

CHAPTER 1

GENERALITIES

1.1 WHY ROBUST PROCEDURES?

Statistical inferences are based only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation. Even in the simplest cases, there are explicit or implicit assumptions about randomness and independence, about distributional models, perhaps prior distributions for some unknown parameters, and so on.

These assumptions are not supposed to be exactly true—they are mathematically convenient rationalizations of an often fuzzy knowledge or belief. As in every other branch of applied mathematics, such rationalizations or simplifications are vital, and one justifies their use by appealing to a vague continuity or stability principle: a minor error in the mathematical model should cause only a small error in the final conclusions.

Unfortunately, this does not always hold. Since the middle of the 20th century, one has become increasingly aware that some of the most common statistical procedures (in particular, those optimized for an underlying normal distribution) are excessively

sensitive to seemingly minor deviations from the assumptions, and a plethora of alternative “robust” procedures have been proposed.

The word “robust” is loaded with many—sometimes inconsistent—connotations. We use it in a relatively narrow sense: for our purposes, *robustness signifies insensitivity to small deviations from the assumptions*.

Primarily, we are concerned with *distributional robustness*: the shape of the true underlying distribution deviates slightly from the assumed model (usually the Gaussian law). This is both the most important case and the best understood one. Much less is known about what happens when the other standard assumptions of statistics are not quite satisfied and about the appropriate safeguards in these other cases.

The following example, due to Tukey (1960), shows the dramatic lack of distributional robustness of some of the classical procedures.

■ EXAMPLE 1.1

Assume that we have a large, randomly mixed batch of n “good” and “bad” observations x_i of the same quantity μ . Each single observation with probability $1 - \varepsilon$ is a “good” one, with probability ε a “bad” one, where ε is a small number. In the former case x_i is $\mathcal{N}(\mu, \sigma^2)$, in the latter $\mathcal{N}(\mu, 9\sigma^2)$. In other words all observations are normally distributed with the same mean, but the errors of some are increased by a factor of 3.

Equivalently, we could say that the x_i are independent, identically distributed with the common underlying distribution

$$F(x) = (1 - \varepsilon)\Phi\left(\frac{x - \mu}{\sigma}\right) + \varepsilon\Phi\left(\frac{x - \mu}{3\sigma}\right), \quad (1.1)$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy \quad (1.2)$$

is the standard normal cumulative.

Two time-honored measures of scatter are the mean absolute deviation

$$d_n = \frac{1}{n} \sum |x_i - \bar{x}| \quad (1.3)$$

and the root mean square deviation

$$s_n = \left[\frac{1}{n} \sum (x_i - \bar{x})^2 \right]^{1/2}. \quad (1.4)$$

There was a dispute between Eddington (1914, p. 147) and Fisher (1920, footnote on p. 762) about the relative merits of d_n and s_n . Eddington had advocated the use of the former: “This is contrary to the advice of most textbooks; but it can be shown to be true.” Fisher seemingly settled the matter

by pointing out that for identically distributed normal observations s_n is about 12% more efficient than d_n .

Of course, the two statistics measure different characteristics of the error distribution. For instance, if the errors are exactly normal, s_n converges to σ , while d_n converges to $\sqrt{2/\pi} \sigma \cong 0.80\sigma$. So we must be precise about how their performances are to be compared; we use the asymptotic relative efficiency (ARE) of d_n relative to s_n , defined as follows:

$$\text{ARE}(\varepsilon) = \lim_{n \rightarrow \infty} \frac{\text{var}(s_n)/(E s_n)^2}{\text{var}(d_n)/(E d_n)^2} = \frac{\frac{1}{4} \left[\frac{3(1+80\varepsilon)}{(1+8\varepsilon)^2} - 1 \right]}{\frac{\pi(1-8\varepsilon)}{2(1+2\varepsilon)^2} - 1}. \quad (1.5)$$

The results are summarized in Exhibit 1.1.

ε	ARE(ε)
0	0.876
0.001	0.948
0.002	1.016
0.005	1.198
0.01	1.439
0.02	1.752
0.05	2.035
0.10	1.903
0.15	1.689
0.25	1.371
0.5	1.017
1.0	0.876

Exhibit 1.1 Asymptotic efficiency of mean absolute deviation relative to root mean square deviation. From Huber (1977b), with permission of the publisher.

The result is disquieting: just 2 bad observations in 1000 suffice to offset the 12% advantage of the mean square error, and $\text{ARE}(\varepsilon)$ reaches a maximum value greater than 2 at about $\varepsilon = 0.05$. This is particularly unfortunate since in the physical sciences typical "good data" samples appear to be well modeled by an error law of the form (1.1) with ε in the range between 0.01 and 0.1. (This does not imply that these samples contain between 1% and 10% gross errors, although this is very often true: the above law (1.1) may just be a convenient description of a slightly longer-tailed than normal distribution.) Thus it becomes painfully clear that the naturally occurring deviations from the idealized model are large enough to render meaningless the traditional asymptotic optimality theory: in practice, we should certainly prefer d_n to s_n , since it is better for all ε between 0.002 and 0.5.

To avoid misunderstandings, we should hasten to emphasize what is *not* implied here. First, the above does not imply that we advocate the use of the mean absolute deviation (there are still better estimates of scale). Second, some people have argued that the example is unrealistic insofar as the “bad” observations will stick out as outliers, so any conscientious statistician will do something about them before calculating the mean square error. This is beside the point: outlier rejection followed by the mean square error might very well beat the performance of the mean absolute error, but we are concerned here with the behavior of the *unmodified* classical estimates.

The example clearly has to do with longtailedness: lengthening the tails of the underlying distribution explodes the variance of s_{π} (d_{π} is much less affected). Shortening the tails, on the other hand, produces quite negligible effects on the distributions of the estimates. (It may impair the absolute efficiency by decreasing the asymptotic Cramér–Rao bound, but the latter is so unstable under small changes of the distribution that this effect cannot be taken very seriously.)

The sensitivity of classical procedures to longtailedness is typical and not limited to this example. As a consequence, “distributionally robust” and “outlier resistant,” although conceptually distinct, are practically synonymous notions. Any reasonable, formal or informal, procedure for rejecting outliers will prevent the worst.

We might therefore ask whether robust procedures are needed at all; perhaps a two-step approach would suffice:

- (1) First clean the data by applying some rule for outlier rejection.
- (2) Then use classical estimation and testing procedures on the remainder.

Would these steps do the same job in a simpler way?

Unfortunately they will not, for the following reasons:

- It is rarely possible to separate the two steps cleanly; for instance, in multi-parameter regression problems outliers are difficult to recognize unless we have reliable, robust estimates for the parameters.
- Even if the original batch of observations consists of normal observations interspersed with some gross errors, the cleaned data will not be normal (there will be statistical errors of both kinds: false rejections and false retentions), and the situation is even worse when the original batch derives from a genuine nonnormal distribution, instead of from a gross-error framework. Therefore the classical normal theory is not applicable to cleaned samples, and the actual performance of such a two-step procedure may be more difficult to work out than that of a straight robust procedure.
- It is an empirical fact that the best rejection procedures do not quite reach the performance of the best robust procedures. The latter apparently are superior

because they can make a smooth transition between full acceptance and full rejection of an observation. See Hampel (1974a, 1985), and Hampel *et al.* (1986, pp. 56–71).

- The same empirical study also had shown that many of the classical rejection rules are unable to cope with multiple outliers: it can happen that a second outlier masks the first, so that none is rejected, see Section 11.1.

Among these four reasons, the last is the crucial one. Its existence and importance had not even been recognized in advance of the holistic robustness approach.

1.2 WHAT SHOULD A ROBUST PROCEDURE ACHIEVE?

We are adopting what might be called an “applied parametric viewpoint”: we have a parametric model, which hopefully is a good approximation to the true underlying situation, but we cannot and do not assume that it is exactly correct. Therefore any statistical procedure should possess the following desirable features:

- **Efficiency:** It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.
- **Stability:** It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly, that is, the latter (described, say, in terms of the asymptotic variance of an estimate, or of the level and power of a test) should be close to the nominal value calculated at the model.
- **Breakdown:** Somewhat larger deviations from the model should not cause a catastrophe.

All three aspects are important. And one should never forget that robustness is based on compromise, as was most clearly enunciated by Anscombe (1960) with his insurance metaphor: sacrifice some efficiency at the model, in order to insure against accidents caused by deviations from the model.

It should be emphasized that the occurrence of gross errors in a small fraction of the observations is to be regarded as a small deviation, and that, in view of the extreme sensitivity of some classical procedures, a primary goal of robust procedures is to safeguard against gross errors.

If asymptotic performance criteria are used, some care is needed. In particular, the convergence should be uniform over a neighborhood of the model, or there should be at least a one-sided uniform bound, because otherwise we cannot guarantee robustness for any finite n , no matter how large n is. This point has often been overlooked.

Asymptotic versus finite sample goals. In view of Tukey’s seminal example (Example 1.1) that had triggered the development of robustness theory, the initial

setup for that theory had been asymptotic, with symmetric contamination. The symmetry restriction has been a source of complaints, which however are unjustified, cf. the discussion in Section 4.9: a procedure that is minimax under the symmetry assumption is almost minimax when the latter is relaxed. A much more serious cause for worry has largely been overlooked, and is still being overlooked by many, namely that 1% contamination has entirely different effects in samples of size 5 or 1000. Thus, asymptotic optimality theory need not be relevant at all for modest sample sizes and contamination rates, where the expected number of contaminants is small and may fall below 1. Fortunately, this scaling question could be settled with the help of an exact finite sample theory; see Chapter 10. Remarkably, and rather surprisingly, it produced solutions that did not depend on the sample size. At the same time, this finite sample theory did away with the restriction to symmetric contamination.

Other goals. The literature contains many other explicit and implicit goals for robust procedures, for example, high asymptotic *relative efficiency* (relative to some classical reference procedures), or high *absolute efficiency*, and this either for completely arbitrary (sufficiently smooth) underlying distributions or for a specific parametric family. More recently, it has become fashionable to strive for the highest possible *breakdown point*. However, it seems to me that these goals are secondary in importance, and they should never be allowed to take precedence over the above-mentioned three.

1.2.1 Robust, Nonparametric, and Distribution-Free

Robust procedures persistently have been (mis)classified and pooled with nonparametric and distribution-free ones. In our view, the three notions have very little overlap.

A procedure is called *nonparametric* if it is supposed to be used for a broad, not parametrized set of underlying distributions. For instance, the sample mean and the sample median are *the* nonparametric estimates of the population mean and median, respectively. Although nonparametric, the sample mean is highly sensitive to outliers and therefore very non-robust. In the relatively rare cases where one is *specifically* interested in estimating the true population mean, there is little choice except to pray and use the sample mean.

A test is called *distribution-free* if the probability of falsely rejecting the null hypothesis is the same for all possible underlying continuous distributions (optimal robustness of validity). Typical examples are two-sample rank tests for testing equality between distributions. Most distribution-free tests happen to have a reasonably stable power and thus also a good robustness of total performance. But this seems to be a fortunate accident, since distribution-freeness does not imply anything about the behavior of the power function.

Estimates derived from a distribution-free test are sometimes also called distribution-free, but this is a misnomer: the stochastic behavior of point estimates is intimately connected with the power (not the level) of the parent tests and depends on

the underlying distribution. The only exceptions are interval estimates derived from rank tests: for example, the interval between two specified sample quantiles catches the true median with a fixed probability (but still the distribution of the length of this interval depends on the underlying distribution).

Robust methods, as conceived in this book, are much closer to the classical parametric ideas than to nonparametric or distribution-free ones. They are destined to work with parametric models: the only differences are that the latter are no longer supposed to be literally true, and that one is also trying to take this into account in a formal way.

In accordance with these ideas, we intend to standardize robust estimates such that they are *consistent estimates* of the unknown parameters *at the idealized model*. Because of robustness, they will not drift too far away if the model is only approximately true. Outside of the model, we then may *define* the parameter to be estimated in terms of the limiting value of the estimate—for example, if we use the sample median, then the natural estimand is the population median, and so on.

1.2.2 Adaptive Procedures

Stein (1956) discovered the possibility of devising nonparametric efficient tests and estimates. Later, several authors, in particular Takeuchi (1971), Beran (1974, 1978), Sacks (1975), and Stone (1975), described specific location estimates that are asymptotically efficient for all sufficiently smooth symmetric densities. Since we may say that these estimates adapt themselves to the underlying distribution, they have become known under the name of *adaptive procedures*. See also the review article by Hogg (1974).

In the mid-1970s adaptive estimates—attempting to achieve asymptotic efficiency at all well-behaved error distributions—were thought by many to be the ultimate robust estimates. Then Kjaassen (1980) proved a disturbing result on the lack of stability of adaptive estimates. In view of his result, I conjectured at that time that an estimate cannot be simultaneously adaptive in a neighborhood of the model and qualitatively robust at the model: to my knowledge, this conjecture still stands.

Adaptive procedures typically are designed for symmetric situations, and their behavior for asymmetric true underlying distributions is practically unexplored. In any case, adaptation to asymmetric situations does not make sense in the robustness context. The point is: if a smooth model distribution is contaminated by a tightly concentrated asymmetric contaminant, then Fisher information is dominated by the latter. But since that contaminant may be a mere bundle of gross errors, any information derived from it is irrelevant for the location parameter of interest.

The connection between adaptivity and robustness is paradoxical also for other reasons. In robustness, the emphasis rests much more on stability and safety than on efficiency. For extremely large samples, where at first blush adaptive estimates look particularly attractive, the statistical variability of the estimate falls below its potential bias (caused by asymmetric contamination and the like), and robustness

would therefore suggest to move toward a less efficient estimate, namely the sample median, that minimizes bias (see Section 4.2). We therefore prefer to follow Stein's original terminology and to classify adaptive estimates not under robustness, but under the heading of efficient nonparametric procedures.

The situation is somewhat different with regard to "modest adaptation": adjust a single parameter, such as the trimming rate, in order to obtain good results. Compare Jaeckel (1971b) and see also Exhibit 4.8. But even there, adaptation to individual samples can be counterproductive, since it impairs comparison between samples.

1.2.3 Resistant Procedures

A statistical procedure is called *resistant* (see Mosteller and Tukey, 1977, p. 203) if the value of the estimate (or test statistic) is insensitive to small changes in the underlying *sample* (small changes in all, or large changes in a few of the values). The underlying distribution does not enter at all. This notion is particularly appropriate for (exploratory) data analysis and is of course conceptually distinct from robustness. However, in view of Hampel's theorem (Section 2.6), the two notions are for all practical purposes synonymous.

1.2.4 Robustness versus Diagnostics

There seems to be some confusion between the respective roles of *diagnostics* and *robustness*. The purpose of robustness is to safeguard against deviations from the assumptions, in particular against those that are near or below the limits of detectability. The purpose of diagnostics is to find and identify deviations from the assumptions. Thus, outlier detection is a diagnostic task, while suppressing ill effects from them is a robustness task, and of course there is some overlap between the two. Good diagnostic tools typically are robust—it always helps if one can separate gross errors from the essential underlying structures—but the converse need not be true.

1.2.5 Breakdown point

The breakdown point is the smallest fraction of bad observations that may cause an estimator to take on arbitrarily large aberrant values. Shortly after the first edition of this book, there were some major developments in that area. The first was that we realized that the breakdown point concept is most useful in small sample situations, and that it therefore better should be given a finite sample definition, see Chapter 11. The second important issue is that although many single-parameter robust estimators happen to achieve reasonably high breakdown points, even if they were not designed to do so, this is not so with multiparameter estimation problems. In particular, all conventional regression estimates are highly sensitive to gross errors in the independent variables, and in extreme cases a single such error may cause breakdown. Therefore, a plethora of alternative regression procedures have been

devised whose goal is to improve the breakdown point with regard to gross errors in the independent variables. Unfortunately, it seems that these alternative approaches have gone overboard with attempts to maximize the breakdown point, disregarding important other aspects, such as having reasonably high efficiency at the model. It is debatable whether any of these alternatives even deserve to be called robust, since they seem to fail the basic stability requirement of robustness. An approach through data analysis and diagnostics may be preferable; see the discussion in Chapter 7, Sections 7.1, 7.9, and 7.12.

1.3 QUALITATIVE ROBUSTNESS

In this section, we motivate and give a formal definition of qualitative asymptotic robustness. For statistics representable as a functional T of the empirical distribution, qualitative robustness is essentially equivalent to weak(-star) continuity of T , and for the sake of clarity we first discuss this particular case.

Many of the most common test statistics and estimators depend on the sample (x_1, \dots, x_n) only through the empirical distribution function

$$F_n(x) = n^{-1} \sum \mathbf{1}_{\{x_i \leq x\}}, \quad (1.6)$$

or, for more general sample spaces, through the empirical measure

$$F_n = n^{-1} \sum \delta_{x_i}, \quad (1.7)$$

where δ_x stands for the pointmass 1 at x . That is, we can write

$$T_n(x_1, \dots, x_n) = T(F_n) \quad (1.8)$$

for some functional T defined (at least) on the space of empirical measures. Often T has a natural extension to a much larger subspace, possibly to the full space \mathcal{M} of all probability measures on the sample space. For instance, if the limit in probability exists, put

$$T(F) \doteq \lim_{n \rightarrow \infty} T(F_n), \quad (1.9)$$

where F is the true underlying common distribution of the observations. If a functional T satisfies (1.9), it is called *Fisher consistent* at F , or, in short, *consistent*.

■ EXAMPLE 1.2

The Test Statistic of the Neyman–Pearson Lemma. The most powerful tests between two densities p_0 and p_1 are based on a statistic of the form

$$\int \psi(x) F_n(dx) = \frac{1}{n} \sum \psi(x_i), \quad (1.10)$$

with

$$\psi(x) = \log \frac{p_1(x)}{p_0(x)}. \quad (1.11)$$

■ EXAMPLE 1.3

The maximum likelihood estimate of θ for an assumed underlying family of densities $f(x, \theta)$ is a solution of

$$\int \psi(x, \theta) F_n(dx) = 0, \quad (1.12)$$

with

$$\psi(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta). \quad (1.13)$$

■ EXAMPLE 1.4

The α -trimmed mean can be written as

$$\bar{X}_\alpha = \frac{1}{1 - 2\alpha} \int_\alpha^{1-\alpha} F_n^{-1}(t) dt. \quad (1.14)$$

■ EXAMPLE 1.5

The so-called Hodges–Lehmann estimate is one-half of the median of the convolution square:

$$\frac{1}{2} \text{med}(F_n * F_n). \quad (1.15)$$

REMARK: This is the median of all n^2 pairwise means $(x_i + x_j)/2$; the more customary versions use only the pairs $i < j$ or $i \leq j$, but are asymptotically equivalent.

Assume now that the sample space is Euclidean, or, more generally, a complete, separable metrizable space. We claim that, in this case, the natural robustness (more precisely, resistance) requirement for a statistic of the form (1.8) is that T should be continuous with respect to the weak(-star) topology. By definition this is the weakest topology in the space \mathcal{M} of all probability measures such that the map

$$F \mapsto \int \psi dF \quad (1.16)$$

from \mathcal{M} into \mathbb{R} is continuous whenever ψ is bounded and continuous. The converse is also true: if a linear functional of the form (1.16) is weakly continuous, then ψ must be bounded and continuous; see Chapter 2 for details.

The motivation behind our claim is the following basic resistance requirement. Take a linear statistic of the form (1.10) and make a small change in the sample, that is, make either small changes in all of the observations x_i (rounding, grouping) or large changes in a few of them (gross errors, blunders). If ψ is bounded and continuous, then this will result in a small change of $T(F'_n) = \int \psi dF'_n$. But if ψ is not bounded, then a single, strategically placed gross error can completely upset $T(F'_n)$. If ψ is not continuous, and if F'_n happens to put mass onto discontinuity points, then small changes in many of the x_i may produce a large change in $T(F'_n)$.

We conclude from this that our vague, intuitive notion of resistance or robustness should be made precise as follows: a linear functional T is robust everywhere if and only if (iff) the corresponding ψ is bounded and continuous, that is, iff T is weakly continuous.

We could take this last property as our definition and call a (not necessarily linear) statistical functional \mathcal{T} robust if it is weakly continuous.

But, following Hampel (1971), we prefer to adopt a slightly more general definition.

Let the observations x_i be independent identically distributed, with common distribution F , and let (T_n) be a sequence of estimates or test statistics $T_n = T_n(x_1, \dots, x_n)$. Then this sequence is called *robust at* $F = F_0$ if the sequence of maps of distributions

$$F \rightarrow \mathcal{L}_F(T_n), \tag{1.17}$$

mapping F to the distribution of T_n , is equicontinuous at F_0 . That is, if we take a suitable distance function d_* in the space \mathcal{M} of probability measures, metrizing the weak topology, then, for each $\varepsilon > 0$, there is a $\delta > 0$ and an $n_0 > 0$ such that, for all F and all $n \geq n_0$,

$$d_*(F_0, F) < \delta \Rightarrow d_*(\mathcal{L}_{F_0}(T_n), \mathcal{L}_F(T_n)) \leq \varepsilon. \tag{1.18}$$

If the sequence (T_n) derives from a functional $T_n = T(F'_n)$, then, as is shown in Section 2.6, this definition is essentially equivalent to weak continuity of \mathcal{T} .

Note the close formal analogy between this definition of robustness and stability of ordinary differential equations: let $y_x(\cdot)$ be the solution with initial value $y(0) = x$ of the differential equation

$$\frac{dy}{dt} = f(t, y).$$

Then we have stability at $x = x_0$ if, for all $\varepsilon > 0$, there is a $\delta > 0$ such that, for all x and all $t \geq 0$,

$$d(x_0, x) \leq \delta \Rightarrow d(y_{x_0}(t), y_x(t)) \leq \varepsilon.$$

1.4 QUANTITATIVE ROBUSTNESS

For several reasons, it may be useful to describe quantitatively how greatly a small change in the underlying distribution F changes the distribution $\mathcal{L}_F(T_n)$ of an es-

timate or test statistic $T_n \equiv T_n(x_1, \dots, x_n)$. A few crude and simple numerical quantifiers might be more effective than a very detailed description.

To fix the idea, assume that $T_n = T(F_n)$ derives from a functional T . In most cases of practical interest, T_n is then consistent,

$$T_n \rightarrow T(F) \quad \text{in probability,} \quad (1.19)$$

and asymptotically normal,

$$\mathcal{L}_F\{\sqrt{n}[T_n - T(F)]\} \rightarrow \mathcal{N}(0, A(F, T)). \quad (1.20)$$

Then it is convenient to discuss the quantitative large sample robustness of T in terms of the behavior of its asymptotic bias $T(F) - T(F_0)$ and asymptotic variance $A(F, T)$ in some neighborhood $\mathcal{P}_\varepsilon(F_0)$ of the model distribution F_0 .

For instance, \mathcal{P}_ε might be a *Lévy neighborhood*,

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid \forall t, F_0(t - \varepsilon) - \varepsilon \leq F(t) \leq F_0(t + \varepsilon) + \varepsilon\}, \quad (1.21)$$

or a *contamination "neighborhood"*,

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid F = (1 - \varepsilon)F_0 + \varepsilon H, H \in \mathcal{M}\} \quad (1.22)$$

(the latter is not a neighborhood in the sense of the weak topology). Equation (1.22) is also called the *gross error model*.

The two most important characteristics then are the *maximum bias*

$$b_1(\varepsilon) = \sup_{F \in \mathcal{P}_\varepsilon} T(F) - T(F_0) \quad (1.23)$$

and the *maximum variance*

$$v_1(\varepsilon) = \sup_{F \in \mathcal{P}_\varepsilon} A(F, T). \quad (1.24)$$

We often consider a restricted supremum of $A(F, T)$ also, assuming that F varies only over some slice of \mathcal{P}_ε where $T(F)$ stays constant, for example, only over the set of symmetric distributions.

Unfortunately, the above approach to the problem is conceptually inadequate; we should like to establish that, for sufficiently large n , our estimate T_n behaves well for *all* $F \in \mathcal{P}_\varepsilon$. A description in terms of b_1 and v_1 would allow us to show that, for each *fixed* $F \in \mathcal{P}_\varepsilon$, T_n behaves well for sufficiently large n . The distinction involves an interchange in the order of quantifiers and is fundamental, but has been largely ignored in the literature. On this point, see in particular the discussion of superefficiency in Huber (2009).

A better approach is as follows. Let $M(F, T_n)$ be the median of $\mathcal{L}_F[T_n - T(F_0)]$ and let $Q_t(F, T_n)$ be a normalized t -quantile range of $\mathcal{L}_F(\sqrt{n} T_n)$, where, for any distribution G , the normalized t -quantile range is defined as

$$Q_t = \frac{G^{-1}(1-t) - G^{-1}(t)}{\Phi^{-1}(1-t) - \Phi^{-1}(t)}, \quad (1.25)$$

Φ being the standard normal cumulative. The value of t is arbitrary, but fixed, say $t = 0.25$ (interquartile range) or $t = 0.025$ (95% range, which is convenient in view of the traditional 95% confidence intervals). For a normal distribution, Q_t coincides with the standard deviation of G ; therefore Q_t^2 is sometimes called pseudo-variance.

Then define the maximum asymptotic bias and variance, respectively, as

$$b(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} |M(F, T_n)|, \tag{1.26}$$

$$v(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} Q_t(F, T_n)^2. \tag{1.27}$$

Theorem 1.1 *If b_1 and v_1 are well-defined, we have $b(\varepsilon) \geq b_1(\varepsilon)$ and $v(\varepsilon) \geq v_1(\varepsilon)$.*

Proof Let $T(F_0) = 0$ for simplicity and assume that T_n is consistent: $T(F_n) \rightarrow T(F)$. Then $\lim_n M(F, T_n) = T(F)$, and we have the following obvious inequality, valid for any $F \in \mathcal{P}_\varepsilon$:

$$b(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} |M(F, T_n)| \geq \lim_n |M(F, T_n)| = |T(F)|;$$

hence

$$b(\varepsilon) \geq \sup_{F \in \mathcal{P}_\varepsilon} |T(F)| = b_1(\varepsilon).$$

Similarly, if $\sqrt{n}[T_n - T(F)]$ has a limiting normal distribution, we have $\lim_n Q_t(F, T_n)^2 = A(F, T)$, and $v(\varepsilon) \geq v_1(\varepsilon)$ follows in the same fashion as above. ■

The quantities b and v are awkward to handle, so we usually work with b_1 and v_1 instead. We are then, however, obliged to check whether, for the particular \mathcal{P}_ε and T under consideration, we have $b_1 = b$ and $v_1 = v$. Fortunately, this is usually true.

Theorem 1.2 *If \mathcal{P}_ε is the Lévy neighborhood, then $b(\varepsilon) \leq b_1(\varepsilon + 0) = \lim_{\eta \downarrow \varepsilon} b_1(\eta)$.*

Proof According to the Glivenko–Cantelli theorem, we have $\sup_x |F_n(x) - F(x)| \rightarrow 0$ in probability, uniformly in F . Thus, for any $\delta > 0$, the probability of $F_n \in \mathcal{F}_\delta(F)$, and hence of $F_n \in \mathcal{P}_{\varepsilon+\delta}(F_0)$, will tend to 1, uniformly in F for $F \in \mathcal{P}_\varepsilon(F_0)$. Hence $b(\varepsilon) \leq b_1(\varepsilon + \delta)$ for all $\delta > 0$. ■

Note that, for the above types of neighborhoods, $\mathcal{P}_1 = \mathcal{M}$ is the set of all probability measures on the sample space, so $b(1)$ is the worst possible value of b (usually ∞). We define the *asymptotic breakdown point* of T at F_0 as

$$\varepsilon^* = \varepsilon^*(F_0, T) = \sup\{\varepsilon \mid b(\varepsilon) < b(1)\}. \tag{1.28}$$

Roughly speaking, the breakdown point gives the limiting fraction of bad outliers the estimator can cope with. In many cases ε^* does not depend on F_0 , and it is often the same for all the usual choices for \mathcal{P}_ε . Historically, the breakdown point was first

defined by Hampel (1968) as an asymptotic concept, like here. In Chapter 11, we shall, however, argue that it is most useful in small sample situations and shall give it a finite sample definition.

■ EXAMPLE 1.6

The breakdown point of the α -trimmed mean is $\varepsilon^* = \alpha$. (This is intuitively obvious; for a formal derivation see Section 3.3.)

Similarly we may also define an asymptotic variance breakdown point

$$\varepsilon^{*v} = \varepsilon^{*v}(F_0, T) = \sup\{\varepsilon \mid v(\varepsilon) < v(F)\}, \quad (1.29)$$

but this is a much less useful notion.

1.5 INFINITESIMAL ASPECTS

What happens if we add one more observation with value x to a very large sample? Its suitably normed limiting influence on the value of an estimate or test statistic $T(F_n)$ can be expressed as

$$IC(x, F, T) = \lim_{s \rightarrow 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}, \quad (1.30)$$

where δ_x denotes the pointmass 1 at x . The above quantity, considered as a function of x , was introduced by Hampel (1968, 1974b) under the name *influence curve* (IC) or *influence function*, and is arguably the most useful heuristic tool of robust statistics. It is treated in more detail in Section 2.5.

If T is sufficiently regular, it can be linearized near F in terms of the influence function: if G is near F , then the leading terms of a Taylor expansion are

$$T(G) = T(F) + \int IC(x, F, T)[G(dx) - F(dx)] + \dots, \quad (1.31)$$

We have

$$\int IC(x, F, T)F(dx) = 0, \quad (1.32)$$

and, if we substitute the empirical distribution F_n for G in the above expansion, we obtain

$$\begin{aligned} \sqrt{n}(T(F_n) - T(F)) &= \sqrt{n} \int IC(x, F, T)F_n(dx) + \dots \\ &= \frac{1}{\sqrt{n}} \sum IC(x_i, F, T) + \dots, \end{aligned} \quad (1.33)$$

By the central limit theorem, the leading term on the right-hand side is asymptotically normal with mean 0, if the x_i are independent with common distribution F . Since it is often true (but not easy to prove) that the remaining terms are asymptotically negligible, $\sqrt{n}[T(F_n) - T(F)]$ is then asymptotically normal with mean 0 and variance

$$A(F, T) = \int IC(x, F, T)^2 F(dx). \tag{1.34}$$

Thus the influence function has two main uses. First, it allows us to assess the relative influence of individual observations toward the value of an estimate or test statistic. If it is unbounded, an outlier might cause trouble. Its maximum absolute value,

$$\gamma^* = \sup_x |IC(x, F, T)|, \tag{1.35}$$

has been called the *gross error sensitivity* by Hampel. It is related to the maximum bias (1.23): take the gross error model (1.22), then, approximately,

$$T(F) - T(F_0) \approx \varepsilon \int IC(x, F_0, T)H(dx). \tag{1.36}$$

Hence

$$b_1(\varepsilon) = \sup_x |T(F) - T(F_0)| \approx \varepsilon \gamma^*. \tag{1.37}$$

However, some risky and possibly illegitimate interchanges of suprema and passages to the limit are involved here. We give two examples later (Section 3.5) where

- (1) $\gamma^* < \infty$ but $b_1(\varepsilon) \rightarrow \infty$ for all $\varepsilon > 0$;
- (2) $\gamma^* = \infty$ but $\lim b(\varepsilon) = 0$ for $\varepsilon \rightarrow 0$.

Second, the influence curve allows an immediate and simple, heuristic assessment of the asymptotic properties of an estimate, since it allows us to guess an explicit formula (1.34) for the asymptotic variance (which then has to be proved rigorously by other means).

There are several finite sample and/or difference quotient versions of (1.30), the most important being the *sensitivity curve* (Tukey 1970) and the *jackknife* (Quenouille 1956, Tukey 1958, Miller 1964, 1974). We obtain the sensitivity curve if we replace F by F_{n-1} and s by $1/n$ in (1.30):

$$\begin{aligned} SC_{n-1}(x) &= \frac{T\left(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\delta_x\right) - T(F_{n-1})}{1/n} \\ &= n[T_n(x_1, \dots, x_{n-1}, x) - T_{n-1}(x_1, \dots, x_{n-1})]. \end{aligned} \tag{1.38}$$

The jackknife is defined as follows. Consider an estimate $T_n(x_1, \dots, x_n)$ that is essentially the "same" across different sample sizes (for instance, assume that it is a

functional of the empirical distribution). Then the i th *jackknifed pseudo-value* is, by definition,

$$T_{ni}^* = nT_n - (n-1)T_{n-1}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n). \quad (1.39)$$

For example, if T_n is the sample mean, then $T_{ni}^* = x_i$. We note that $T_{ni}^* - T_n$ is an approximation to $IC(x_i)$; more precisely, if we substitute F_n for F and $-1/(n-1)$ for s in (1.30), we obtain

$$\begin{aligned} & \frac{T\left(\frac{n}{n-1}F_n - \frac{1}{n-1}\delta_{x_i}\right) - T(F_n)}{-1/(n-1)} \\ &= (n-1)[T_n - T_{n-1}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)] \\ &= T_{ni}^* - T_n. \end{aligned} \quad (1.40)$$

If T_n is a consistent estimate of θ , whose bias has the asymptotic expansion

$$E(T_n - \theta) = \frac{a_1}{n} + \frac{a_2}{n^2} + O\left(\frac{1}{n^3}\right), \quad (1.41)$$

then

$$T_n^* = \frac{1}{n} \sum_i T_{ni}^* \quad (1.42)$$

has a smaller bias:

$$E(T_n^* - \theta) = -\frac{a_2}{n^2} + O\left(\frac{1}{n^3}\right). \quad (1.43)$$

■ EXAMPLE 1.7

If $T_n = 1/n \sum (x_i - \bar{x})^2$, then

$$T_{ni}^* = -\frac{n}{n-1}(x_i - \bar{x})^2,$$

and (1.42) produces an unbiased estimate of σ^2 :

$$T_n^* = \frac{1}{n-1} \sum (x_i - \bar{x})^2.$$

Tukey (1958) pointed out that

$$\frac{1}{n(n-1)} \sum (T_{ni}^* - T_n^*)^2 \quad (1.44)$$

(a finite sample version of (1.34)) is usually a good estimator of the variance of T_n . It can also be used as an estimate of the variance of T_n^* , but actually it is better matched to T_n .

In some cases, namely when the influence function $IC(x; F, T)$ does not depend smoothly on F , the jackknife is in trouble and may yield a variance that is worse than useless. This happens, in particular, for estimates that are based on a small number of order statistics, like the median.

1.6 OPTIMAL ROBUSTNESS

In Section 1.4, we introduced some quantitative measures of robustness. They are certainly not the only ones. But, as we defined robustness to mean insensitivity with regard to small deviations from the assumptions, any quantitative measure of robustness must somehow be concerned with the maximum degradation of performance possible for an ε -deviation from the assumptions. An *optimally robust* procedure then minimizes this maximum degradation and hence will be a minimax procedure of some kind. As we have considerable freedom in how we quantize performance and ε -deviations, we also have a host of notions of optimal robustness, of various usefulness, and of various mathematical manageability.

Exact, finite sample minimax results are available for two simple, but important special cases: the first corresponds to a robustification of the Neyman–Pearson lemma, and the second yields interval estimates of location. They are treated in Chapter 10. Although the resulting tests and estimates are quite simple, the approach does not generalize well. In particular, it does not seem possible to obtain explicit, finite-sample results when there are nuisance parameters (e.g., when scale is unknown).

If we use asymptotic performance criteria (e.g., asymptotic variances), we obtain *asymptotic minimax estimates*, treated in Chapters 4–6. These asymptotic theories work well only if there is a high degree of symmetry (left–right symmetry, translation invariance, etc.), but they are able to cope with nuisance parameters. By a fortunate accident, some of the asymptotic minimax estimates, although derived under quite different assumptions, coincide with certain finite sample minimax estimates; this gives a strong heuristic support for using asymptotic optimality criteria.

Multiparameter regression, and the estimation of *covariance matrices* possess enough symmetries that the above asymptotic optimality results are transferable (Chapters 7 and 8). However the value of this transfer is somewhat questionable because of the fact that in practice the number of observations per parameter tends to be uncomfortably low. Other, design-related dangers, such as leverage points, may become more important than distributional robustness itself.

In problems lacking invariance, for instance in the general one-parameter estimation problem, Hampel (1968) has proposed optimizing robustness by minimizing the asymptotic variance at the model, subject to a bound on the gross-error sensitivity

γ^* defined by (1.35). This approach is technically straightforward, but it has some conceptual drawbacks: reassuringly, it again yields the same estimates as those obtained by the exact, finite sample minimax approach when the latter is applicable. For details, see Section 12.2.

1.7 PERFORMANCE COMPARISONS

In robustness, optimality (i.e., minimaxity) of a given procedure is an important aspect, but it must be regarded as part of a larger picture. In particular, it must be complemented by *performance comparisons*—for different sample sizes and underlying situations, and with other procedures. The so-called Princeton robustness study was a first, and exemplary, investigation of this kind, see Andrews *et al.* (1972).

The Princeton study showed up some intrinsic drawbacks of empirical sampling studies. The main one is that they only can give a collection of punctuated spotlights, since each simulation is done for one specific procedure and one specific situation (sample size and distribution). Even worse, the Monte Carlo sampling variability at each such spotlight may exceed the performance differences one is interested in (e.g., between the effects of the underlying distributions), for all practicable Monte Carlo sample sizes. The Princeton study managed to overcome this in part—that is, for suitably structured families of distributions—by Tukey’s “Monte Carlo Swindle”: utilize information available to the person conducting the Monte Carlo simulation, but not to the statistician applying the procedure. This “swindle” permits one to reduce the differential sampling variability. After the Princeton study, Tukey proposed an even more sophisticated approach based on the idea that any particular sample configuration can occur under any underlying distribution (provided the latter has a strictly positive density), but its probability of occurrence depends on the latter. This is the basis of the so-called *configural polysampling* method, see Morgenthaler and Tukey (1991).

Another approach to the investigation of the small sample behavior of robust estimates, avoiding empirical sampling altogether, is based on the so-called *small sample asymptotics*. This will be discussed in Chapter 14.

1.8 COMPUTATION OF ROBUST ESTIMATES

In many practical applications of (say) the method of least squares, the actual setting up and solving of the least squares equations occupies only a small fraction of the total length of the computer program. We should therefore strive for robust algorithms that can easily be patched into existing programs, rather than for comprehensive robust packages.

This is in fact possible. Technicalities are discussed in Chapter 7; the salient idea is to achieve robustness by modifying deviant observations.

To fix the ideas, assume that we are doing a least squares fit on observations y_i , yielding fitted values \hat{y}_i and residuals $r_i = y_i - \hat{y}_i$. Let s_i be some estimate of the standard error of y_i (or, even better, of the standard error of r_i).

We metrically Winsorize the observations y_i and replace them by pseudo-observations y_i^* :

$$y_i^* = \begin{cases} y_i & \text{if } |r_i| \leq cs_i, \\ \hat{y}_i - cs_i & \text{if } r_i < -cs_i, \\ \hat{y}_i + cs_i & \text{if } r_i > cs_i. \end{cases} \quad (1.45)$$

The constant c regulates the amount of robustness: good choices are in the range between 1 and 2, say $c = 1.5$.

We then use the pseudo-observations y_i^* in place of the y_i to calculate new fitted values \hat{y}_i , new residuals $r_i = y_i - \hat{y}_i$, and new s_i . We then use (1.45) to produce new pseudo-observations, and iterate to convergence.

If all observations are equally accurate, the classical estimate of the variance of a single observation would be

$$s^2 = \frac{1}{n-p} \sum r_i^2, \quad (1.46)$$

where $n-p$ is the number of observations minus the number of parameters, and we can then estimate the standard error of the residual r_i by $s_i = \sqrt{1-h_i}s$, where h_i is the i th diagonal element of the hat matrix $H = X(X^T X)^{-1}X^T$, see Chapter 7, Sections 7.2 and 7.9.

If we use modified residuals $r_i^* = y_i^* - \hat{y}_i$ instead of the r_i , we clearly would underestimate scale; we can correct this bias (to a zero order approximation), if we replace (1.46) by

$$s^2 = \frac{1}{\frac{n-p}{(m/n)^2}} \sum r_i^{*2}, \quad (1.47)$$

where m is the number of unmodified observations ($y_i^* = y_i$).

More elegantly, we can use the classical analysis of variance formulas if we move the correction factor into the residuals, that is, if we use boosted pseudo-residuals $(n/m)r_i^*$. In detail, this approach works as follows: we first determine robust fitted values \hat{y}_i as above and iterate to convergence. Then we determine the number m of unmodified residuals and boost all pseudo-residuals (whether or not they are affected by metric Winsorization). Finally, we apply the classical analysis of variance formulas to the boosted pseudo-observations

$$y_i^* = \hat{y}_i + (n/m)r_i^*. \quad (1.48)$$

This will give approximately correct results also for the estimated variances. See Section 7.10 for higher order bias corrections.

It is evident that this procedure deflates the influence of outliers. Moreover there are versions of this procedure that are demonstrably convergent; they converge to a reasonably well-understood M -estimate.

These ideas yield a completely general recipe to robustize any statistical procedure for which it makes sense to decompose the underlying observations into fitted values and residuals. Of course, such a recipe will work only if the fitted values are noticeably more accurate than the observations; see Section 7.9 for a discussion of the latter point. We first “clean” the data by pulling outliers towards their fitted values in the manner of (J.45) and re-fit iteratively until convergence is obtained, that is, until further cleaning no longer changes the fitted values. Then we apply the statistical procedure in question to the (boosted) pseudo-observations y_i^* . Compare Bickel (1976, p. 167), Huber (1979), and Kleiner *et al.* (1979) for nontrivial early examples.

1.9 LIMITATIONS TO ROBUSTNESS THEORY

Perhaps the most important purpose of robustness is to safeguard against occasional gross errors. Correspondingly, most approaches to robustness are based on the following intuitive requirement:

A discontant small minority should never be able to override the evidence of the majority of the observations.

We may say that this is a frequentist approach that makes sense only with relatively large sample sizes, since otherwise the notion of a “small minority” would be meaningless. It works well only for samples that under the idealized model derive from a single homogeneous population, and for statistical procedures that are invariant under permutation of the observations. In particular, one has to make sure that a small minority should not be able to overcome its smallness and to exercise undue power either by virtue of *position* or through *coalitions*. In order to prevent this in a theoretically clean and clear-cut way, we are practically forced to make an exchangeability requirement: the statistical problem (or at least the procedures used for dealing with it) should be invariant under arbitrary permutations.

Exchangeability does not sit well with structured problems. Very similar difficulties occur also with the bootstrap. Only partial remedies are possible. For example, in time series problems, it seems at first that it should be possible to satisfy the exchangeability requirement, since state space models permit one to reduce the ideal situation to i.i.d. innovations. However, some of the most typical corruptions against which one should safeguard in time series problems are clumps of bad values affecting contiguous observations. That is, one runs into problems with “coalitions” of bad observations. How should one formalize such coalitions? Moreover, in state space models, gross errors can enter the picture in several different places with quite different effects. The lack of convincing models is a very serious obstacle against developing a convincing theory of robustness in time series.

In regression, we encounter the other problem: high influence through position, see Chapter 7, in particular Sections 7.1, 7.9, and 7.12. In that case, the situation is very delicate. In my opinion, dealing with high positional influence requires what-if analyses and human judgment rather than a blind, automated robustness approach.

An approach to robustness that does not depend on sample size might be based on the following, admittedly vague, intuitive idea:

Make sure that uncertain parts of the evidence never have overriding influence on the final conclusions.

Such an approach, at least in principle, clearly applies also to small samples, and in particular, it permits one to formalize robustness with regard to uncertainties in a Bayesian prior (cf. Chapter 15). But it does not resolve the technical problems, and serious technical difficulties persist with small sample robustness theory, as well as with lack of exchangeability and with coalitions. Also, nuisance parameters continue to present a serious obstacle.

As a final remark, I should emphasize once more that robustness theory, as conceived here, is concerned with small deviations from a model. Thus two important limitations of that theory are that we need (i) a *model* and (ii) a notion of *smallness*. Unfortunately, much of the literature, in particular on robust regression, is sloppy with respect to model specification. Also, the currently fashionable (over-)emphasis of high breakdown points, that is, safeguarding against deviations that are not small in any conceivable sense of the word, transmits a wrong signal. A high breakdown point is nice to have, if it comes for free, but otherwise the strife for the highest possible breakdown point may be overly pessimistic. The presence of a substantial amount of contamination usually indicates a mixture model and calls for data analysis and diagnostics, whereas a thoughtless application of robust procedures might only hide the underlying problem. Moreover, all attempts to maximize the breakdown point seem to run into the notorious instability problems of “optimal” procedures (cf. Section 7.12). See Huber (2009) for the pitfalls of optimization.

