Wozu brauchen wir Robuste Statistik?

Frank Hampel

Research Report No. 75
September 1995

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich, Schweiz

Wozu brauchen wir Robuste Statistik?

Frank Hampel
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich

September 1995

Manuskript des Vortrags vom 26.9. am Seminar 1995 der
Internationalen Biometrischen Gesellschaft Region Oesterreich Schweiz (ROeS)
Rapperswil, 25. -29. September 1995

1 Einleitung

Robuste Statistik (vgl. Hampel et al. 1986, Ch. 1) lässt sich auffassen als die Stabilitätstheorie statistischer Verfahren. Sie untersucht die Effekte kleinerer (Huber 1981) und grösserer (vgl. auch Tukey l977) Abweichungen von exakten Modellannahmen und füllt damit eine Lücke in der logischen Rechtfertigung statistischer Methoden (vgl. Hampel et al. 1986, p. 7, p. 55). Inwieweit diese Lücke bedeutsam ist, werden wir noch kurz diskutieren.

Bekanntlich gibt es Abweichungen vom Modell der Normalverteilung, die daher für jede parametrische statistische Theorie, auch die der Bayesianer, potentiell relevant sind. Die Robustheitstheorie untersucht jedoch allgemein Abweichungen von jedem beliebigen („vernünftigen") parametrischen Modell, etwa auch dem der Rechteck- oder Gleichverteilung (Hampel et al. 1986, p. 22, p. 54), wobei freilich die Einzelheiten für die Praxis noch längst nicht immer gut ausgearbeitet sind. Überdies untersucht sie auch die Verletzung der Unabhängigkeitsannahme (Hampel et al. 1986, Ch. 8.1), die auch in der nichtparametrischen Statistik oft gemacht wird.

2 Was für Abweichungen von exakten parametrischen Modellen kommen in der Praxis vor?

Wir können generell vier Typen von Abweichungen unterscheiden.

a) Das Auftreten grober Fehler (sowie auch Abweichungen vom Modell in nur einem Teil der Daten). – Wie viele Beispiele zeigen (vgl. z.B. Hampel et al. 1986, Ch. 1.2c), ist ihr Anteil in wissenschaftlichen Routinedaten typischerweise zwischen 1% und 10%; in medizinischen Daten scheint er eher zwischen 10% und 25% zu liegen (gemäss den sarkastischen Bemerkungen einiger Spezialisten). Daten hoher Qualität, mit keinen oder praktisch keinen Ausreissern auch in grossen Datensätzen, existieren, aber sind relativ selten. In strukturierten Daten werden Ausreisser öfters auch von namhaften Statistikern übersehen (vgl. die Varianzanalyse-Beispiele in Daniel 1976).

b) Die Tatsache, dass alle Messdaten nur eine endliche Genauigkeit haben und meist diskretisiert werden. – Dieser Effekt scheint harmlos und ist es oft auch, kann aber nicht nur in der Theorie (etwa für die Rechteckverteilung, vgl. oben), sondern auch für den Computer (wie ein Teil der numerischen Stabilitätstheorie zeigt) zu dramatischen Ergebnissen führen.

c) Oft ist das Modell von vornherein nur eine Approximation, etwa auf Grund des Zentralen Grenzwertsatzes (oder wegen seiner mathematischen Einfachheit).

d) Die Erfahrung zeigt, dass selbst Musterbeispiele für Reihen „unabhängiger" Messungen aus den exakten Naturwissenschaften nicht nur korreliert, sondern sogar langfristig korreliert sind (vgl. Hampel et al. 1986, Ch. 8.1a), ein Phänomen, das auch unter dem Namen „semisystematische Fehler" beschrieben wird. Es war erstklassigen Anwendern wohl schon immer wenigstens qualitativ bekannt, wurde jedoch bis vor kurzem praktisch nie modelliert und nur sehr selten explizit in der Literatur erwähnt.

3 Wie effizient sind Kleinste Quadrate wirklich?

Viele Statistiker glauben, dass der Effizienzverlust der Methode der Kleinsten Quadrate unter den üblichen Abweichungen von der Normalverteilung nur einige Prozent beträgt. Sie berufen sich dabei zum Teil, direkt oder indirekt, auf Cox & Hinkley (1968), die, ohne es explizit zu behaupten, einen maximalen Effizienzverlust von 10% suggerieren, und zwar durch eine formal korrekte, aber lückenhafte und entstellende Interpretation von Ergebnissen in Jeffreys (l939, Ch. 5.7). Vereinfacht gesagt, folgt aus Jeffreys jedoch nicht ein Verlust von höchstens 10%, sondern von mindestens 10% Effizienz (und meist weit mehr).

Eine andere irreführende Arbeit ist die von Stigler (1977). Er vergleicht Schätzwerte, die auf alten Messungen naturwissenschaftlicher Konstanten beruhen, linear (mit Absolutbetrag) mit deren modernen Werten und kommt zum Schluss, dass die Güte des arithmetischen Mittels im Schnitt nur wenige Prozent unterhalb der der besten robusten Schätzer liegt. Er übersieht dabei die systematischen und semisystematischen Fehler, die zwischen den alten und den modernen Messungen liegen, so dass die modernen Werte meist weit ausserhalb der durch die alten Daten suggerierten Wertebereiche liegen. Tatsächlich sind typischerweise die meisten robusten Schätzwerte dicht beisammen, während das arithmetische Mittel relativ weit draussen mal auf der „richtigen", mal auf der „falschen" Seite liegt, im (linearen) Schnitt also etwa „gleich gut" wie die robusten Schätzer ist. Dieser Punkt wurde von Pratt in der Diskussion der Arbeit (loc. cit.) bemerkt. Stigler versuchte Pratt in seinem „rejoinder" mit einem Theorem zu widerlegen, das zwar an sich richtig war, das er aber falsch interpretierte, indem er „(strikt) steigend" und „nichtfallend" verwechselte. Damit wurde (vermutlich von vielen Lesern unbemerkt) sein Rechtfertigungsversuch hinfällig. Gute Anwender wie Newcomb und Eisenhart zogen übrigens einen „robusten" Wert dem Mittel vor (loc. cit.).

Manche Statistiker glauben auch, sie könnten wegen des Gauss-Markoff-Theorems („Kleinste Quadrate liefern, auch ohne Normalverteilung, die beste lineare erwartungstreue Schätzung") ganz auf die Normalverteilung verzichten. Sie übersehen dabei, dass ausserhalb einer sehr „kleinen" Umgebung der Normalverteilung (etwa im Raum der Pearson-Kurven) alle linearen Schätzungen schlecht (ineffizient) sind, wie bereits Fisher (1922), wenn auch mit anderer Stossrichtung, demonstrierte.

Berühmt wurde der Streit zwischen Fisher (1920) und dem Astronomen Eddington um die beste Skalenschätzung. Eddington behauptete, dass die „mean deviation" (durchschnittliche Abweichung) in der Praxis besser sei als die Standardabweichung, während Fisher bewies, dass für strikt normalverteilte Daten (nur) die Standardabweichung optimal ist. Beide hatten recht. Erst Tukey (l960) und Huber (1981, p. 3) zeigten, für welche winzigen Abweichungen von der Normalverteilung (0.18% in Tukey's Modell) die durchschnittliche Abweichung bereits besser ist. Dabei ist diese nur „weniger nichtrobust“ als die Standardabweichung und daher auch nicht allgemein zu empfehlen.

Wenn man versucht, die in der Praxis vorkommenden „Normalverteilungen“ (ev. nach Transformation) genauer durch t-Verteilungen zu modellieren, so passen meist 3 bis höchstens 9 Freiheitsgrade am besten, und 3 (bis 5) Freiheitsgrade sind recht typisch (Jeffreys 1939, Huber 1981). Auf die Varianzkomponenten eines berühmten Datensatzes (Bennett 1954) passt sogar auf einer Stufe am besten die Cauchy-Verteilung, also t mit einem Freiheitsgrad (Hampel, unpubliziert). Unter t mit 9 bzw. 5 bzw. 3 Freiheitsgraden ist nun die asymptotische Effizienz des arithmetischen Mittels 93%, 80%, 50%, die der Standardabweichung 83%, 40% und 0%. In diesem Bereich bewegt sich also typischerweise die Effizienz der Methode der Kleinsten Quadrate.

Das heisst nun nicht, dass man Kleinste Quadrate nicht mehr verwenden darf. Die Einfachheit und Eleganz der Methode ist ein grosser Vorzug (teilweise auch ihr Bekanntheitsgrad), und „Effizienz“verluste durch schlechte Modellierung oder unpassende Daten können noch viel grösser sein. Aber es heisst, dass man realistischerweise Effizienzverlusten von Dutzenden von Prozent und nicht nur von einzelnen Prozenten ins Auge sehen muss, wenn man keine nahezu optimalen robusten Methoden verwenden will.

Während Tukey (1960) unter anderem auch zeigte, dass der Median in der Praxis oft besser als das Mittel ist, wies E.S. Pearson (1929, 1931) schon viel früher nach, dass beim Chiquadrat- und F-Test für Varianzen üblicherweise nicht einmal das Niveau auch nur näherungsweise stimmt; demgegenüber ist das Niveau beim F-Test in der Varianzanalyse mit festen Effekten (Modell I) auch unter mittelgrossen Abweichungen von der Normalverteilung bemerkenswert stabil. Was dagegen oft übersehen wird (und mit der Effizienz des Mittels zusammenhängt), ist die Tatsache, dass t- und F-Test (in derselben Situation) rasch an Macht verlieren. Wiederum mag man freiwillig (statt versehentlich und unbemerkt) auf Macht verzichten, solange das Niveau wenigstens einigermassen stimmt. (Übrigens kann die Robustheitstheorie auch die Robustheitseigenschaften von Rang- und anderen nichtparametrischen Methoden in verschiedenen Situationen prägnant beschreiben.)

4 Inwieweit genügen Methoden zur Verwerfung von Ausreissern?

Ausreisserregeln und auch subjektive Verwerfung von Ausreissern können als (z.T. informelle) robuste Methoden aufgefasst werden, die hauptsächlich einfach und durchsichtig sein und das Schlimmste verhüten sollen. Diese Zwecke erfüllen sie auch oft (um den Preis von Effizienzverlusten), aber es gibt auch Unterschiede zwischen den Methoden, bis hin zu Versagern.

Die gründliche Analyse einer Monte-Carlo-Studie (Hampel 1985, vgl. auch Hampel 1980 und Hampel et al. 1986, Ch. 1.4, zur Ausreisserbehandlung) ergibt, dass sich das Verhalten vieler Verwerfungsregeln sehr präzise durch den Bruchpunkt („breakdown point“) der kombinierten Verwerfungs- und Schätzmethode beschreiben lässt. Dabei zeigt sich, dass eine bestimmte „Verwerfungsregel“ nicht einmal einen einzigen weit entfernten Ausreisser unter 20 Werten entdecken kann; und die wohl verbreitetste Regel, das grösste studentisierte Residuum, versagt bereits bei (je nach Testniveau) rund 10% Ausreissern, die, wie wir oben gesehen haben, ohne weiteres vorkommen können. Andererseits gibt es neuere Verwerfungsregeln mit Bruchpunkt 50%, bei denen Mittel und Standardabweichung (da nicht robust) durch Median und Median-Abweichung („median (absolute) deviation, MAD") ersetzt werden: „Man verwerfe alle Daten, die vom Median mehr als k Median-Abweichungen entfernt sind, und nehme das Mittel der übrigen", wobei ungefähr k=5 für viele Zwecke, auch für verschiedene Stichprobengrössen, ein brauchbarer Kompromiss zu sein scheint (Schweingruber 1980, Hampel 1985 und unpubliziert).

Alle, auch die besten (zuverlässigsten und effizientesten) Verwerfungsregeln haben jedoch gegenüber besseren robusten Methoden in häufigen Situationen (etwa t mit 3 Freiheitsgraden als Faksimile für Daten hoher Qualität) einen Effizienzverlust von mindestens etwa 15-20%. Dieser hängt mit der künstlichen und sprunghaften Einteilung der Daten in „gute" und „schlechte" Daten zusammen. (Unter ungünstigen, wenn auch unwahrscheinlichen Situationen kann der Effizienzverlust beliebig hoch sein.)

Es ist bemerkenswert, dass ähnliche Effizienzverluste von etwa 15-20% (allerdings offenbar unter mehr Verteilungstypen) auch bei subjektiver Verwerfung von Ausreissern auftreten. Dies wurde von Relles & Rogers (1977) in einer ungewöhnlichen Monte-Carlo-Studie mit 5 subjektiv verwerfenden Statistikern gezeigt. Man mag die Verluste wieder als gross oder klein ansehen, je nach Geschmack und Massstab. Ein Vorteil der subjektiven Begutachtung von Daten ist der engere Zusammenhang mit der Interpretation der Werte, die grundsätzlich nie vernachlässigt werden sollte, auch nicht nach „objektiver" Verwerfung oder anderer robuster Behandlung. Nachteile sind das Fehlen objektiver Reproduzierbarkeit und Probleme bei komplizierten und umfangreichen Datensätzen.

Wie sinnvoll und wichtig in der Praxis die Interpretation (und nicht nur die blosse Identifizierung) von Ausreissern ist, wird sehr schön an den Beispielen in Daniel (1976) veranschaulicht.

Leider hat nicht nur subjektive Verwerfung Mühe mit komplexen, insbesondere hochdimensionalen Datenstrukturen. Der „Fluch der Dimension" bewirkt im allgemeinen grosse Identifikations-, Effizienz- und Berechnungsprobleme.

Insgesamt ermöglicht die Robustheitstheorie ein besseres Verständnis der Ausreisserregeln, insbesondere ihrer unvermeidlichen Schwächen, die Entwicklung neuer, besserer Ausreisserregeln und, wenn gewünscht, den Ersatz „harter" Verwerfungsmethoden durch Verwerfung mit einer Übergangszone stetig zunehmenden Zweifels, die praktisch keine vermeidbaren Effizienzverluste mehr hat.

5 Was liefert Robuste Statistik für strukturierte Datensätze?

Strukturierte Datensätze – als Prototyp sei die Multiple Regression genannt – lassen sich prinzipiell ebenfalls mit robusten Methoden behandeln, auch wenn diese noch längst nicht für alle Modelle und Problemstellungen voll ausgearbeitet sind. Die nachfolgenden Vorträge an der gegenwärtigen Tagung werden einzelne Punkte näher diskutieren.

Grundsätzlich liefern robuste Methoden eine zuverlässigere, direktere und effizientere Modellanpassung an die Mehrheit der Daten, im Gegensatz zu einem oft verschwommenen Kompromiss für alle Daten. Welches Modell für welche Mehrheit am besten passt, ist freilich nicht immer eindeutig durch die Daten festgelegt und bedarf daher der nachfolgenden Interpretation der Ergebnisse. Die Hoffnung auf eine „blinde", gedankenlose robuste Routinelösung lässt sich im allgemeinen nur selten erfüllen, und der Glaube, dass man bei Anwendung robuster Methoden auf eine Dateninterpretation verzichten könne, ist ein Missverständnis. Gefördert wurde dieses Missverständnis vermutlich durch die Prominenz rein technischer Probleme in der Literatur, wie dem Konflikt zwischen niedrigem Bruchpunkt und niedriger Effizienz oder dem Problem des Rechenaufwandes, für den selbst heutige Computer oft nicht hinreichen (vgl. z.B. Rousseeuw & Leroy 1987 oder die beiden Sammelbände von Stahel & Weisberg 1991, in denen auch der Standpunkt von „regression diagnostics“ vertreten ist). Die technische Forschung ist notwendig und schwierig und macht weitere Fortschritte, ebenso wie die Computerimplementierung robuster Methoden (vgl. Marazzi 1993). Aber statt einen Gegensatz zwischen „regression diagnostics“ und Robuster Statistik zu konstruieren, sollte man beide Ansätze vielmehr integrieren.

Ein sehr schönes Beispiel für den Erfolg einfacher und flexibel gehandhabter robuster Methoden in einem grossen Datensatz liefert die Analyse des Spektrums des Tritium-Moleküls (Ruckstuhl 1995, Stahel et al. 1994). Naive einfachste robuste Regression (mit Absolutbeträgen statt Quadraten) reduziert den Standardfehler der Residuen von naiven, routinemässigen Kleinsten Quadraten um den Faktor 10 000 (!); und selbst „intelligent benutzte“ Kleinste Quadrate, mit sorgfältigen Datendiskussionen unter Einbezug aller Erfahrung der Fachwissenschaftler, liefern noch ein deutlich ungenaueres Ergebnis als intelligent benutzte robuste Methoden. Solche Beispiele mögen bisher selten sein, aber sie zeigen doch das zusätzliche Potential (zusätzlich zu sorgfältiger Datenanalyse), das in robusten Methoden steckt.

6 Wie steht es mit der Verletzung der Unabhängigkeitsannahme?

Noch grundlegender und verbreiteter als die Annahme der Normalverteilung (oder irgendeiner anderen Marginalverteilung der Daten), aber viel seltener diskutiert, ist die Annahme der Unabhängigkeit. Inwieweit ist diese Annahme zu rechtfertigen?

Die Antwort ist komplex (vgl. Hampel et al. 1986, Ch. 8.1, Hampel 1987, Künsch et al. 1993). Zunächst zeigt die Erfahrung (qualitativ bei vielen erstklassigen angewandten Statistikern, und quantitativ bei langen Messreihen höchster Qualität aus Physik, Chemie und Astronomie), dass es Folgen von „unabhängigen, identisch verteilten Zufallsvariablen“ (ein Hauptparadigma der Stochastik) praktisch nirgends gibt. Überdies klingen die beobachteten seriellen Korrelationen so langsam ab, dass die herkömmlichen Zeitreihenmodelle, wie ARMA-Prozesse, und auch die Theorie stark mischender stochastischer Prozesse ebenfalls nicht mehr passen. Die einfachsten geeigneten Modelle, die erst in neuerer Zeit näher erforscht wurden, sind selbstähnliche Prozesse und fraktionelle ARIMA-Prozesse. Als (empirisch überprüfbare) Konsequenz sinkt die Varianz des arithmetischen Mittels von n Beobachtungen nicht wie $1/n$, sondern (beliebig langsam, je nach Datensatz) wie eine andere Potenz von n zwischen 0 und -1. Wegen dieser „semisystematischen Fehler“ (und natürlich auch wegen der bekannteren systematischen Fehler) ist dem Standardfehler des arithmetischen Mittels besonders aus grossen Messreihen nicht zu trauen, in Einklang mit dem gesunden (wenn auch vielleicht übertriebenen) Misstrauen mancher Physiker. (Ähnliche Phänomene liegen offenbar der Diskussion von Kontingenztafeln bei Berkson 1938 bis hin zu den modernen Korrekturfaktoren für die Variabilität von Zähldaten zugrunde.)

Bei kleinen Stichproben sind diese Effekte freilich noch nicht sehr ausgeprägt, und in Ermangelung der Möglichkeit, für weniger als 50-100 Beobachtungen die Intensität der langfristigen Korrelationen überhaupt informativ zu schätzen, bleibt uns in diesem Fall auch kaum etwas anderes übrig, als so zu tun, als ob unsere Daten unabhängig wären (ausser natürlich bei zusätzlichen kurzfristigen Korrelationen); allerdings sollten wir die Existenz des unbekannten, aber mit n rasch wachsenden Korrekturfaktors für den Standardfehler des Mittels nicht vergessen.

Zum Glück ist dies noch nicht das vollständige Bild. Die laut angewandter Folklore erfolgreichsten Anwendungsmethoden der Statistik betreffen vor allem Effekte in der Varianzanalyse und Steigungen in der Regression, beides Beispiele für Kontraste (im Gegensatz zu absoluten Konstanten). Hier gilt nun (Künsch et al. 1993): Wenn die Versuchsbedingungen gut durchmischt sind (also z.B. in zufälliger Reihenfolge gewählt wurden), ist der unter der Unabhängigkeitsannahme hergeleitete Standardfehler der Kontraste auch unter langfristigen Korrelationen asymptotisch korrekt, ebenso wie das Niveau der t- und F-Tests im linearen Modell, soweit sie nur Kontraste betreffen. Dieser Satz gibt eine späte tiefere Begründung für die grossen Erfolge der Statistik in vielen Bereichen, trotz mangelnder Übereinstimmung zwischen den hergebrachten mathematischen Modellen und der Wirklichkeit.

Aber auch hier müssen einige Feinheiten berücksichtigt und einige Einschränkungen gemacht werden. Obiger Satz gilt nicht für das Gesamtmittel in der Varianzanalyse und den Achsenabschnitt in der Multiplen Regression, und es ist eine bemerkenswert weise Sitte, diese aus der Varianzanalysetabelle üblicherweise wegzulassen. Natürlich gilt der Satz auch nicht für den Einstichproben-t-Test, der sich somit in der Anwendung (ausser bei sehr kleinen Stichproben, für die er ursprünglich konzipiert war) als viel problematischer erweist als der Zweistichproben-t-Test. Der Satz gilt auch nicht bei (zeitlich, örtlich oder anderweitig) schlecht durchmischten Versuchsbedingungen. Schliesslich muss leider auch festgestellt werden, dass die „naiven“ Methoden unter den genannten Bedingungen noch einen (oft recht grossen) Effizienz- und Machtverlust aufweisen.

Dieser Effizienzverlust kann hinwiederum oft stark reduziert werden durch die Bildung möglichst kleiner (zeitlich etc. homogener) Blöcke, ohne dass die Effekte von langfristigen Korrelationen explizit modelliert und geschätzt werden müssten; sie sind dann weitgehend mit den Blockeffekten verquickt oder vermengt („confounded“) und fallen aus der Analyse heraus. (Hier zeigt sich wieder, wie wertvoll die Prinzipien der Randomisierung und Blockbildung sein können.)

7 Robuste Statistik als Brücke zwischen mathematischer Modellbildung und statistischer Praxis

Die obige Diskussion zeigt deutlich, wie unrealistisch und teilweise sogar irreführend die Modellierung von Daten durch exakte mathematische Modellannahmen (wie Normalverteilung und Unabhängigkeit) sein kann, so nützlich diese Modellannahmen als Anregung und erste Schritte für die Praxis auch sein mögen. Daneben hat sich ein Erfahrungsschatz guter angewandter Statistiker herausgebildet, der teilweise im Widerspruch zu den Behauptungen der mathematischen Theorie stand (sofern diese, wie üblich, auf wirkliche Daten übertragen wurden). Die Robuste Statistik, die explizit und mathematisch präzise ganze Umgebungen von „punktförmigen“ Modellannahmen (also z.B. „ungefähre Normalverteilung“) betrachtet, kann viele dieser Widersprüche auf einleuchtende Art auflösen. Sie liefert damit nicht nur neue Begriffe zur Beschreibung wichtiger statistischer Sachverhalte, eine umfassendere Beschreibung für und tiefere Einsicht in die Eigenschaften herkömmlicher statistischer Methoden und eine Fülle neuer robuster Methoden, die als nützliches Handwerkszeug gebraucht werden können, sondern sie ermöglicht auch für den Bereich der statistischen Datenanalyse eine vertiefte Verbindung zwischen der modellhaften Welt der Mathematik und der bunten Welt aussermathematischer Realität.

Literatur

Bennett, C. A. (1954). Effect of measurement error inchemical process control, *Industrial Quality Control* **11**: 17–20.

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

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

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, Wiley, New York.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error, *Monthly Not. Roy. Astr. Soc.* **80**: 758–770.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philos. Trans. Roy. Soc. London Ser. A,* **222**: 309–368.

Hampel, F. R. (1980). Robuste Schätzungen: Ein anwendungsorientierter Überblick, *Biometrical J.* **22**: 3–21.

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

Hampel, F. R. (1987). Data analysis and self-similar processes, *Proc. 46th Session of the ISI, Tokyo*, Vol.LII, Book 4, pp. 255–264 (Discussion: pp. 255–264).

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

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

Jeffreys, H. (1939). *Theory of Probability*, Clarendon Press, Oxford.

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

Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics*, Wadsworth, Pacific Grove, Cal.

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

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

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

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

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

Schweingruber, M. (1980). Das Monte Carlo Verhalten einiger Verwerfungsregeln, *Diploma thesis*, Fachgruppe für Statistik, ETH, Zurich.

Stahel, W. A. and Weisberg, S. (1991). *Directions in Robust Statistics and Diagnostics*, Vol.1, 2, Springer, N. Y.

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

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

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

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