Some additional notes on the "Princeton Robustness Year"

by

Frank Hampel

Research Report No. 76
October 1995

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Some additional notes on the "Princeton Robustness Year"

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

October 1995

Abstract

About 25 years ago, Peter Bickel, Peter Huber and the author were invited to join John Tukey and spend the academic year 1970/71 at Princeton University in order to make a cooperative effort for progress in robust statistics. What later was called the "Princeton robustness year," became widely known for its big Monte Carlo study of "robust location estimates under symmetric longtailed distributions" published in the book by Andrews et al. (1972).

However, this "public" view of what happened is incomplete in several respects.

First, the four carriers of the Princeton robustness seminar spent half the time, namely the entire fall semester, discussing foundations of and approaches to robustness for a start, and then discussing the "really important" practical problems of robustness, mainly robust regression under widely different aspects. Especially Tukey gave a broad exposition of the pertaining challenges arising from data analysis. Many of the results and tentative results in the seminar were worked out and published only years later.

Secondly, the "Princeton robustness book," although containing a wealth of interesting information hard to digest and difficult to appreciate properly, still has various gaps concerning the Monte Carlo study. Some are only anecdotal, such as Tukey's finding a single outlier among several thousand numbers (Takeuchi in one situation) which was actually due to a copying error. The graphs for a highly condensed analysis of the data, reducing all numbers to 2-4 "parameters" per estimator and 1-2 per situation (Ch. 7C6 of the book), unfortunately had to be published elsewhere. A highly accurate numerical relationship between some of the Monte Carlo variances (nicely explainable by means of the change-of-variance sensitivity) was only discovered much later and is also briefly published elsewhere. A regrettable omission is the lack of an analysis of the few asymmetric situations in the study, which might have mitigated some of its later criticisms.

Another omission, which only seems to have become more interesting with time, is the brief history of the introduction of redescending M-estimators. Apart from purely historical and anecdotal aspects, I believe that it may help to give a psychological explanation for the gap – persisting through decades – between vigorous advocates of, and equally vigorous warners against their (uncritical) use.

Thirdly, the book by Andrews et al. contains only "wave A" of a much bigger Monte Carlo study subsequently done by Tukey and coworkers and extending apparently at least till "wave N." While it started out as an extension and "fine-tuning" of wave A, wave F and especially wave H, left by Tukey to be filled by Bickel, Huber and the author, contained different types of estimators for comparison, such as the mean combined with different hard rejection rules, various types of rank-based estimators

(both published elsewhere), and a new simple type of estimators (still waiting for analysis) called "trapped trimmed means."

1 Introduction.

For the academic year 1970/71, G.S. Watson, then chairman of the Princeton statistics department, invited P.J. Bickel (University of California, Berkeley), P.J. Huber (ETH Zurich) and the author (University of Zurich) to join J.W. Tukey (Princeton University and Bell Laboratories) for a cooperative effort for progress in robust statistics. The best–known, and virtually the only known outcome of this intensive research year is the book by Andrews et al. (1972) on the extensive Monte Carlo study of "robust estimates of location under symmetric longtailed distributions." On the one hand, the book contains a wealth of new information, methods, results, and insights, though it is hard to read and probably still not sufficiently appreciated. On the other hand, it has narrowed down the public conception of robustness, at least for a long time, to consist of the phrase in quotation marks above. This was unfortunate, for the book does not represent at all the broad scope of discussions on robustness during (and around) the "Princeton robustness year." Rather, it centers on "hard" empirical numbers in the simplest possible situation in order to provide a firm foundation for the discussion of more elusive, but also more powerful and penetrating theoretical concepts. Since little, apart from the book, is known outside the small circle of participants about the discussions before, during and after the Monte Carlo study, and since especially young statisticians express an increasing interest in learning more about this period, Tukey's 80^{th} birthday and the 25^{th} anniversary of the "Robustness year" may provide a suitable occasion to present some personal memories and notes from my files on this stimulating and exciting period.

2 The "Princeton Robustness Seminar."

During the Princeton robustness year, Bickel, Huber, Tukey, and the author (chairman) constituted a research seminar on robustness. There was usually one weekly session, with overflow sessions at Tukey's other seminar time or in one of the two weekly colloquia (outside speakers' seminars). In addition, there was during the first semester a special evening seminar where (mainly) students talked about (mainly) robustness in time series, including talks by A. Hoffman, Th. Gasser, R. Madden, P. Feder and G. Gruvaeus on clipping (smooth and hard limiting), bispectra, interpolation of time series, and on papers by Mandelbrot and Hannan, respectively.

The robustness seminar started on Sept. 21, 1970 with talks by Hampel (Sept. 21 and 23) on basic concepts of robustness, Huber (Sept. 28) on asymptotic aspects of robustness, Tukey (Sept. 30) on least squares, generalized variance etc., Bickel (Oct. 5) on estimators derived from tests, and on one–step estimators, Tukey (Oct. 12) on the jackknife, and Louis Jaeckel from Bell Labs (Oct. 14) on his Ph.D. thesis, in order to familiarize the other participants with one's own ideas and approaches. On Oct. 19, there were 2 short talks: by Bickel on asymmetry, and by David Andrews, who also visited Princeton during this year, on his personal view. There were already some smaller discussions, such as about estimating the center of a distribution symmetric in the middle, but asymmetric in the tails.

The handwritten original list of participants of the robustness seminar still exists (Fig. 1). It includes (apart from the chairman) 22 names (some not quite legible), namely:

P. Bickel, P. Huber, P. Feder, D.F. Andrews, L.M. Steinberg, H.J. Hoffman, J.R. Traversy (?), Stan, G. Gruvaeus, A. Nowlin, G.S. Watson, J.W. Tukey, Tom Corvin, William F. Eddy, Robert Epp, William Rogers, Dennis Ash, Wayne Larsen, Louis Jaeckel, Maud? Bin.....?, Theo Gasser, Rick Landau. Later there were also other participants, including notably Colin Mallows (Bell Labs).

After the first half of the semester there were gaps due to the midterm break and also to a period when both Tukey and I had to stay in bed. This gave us the unique occasion for hour–long telephone conversations, during which I tried to convince Tukey that the Normal, as opposed to, for example, the Cauchy, plays a distinguished role in robustness theory. But the result was that Tukey became convinced of the "notoriousness" of the Cauchy, of its "unrealistic peakedness in the middle," which led him to the introduction of Normal/Uniform and related distributions (cf. Andrews et al. 1972, Ch. 6A3 and 6A4).

The main discussions in the seminar, often vivid and even heated, took place during the second half of the fall semester. This was in particular true when Tukey (on Nov. 4) talked about resistance, probability–free approach, fixed sample sensitivity curves ($\equiv$ finite sample influence curves) and other inspiring and memorable topics, and when (on Nov. 9) he discussed "Robust regression in adequately messy situations:" data with many holes (in particular many scattered holes in single coordinates of highly correlated "independent" variables), sour values (also in the "independent" variables), mixed–up factors, selection of independent variables, identification of outliers in single coordinates (where they are often more likely to occur) etc. Here we were inmidst realistic applied regression, and I don't know how much of these topics has been worked out in detail later on.

Regression was also the main subject of other talks. Huber (on Nov. 30) gave conditions for the asymptotic normality of least squares estimates under nonstandard conditions, outlined his approach to robustified least squares and brought some new results about its properties. In a rudimentary form, there was already much of his 1972 and 1973 papers on robust regression. Bickel (on Dec. 7) pursued the approach to regression via analogues of linear functions of the order statistics and gave some examples of such estimators. The author (on Dec. 7 and 14) described the first (asymptotic) influence curve for a skipped estimator, clarified the relation between jackknife and influence curve, and introduced the "influence of position" in factor space (in locally linear designs) with the associated questions of robustifying (also) with regard to position, and of "robustness of design." In later years, this gave rise to bounded–influence, in particular Mallows–type regression estimators, while robustness properties of nonlinear models and of designs are probably still not sufficiently worked out.

The last seminar hour of the first semester (on Dec. 14) was devoted to a vehement debate about the future of the seminar. Some participants, especially Mallows and the author, voted strongly for more work in badly needed robust regression. On the other hand, there was also some partial breakdown of communications among the widely divergent carriers and participants of the seminar (cf. also the Preface to Andrews et al. 1972), and so David Andrews came with his (typed) modest–looking "proposal of order $n > 3$" (see Appendix I). (Only this year Dave and I clarified that the mysterious "$n > 3$" should have been "n^3.") The proposal won easily the majority of the participants and was already augmented during the discussion. Tukey was still undecided. But when the long seminar–free time from Christmas till semester break was over, Tukey was already fully engaged, and the large Princeton Monte Carlo Project (cf. Andrews et al. 1972) started to roll and gain momentum.

In the spring semester, most sessions of the robustness seminar were spent in preparation of the Monte Carlo project, and the new "Monte Carlo committee" included among

Figure 1: List of participants of the Princeton robustness seminar (Sept. 21, 1970)

its six members also David Andrews and Bill Rogers, then an undergraduate student at Princeton who was in highly competent and reliable charge of the large–scale computing. But there were still other topics in the robustness seminar. A highlight was the visit by Cuthbert Daniel (on March 29, 1971) for a "Discussion on applied robustness," in which he talked about an applied regression problem, after commenting on a brief survey of general practical aims and tools of robustness. In a different subfield, Huber (on May 3) generalized the Neyman–Pearson lemma from probabilities to Choquet capacities (cf. Huber & Strassen 1973). C. Mallows (on May 10) spoke about "Hoeffding, Tukey, Hájek and von Mises," uncovering the various relations between U–statistics, the jackknife, Hájek's projection method and von Mises functionals which had become important and popular only with the introduction of M–estimators. Unfortunately, the relations and resulting insights are still unpublished, as far as I know. There is a page on them in the unpublished first version of the "Influence curve" manuscript by the author, see Appendix II. There were also several robustness–related talks in the outside speakers' seminar. In particular, already on Dec. 14, 1970, B.B. Mandelbrot spoke on R/S–analysis, which inspired the author later to his work on robustness against the violation of the independence assumption (cf. Hampel et al. 1986, Ch. 8.1). In his course on multivariate statistics in spring 1971 at Princeton, R. Gnanadesikan (Bell Labs) considered also multivariate outliers and other deviations from multivariate models (cf. Gnanadesikan 1977).

By end–May and June 1971, the members of the Princeton Robustness Seminar started to disperse, and by early August 1971 the manuscript of the book was completed.

3 Some gaps in the "Princeton robustness book."

The book (Andrews et al. 1972) was written in an amazingly short time, thanks to the organizational talent and insistence of David Andrews. Otherwise it probably would have never been finished. On the other hand, the pressure of time and communication problems, especially across the Atlantic, caused a number of gaps and uncontrolled cuts. Some of these gaps will be discussed in this and the following chapters.

The fist gap is tiny, but it throws some light both on Tukey's ingenuity in data analysis and on the existing communication problems. Apart from the mysterious behavior of Cauchy 10A, Tukey found one single outlier among several thousand numbers — truly a remarkable feat —, namely TAK = Takeuchi under Cauchy 20. In his analysis, he commented upon it several times in bewilderment, namely on p. 147, 157, 209, and especially on p. 155 (Ch. 6E1): "... this leaves us with a challenge: What is going on?" The challenge had been answered before the book was finished: it was merely a gross error, most obviously a copying error, but this simple explanation somehow could not find the way into the book.

Another tiny gap (literally), which also others will have noticed and which can be filled in from other tables, occurs in Exhibit 5–4A, the very first table with Monte Carlo results, where estimators 42–44 are missing.

Another gap might have more to do with the time pressure for the analysis. None of the analysts then noticed a highly accurate linear relationship among some of the log variances against ϵ. These relationships, as well as their explanation and reproduction by means of the change–of–variance sensitivity, were only discovered much later and shown in a tiny table (Hampel 1983, p. 214; cf. also Hampel et al. 1986, p. 49ff, for more details on the model).

A more important gap was the omission of the graphs for a highly condensed analysis of the Monte Carlo variances (Ch. 7C6) which described all variances quite well by only

2 (or, more refined, at most 4) "parameters" per estimator (and 1 per situation, in an "additive" model). The verbal description alone is, of course, much less striking to read. Only two summarizing graphs were later published elsewhere (Hampel 1977; see Figure 2); to understand that and how they contain most of the information of about a thousand or more numbers, some diligent study of 7C6 (p. 248ff) will be required. (The summarizing graphs contain also highly concentrated information on the behavior of hard rejection rules studied in "wave H" (see below), cf. Hampel 1977, 1985).

4 An analysis of the asymmetric situations.

Perhaps the most regrettable gap in the book is the lack of any analysis of the two asymmetric situations (Ch. 5I, p. 109–111). It was hard enough for Huber and myself to get at least these two situations included into the simulations, although the whole layout of the computer outputs was designed to include asymmetric cases. But when it came to the analysis, the computer outputs for these cases (many cases were only run after the end of the spring semester and then mailed to the participants) arrived in Zurich after the deadline for the manuscript. Nobody else did an analysis; and this despite the nice–sounding words in the introduction to the asymmetric situations' tables (p. 109): "Except in a few instances there may be no reason to believe the underlying distribution is symmetric .... Much more must be learned but a beginning has been made."

The omission is regrettable for at least three reasons: the analysis might have mitigated some of the criticisms of the study which maintained that symmetry is plainly unrealistic; it fits very nicely also the rest of the analyses; and it broadens their horizon by a new dimension, making the rest look one–dimensional and highly restricted in scope. After all, we all have learned that there is variance **and** bias (and bias dominates in larger samples).

Let us now consider a brief and simple analysis of the two asymmetric situations. They are 18N(0,1)&2N(2,1) and 18N(0,1)&2N(4,1), and together with 20N(0,1) they provide a cross–section through "close" to "moderately distant" contamination of a realistic amount. When we shall talk about bias, it should be clear that the bias in general cannot be uniquely determined from the data, but that there is an unavoidable but hopefully limited range of possible bias, as I have tried to explain in many of my papers. Rather we shall start from the "true," ideal model and shall study how much bias the contamination can cause. This in turn gives us the upper bounds on bias (as a function of ϵ only) needed in practice.

Thus, for each estimator and situation, we obtain two numbers, the bias b (called "Average" in the book) and the variance. Table 1 gives squared bias b^2, variance var and their sum, the mean squared error (MSE). It also indicates when bias and variance, respectively, go down with increasing distance of the contamination; obviously, the former are essentially the redescenders (including rejection rules), while only rarely also the variance does go down.

Figure 3 shows abbreviated stem–and–leaf diagrams for bias and MSE in the two situations. (Similar plots could be done for the variance, but are somewhat less interesting.) The diagrams are hopefully self–explanatory. Since Tukey invented a new kind of tallying in groups of 10 (cf. Tukey 1970/71, 1977), I might mention that I presently prefer and use the Vietnamese way of tallying (4 sides of a square and one diagonal); there is also a Japanese way of tallying with 5 lines, which is closely related to the Kanji character for "stop" (a top line and the character).

Table 1 and Figure 3 show that there are large differences in variance and large differences in bias (both predictable by the influence function!), but much smaller differences in MSE. It may be noticed that for 2N(2,1), there is only one estimator with small (but

Figure 2: Graphical summary of most numbers in Andrews et al. (1972) and Hampel (1985), taken from Hampel (1977). See text.

			$18\,N(0,1)$ &					
			$2\,N(2,1)$			$2\,N(4,1)$		
	bias	var	b^2	var	MSE	b^2	var	MSE
M			.870	1.000	1.870	3.201	1.000	4.201
5%			.669	1.039	1.708	1.713	1.051	2.764
10%			.566	1.092	1.658	.797	1.153	1.950
15%			.508	1.145	1.653	.624	1.192	1.816
25%			.434	1.264	1.698	.468	1.301	1.769
50%			.426	1.605	2.031	.430	1.655	2.085
GAS			.426	1.301	1.727	.456	1.345	1.801
TRI			.491	1.212	1.703	.573	1.258	1.831
JAE			.504	1.175	1.679	.540	1.263	1.803
BIC			.566	1.130	1.696	.872	1.229	2.101
SJA			.524	1.163	1.687	.709	1.255	1.964
JBT			.490	1.177	1.667	.569	1.331	1.900
JLJ	$\searrow$		.466	1.267	1.733	.347	1.544	1.891
H20			.738	1.023	1.761	1.583	1.115	2.698
H17			.659	1.048	1.707	1.124	1.139	2.263
H15			.607	1.071	1.678	.910	1.152	2.062
H12			.530	1.118	1.648	.687	1.180	1.867
H10			.483	1.167	1.650	.575	1.217	1.792
H07			.430	1.266	1.696	.466	1.308	1.774
M15			.591	1.084	1.675	.852	1.192	2.044
D20			.702	1.042	1.744	1.306	1.159	2.465
D15			.591	1.082	1.673	.882	1.172	2.054
D10			.483	1.161	1.644	.584	1.212	1.796
D07			.440	1.233	1.673	.489	1.273	1.762
A20			.692	1.044	1.736	1.254	1.155	2.409
A15			.584	1.092	1.676	.843	1.179	2.022
P15			.582	1.092	1.674	.841	1.181	2.022
HMD	$\searrow$		.346	1.369	1.715	.068	1.453	1.521
25A			.593	1.103	1.696	.616	1.346	1.962
22A	$\searrow$		.476	1.209	1.685	.194	1.455	1.649
21A	$\searrow$		.517	1.155	1.672	.372	1.380	1.752
17A	$\searrow$		.453	1.212	1.665	.268	1.374	1.642
12A	$\searrow$		.403	1.291	1.694	.204	1.417	1.621
ADA	$\searrow$		.548	1.177	1.725	.454	1.446	1.900
AMT	$\searrow$		.534	1.148	1.682	.319	1.455	1.774
OLS	$\searrow$		.394	1.435	1.829	.294	1.469	1.763
TOL	$\searrow$		.329	1.533	1.862	.055	1.632	1.687
MEL	$\searrow$	$\searrow$	.255	1.597	1.852	.048	1.349	1.397
CML	$\searrow$	$\searrow$	.293	1.833	2.126	.100	1.763	1.863
LJS			.694	1.050	1.744	1.080	1.209	2.289

Table 1: Squared bias, variance, and mean squared error in the two asymmetric situations. The falling arrows after the code of the estimator indicate that bias or variance, resp., goes down with the more distant contamination.

	bias	var	18 $N(0,1)$ & 2 $N(2,1)$ b^2	var	MSE	2 $N(4,1)$ b^2	var	MSE
SST	↘		.264	1.725	1.989	.007	1.753	1.760
CST	↘		.442	1.285	1.727	.162	1.536	1.698
33T	↘		.764	1.540	2.304	.331	2.137	2.468
3T0	↘		.540	1.227	1.767	.171	1.615	1.786
3T1	↘		.476	1.225	1.701	.142	1.525	1.667
5T1	↘		.472	1.239	1.711	.134	1.571	1.705
5T4	↘		.469	1.237	1.706	.132	1.561	1.693
CTS	↘		.272	1.595	1.867	.011	1.620	1.631
TAK			.701	1.100	1.801	.783	1.337	2.120
DFA	↘		.464	1.533	1.997	.317	1.714	2.031
JOH	↘		.524	1.211	1.735	.364	1.456	1.820
HGD			.815	1.083	1.898	1.355	1.565	2.920
HGL			.676	1.131	1.807	.719	1.485	2.204
THP	↘		.570	1.271	1.841	.205	1.656	1.861
THL	↘		.542	1.256	1.798	.166	1.649	1.815
JWT	↘		.464	1.284	1.748	.184	1.717	1.901
H/L			.579	1.106	1.685	.848	1.187	2.035
BH			.654	1.078	1.732	1.067	1.188	2.255
2RM			.769	1.029	1.798	1.633	1.152	2.785
3RM			.821	1.010	1.831	3.066	1.032	4.098
4RM			.845	1.002	1.847	2.945	1.004	3.949
3R1			.612	1.067	1.679	.984	1.148	2.132
3R2			.529	1.130	1.659	.677	1.188	1.865
CPL	↘		.347	1.438	1.785	.163	1.477	1.640
SHO	↘	↘	.100	4.836	4.936	.004	4.320	4.324

Table 1: Continued.

b	$18\,N(0,1)\&2\,N(2,1)$	
1.0		
.95		
.9	(4)	
.85	(3)	
.8	(8)	
.75	(11)	
.7	(13)	
.65	(16)	
.6	28,35	OLS, 12A
.55	74,88,89	TOL,HMD,CPL
.5	05,14,22,41	MEL,SST,CTS,CML
.45		
.4		
.35		
.3	(1)	SHO

b	$18\,N(0,1)\&2\,N(4,1)$
1.7	M, 3RM, 4RM
1.6	
1.5	
1.4	
1.3	5%
1.2	H20, 2RM
1.1	D20, A20, HGD
1.0	H17, LJS, BH
.9	(8)
.8	(6)
.7	(7)
.6	(8)
.5	(6)
.4	(8)
.3	CML, 5T4, 5TI, 3TI
.2	MEL, TOL, HMD
.1	CTS
0.0	SHO, SST

MSE	$18\,N(0,1)\&2\,N(2,1)$
4.9	SHO
⋮	
2.3	33T
2.2	
2.1	CML
2.0	50%
1.9	SST, DFA
1.85	(5)
1.8	(6)
1.75	(5)
1.7	(17)
1.65	(24)
1.6	(2)

(detail:)	
1.67	(9)
1.66	17A, JBT
1.65	H10, 15%, 10%, 3R2
1.64	D10, H12

MSE	$18\,N(0,1)\&2\,N(4,1)$
SHO 4.3, M 4.2, 3RM 4.0, 4RM 3.9	
HGD 2.9, 5%, 2RM 2.7, H20 2.6,	
D20, A20, 33T 2.4	
2.2	(4)
2.1	(3)
2.0	(8)
1.9	(6)
1.8	(11)
1.7	(11)
1.6	(9) 12A 1.621; CTS, CPL, 17A; 22A 1.649; 3T1 1.667, TOL 1.687, 5T4, CST
1.5	(1) HMD 1.521
1.4	
1.3	(1) MEL 1.397

Figure 3: Abbreviated stem–and–leaf diagrams for bias and MSE in the two asymmetric situations.

Figure 4: Variance against squared bias for 18 $N(0,1)$ & 2 $N(2,1)$

Figure 5: Variance against squared bias for $18\,N(0,1)$ & $2\,N(4,1)$

nonnull) bias (namely the shorth), corresponding to the unidentifiability of nearby contamination, while for 2N(4,1), there is a whole group with small biases ranging all the way down to zero.

The MSE, as the most important one–number summary (— of course we should use two numbers if we can —), shows a different pattern. For 2N(2,1), most estimators have about the same MSE (34 out of 65 estimators have MSE $\leq$ min (MSE) $\times$ 1.05 !). On the other hand, for 2N(4,1), MEL, and also HMD, are completely isolated as the best estimators. Only then begins a thin grouping, starting with 12A, of other strongly rejecting and yet decently efficient estimators. This shows that there is a whole region of estimators around MEL and HMD (which should include some tanh–estimators) waiting to be explored! Many of them will be rather inefficient close to the normal, but as HMD shows, there may be still unknown estimators which are very useful for rather "dirty" data and still usable near the normal.

If we now look across the 3 situations (including N) simultaneously and ask for the best compromises with at most 10% increase of MSE over the best one in each situation (N(4,1): over the third best), the only solutions are 21A and AMT. If we allow 20% increase for N(4,1) (this time from the best value) and 10% for the two others, then only 22A is almost a solution. Further competitors for optimal compromises are 17A, 3T1, 5T4, 5T1. These 7 estimators are selected in excellent agreement with the results of the other situations in the book (cf. Ch. 7C5); they are also among the overall best ones, with perhaps AMT and 3T1 overall a bit weaker, and 17A overall best.

We obtain a new perspective if for each asymmetric situation we consider variance and squared bias simultaneously in a two–dimensional plot (Fig. 4 and 5) and ask for the optimal estimators for different values of n. Since the bias stays roughly constant, and the variance goes roughly down like $1/n$ (assuming independence! Reality outside Monte Carlo studies is different, cf. Hampel et al. 1986, Ch. 8.1), we can put in first, crude, but usable approximation $\mathrm{MSE}(n) = b^2 + \mathrm{var}(1)/n$ (cf. also the analogous theoretical computations in Hampel et al. 1986, p. 49–52). We can replace var(1) by 20var(20) and hence put tangents from the lower left in different angles at the plot and thus get the optimal (w.r. to MSE) estimators for every n.

For 2N(2,1), we find D10 optimal from about $n = 20$ to about $n = 30$ (where 17A and 12A are about equally good), HMD for about $30 \leq n \leq 50$, MEL for about $50 \leq n \leq 400$, and SHO for about $n \geq 400$.

For 2N(4,1), we find MEL optimal for about $5 \leq n \leq 150$, CTS for about $150 \leq n \leq 650$, SST for about $650 \leq n \leq 15'000$, and SHO for about $n \geq 15'000$. Of course, these numbers, especially the latter ones, are bold extrapolations; they should not be taken literally, but they show the kind of results one can tentatively derive even from a single n, and one can expect at least semi–quantitatively to be obtained by a more comprehensive study with different values of n (and asymmetric contamination).

For very small n, in both situations some Hubers or trimmed means will do quite well. With increasing n, the numbers show clearly the increasing importance of bias and hence a shift from trimming/Huberizing towards stronger and stronger (smooth and hard) rejection, until finally estimators like shorth and shordth (Hampel 1975) are optimal (out of a finite selection; or "asymptotically" optimal). The numbers also give a crude indication how large n has to be for this to happen (in dependence on the competitors, of course).

A more refined analysis would be possible after filling the region around MEL and HMD with new estimators, and by asking for the range of n in which each estimator is no more than, say, 5% worse than the optimal one. Further results can be obtained along the lines above, with simultaneous compromises over many different situations for each n.

The qualitative part of these results was long known to at least some people. It is important for statistical practice, but not generally stressed or even mentioned, and it is the more regrettable that such an analysis is missing in the book.

5 The introduction of redescending M-estimators.

Smoothly redescending M-estimators, such as HMD or CML, also have a prehistory. Thus, Jeffreys (1939) gave a numerical algorithm for the maximum likelihood estimator for a particular t–distribution (for desk calculators!). But a real understanding came only with robustness theory, when concepts such as M–estimators, von Mises functionals, influence curves, breakdown points, and rejection points were investigated. Yet the early robustness theory, with its stress on convex ρ–functions (Huber 1964), even provided a certain obstacle, although Huber in passing also considered nonmonotone ψ–functions in theory (but not in practice).

In February 1971, I proposed two estimators for the study: a Huber with the median (absolute) deviation scale (first discussed in Hampel 1968), which was called A15, and a new estimator that was to do "what every reasonable and good data analyst would want it to do" according to my conviction: behave like the arithmetic mean in the middle, reject clear outliers completely, and allow for a "zone of doubt" in between. To keep things simple, I chose a linear influence curve also for the zone of doubt. This estimator was called HMD (Hampel M–descending).

Nobody believed it would work. There was a feeling in the air that something catastrophic would happen. The student (not Bill Rogers) who programmed it even changed the second constant from 6 (a multiple of 2) to 5.5 "to give it a slightly better chance at least to survive," as he said.

The first computer outputs were puzzling. Contrary to everybody else's expectation, HMD did work, but not as I had expected. It was never very bad, but never very good either. In fact, it behaved roughly like the median.

I decided to make another, more cautious attempt, added a constant piece for safety and tried out a (for me) reasonable looking range of parameter values. These were called the three–part redescending M-estimators 25A to 12A (since Hxy was used for Huber, and Axy, the next letter in Hampel, was used in A15, the A was put at the end).

I also tried a mildly adaptive ad hoc version ADA, since adaptive estimators were much discussed these days.

The next batch of distributions run on the computer were all very longtailed distributions, though quite a variety of them. The computer outputs were ordered by the variances, not according to the standard sequence of estimators as in the book. I still remember the atmosphere when I came to Fine Hall that morning (I believe it was early May) and saw the others bent over the output packages. In every situation run, all 6 of my estimators (though designed for quite different situations, with very different parameters) were in the top 6 positions!

This historical accident — there was no competition yet in the study dealing with distant outliers — left a deep impression on a number of participants. The original extreme scepticism switched into extreme enthusiasm, so extreme in fact that it was overlooked that I had built a number of precautions into my estimators, such as: starting with median and median deviation, thus always having at least half of the data on the ascending linear piece in the beginning (by the choice of the first parameter), having a constant buffer, and never descending too quickly — 22A was an extreme experimental piece, as Tukey partly rediscovered — so that practically always the central part of the data would dominate and determine the estimate (in an also numerically stable way). Furthermore, some properties of redescending estimators, namely the local uniqueness of the proper solution, follow from a theorem cited in my thesis (Hampel 1968, p. 54 and, applied to Cauchy location, p. 93), and some were much later worked out by Brenton Clarke (1986). It is also still often overlooked that distant outliers are actually easy to treat; the main problem, with bias and efficiency, lies with almost "hidden" contamination in the "flanks" of the distribution. Cf. also the challenges and experiences with the asymmetric situations above. In all, redescenders are a valuable but rather delicate tool.

By contrast, the follow–up redescenders such as Andrews' sine and Tukeys' biweight dropped already some precautions, notably the slow speed of redescending, though they will probably encounter problems only under rather special circumstances. Other redescenders were even more carelessly defined, and this naive adaption of a tool formally considered unusable and in fact requiring quite a few additional thoughts, together with the exaggerated switch of opinion, then provoked the counterreaction of statisticians like Peter Huber (cf. Huber 1981) who warned against their indiscriminate and thoughtless use. In order to make the point, these warnings then tended to downplay the non–negligible advantages which the redescenders actually do possess (such as typically 15–20% efficiency gain against 1–3% efficiency loss compared with Hubers in various realistic situations).

And what happened to HMD? Around the same time, the first finite–pseudosample influence curves were plotted on the computer, and the first use of these plots was the discovery that HMD contained a programming error: instead of going down from (2, 2) to (5.5, 0), the second line piece went up again from (2, 0) to (5.5, 2)! No wonder it behaved so strangely. In fact, with its two jumps, it was worse than a quick redescender or a hard rejection rule in some ways, and it was the more remarkable that it came out as decently as it did. The error was corrected, and the new HMD behaved as it should, a bit too inefficient at the Normal (also partly because of the parameter change from 6 to 5.5), but in some other situations nearly optimal.

6 After the "Princeton robustness book."

All the computations in the book by Andrews et al. (1972) were only "wave A" of a much bigger Monte Carlo study which Tukey continued with Alan Gross in charge of the computations. I received computer outputs till "wave I," and a short description of all estimators till "wave N" in Gross & Tukey (1973). I am wondering what happened to most of these waves.

Of special interest to me were parts of wave F, and especially wave H, which Tukey generously left to Bickel, Huber and myself to fill it with our proposals. Wave F and H contained, among others, rank–based estimators, including "redescending" R–estimators (as well as that can be done at all), one–step normal scores, and midpoints of confidence intervals derived from rank tests. A highly condensed, but thorough analysis is given in Hampel (1983).

Wave H contained also the mean combined with various rejection rules as robust estimators, whose comparison with other robust estimators was deemed overdue. Analyses are given in Hampel (1976, 1977, 1985), cf. also Hampel et al. (1986, Ch. 1.4). It is interesting to compare the behavior of these "objective" rejection rules with the results of the Monte Carlo study on subjective rejection run by Relles & Rogers (1977).

There is another group of estimators run in wave H, namely the "trapped trimmed means." The basic idea was to use symmetrically trimmed means, but trimmed such that all untrimmed values are within a certain multiple of the median deviation from the median. This idea can also be combined with hard rejection in various ways. It is an open question how good these simple estimators are. I did not yet get around to doing an analysis, and I wonder whether and when an analysis will be done.

References

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

Clarke, B. R. (1986). Asymptotic theory for description of regions in which Newton-Raphson iterations converge to location $M-$estimators, *J. Statistical Planning Inference* **15**: 71–85.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, N. Y.

Gross, A. M. and Tukey, J. W. (1973). The estimators of the Princeton robustness study, *Tech. Rep. 38, Ser. 2*, Dept. of Statistics, Princeton University, Princeton, N.J.

Hampel, F. (1968). *Contributions to the Theory of Robust Estimation*, PhD thesis, University of California, Berkeley.

Hampel, F. (1975). Beyond location parameters: Robust concepts and methods, *Bull. Internat. Statist. Inst.,* **46**: 375–391, (with discussion).

Hampel, F. R. (1976). On the breakdown points of some rejection rules with mean, *Res. Rep. 11*, Fachgruppe für Statistik, ETH, Zurich.

Hampel, F. R. (1977). Rejection rules and robust estimates of location: An analysis of some Monte Carlo results, *in* D. Reidel (ed.), *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, 1974 European Meeting of Statisticians*, Vol. A, Dordrecht, Hingham, MA, pp. 187–194.

Hampel, F. R. (1983). The robustness of some nonparametric procedures, *in* P. J. Bickel, K. Doksum and J. L. Hodges Jr. (eds), *A Festschrift for Erich L. Lehmann.*, Wadsworth, Belmont, pp. 209–238.

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

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

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

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

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

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

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

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

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

Tukey, J. W. (1970/71). *Exploratory Data Analysis. Preliminary edition*, Addison-Wesley, Reading, Mass.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.

Appendix I

Proposal by D.F. Andrews of Dec. 14, 1970, that started the Princeton Monte Carlo project.

A Proposal of order $n > 3$ for the Robustness Seminar

A group of n interested persons shall gather and after consultation with the others each shall propose

(i) at most 1 estimate of location in computable form.
(ii) at most 1 measure of desirability of an estimate in computable form
(iii) at most 1 set of assumptions for which these estimates should be assessed.

There will be at most n^3 evaluations of the desirability of estimates, some will be capable of calculation. It is assumed that at least n will be readily available.

Work will proceed in groups of at most 2 on the further completion of this table. A chart will be provided recording who is working on what and what has been completed.

In addition a small number of 1 hour seminars will be given by group members presenting concisely some related results in the literature. These will lead to a bibliography of comparative robustness to be included in the report.

Reporting results. At a second meeting numbers will be presented and the estimates ranked for each set of assumptions and measure of desirability.

Note — there will be many missing observations.

This table may be further analyzed and reported.

On May 1, a short paper will be written and on June 1 this will be reproduced and the proposal will be considered completed.

Appendix II

Notes on the talk by C.L. Mallows of May 10, 1971, on the relations between various important robustness-related concepts, taken from Hampel (1972): The Influence Curve and Its Role in Robust Estimation, first (unpublished) version.

The relation between von Mises derivatives and U–statistics has been investigated by C. Mallows who had been a frequent visitor of the Princeton Robustness Seminar. A few of his results are briefly sketched here with his kind permission. The two concepts are indeed very closely related, but for some peculiar reasons, many mathematical statisticians tend to prefer one of them and all but ignore the other. Mallows has shown that a U–statistic can be expanded into a finite series whose terms correspond to the terms of a von Mises expansion. However, for finite sample size, a U–statistic in general is not just a finite von Mises series. It differs essentially by lacking the "diagonal" terms. Asymptotically, this makes no difference; in finite samples, with U–statistics we gain uncorrelated terms, unbiasedness etc., but lose independence from the sample size and "Fisher-consistency" (see below) and may end up with plainly inadmissible estimators. (Mallows also showed that, starting with a finite von Mises expansion, a U–statistic can simply be obtained by applying a higher–order jackknife to it.) In viewing these results, it seems to this author that the more elegant U–statistics are more closely connected with the concepts and tools of probability theory, while the more general von Mises expansions lean more toward the theories of (differentiable) functions and functionals, without any additional probabilistic features. Another disguised reason for the more elegant theorems and the greater popularity of U–statistics seems to be that the functional which is the asymptotic limit of a U–statistic has an infinite von Mises expansion with only finitely many terms differing from zero. Thus it seems that among the class of all statistics, U–statistics in a way play the role which polynomials play among real functions, being simple, useful, elegant and suitable for approximations, but still rather special.