansion. If the optimal functional T_L^{opt} is applied to real data or to data from some other distribution the results will be nonsensical. Statistical modelling is an ill-posed problem.

Given any data set there are an infinite number of models which can be used to model it. Applying a goodness-of-fit test will not help as indicated by (15). Optimal methods tuned to a precise model will produce nonsense unless the problem is regularized in some manner. Minimizing the Fisher information is one form of regularization and results in the arithmetic mean which, although not perfect, will often give excellent results. There is as far as I am aware no theorem which states that the optimal functional for the regularized model will work well although in some specific cases it does (Huber (1981)).

3 Robust statistics

3.1 Folklore and small print

Although statistics makes use of mathematics and has its own mathematical theorems it not an exact science in the sense that mathematics is. All theorems are proved under assumptions which for the purpose of this paragraph we shall refer to as the small print. As statistical theorems are no more than mathematical theorems we may assume that they are correct but, as with all applied sciences, they must be interpreted to be of use to a practising statistician. When discussing the results of a theorem the small print is sometimes not mentioned and maybe even forgotten. As an example we mention the folklore that the MAD has the highest possible finite sample breakdown point for a scale functional. As it stands this statement is false as there are scale functionals whose finite sample breakdown point is never less than that of the MAD and which is at some distributions strictly higher. In Davies and Gather (2002) the highest possible breakdown point for a scale functional at any distribution is calculated (see also Davies (1993)) and they also give a scale functional which attains this at every distribution. The small print for the MAD is that all the sample values are different. If this is not the case the finite sample breakdown point of the MAD can be zero. Although it is rarely stated the finite sample breakdown point of a functional is a local property of the functional. One should speak of the finite sample breakdown point of the functional at a particular distribution. Sometimes this is done explicitly when reference is made to points "in general position" but often it is not. In other cases the folklore exists because of the lack of a theorem. It has for example been claimed that the middle of the shortest half-sample does not have an influence function. This is a very vague claim which may be interpreted as that there does not exist any distribution F at which the functional has an influence function. In this sense the claim is false. The influence function exists for Gaussian and other distributions (see Davies (1993)). A similar calculation for the Hampel-Rousseeuw least median of squares estimator for certain regression models can be made (Davies (1993)). In spite of the formal nature of such calculations they do seem to aid the understanding of the behaviour of the LMS-functional with respect to inliers (see Ellis (1998) and Sheather, McKean and Hettmansperger (1998)). Although every discipline has folklore and probably cannot do without it, it does pose a danger. It is easier to make plausibility statements than prove results and if this becomes an accepted practice in statistics it can lead to a situation where we no longer know what always holds, what often holds, what occasionally holds and what never holds and under what conditions these various alternatives are correct. On the other hand the danger with mathematical theorems is that the small print is often not taken very seriously. Statistics is not the only discipline with this problem. In Hooft (1997) the Physics Nobel prize winner Gerard 't Hooft describes his problems with so called "no-go theorems" in physics. He writes

> One often forgets to mention the small-print, so that such theorems sometimes unjustifiably keep us from investigating important possibilities.

3.2 Boundedness, continuity and differentiability

Given a functional T on the space of probability measures the modulus of continuity or bias of T at the point F with respect to the metric $d_{\mathcal{C}}$ is defined by

$$b(T, F, \varepsilon, d_{\mathcal{C}}) = \sup\{\|T(G) - T(F)\| : G \in B(F, \varepsilon, d_{\mathcal{C}})\}. \tag{22}$$

In (22) we have assumed that the functional is uniquely defined at each distribution at least in the appropriate ball centred at F. In fact we do not require this as we can take the supremum over all possible values and if the functional is not defined at all for some G in $B(F, \varepsilon, d_{\mathcal{C}})$ then we set $b(T, F, \varepsilon, d_{\mathcal{C}}) = \infty$. If T is locally bounded, that is $b(T, F, \varepsilon, d_{\mathcal{C}}) < \infty$ for some $\varepsilon > 0$ then it has a non-zero breakdown point. Such a functional is useful at least in exploratory data analysis as it can in principal be used to detect outliers or observations which do not conform to the structure of the majority of the observations, assuming such a structure to exist. The usefulness of T will depend on the size of $b(T, F, \varepsilon, d_{\mathcal{C}})$ and this will in general be larger if T is set-valued at some G in the ball. If T is continuous at F then

$$\lim_{\varepsilon \downarrow 0} b(T, F, \varepsilon, d_{\mathcal{C}}) = 0$$

This will in general be of help in exploratory data analysis but it is of little help in confirmatory data analysis by which we mean attempting to answer questions of significance and the size of approximation intervals. For this a much stronger property is required namely that the functional should be locally uniformly Fréchet differentiable. In other words for all F there should exist functions $I(x, T, G)$ and positive constants C and ε satisfying

$$\sup\{\|I(x, T, G)\| : x, G \in B(F, \varepsilon, d_{\mathcal{C}})\} < C$$

and such that for every $\delta > 0$ there exists an $\eta > 0$ for which

$$\|T(H) - T(G) - \int I(x, T, G)\, d(H - G)\| \leq \delta d_{\mathcal{C}}(H, G)$$

for all $H \in B(G, \eta, d_{\mathcal{C}})$ (see Davies (1998) and Clarke (2000)). If this form of locally uniformly Fréchet differentiability does not hold then one can expect some form of non-smooth behaviour such as non-Gaussian limiting distributions or non-uniform asymptotics. In the one dimensional location-scale problem it is quite easy to produce functionals which are locally uniformly Fréchet differentiable (Davies (1998)). In higher dimensions it is more difficult. Kent and Tyler (1991) have shown that M-functionals based on the multivariate t-distributions have a non-zero breakdown point and, although no-one seems to have done the necessary calculations, it seems quite clear that their functional is locally uniformly Fréchet differentiable. The conditions for the existence and uniqueness of their functional are weak in that they can be expressed in terms of a weak metric. They involve only the amount of mass on lower dimensional hyperplanes. The only drawback is that the breakdown point is at most $1/(1+k)$ for dimension k. It seems to be a very difficult problem producing high breakdown location and scatter functionals which are well defined at each non-degenerate distribution. This has been done by Dietel (1993) for the multidimensional location and scatter and the linear regression problems. He even managed to obtain a locally uniform Lipschitz condition but not locally uniform Fréchet differentiability. No progress seems to have been made since his work.

3.3 The gross error model

The gross error neighbourhood of size $\varepsilon, 0 \le \varepsilon \le 1$ of a distribution F is defined by

$$GE(F, \varepsilon) = \{G : G = (1-\varepsilon)F + \varepsilon H \ H \text{ arbitrary}\}. \tag{23}$$

The interpretation is that we have a model F which is not thought to hold exactly and to allow for this one considers an amount of contamination ε which is represented by the distribution H. If samples are generated from a G in $\mathrm{GE}(F, \varepsilon)$ then on average a proportion $1-\varepsilon$ of the observations will come from the model F and a proportion ε of observations, the so called contaminants, will come from the distribution H. The gross error model is therefore often regarded as a simple and reasonable way of modelling deviations from a purported model F and can therefore be used to investigate and evaluate the behaviour of functionals under deviations. Its main advantage from the point of view of research is that it is relatively easy to analyse, compared with other ways of considering deviations such as those defined in terms of metrics. We refer to the discussion on pages 400-401 of Hampel, Rousseeuw, Ronchetti and Stahel (1986). This programme is only justified if the results obtained from the gross error model offer an insight into the behaviour of functionals when applied to real data. In this section we argue that this is sometimes but not always the case.

The use of the gross error model for generating outliers or contaminants was mentioned above. It has been criticized by Tukey (1960) and Gather (1990). The data are usually generated as i.i.d. using a distribution $G \in \mathrm{GE}(F, \varepsilon)$. Although the i.i.d. random variables may be a reasonable approximation for the main set of data there is no reason why this should be the case for the outliers or exotic observations. An example is the balloon data of Davies and Gather (1993) where the outliers are clearly dependent and cannot be adequately approximated by i.i.d. random variables. A further criticism is that the number of contaminants in the sample is itself a random variable and that they may lie within the body of the data set and not be extreme values. Tukey (1960) (see also Gather (1990)) considered the model G with $F = N(0,1)$, $H = N(0,9)$ and $\varepsilon = 0.1$. A sample of size $n = 1000$ generated using this model has on average about 12 observations from the $N(0,1)$ distribution which lie outside the interval $(-2.5, 2.5$ whereas only about 4 of the observations from the $N(0,9)$ distribution lie outside of it. A better way of generating outliers is described in Davies and Gather (1993). The main criticism of the gross error neighbourhood is that it is too small. The use of strong metrics such as total deviation was criticized above. In one sense the gross error model is even worse as for any $\varepsilon > 0$ we have

$$GE(F,\varepsilon) \subset B(F,\varepsilon,d_{tv}) \tag{24}$$

In particular for any continuous distribution F no empirical measure G_n deriving from any $G \in GE(F,\varepsilon)$ lies in $GE(F,\delta)$ for any $\delta < 1$. This means that results which apply for all $G \in GE(F,\varepsilon)$ are not guaranteed to hold for any G_n, not even approximately. This is not to say that a particular result does not hold, it may well as is shown by Huber's proof of the fact the median minimizes the maximum bias over the gross error neighbourhood GE$(N(0,1),\varepsilon)$ for any $\varepsilon < 0.5$. His proof however can be carried over to the Kolmogoroff neighbourhood $B(N(0,1),\varepsilon,d_{ko})$. Indeed the idea of the proof can be used to give a lower bound for the maximum bias over $B(F,\varepsilon,d_{ko})$ for any F, even empirical distributions, which can then be compared with the behaviour of the median. Another example of where results based on the gross error neighbourhood are applicable is Martin and Zamar (1993). The authors consider the class of M-functionals and then within this class they determine that functional which minimizes the bias over a gross error neighbourhood. The restriction to M-functionals means that the resulting functional can be applied to real data sets and its performance compared with those of others. Finally we give an example where considerable care in interpreting the result is required to avoid misleading conclusions. As in He and Simpson (1993) we set

$$b(T,F,\varepsilon,GE) = \sup\{\|T(G) - T(F)\| : G \in B(F,\varepsilon,GE)\} \tag{25}$$

and consider a parametric family F_θ of distributions. The minimum distance functional T_{mdtv} based on the total deviation metric d_{tv} is defined by

$$T_{mdtv}(G) = \text{argmin}_\theta\, d_{tv}(G,F_\theta). \tag{26}$$

Suppose now that F_θ is the multinormal family $N(\theta, I_p)$ which is the final example in Section 2.4 of He and Simpson (1993). It follows from Theorem 2.1 and Corollary 2.2 of He and Simpson (1993) that for small ε

$$c_1\varepsilon \le b(T_{mdtv},F_\theta,\varepsilon,GE) \le c_2\varepsilon \tag{27}$$

where c_1 and c_2 are independent of the dimension p (see Section 2.2 of He and Simpson (1993)). If T_M denotes an M-functional with a bounded and sufficiently smooth ψ-function then T_M is locally linear and Theorem 2.2 of He and Simpson (1993) gives

$$b(T_M, F_\theta, \varepsilon, GE) \le c_3\sqrt{p}\varepsilon. \tag{28}$$

Consider now the metric D_{hsp} on the set of probability measures on $\mathbb{R}^p$

$$d_{hsp}(P, Q) = \sup\{|P(H) - Q(H)| : H \text{ half-space}\}. \tag{29}$$

It follows from results of empirical process theory that if $\varepsilon_n = C/\sqrt{n}$ then with probability one as $C \to \infty$ the empirical measures $F_{\theta,n}$ deriving from n i.i.d. random with the distribution F_θ will lie in $B(F_\theta, \varepsilon_n, d_{hsp})$. If T_M is based on a sufficiently smooth ψ-function (28) continues to hold with d_{hsp} in place of GE and with $\varepsilon = \varepsilon_n$. Suppose that (27) also continues to hold. In this case we can deduce

$$\|T_{mdtv}(F_{\theta,n}) - \theta\| \le c_4/\sqrt{n} \tag{30}$$

with c_4 independent of the dimension p. In other words we can estimate the mean of a normal distribution in p dimensions with the same order of accuracy as in one dimension. Of course if the model is continuous the one cannot apply the minimum distance functional to empirical distributions using the total variation distance. He and Simpson (1993) therefore suggest smoothing the empirical distributions using say a kernel method so that a comparison is now possible. Apart from the considerable difficulties of doing this it would still not alter that fact that (30) is not possible. In other words (27) cannot be applied to data situations in contrast to the result on the bias optimality of the median. The reason why (27) holds is because of the extremely small size of the gross error neighbourhood. Other results in He and Simpson (1993) which are based on the gross error model can be carried over to data situations, in particular when the measures involved are discrete. The point is however that it is not immediately clear when this is possible and when it is not possible. The best advice is not to use the gross error model and to replace it by neighbourhoods defined by weak metrics or, if it is used, to indicate to what extent the results are relevant to the analysis of data.

3.4 Densities

The relationship between robustness and densities with respect to Lebesgue measure is somewhat ambiguous. Two examples where densities are of little import are the following. Consider the bias $b(T, F, \varepsilon, d)$ of a functional T at the distribution F with respect to the metric d as defined by (22). If $d = d_{\mathcal{C}}$ is a weak metric of the form (1) then it is immaterial as whether F has a density or not. A second example is where we calculate the efficiency of a robust functional on some test bed. As argued above it only makes sense to do this for bland models and as blandness implies smoothness we are lead to investigating the behaviour of robust functionals on test beds with densities. Again this is innocuous.

In other situations densities are important but not innocuous. Davies (1987) considered S-functionals in the context of the k-dimensional location-scale problem. In order to show that the functional is well defined at some distribution F he assumed that F had a density of the form $cf(\|x-\mu\|_\Sigma)$ where f is a smooth decreasing function on $\mathbb{R}_+ = [0, \infty)$. These assumptions have since been weakened by Tatsuoka and Tyler (2000) but it remains the case that it must be assumed that F has a density which satisfies some regularity conditions. Similar assumptions are required in other situations such as linear regression. Unfortunately there is a tendency to dismiss these and similar assumptions with phrases such as "under weak conditions" or "under general conditions". Such conditions are at variance with the aims of robust statistics which are to provide tools which will help to stabilize the analysis of data. If in any neighbourhood of a distribution there are distributions at which the functional is not uniquely defined then the suspicion is that it will not be stable. The gold standard for everyday stability of analysis is locally uniform Fréchet differentiability and this cannot hold if the functional is not even well defined within any non-trivial neighbourhood. This does not mean that such functionals are of no use. They may indeed be very valuable but their use will be restricted to exploratory data analysis.

3.5 Calculability

Apart from the problems of definability and uniqueness mentioned above the difficulty in calculating multidimensional location and scatter functionals is one of their main weaknesses. As far as I am aware the only multidimensional location and scatter functionals with a non-zero breakdown point and which can be easily calculated using a convergent algorithm are those of Kent and Tyler (1991). Other functionals can be calculated such as the Hampel-Rousseeuw least median of squares (Hampel (1975) and Rousseeuw (1984)) but the only known algorithm which yields the correct solutions is that of Stromberg (1993). This is however of such a complexity that it is not possible to compute it in reasonable time for samples of size $n = 500$ even in the case of a simple linear regression. Fortunately it is not the case that the lack of exact calculability makes the functionals non-applicable. There exist some ingenious algorithms which perform very well in practice although they do not yield the exact solution (Rocke and Woodruff (1996)). They are an invaluable tool in exploratory data analysis and for detecting outliers in high dimensions.

3.6 Breakdown

The most successful area of robust statistics and the one which seems to have been the subject of the most research is the location-scatter problem in Euclidean space. Apart from the fact that it is an important and difficult problem the reason for this attention may be the fact that it is possible to define affinely equivariant functionals which attain at least asymptotically the highest possible breakdown point. This emphasis on functionals with the optimal breakdown point may have been taken too far and lead to a situation where other approaches have not been sufficiently investigated. We mention two examples. Firstly the emphasis on breakdown *point* is too pessimistic. There are examples where it makes more sense to talk of breakdown *patterns* as the breakdown of a functional may depend not only on their number but also on their position in the data set. We mention the two-way table where the breakdown patterns for the case of one observation per cell were characterized in Terbeck and Davies (1998). Thus in the two-way table where each factor has five levels the breakdown point is 0.12. However the L^1- functional can withstand up to 6 aberrant observations depending on their location and this is optimal. A more general result for L^1-regression is given by Ellis and Morgenthaler (1992)). The existence of a highest possible breakdown point is only then of importance if it is non-trivial. In Davies and Gather (2002) it is argued that this is intimately connected with a large group of transformations which leave the problem unchanged. In the location-scatter problem this is the group of affine transformations. If the group of transformations which leave the problem unchanged is not sufficiently large then the breakdown point will be 1 and this can often be attained by the constant functional. As an example we mention an autoregressive process of order p. This problem remains invariant under the shift operator and non-zero scalar multiplication. However the constant functional $T_0(\mathbb{P}) = (0, \ldots, 0)^t$ is consistent with this group of transformations and has breakdown point 1. This functional is not Fisher consistent but may be modified to be so as follows: put $T_m(\mathbb{P}) = T(\mathbb{P})$ if $T(\mathbb{P})$ defines a stationary process and $T_m(\mathbb{P}) = (0, \ldots, 0)^t$ otherwise. Then T_m is Fisher consistent at Gaussian models and consistent in the usual sense at empirical Gaussian data derived from a Gaussian model. The breakdown point remains 1. It seems that Fisher consistency alone is not sufficiently stringent in order to be able to give a useful definition of breakdown point. Rather it is the lack of a sufficiently large group of transformations which has defeated all attempts to provide a satisfactory definition of breakdown in time series. This would also seem to apply to other structured situations such as those defined by graphical models. Some other ideas would seem to be required.

References

[Bednarski, 1993] Bednarski, T. (1993). Féchet differentiability and robust estimation. In Mandl, P. and Husková, M., editors, *Asymptotic Statistics: Proceedings of the Fifth Prague Symposium*, Springer Lecture Notes, pages 49–58. Springer.

[Bednarski and Clarke, 1998] Bednarski, T. and Clarke, B. R. (1998). On locally uniform expansions of regular functionals. *Discussiones Mathematicae: Algebra and Stochastic Methods*, 18:155–165.

[Clarke, 2000] Clarke, B. (2000). A review of differentiability in relation to robustness with an application to seismic data analysis. *Proceedings of the Indian Science Academy, Series A, Physical Sciences*, 66(5):467–482.

[Cohen, 1991] Cohen, M. (1991). The background of configural polysampling: a historical perspective. In Morgenthaler, S. and Tukey, J. W., editors, *Configural Polysampling: A Route to Practical Robustness*, chapter 2. Wiley, New York.

[Davies and Gather, 2002] Davies, P. and Gather, U. (2002). Breakdown and groups. Technical Report 10/2002, SFB 475, University of Dortmund, Dortmund, Germany.

[Davies, 1987] Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15:1269–1292.

[Davies, 1993] Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics*, 21:1843–1899.

[Davies, 1998] Davies, P. L. (1998). On locally uniformly linearizable high breakdown location and scale functionals. *Annals of Statistics*, 26:1103–1125.

[Davies and Kovac, 2001] Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Annals of Statistics*, 29(1):1–65.

[Dietel, 1993] Dietel, G. (1993). *Global location and dispersion functionals*. PhD thesis, University of Essen.

[Donoho and Liu, 1988] Donoho, D. and Liu, R. (1988). Pathologies of some minimum distance estimators. *Annals of Statistics*, 16(2):587–605.

[Donoho et al., 1995] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society*, 57:371–394.

[Ellis, 1998] Ellis, S. P. (1998). Instability of least squares, least absolute deviation and least median of squares linear regression. *Statistical Science*, 13(4):337–350.

[Ellis and Morgenthaler, 1992] Ellis, S. P. and Morgenthaler, S. (1992). Leverage and breakkdown in l_1 regression. *Journal of the American Statistical Association*, 87:143–148.

[Gather, 1990] Gather, U. (1990). Modelling the occurence of multiple outliers. *Allgemeines Statistisches Archiv*, 74:413–428.

[Hampel et al., 1986] Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[Hampel, 1975] Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods (with discussion). In *Proceedings of the 40th Session of the ISI*, volume 46, Book 1, pages 375–391.

[He and Simpson, 1993] He, X. and Simpson, D. G. (1993). Lower bounds for contamination bias: globally minimax versus locally linear estimation. *Annals of Statistics*, 21(1):314–337.

[Hewitt and Stromberg, 1969] Hewitt, E. and Stromberg, K. (1969). *Real and Abstract Analysis*. Springer, Berlin, Heidelberg.

[Hooft, 1997] Hooft, G. t. (1997). *In search of the ultimate building blocks*. Cambridge University Press, Cambridge, U.K.

[Huber, 1981] Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

[Kent and Tyler, 1991] Kent, J. T. and Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics*, 19:2102–2119.

[Martin and Zamar, 1993] Martin, R. D. and Zamar, R. H. (1993). Bias robust estimation of scale. *Annals of Satistics*, 21(2):991–1017.

[Rocke and Woodruff, 1996] Rocke, D. M. and Woodruff, D. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061.

[Rousseeuw, 1984] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

[Sheather et al., 1997] Sheather, S. J., McKean, J. W., and Hettmansperger, T. P. (1997). Finite sample stability properties of the least median of squares estimator. *Journal of Statistical Computing and Simulation*, 58(4):371–383.

[Stromberg, 1993] Stromberg, A. J. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal of Scientific Computing*, 14(6):1289–1299.

[Tatsuoka and Tyler, 2000] Tatsuoka, K. S. and Tyler, D. E. (2000). On the uniqueness of s-functionals and m-functionals under non-elliptic distributions. *Annals of Statistics*, 28(4):1219–1243.

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

[Tukey, 1960] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Olkin, I., editor, *Contributions to Proabaility and Staistics*. Stanford University Press, Stanford, California.

[Tukey, 1993] Tukey, J. W. (1993). Exploratory analysis of variance as providing examples of strategic choices. In S.Morgenthaler, E.Ronchetti, and W.A.Stahel, editors, *New Directions in Statistical Data Analysis and Robustness*, Basel. Birkhäuser.