Is statistics too difficult?

by

Frank Hampel

Research Report No. 81
May 1997

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Is statistics too difficult?

Frank Hampel

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

May 1997

Key words and phrases: Uses and misuses of statistics, mathematical and applied statistics, statistical paradigms, interpretation of mathematical results, asymptotics, model choice, robust statistics, long–range correlations, field ornithology, astronomy, statistics as subject matter science.

AMS 1991 subject classifications: 62A99, 62-07, 00A69, 92B15, 85A35.

Abstract

By means of several historical examples, it is shown that it does not appear to be easy to build bridges between rigorous mathematics and reasonable data–analytic procedures for scientific measurements.

After mentioning of both some positive and some negative aspects of statistics, a formal framework for statistics is presented which contains the concept formation, derivation of results, and interpretation of mathematical statistics as three essential steps. The difficulties especially of interpretation are shown for examples in several areas of statistics, such as asymptotics and robustness. Some problems of statistics in two subject matter sciences are discussed, and a summary and outlook are given.

Résumé

Au moyen de plusieurs exemples historiques on démontre qu'il ne semble pas facile de construire un pont entre les mathématiques rigoureuses et des méthodes raisonables de l'analyse des données.

Après mentionner quelques aspects positifs comme negatifs de la statistique, on présente un cadre formel pour la statistique contenant la formation du concept, la dérivation des résultats et l'interprétation de la statistique mathématique en trois pas essentiels. Les difficultés de l'interprétation en particulier sont exposées pour des exemples dans des domaines diverses de statistique comme l'analyse asymptotique et la robustesse. On discute quelque problèmes de la statistique en deux sciences spécifiques et présente un sommaire et une perspective.

1 Introduction

This paper is a written version of the talk given on August 16, 1996 at the Symposium on Statistics and the Sciences in Halifax, beautifully organised by Luisa Fernholz, Chris Field and David Tyler with the aid of the Minerva Foundation and Dalhousie University. After stressing the positive sides of statistics and its interaction with the sciences, as they were also nicely shown at this symposium, I analyze in more detail various difficulties statistics faces, partly because of its unique position between pure mathematics and scientific data. Four areas of statistics are singled out to describe in more detail some problems of applications of statistics and of the dialog between statistics and the sciences. Some special remarks pertain to statistics and field ornithology, and to statistics and astronomy, before a data–analytic outlook is given.

To stress astronomy seems especially appropriate in honor of a genius loci, Simon Newcomb (1835–1909), a son of Nova Scotia, born in Wallace, who was the leading American astronomer of his time, and who was at the same time a most creative statistician, giving some of the deepest analyses of statistical errors in actual scientific data I know of. In his pioneering work, he described at length the "semisystematic errors" (the term he used; an important kind of violation of the independence assumption) occurring in real data (Newcomb 1895); and he proposed already mixtures of normals to model the deviations from normality occurring in real data (Newcomb 1886). He thus foreshadowed theories which were developed in statistics only about 80 years later, and his work can be considered a prime example for a connection on a very deep level between statistics and the sciences.

2 Statistics and the sciences: A challenge for statisticians

This subtitle was suggested by Luisa Fernholz; and it very appropriately summarizes the beauty and attraction which statistics has for many of us statisticians and data analysts. We find many interesting problems to work on, and we get into contacts with many different scientific fields (which puts high demands on interdisciplinary thinking and communicating for both sides). We are able to make helpful and often needed contributions to other fields, which means we can do service to a greater community, besides working for ourselves. And if our eyes are open, we can find all kinds of versatile and difficult challenges for research and problem–solving, for thinking and intuition, anywhere on the full scale between pure mathematics and data analysis.

In fact, in my extensive former consulting work, mainly for biologists, I rarely found a problem which could be satisfactorily dealt with by mere known statistical routine methods. Most consulting problems required some ad hoc modifications or extrapolations of existing methods or even the sketchy and heuristic (because of lack of time) development of new methods. Many of these could in principle be worked out mathematically to new statistical methods and theories. A collection of examples, some in reasonable detail, some only sketched, are given in Hampel (1987b).

I might also mention that one of the reasons that brought me into statistics was my fascination with the statistical problems arising out of my early work in field ornithology, especially bird migration (cf. also Section 9).

My statistical ideas in ornithology range from a simple but basic and general model to more sophisticated special ones (cf. Hampel 1987b, p. 109 ff). In recent years I noticed that one of the vague new concepts I had arrived at — but never worked out — while pondering over my data seems to be also the basic idea of the "possibility theory" ("théorie des possibilités") by Dubois and Prade (1988), which shows how deep thinking in data analysis can lead to new horizons.

Ironically, my main sophisticated model was much easier accepted by ornithologists than my simple one. This points to the possibility of nontrivial challenges in communication between statisticians and subject matter scientists (cf. Section 9 below).

Returning to sciences in general, we notice that many of our greatest statisticians were simultaneously working in some science (e.g. Gauss and Newcomb in astronomy, Gosset

("Student") in chemistry, Fisher in genetics, Jeffreys in geophysics, to mention a few), and also quite a few of our best contemporary data analysts (including J. W. Tukey) have in addition studied some field of science. I strongly believe this is no accident. Statistics is most alive when it reacts to the needs of applications. To be sure, mathematical (and philosophical) clarification and extension is also needed, but when it loses the connection to applications, it is in danger of becoming sterile. As Tukey said, "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

If we look at statistics as a whole, we see that it is not only useful and partly necessary in the various sciences, but that it is an integrated and in some areas necessary part of our civilization. Fields like quality control in mass production, medical drug testing, and survey sampling show this very clearly. Compare also Wallis and Roberts (1956), Tanur et al., eds. (1989) or Atkinson and Fienberg, eds. (1985).

3 The other side of statistics

Besides all its impressive successes, statistics also has another, darker side, and not quite accidentally often has a bad reputation. Let us briefly look at some of its various facets.

"Students frequently view statistics as the worst course taken in college" (Hogg 1991, Iman 1994).

The field of statistics is in a "crisis," and the subject has become "irrelevant to much of scientific enquiry" (Box 1995).

Many physicists look with disdain upon statistics and comment, for example, on the standard error of the mean: "It is merely a statistical result and has no correspondence with physical reality" (Jeffreys (1939) 1961, p. 301; p. 270). Compare also Section 8.

Statistics, overall, is used most prominently, and most superficially, in the "least scientific" fields of research (such as medicine, psychology, sociology.....). One may wonder about the reasons. (To be sure, there is also excellent statistical work in these fields; I am talking about an overall impression.)

Even if statistics has played a major (auxiliary) role in scientific progress (as it has in astronomy, cf. Section 10, or geology), this role is usually downplayed or ignored. Why?

Misuses ("How to lie with statistics") and misunderstandings, and plain lack of understanding of statistics ("statistical illiteracy") are so widespread they have become proverbial and need no further comment here (cf. also Huff 1954, Bibby 1983, Dewdney 1993).

Statistics is still a field with no agreed–upon foundations; apart from smaller groups, two main schools fight back and forth over centuries and do not seem to get together, nor does anyone of them offer a complete solution.

Modern outgrowths of statistics use different names ("artificial intelligence," "neural networks," "image analysis," "belief function theory"), and the suggestive term "informatique" in French or "Informatik" in German is occupied by computer science, which deals with storing and processing, but not directly with obtaining information. One may also think about the intended meaning(s) and actual uses of "data analysis" compared with "applied statistics."

A perhaps minor but bothersome point is that clinical trials (to my knowledge the only field of application where statisticians are required by law) are considered unethical by statisticians with different viewpoints on foundations.

Several ones of these aspects are of course connected.

I think we have to face these facts, observations, opinions, beliefs, convictions, experiences ..., as well as the positive ones about statistics, in order to get a more balanced view, and to avoid that neglected and repressed aspects rise unexpectedly "out of the shadow" and spoil our efforts towards better statistics.

4 Statistics between pure mathematics and the sciences

Besides difficulties which may occur, and sometimes do occur, also in other fields of research, one specific problem for statistics is its peculiar situation in the field of tension between the rigor of pure mathematics and the need for reasonable pragmatic solutions in the context of complicated and not fully formalizable "real life" situations in the fields of application.

As a framework for discussing this situation, I suggest the following 4–point, 3–step scheme. First, there is a vague, ill–defined but clearly existent data–analytic problem in some (or all) field(s) of application. A first step towards its solution is its mathematical formalization. This step is rare (and not very often needed), but very important, as it defines the new mathematical concepts and quality criteria which are then used, routinely and often uncritically, by mathematicians (mathematical statisticians) in their step (the second step) from a well–defined mathematical problem or paradigm to rigorous mathematical solutions. Mathematically well–chosen concepts (whether well applicable or not) can give rise to large theories in mathematical statistics, and clearly most of the mathematics–oriented work and publishing activity in statistics is in this second step.

But in observing statistics over the decades, I became more and more convinced that we often explicitly need a third step: the interpretation of the mathematical solutions in the light of real data and of the situations which first stimulated the preceding investigations. Only this interpretation shows to what extent a mathematically optimal solution derived under rigorously specified conditions is also a useful practical solution under the different conditions of real data in scientific research. I think this step, usually overlooked, is often most sorely needed. Interestingly, it is also missing as an explicit step in the short but worthwhile related paper by Huber (1975), though it is indirectly briefly alluded to in the paragraph on p. 84/85 (loc.cit.).

An example for the 3 steps is provided by the ("classical") robustness problem (cf. also Section 7). In the beginning, there was just a vague problem of the stability of statistical procedures under deviations from idealized modelling assumptions. Then more and more isolated examples became known showing its urgency. But only with the coining of precise robustness concepts, starting in 1964, began fruitful systematic research, building a formal robustness theory, which allowed a broad and encompassing understanding of the problem. The ensuing research was carried away by its own momentum, and I think it is precisely because of lack of interpretation and understanding that at some time ridiculously many (mostly superfluous) robust estimates of location were included into large computer packages, while strongly needed robust regression estimators were not. The robustness example also shows that mathematical optimality is not always practical optimality: hardly anybody uses a tanh-estimator (which is in some ways "optimally robust"), but Tukey's biweight or Hampel's 3–part redescenders are rightly often used although they possess no (or no important) optimality property. The mathematical robustness theory has led both to the development of new robust procedures, and to a much better and deeper understanding of the properties of old, more or less robust, procedures. Interestingly, this was not only done by interpretation of and comparison with optimal robust procedures, but sometimes directly by the application of the concepts of robustness, as with the clarification of the behavior of rules for rejection of outliers by means of the breakdown point (which needed only a little mathematical research to work out the breakdown points).

This suggests in addition sometimes a direct connection from point 2 to point 4, largely skipping step 2.

Besides the general framework described above, two alternative schemes seem sometimes possible or even necessary (partly in combination with the first one or with each other). One is to go from the vague applied problem to the shortcut of non–rigorous mathematics and heuristics, which then may lead to semi–intuitive solutions. These then need again an intuitive interpretation in the light of real data. This shortcut, with its danger of making formal mistakes (the only ones that count in pure mathematics), needs a good mathematical intuition which seems to be not very common. On the other hand, practical solutions cannot always wait for a full–fledged mathematical theory (mathematicians themselves tell jokes about this); and for example the Dirac delta function, invented by physicists with a good grasp of mathematics, was first laughed at by mathematicians (it still is), and then they built a rigorous theory based on its idea.

Another scheme works via a computer simulation, making statistics partly an experimental science. However, already for the Monte Carlo setup, we need at least some models or theories. Next, the computer runs yield many isolated numbers, and it is often a nontrivial task to obtain meaningful patterns by a purely empirical analysis which then allow a limited extrapolation beyond the situations tried. The full power of a Monte Carlo study comes when it is combined with a good theory. The analysis in the light of the theory allows a much deeper understanding of the observed patterns, as well as the discovery of many new ones, and it permits a much broader extrapolation (cf. also the remarks on the study in Andrews et al. 1972 below). Again, the last step needed is the interpretation of what has been learned for the purposes of real life applications.

5 The problems of asymptotic theory

This section, as well as some of the following ones, describe in more detail some difficulties of interpreting mathematical results in several different areas.

The first one chosen is asymptotics. It comprises a large part of mathematical statistics, probably with the largest amount of applications (just think of asymptotic normality), and it can be eminently useful for practical statistics.

However, taken by itself, it is (usually) totally irrelevant. It only tells us what happens if something, usually some sample size n, "goes to infinity"; it does not tell us anything for any fixed finite n. Yet, in practice we need answers for finite n. It is then suggested that asymptotic theory provides, or may provide, approximations for finite n, but which n is big enough so that the approximation is sufficiently accurate for the purpose at hand? $n = 20$? 100? 1000? one million? 4? 1? All these answers can be true under specific circumstances, but my point is that the purely mathematical theory alone does not tell us which one.

I think that (the few cases with usable strict error bounds excepted) we need either intuition and insight into the mathematical formalism (a "feeling" for the formulas and structures used), and/or the underpinning by means of a Monte Carlo study or other (exact or sufficiently accurate) finite sample results, in order to "anchor" the asymptotic results derived at fleeting infinity in the finite realm. Best would be a combination of all three approaches, as done in some, but not all, analyses of the large Monte Carlo study in Andrews et al. (1972; cf. also Hampel 1995) which would not have succeeded without the combination of powerful asymptotic theory, deep intuition about its interpretation and application, and penetrating analysis of the many finite sample results. Mathematical intuition is possible, on various levels of mathematical sophistication (cf. Cuthbert Daniel's admirable intuition), but it is normally not cultivated. Apparently most mathematical statisticians have learned to do proofs, but cannot understand in a deeper sense what they have achieved, and what not.

A rather extreme example of lack of mathematical feeling is the old "large deviations" theory as described, for example, in Feller (1966). According to L. LeCam, it was considered the "messiest area of mathematical statistics" at least up to the sixties. Now the proofs were all mathematically correct, but (rare?) numerical applications for small and medium sample sizes were extremely bad. One needed exorbitant sample sizes (as they sometimes occur in physics, like statistical mechanics, for example) to get good results.

The reasons, as I see them, are the following: Experience shows that in what I consider the most natural asymptotic expansion, one needs the first two terms in practice (and hardly more). Using the first term only is (even literally) like forcing every straight line through the origin; but it is common intuitive knowledge (based on a mathematical background) that one can approximate any sufficiently smooth curve locally by a (general!) straight line. In traditional large deviations theory, one has taken the first term only, and then expanded it into another infinite sequence (in a different "direction"), of which the first, or the first few terms were to be used. From the foregoing follows that even the full infinite sequence would give very bad results, since all the other relevant terms had been "lost" before on the rather peculiar path towards infinity.

It is interesting to observe that the first term of the saddlepoint approximation, a specific technique for such asymptotic results which has become quite popular in statistics fairly recently, corresponds essentially to the first two terms in the basic expansion mentioned above. Only essentially, because the first term of the saddlepoint approximation contains in addition an artificial normalisation, which is the best one can do without using any more information, but which is often abandoned by users of the saddlepoint technique themselves in favor of an empirical normalisation (or a higher order expansion). The full saddlepoint expansion contains also an expansion of an exponential with small exponent into one plus a sum, and although the difference is slight, this expansion empirically leads again to a slightly worse approximation.

There exist a number of further, different insights into this realm (including how to resolve the existence problem for the moment generating function very generally). For a publication, describing also what was said about large deviations (besides many new mathematical facts), see Field and Hampel (1982). But apparently many specialists on the saddlepoint technique did not notice, or not care about such nonmathematical trivia; the publication of this paper was delayed by four years, outlooks on possible future research were eliminated by the editor, and a reviewer described the paper, in the context of other papers on the saddlepoint method, by something like: "More of the same." This may or may not have been an exception, but it is not the only example to show that papers which go beyond mathematical formalism into a deeper interpretation of mathematics may encounter special difficulties.

Another example for problems with the interpretation of asymptotics are Bayes procedures. True, usually they are "asymptotically" equivalent to maximum likelihood procedures and hence "asymptotically optimal"; but for any fixed n they may be arbitrarily far away from their "asymptotic" behavior. Their convergence is not uniform with respect to the prior distribution. Hence, strictly theoretically speaking, any Bayes solution is worthless for any fixed n (unless, of course, one believes in the assumed prior, as most Bayesians pretend to do). I suggest that in order to use asymptotic theory for approximating the behavior of a Bayes estimator for a given fixed n, one should embed the estimator into a different sequence of estimators which is more informative. For example, some simple

Bayes estimators are of the form $\overline{X}_n \cdot n/(n+m) \ + \ c \cdot m/(n+m)$, with c and m known and fixed. "Asymptotically," for $n \to \infty$, these estimators behave like $\overline{X}_n$, but for the given n, it would be much more instructive to use whatever asymptotic approximation for the distribution of $\overline{X}_n$ and plug it into this formula; and the simplest and most informative sequence of estimators which "crosses" the sequence of Bayes estimators at this point, is $a\overline{X}_n + b$, with a and b fixed and chosen to match the above formula for the given n. Of course, mathematical statisticians may claim that they are not interested in the behavior of the sequence of estimators for *any* fixed n, only in their "limit" behavior (and they have a right to do so, as a first step (!), in their search for simple and interpretable (!) structures); but ignoring the path towards infinity totally can lead to misleading, bewildering, and stupid claims, as in the discussion of a certain type of "superefficient" estimators (e.g. for the normal distribution) where, by the way, Fisher's vague intuition about ''consistent" estimators appears to have been at least as important as the rigorous mathematical formalism (and optimal is the combination of both). Other examples where there are several, more or less natural (or else, practical) ways of embedding a given estimator into sequences going to infinity, are α–trimmed and α–Winsorized means, and various rules for the rejection of outliers (cf., e.g., Hampel 1985 for some special forms of "asymptotic" argument).

We see that in practical applications of asymptotic theory, we have not only the question: How to extrapolate back from infinity to a given n, and what is the accuracy of that extrapolation? We also have the big question: How to embed a given practical situation (for fixed n) reasonably and meaningfully into one of the many asymptotic sequences leading through it to infinity? Again, intuition for mathematical formalisms is needed, not merely mastery of mathematical proof techniques, if the results of mathematical theory are to be fruitful for applications.

6 The problem of "specification" or model choice

This problem is perhaps farther away from any satisfactory theory, especially from any good and encompassing mathematical theory, than most other problems in applied statistics. We are led to such innocuous–looking questions as: What is a "simple" model? What is a "good" model? I don't know any general answers.

It is interesting to note that Fisher left the problem of "specification," as he called it, explicitly to the practical statistician (an attitude which looks somewhat evasive to me).

It is true that there are some mathematical theories solving parts of the problem area. In particular, there are many methods for comparing models within a given set of hierarchically structured models. But even then, there are questions about whether to take a fixed subset of variables fully, whether to mix variables, to transform variables, or to introduce new, "latent" variables, and whether to do the same or different things for different purposes (parameter estimation and testing, prediction in various contexts).

On a higher level are the methods based on information theory (cf., e.g., Rissanen's work) to describe the quality of any parametric model, both with regard to the goodness of its fit and to the complexity or simplicity of its form. I think this may be a fruitful road, but one problem I see is the arbitrariness in the choice of the program used to describe the model and the data. This arbitrariness seems to be related to the problem in the foundations of probability of describing the degree of randomness or nonrandomness of a finite sequence of, for example, zeroes and ones.

Another problem is that almost all models are only approximate, and that one needs to incorporate this approximate character into information theory. Some research in this direction does already exist. In a different framework, this problem of inaccurate models has been treated and largely solved by robustness theory (see Section 7 below).

A further complication is that a good model choice is determined not only by the data (often in several iterations), but also by experience with similar data (e.g. "borrowing strength"), by general statistical experience (about which models tend to be useful), and by the background knowledge of the subject matter science, including the existing formalizations and theories about these and related phenomena.

It seems difficult to formalize the intuition about model choice that exists already in applied statistics. Not only are there statements like: "All models are wrong; but some are more useful than others." But there is also experience that the "best" model (which is so "simple" that it allows also extrapolation for "distant" populations or new situations) is often not among the ones statistically compatible with the data at hand, but rather "significantly wrong"! I believe this has to do with the "long–range correlations" in virtually all data (and hence models, if not explicitly incorporated), with the "semisystematic" errors and heterogeneities somehow due to slowly changing background conditions, which eventually lead all our models based on independence and short–term correlations astray (cf. Section 8 below).

Given a data set, we have to ask up to what extent it supports a ("tentatively entertained") model; once we arrived at a model, we have to ask up to what extent it describes real data. In data analysis, we ought to do "fitting equations to data" and not the other way around.

7 "Classical" robustness theory

Robustness theory studies the deviations from exact parametric models and their consequences. It considers both the deviations from the assumed marginal distributions ("classical" robustness theory) and from the assumed correlation structure, violation of the normality and of the independence assumption being the most important special cases. It may be (and might have been) called the stability theory of statistical procedures; but unlike stability theories in other parts of applied mathematics, it met with much resistance among parts of the statistical community.

One (minor) problem was the introduction of a new paradigm. I still remember how in the traditional way of doing mathematical statistics there was simply no space for a full neighborhood of potentially true distributions (in a way, they were "beyond infinity"); and I also remember a talk on differential geometry in statistics where the speaker missed one little but decisive step; namely that he had complete freedom to choose the most beautiful and elegant probability distributions outside the parametric model on which they where fixed.

Another problem was probably a certain prejudice, an attitude that "we don't need all this" because the improvements brought by robustness theory were deemed unimportant. Apparently rather influential was a paper by Cox and Hinkley (1968) which on the basis of facts given by Jeffreys created the impression that least squares loses in practice usually less than 10% efficiency under nonnormality (cf., e.g., McCullagh and Nelder 1983) without ever making such a statement, while Jeffreys' facts lead to the opposite conclusion. Another misleading paper was Stigler (1977) which contained the basic conceptual flaw of ignoring systematic and semisystematic errors and their consequences; when the flaw was pointed out in the discussion of the paper, the author ended the discussion with a mathematical error (cf. Hampel et al. 1986, p. 31f). Some reactions were obviously all too human, and they were facilitated by the facts that models cannot be fully trusted, and that all empirical statistical results contain random variability, so that it is hard to assess reliably the quality of any statistical procedure in practice (as opposed to other parts of mathematics).

Again another problem was misinterpretation of mathematical theorems. The Gauss–Markov theorem says that least squares estimates are optimal among all linear, unbiased estimates, even without any assumption of normality; it neglects to say that outside a very small neighborhood of the normal distribution, *all* linear estimates are rather inefficient — a fact known and stressed already by Fisher (1922) in a slightly different context. An example where empirical evidence and mathematical reasoning seemed to clash was the dispute between Fisher (1920) and Eddington (1914, and footnote in Fisher, loc.cit.) on the better scale estimator: Fisher proved that under strict normality the standard deviation was optimal, while Eddington maintained that in his astronomical data the mean deviation was empirically better. Interestingly, the mean deviation, not the standard deviation, all but disappeared from statistical practice, apparently showing a stronger belief in, and reliance in practice on isolated pieces of mathematical theory proved under unwarranted assumptions, than on empirical facts and observations. The reconciliation of both lines of thought (which always ought to exist, of course) came much later with the work by Tukey (1960) and Huber (1981) who showed also mathematically that Eddington was in fact right (under more realistic models than Fisher's), although neither estimator should now be used in practice (while the standard deviation still plays an important theoretical background role).

Practitioners of statistics have noticed for more than one and a half centuries that even high–quality measurement data are not exactly normally distributed, but typically longer–tailed. The tacit hope that small deviations would have only small effects on statistical procedures was shattered by Tukey (1960). Practitioners also know that real data contain gross errors or blunders; in fact, for scientific routine data (i.e. data not taken with special care), 1–10% gross errors are the rule rather than the exception and should be faced realistically (some fields, like medicine, tend to have even higher rates of gross errors). Gross errors often show up as outliers, but not all outliers are gross errors: some may be by far the most valuable observations of the data set, even worth a scientific prize or a patent. Statistical rules for rejection of outliers (better: for separating the outliers from the rest) were partly so bad that they could not even reject a single distant outlier, and the mathematical theory about them was so limited in scope as to be pretty useless in practice. The theory of robustness offers a deeper understanding of the properties of these rules, better rejection rules, and better alternatives to rejection rules (Hampel 1985); but many statisticians were rather reluctant to accept the new tools.

For more details on robust statistics, including more discussions of misunderstandings, see for example the books by Huber (1981) and Hampel et al. (1986). The latter book contains a whole subchapter (8.2), with further references, on misapprehensions of robustness, including more than 5 pages just on misunderstandings of Huber's (1964) pioneering first paper in the field, reprinted with an introduction (also for applied statisticians) in Kotz and Johnson, eds. (1992).

8 Semisystematic errors, and why statistics still sometimes works

There is hardly any paradigm more deeply entrenched in statistical theory and usage than that of independence. Virtually every statistics book starts with the assumption of "independent identically distributed" random variables, observations, measurement errors, or the like, and hardly any book questions this assumption seriously, except in contexts like time series analysis where some short–range correlations are obvious. But the truth is that really independent observations practically do not exist, and that even measurement errors of high–quality data in the hard sciences like physics, astronomy, and chemistry are not only dependent, but even long–range dependent (which means they cannot be modelled by any strongly mixing process such as ARMA processes).

This fact was qualitatively known to eminent statisticians, and published by them, for over a century. The first reference I know is Simon Newcomb (1895) who gave a penetrating analysis of astronomical measurement errors. Gosset ("Student" 1927) gave a similarly deep analysis of the structure of errors in chemistry. Karl Pearson (1902) and some of his co–workers did extensive empirical investigations with pseudo–astronomical observations made by them; "K.P." found not only qualitatively long–range correlations in time for each observer and tried to model them by ad hoc models; he also discovered surprising cross-correlations between different observers which to my knowledge later have never been studied again. Jeffreys (1939) reanalyzed these and other series and in addition to the long–range correlations in time found a strong negative correlation between the intensity of the correlations and the long–tailedness of the data, another clue that to my knowledge has not yet been pursued further.

Yet these studies (and a few others) were completely ignored by the mainstream of statistics. People continued (and mostly still continue) to believe in the "declaration of independence" (Box et al. 1978) instead of "hunting out the real uncertainty" and noticing "how $\sigma/\sqrt{n}$ can mislead" (Mosteller and Tukey 1977). This attitude dominates statistics virtually uncontested, although also in our times, top applied statisticians like Cuthbert Daniel (cf. Daniel and Wood 1980, and the two books just cited), and in all probability also experienced and realistic subject matter scientists, are very much aware of at least the qualitative unsuspected correlation problem.

The mathematical modelling of long–range correlations came rather late. It was B. B. Mandelbrot who in a series of papers around the late sixties reanalyzed the Nile data (studied by Hurst for the construction of the Aswan Dam), which had puzzled even top probabilists like Feller, and showed that despite their "nonstationary" looks, they could be modelled by a stationary though long–range dependent process, namely a self–similar process invented much earlier in an abstract setting by Kolmogorov (1940); moreover he showed that not only hydrological, but virtually all geophysical processes he studied (river flows, geological layers, tree ring indices, annual precipitations, earthquake frequencies ...) are long–range correlated (Mandelbrot and Wallis 1969).

There was a big debate, particularly in hydrology, pro and against Mandelbrot's model. Now I do not think that the model of self–similar processes is the only reasonable model (e.g. fractional ARIMA processes, invented later, behave very similarly), and if reasons for nonstationarity are known, they can clearly improve the model; but I do think that any model which does not account for the features of long-range correlations observed in real data is doomed to failure in the long run.

Mandelbrot's statistical tools were still rather coarse. They were optimized and expanded in our group (Graf et al. 1984; Beran 1989, 1992), and applied not only to further geophysical series (such as the famous Arosa ozone series), but also to pure measurement series in chemistry (taken from Student 1927) and physics (about 3000 measurements of the velocity of light, taken from Michelson et al. 1935; sets of hundreds of measurements of the 1 kg check standard weight done with utmost accuracy by the National Bureau of Standards in Washington, D.C.). Even those segments that looked "under statistical control" (Shewhart) were not independent, but highly significantly long–range correlated. Why, is still a mystery to me, although a part of the phenomenon might be explained by the discussions in Newcomb (1895) and Student (1927): it just appears to be a fact of life.

The consequences of long–correlation models such as self–similar processes are serious. Apparent "trends" and "cycles" ("hidden periodicities") may just be artefacts of such a process. Naive estimates of the variance are highly biased and inefficient. The variance of the mean of a sample of size n goes down to zero at a slower rate than $1/n$ (like $n^{-\alpha}$, with α anywhere between zero and one, depending on the particular series), and hence the standard error of the mean and the lengths of confidence intervals and of acceptance regions of tests are wrong by a factor tending to infinity with n. For example, for Student's $n = 135$ measurements, the true variance of the mean is already about 20 times the variance derived under the independence assumption commonly used for such data. For more details, see Hampel et al. (1986, Ch. 8.1), and Hampel (1987a).

The reaction of statisticians to these facts can be quite interesting. I once was told about the reaction to an invited talk on this topic for a broad range of statisticians. The first day, the audience was shocked. The next day, they tried to suppress and forget it, and do "business as usual."

Now, in view of the above facts, we must wonder why statistics sometimes works, as it does, according to the folklore of applied statistics. The answer is very interesting, and in full accordance with experience and intuition of applied statistics. Statistics does not work, except with rather small samples, for absolute constants (in accordance with the "skeptical physicist" of Jeffreys 1939, cf. also Youden 1972). But it does work (in a limited sense) for contrasts under well–mixed experimental conditions (Künsch et al. 1993), because then the first-order effects of long–range correlations cancel out, and the levels of confidence intervals and tests are approximately correct. Still, there may be large losses of power and efficiency; but these can in turn be greatly reduced by using small blocks of experimental units. Now, areas like regression and analysis of variance are generally counted as main success stories of applied statistics ; but slopes in regression and effects in ANOVA are prime examples of contrasts (and blocking is a highly recommended tool). On the other hand, intercept and grand mean are not contrasts, and it is a remarkably wise custom, now deeply justified, that they are left out of the usual analysis of variance table. Despite these results, which create a beautiful bridge between deep intuition of applied statistics and high–level mathematics and which would appear to be of rather basic importance for all of statistics (as they investigate a decently realistic model of "random errors"), it was very hard to get the paper published.

It is amazing to see how long the tradition of mathematical statistics, as reflected in all statistical textbooks, has ignored basic knowledge of the tradition of good applied statistics, being either not able or not willing to model salient features of statistical data and, worse, not being willing to admit that there is an open problem, thereby discrediting statistics as a science in the eyes of experienced subject matter scientists.

9 Some personal experiences in field ornithology

It is not only statisticians who may have problems understanding statistics. When I now mention some personal experiences with field ornithology, the "scientia amabilis" with which I had rather early contacts, I hasten to stress that there is also excellent statistical work in this field, for example the first large bird survey of a whole country (Merikallio 1958) with its blend of well–used statistical techniques and profound ornithological intuition and experience. My experiences may not be at all representative, but I mention them as examples of what *might* happen, presumably also in other fields.

A basic problem was that the ornithologists I spoke to essentially could not imagine that a bird might be in a certain place even though it was not observed (simply because there was no observer there), and that one could genuinely estimate the number of unobserved birds from the number of observed birds, given a suitable data base which in addition to the usual data included nothing more than an "excursion list," or a description of how well a certain area was covered at what times. Many amazing differences in bird occurrences could largely be explained by the different intensities of bird watching (as well as different abilities to identify birds in some cases), and a "rare bird," that elusive romantic "blue flower" of many hobby ornithologists, became just a bird (— how sober! —) with low probability of encounter ("Antreff–Wahrscheinlichkeit", cf. Hampel 1965, 1966). Seeming contradictions just dissolved, like the fact that we hardly ever saw wild swans on a certain lake while a permanent resident claimed to see them practically every winter; we just had to count the number of excursions we had done around this time of year. To extrapolate from the observed to the unobserved would seem to be one of the main tasks of statistics.

Some years later, I learned what modern "scientific" field ornithology was supposed to be. In order to solve the problem of gaps in the excursion list, only daily observations for a considerable length of time were accepted and evaluated, the great majority of data (obtained with considerable enthusiasm, and containing lots of precious information) was discarded as being "scientifically useless." Instead of adapting the method of evaluation to the data available, the ornithologists were forced to adapt their behavior (for the sake of "science") to the most primitive evaluation method which was the only one considered or known, or else throw their data away. To be sure, there are some advantages in having long uninterrupted sequences of data, and it may have been good to encourage people to obtain them where feasible, but there was no need to waste most available information. As a side point, averages were computed over rather unintuitive but equally long periods of time, instead of, for example, approximate thirds of months, which just would have required weighted averages with slightly nonconstant weights for further evaluation.

Again later, I heard that a big fuss and many publications were made about the so–called "weekend effect," namely the fact that most rare birds show up on weekends. With something like an excursion list, the weekend effect would simply disappear (or else what small — positive or negative — effects would remain would be due to rather subtle effects, such as different average qualities and eagerness of weekday and weekend observers, and sometimes to different disturbance of the birds).

I still think it would be a great statistical challenge to try to model bird migration quantitatively (the numbers, routes, speeds, and resting patterns of birds), using among several other types of data the large numbers of recoveries of banded birds and counting them not only absolutely, as has been commonly done, but trying to estimate (at first relatively, then crudely, then perhaps using nonparametric smoothing techniques) the chance that a banded bird in a given locality or area is caught again, shot or found, and reported to a bird observatory. Already before that step, relative frequencies of ecologically similar species could be estimated for different areas. We might have to acknowledge how little we still know quantitatively, and might have to cope with systematic errors, large random variability, long–range correlations in time and space, mavericks, gross errors, and possibly trends and changing patterns; but statistics, used properly, would be a central tool for bringing light from the known into the unknown.

Now the problem is not that statistics is not used in field ornithology. Already during my active time in the sixties, statistical methods were proposed to be used (in order to be "modern" and "scientific"), but the problem was that they were just superficially superimposed on the ornithological data, without exploring what kind of evaluation the data themselves were calling for. (For examples, though in very different fields, of what I mean by fitting equations to data, and not fitting data to equations and preconceived models, see the books by Daniel and Wood 1980 and Daniel 1976.) Clearly, most ornithologists had no intuition about the underlying assumptions and the strengths and limitations of statistical methods, and statisticians in general had no or little understanding of the peculiarities and rich and sophisticated background of ornithological data. It needs high competence in both fields, or at least a broad overlap of competences of two scientists somewhere in between, with a good ability to communicate, in order to do really good applications of statistics in a subject matter science.

10 Statistics and astronomy

Statistics and astronomy have very old and close ties. One of the historical roots of statistics, the theory of errors ("Fehler- und Ausgleichsrechnung"), goes back to astronomy and geodesy. Statistics was not only useful, but a necessary and central tool for all astronomy outside a small neighborhood of the solar system, including all we know, or believe to know, about extragalactic astronomy, the structure of the universe, cosmology, the "big bang," and other related theories. I once read the fitting description that the astronomical view of the world has been obtained by putting pyramid after pyramid on top of each other, each pyramid resting on its tip; and all but the first pyramid (direct parallaxes of neighboring stars) are based on inaccurate and laborious statistics, from the proper motions of nearby stars to the period–luminosity relationship of delta Cephei–type variable stars and the redshift–distance law of galaxies. A nice description of this bold construction is given in Rowan–Robinson (1985). Yet I have the impression that in most popular accounts of astronomy, the central role of statistics is downplayed or not even mentioned, and astronomers themselves seem to push it aside (as a necessary nuisance? or as a scientifically uninteresting tool? or what else?).

By the way, I also get the impression that quite a bit of the cumbersome and perhaps boring statistical groundwork in astronomy was done by women, while most of the fame went to men who drew further conclusions from their findings. At least I read rarely anything about Annie Jump Cannon and Antonia Maury who laid the foundations for the famous Hertzsprung–Russell diagram, or about Henrietta Swan Leavitt who discovered the statistical period–luminosity relation of Cepheid variables (cf., e.g., Rowan–Robinson 1985). One might speculate whether in some sense both women and statistics are the Cinderella of sciences, doing a lot of the necessary "dirty" work without too much recognition.

It may be that connections between statistics and astronomy are presently again increasing. For example, there are two meetings on both topics at the I.S.I. conference 1997 in Istanbul, and there is also a recent book with the programmatic title "Astrostatistics" (Babu and Feigelson 1996). In this book, both fields are presented to scientists of the other field, with a lot of potentially useful and stimulating information. However, in glancing through the book, I got the impression that again statistical methods are looking for data to be applied to, rather than being developed for the needs of the data. This may be very helpful to some extent, but it can also lead to the type of superficial statistics that is rightly despised by good scientists. On the other hand, the book also points out many needs of astronomical data, which should and hopefully will stimulate statistical research. The mathematical–logical arguments in the book are usually of the fragmentary type commonly found in statistics books, giving isolated islands of rigorous conclusions based on rigorous assumptions without embedding them in the ocean of experience and background knowledge concerning the application of mathematical models to the real world. This makes most conclusions only half–truths (or partial truths) in applications. For example, if the sentence "The entire [parametric] analysis will be a futile exercise if the underlying model is incorrect" (loc.cit., p. 85) were taken at face value, the authors could have spared writing large parts of the book (because parametric models are practically never correct); but quite rightly they did write them, because the mathematical dichotomy "correct – incorrect" is not appropriate in practice. Or the statement "There is no theoretical reason why one should prefer least squares method over methods that minimize sum of absolute distances or some other metric distances" (loc.cit., p. 60) floats around in isolation and shows the helplessness of purely mathematical reasoning regarding applications; a lot can be said about this topic, combining theory, insight, and experience (cf. also Section 7). Some minor surprises for me were that at cursory glancing through the book I did not see Simon Newcomb mentioned, one of the astronomers who thought most deeply about the connection between astronomy and statistics and whose thoughts are still highly relevant; I found Eddington only negatively mentioned as having no connection with statistics (loc.cit., p. 4; but cf. Section 7 above); and in the historical section I missed even Gauss (except in the term "Gaussian distribution") and his famous story about the rediscovery of Ceres, undoubtedly one of the greatest triumphs of astrostatistics ever. But these remarks should not distract from the value of the book in bringing together, in a concise and readable way, a lot of technical information about statistics, as well as about astronomy. I only try to point out that above the (necessary) mathematical–technical level of statistics there exists still another level, where for example the usefulness and limitations of statistical methods in astronomy are assessed, and which some of the greatest astronomers were clearly aware of.

11 Summary and outlook

While almost all of statistical research is done on the purely mathematical–technical level (and this may often be a necessity for young people trying to get recognition), and while even this level sometimes has its intrinsic problems (cf. Section 5), I tried to show that there are two other steps which are necessary for applying statistics properly (cf. Section 4), and which are on a different, in some sense higher level than mathematical statistics. I tend to distinguish three levels of mathematics and its use: in the middle the core or skeleton of pure mathematics, the center within which most mathematicians are working; above it the intuition ("blood, flesh, skin...") about mathematics, which becomes important for the most creative geniuses in mathematics, and which is also central for defining new concepts and paradigms and for applying mathematical results to the non-mathematical world, as in good applied statistics; and below it the use of mathematics (in our case, statistical formulae) with neither formal nor intuitive understanding (e.g., "cookbook statistics" in the bad sense). We might also say the middle level is characterized by analytic thinking, and the upper level by synthetic thinking. In talking to my colleagues in pure mathematics, I got the impression that some of them are not even aware of the existence of the upper level, or try to repress it: this may be also a reason for the difficulties good applied statisticians have in getting recognition by mathematicians.

But I think the "logic" of good applied statistics is different from, or at least much more sophisticated than the logic of pure mathematics. It takes the approximate character of all models, realistic situations, and various background knowledge and experience into account, things which are hard or impossible to formalize logically. (Some aspects may be formalized and then lead to new research in mathematics; but in a data–analytic situation we also have to take care of the unformalized rest.) Two simple examples for the different ways of thinking in pure mathematics and in applied statistics are due to J. W. Tukey: Consider the distributions $F(x) = (1 - 10^{-10})$ Standard–Normal $+10^{-10}$ Standard–Cauchy, and $G(x) =$ Standard–Cauchy restricted to $[-10^{100},\ 10^{100}]$. $G(x)$ has all moments, the Central Limit Theorem applies for i.i.d. observations from it, etc.; but for almost all practical purposes it behaves like a Cauchy distribution. On the other hand, the mean of i.i.d. observations from $F(x)$ behaves "normally" except for exorbitantly large samples, although the expectation does not even exist.

The requirements of real data analyses, with the additional restrictions of time, money, etc., are different from those of mathematical statistics, although the door to the latter should always remain open. I may cite a paragraph (Hampel 1987b, p. 95) describing my own "philosophy" in consulting:

"The data were analyzed as they came in. The idea was to do what was required by the data (and their scientific background), *no less and no more.* Doing no less implies that there was hardly any interesting data set in consulting which the writer could treat by mere routine methods... Doing no more implies, for example, that many approximations could be improved or generalized, but this was not deemed necessary for the particular data at hand... though some readers may find some germs of general 'theories' or 'methods' in some of [the analyses]."

To give an example for the bridge to mathematics: I was told that much later there were 2 Ph.D. theses on the nonparametric density estimates for the distribution of the resting periods of migratory birds (loc.cit., p. 111f), on their rates of convergence etc. using mathematical tools which did not even exist when I analyzed the data; but these data, precious as they were, contained hardly more information than for a crude and biased estimate of the mean and possibly the variance.

The above discussions (especially Sections 3, 7, 8) show that statistics should perhaps study more open–mindedly what real data look like (rather than being caught too early in preconceived model assumptions). With many and large data sets, including graphical displays, easily available now on the computer, and with the tools of simulation and Monte Carlo studies having become convenient ("Student" 1908 did his simulations still by hand!), statistics might become more of an experimental science (like physics; with actual data as its subject of study), and only one part of it (mathematical statistics) would (and should) try to make it mathematically rigorous (the "skeleton"). We still can (and have to) learn quite a bit about the general properties of real data. A few open problems have been mentioned, and I am sure there is more in the oral traditions of experience of subfields like quality control, biostatistics, and survey sampling, including governmental statistics; we might even learn from fields like information theory, computer science, different kinds of logic, linguistics, psychology, and philosophy, as "data" are subject to many kinds of distortions, misconceptions, correlations, systematic blunders, and faulty or shaky definitions, in addition to classical random variability.

Statistics does often need *high–level mathematics.* But it needs also *intuition about mathematics, intuition and experience about statistics, intuition and high standards about computer experiments, intuition and experience about real data, intuition and background knowledge about the data's subject matter area.*

Let us face reality and acknowledge what is still unknown.

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.

Atkinson, A. C. and Fienberg, S. E. (eds) (1985). *A Celebration of Statistics; The ISI Centenary Volume*, Springer-Verlag, NY.

Babu, G. J. and Feigelson, E. D. (1996). *Astrostatistics*, Chapman & Hall, London, UK.

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

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

Bibby, J. (1983). *Quotes, Damned Quotes, and...; An anthology of sayings, epithets, and witticisms – several of them something to do with statistics*, Demast Books, Halifax.

Box, G. E. P. (1995). "Scientific Statistics – The Way Ahead (abstract)", *1995 Abstracts: Summaries of Papers Presented at the Joint Statistical Meetings*, American Statistical Association, Alexandria, VA: 38.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters*, Wiley, New York.

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

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

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, New York. *1st ed. 1971.*

Dewdney, A. K. (1993). *200 % of Nothing; From "Percentage Pumping" to "Irrational Ratios"*, Wiley, New York.

Dubois, D. and Prade, H. (1988). *Theory of Possibility*, Plenum, London, UK. *Original Edition in French (1985) Masson, Paris.*

Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*, Macmillan, London, UK.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, New York.

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

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. *Reprinted in Contributions to Mathematical Statistics. Wiley, New York 1950.*

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philos. Trans. Roy. Soc. London.* **A 222**: 309–368. *Reprinted in Contributions to Mathematical Statistics. Wiley, New York 1950.*

Graf, H., Hampel, F. R. and Tacier, J.-D. (1984). The problem of unsuspected serial correlations, *in* J. Franke, W. Härdle and R. D. Martin (eds), *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statist. 26, Springer, pp. 127–145.

Hampel, F. (1965). Artenliste vom Seeburger See 1955-1964 (unter knapper Berücksichtigung des Raumes um Göttingen), mimeographed manuscript, Göttingen, 23 pp.

Hampel, F. (1966). Ueberwinterung und Verhaltensweisen der Beutelmeise (Remiz pendulinus) am Seeburger See, *J. für Ornithologie* **107**(Vol. 3/4): 359–360.

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

Hampel, F. (1987a). Data analysis and self-similar processes, *Bull. Internat. Statist. Inst., Tokyo* **52**: (Book 4) 235–264, (with discussion).

Hampel, F. (1987b). Design, modelling, and analysis of some biological data sets, *in* C. L. Mallows (ed.), *Design, Data, and Analysis, by some friends of Cuthbert Daniel*, Wiley, New York., pp. 93–128.

Hampel, F. (1995). Some additional notes on the "Princeton Robustness Year", Research Report 76, Seminar für Statistik, ETH Zürich. *To appear.*

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.

Hogg, R. V. (1991). Statistical education: Improvements are badly needed, *The American Statistician* **45**: 342–343.

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

Huber, P. J. (1975). Applications vs. abstraction: The selling out of mathematical statistics?, *Suppl. Adv. Appl. Prob.,* **7**: 84–89.

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

Huff, D. (1954). *How to Lie With Statistics*, Lowe and Brydone, London, UK.

Iman, R. L. (1994). The importance of undergraduate statistics, *Amstat News* **215**: 6.

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

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

Kotz, S. and Johnson, N. L. (1992). *Breakthroughs in Statistics, Vol. II: Methodology and Distribution*, Springer-Verlag, New York.

Künsch, H., Beran, J. and Hampel, F. (1993). Contrasts under long-range correlations, *The Annals of Statistics,* **21**(2): 943–964.

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

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*, Chapman and Hall, London, UK.

Merikallio, E. (1958). *Finnish Birds, Their Distribution and Numbers*, Societas Pro Fauna et Flora Fennica, Fauna Fennica V, Oy Tilgmann AB, Helsinki–Helsingfors.

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

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

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

Newcomb, S. (1895). Astronomical constants (the elements of the four inner planets and the fundamental constants of astronomy), *Supplement to the American Ephemeris and Nautical Almanac for 1897.*

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

Rowan-Robinson, M. (1985). *The Cosmological Distance Ladder: Distance and Time in the Universe*, Freeman, NY.

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

"Student" (1908). The probable error of a mean, *Biometrika* **6**: 1–25.

"Student" (1927). Errors of routine analysis, *Biometrika* **19**: 151–164.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Lehmann, E. L., Link, R. F., Pieters, R. S. and Rising, G. R. (eds) (1989). *Statistics: A Guide to the Unknown*, Wadsworth and Brooks/Cole, Pacific Grove, California.

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.

Wallis, W. A. and Roberts, H. V. (1956). *Statistics: A New Approach*, The Free Press, Glencoe, Illinois.

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