
DEPARTMENT OF STATISTICS

The University of Wisconsin
Madison, Wisconsin 53706

TECHNICAL REPORT NO. 319

December 1972

SIMON NEWCOMB, PERCY DANIELL, AND THE HISTORY OF ROBUST ESTIMATION 1885-1920

by

Stephen M. Stigler*
The University of Wisconsin, Madison**

Typist: Bernice R. Weitzel

*This research was partially supported by Research Grant No. NSF GP 32037x from the Division of Mathematical, Physical and Engineering Sciences of the National Science Foundation at the Department of Statistics, University of Chicago, and by the U.S. Navy through the Office of Naval Research under Contract No. ONR-N00014-67A-0128-0017 at the Department of Statistics, University of Wisconsin.

**On leave 1972-3 at the Department of Statistics, University of Chicago.

SIMON NEWCOMB, PERCY DANIELL, AND THE HISTORY OF ROBUST ESTIMATION 1885-1920.

by

Stephen M. Stigler[*]
The University of Wisconsin, Madison[**]

0. Introduction

In the eighteenth century, the word "robust" was used to refer to someone who was strong, yet boisterous, crude, and vulgar. By 1953 when Box first gave the word its statistical meaning, the evolution of language had eliminated the negative connotation: robust meant simply strong, hardy, healthy. The subject of robust inference, just like the word "robust", has a long and varied history. It is the aim of this present study to examine a part of this history and its relationship to current work.

The scope of this paper will be rather narrow - we shall only be concerned with the mathematical background and development of robust estimation up to 1920. Thus we shall be less concerned with the first appearances of estimators such as the median and trimmed mean than with

[*] This research was partially supported by Research Grant No. NSF GP 32037x from the Division of Mathematical, Physical and Engineering Sciences of the National Science Foundation at the Department of Statistics, University of Chicago, and by the U.S. Navy through the Office of Naval Research under Contract No. ONR-N00014-67A-0128-0017 at the Department of Statistics, University of Wisconsin.

[**] On leave 1972-3 at the Department of Statistics, University of Chicago.

the first mathematical analyses of their behavior and properties. The main emphasis will be on the period 1885-1920, and particular attention will be given to work which is not widely known, yet is relevant to modern lines of thought. Section two discusses the contributions of Simon Newcomb to robust estimation, and to the use of normal mixtures as models for heavy-tailed distributions; section three is concerned with the history of the mathematical analysis of order statistics in relation to robust estimation, with due attention to the works of Laplace, Sheppard, and Percy Daniell; and section four contains some brief remarks on "M-estimators".

The reader may be as surprised as the author was to find to what extent priorities in these areas have been misassigned. While many other points will be touched upon in the paper, our major findings are as follows: Laplace (1818) and Sheppard (1899) seem to have been the first to present a large sample theory for one or two order statistics. Simon Newcomb (1886) provided the first sound, modern approach to robust estimation, including the first use of mixtures of normal densities as representing heavy-tailed distributions. Percy Daniell (1920) should be credited with the first mathematical analysis of the class of estimators which are linear functions of order statistics, including the derivation of the optimal weighting functions for estimating scale and location parameters (the so-called "ideal" linear estimators) and the first mathematical treatment of the trimmed mean. Some of Newcomb's work has been commented upon recently by Huber (1972), but much of the remainder of the work discussed in this paper, including that due to Edgeworth, Galton, Laplace, Sheppard, and Daniell, has been largely ignored in recent years.

We shall begin with a brief overview of the situation prior to 1885.

1. The Situation before 1885.

Scientists have been concerned with what we would call "robustness" - sensitivity of procedures to departures from assumptions, particularly the assumption of normality - for as long as they have been employing well-defined procedures, perhaps longer. For example, in the first published work on least squares, Legendre (1805) explicitly provided for the rejection of outliers:

> "If among these errors are some which appear too large to be admissible, then those equations which produced these errors will be rejected, as coming from too faulty experiments, and the unknowns will be determined by means of other equations, which will then give much smaller errors".

Yet most of the early work in mathematical statistics was obsessed with "proving" the method of least squares, either starting with the assumption that the sample mean is the best estimate of the mean and deriving the normal distribution, as Gauss did in his first proof in 1809, or starting with the normal law or the central limit theorem, as did Laplace in 1812. The first mathematical work on robust estimation seems to have been that of Laplace (1818) on the distribution of the median. We shall defer a discussion of Laplace's work until section three, where it will be considered with later work on linear functions of order statistics.

The next statistical problem connected with robust estimation to receive mathematical treatment was the rejection of outliers. In 1852, the first proposal of a criterion for the determination of outliers was published by Benjamin Peirce, the Harvard mathematician-astronomer and father of logician-philosopher C. S. Peirce. Peirce's paper and most others on this subject* are not really about robust estimation, as their authors did not concern themselves with the properties of the resulting estimators; rather, they implicitly assumed that after the outlier test was performed the estimation could be done with no thought given to what had gone before, nor what information might be lost. This narrowness of view did not go unnoticed at the time. The first paper proposing an outlier criterion (Peirce, 1852) was soon followed by the first paper criticizing the use of outlier criteria (Airy, 1856). Airy, the Astronomer Royal, wrote:

> "And I have, not without surprize to myself, been led to think that the whole theory is defective in its foundation, and illusory in its results; that no rule for the exclusion of observations can be obtained by any process founded purely upon a consideration of the discordance of these observations".

A lively debate ensued, with the participants not always expressing themselves with Airy's restraint. For example, Glaisher (1872) wrote "Professor Pierce's [sic] criterion for the rejection of doubtful observations seems to me to be destitute of scientific precision".**

One of the more interesting papers of this time (and one of the most unusual statistical papers of all time) appeared in the Report of the Superintendent of the U.S. Coast Survey for 1870. It is by C. S. Peirce,

*See Anscombe (1960) and Rider (1933) for historical surveys of outlier techniques.

**At one point an exchange in print between the mathematician Glaisher and the astronomer Stone became so heated that one of Glaisher's papers was itself rejected by the Monthly Notices of the Royal Astronomical Society due to the personal nature of his comments; see Glaisher (1874).

written while he was an Assistant to the Coast Survey (at the time his father was Superintendent of the Survey!). In the paper Peirce presented the then standard material of the theory of errors, but in the language and notation which he had developed for the logic of relations, for which he later became famous. Thus we find, regarding averages,

> "Since [m] denotes all men, we may naturally write $\frac{[m]}{m}$ to denote what all men become when that factor is removed which makes [m] refer to *men* rather than to anything else; that is to say, to denote the *number* of men. We may write this for short **[m]** with heavy brackets. Then t being a relative term ("a tooth of,") by **[tl]** will be denoted the total number of teeth in the universe. But **[t]** will be used as equivalent to $\frac{[tl]}{[l]}$, or the average number of teeth that anything has."

Peirce included a sensible - one is tempted to say "logical" - defense of his father's outlier criterion in the paper (p. 210). By 1885 a number of rejection criteria were in use, often only by the proposer and his employees.

But techniques other than simply "reject outliers, then use the sample mean" were also employed. A variety of weighted means had been used prior to 1885. For example, in 1763 James Short (an English astronomer and noted manufacturer of telescopes) had estimated the sun's parallax based on observations of the transit of Venus of 1761 by averaging three means: the sample mean, the mean of all observations with residuals less than one second, and the mean of those with residuals less than half a second. The median and the midrange had appeared even earlier (Eisenhart (1971)).

By the last half of the nineteenth century, weighted least squares had become a standard topic in the literature of the theory of errors, and it was a frequent practice (at least in astronomical investigations) to weight observations differently, depending upon the statistician's (often subjective) estimate of the "probable error"* of the observation. The estimate of the probable error was supposed to be based solely on external evidence: scientists were warned of the possible biases if the magnitude of the observation were allowed to influence its weight (see Jevons (1874, p. 450), for example), but it is doubtful that this advice was faithfully adhered to. We shall discuss the use of these weighted means further in the next section, in connection with the contributions of Simon Newcomb.

Other estimators were proposed in this period. In particular, De Morgan (1847, p. 456) had outlined a scheme for discounting the more extreme observations. This method, more fully developed by Glaisher (1873), amounted to starting with the sample mean, then assigning different probable errors to the different observations based on the value of the likelihood function at those observations, and iterating this process. Glaisher's estimate was criticized by both Stone (1873) and Edgeworth (1883), who both (independently) proposed an alternative based on looking at a local maximum of the likelihood function (without assuming equal probable errors). Edgeworth later became disenchanted with this alternative (Edgeworth, 1887a).

At about this time, Francis Galton was making much use of the median (Galton, 1875), although his motivation was less suspicion of the normal distribution, which he considered a good representation of many real phenomena, than an appreciation of the simplicity, ease of calculation,

*The probable error of a symmetric distribution is half the interquartile range; for normal distributions p.e. = $(.6745)\sigma$.

and ease of interpretation of the median. Also, various formulae for index numbers were developed during this period; these included weighted averages and geometric means, each designed for a specific purpose.

However, it can still be said that by 1885, the conventional wisdom (but by no means the unanimous view) was that for purposes of estimation, the cautious use of the sample mean was recommended - sometimes weighted, sometimes after discarding outliers, but still the sample mean.

2. Simon Newcomb and mixtures of normal densities

1885 can be conveniently taken as the start of one of the most active and innovative periods in the history of mathematical statistics. The story of the development of mathematical statistics into a subject in its own right through the work of such men as Edgeworth, Karl Pearson, Gosset, and Fisher has been told by E. S. Pearson (1967). Our present, rather narrow purpose is to describe how the modern theory of robust estimation developed over this period. To this end, we shall place particular emphasis on the introduction of mixtures as models for the heavy-tailed distributions which scientists had encountered in practice, and on the use of linear functions of order statistics as robust estimators of location parameters.

Simon Newcomb appears to have been the first to introduce a mixture of normal densities as a model for a heavy-tailed distribution, and to exploit this model to get an estimator of location which was more robust than the sample mean. (Francis Galton and Karl Pearson had modeled measurements of natural populations by normal mixtures about the same time, but with a completely different object in mind, namely to demonstrate how a single population could be broken down into components.) While Newcomb's name may be unfamiliar to present day statisticians, it should not be so to astronomers, applied mathematicians, and economists.

Simon Newcomb (1835-1909) was born in Nova Scotia, attended Harvard, and spent most of his adult life (1861-1897) as a professor of mathematics in the U.S. Navy, working for the U.S. Nautical Almanac Office. He is generally regarded as the greatest American astronomer of the nineteenth century, and was responsible for many of the determinations of astronomical constants which are still accepted today. In addition, he was a powerful applied mathematician, co-founded and for many years edited the American Journal of Mathematics, and as an avocation wrote Principles of Political Economy (1885), a book which has established him as a major American economic theorist, and which contains one of the earliest modern mathematical statements of the quantity theory of money.

As was the practice in astronomy at the time, Newcomb made frequent use of weighted means in his estimation of astronomical constants. The relative weights were usually thought of in terms of "probable errors", and were assigned somewhat subjectively on the basis of Newcomb's judgment of the relative accuracy of the process which produced the observation. For example, after assessing some data on eclipses collected by Ptolemy in the second century A.D., he remarked (Newcomb, 1878, p. 41):

> "the [assigned] probable errors are the result of judgment from the terms of [Ptolemy's] description rather than of calculation; they were estimated without any knowledge of the way the comparison with theory would come out, and are printed without subsequent alteration".

With more contemporary data, Newcomb would base his choice of weights upon "the quality of the image and the generally satisfactory way in which the image was kept on the crosswires" (Newcomb, 1891a, p. 170) in the case of an experiment he was peronsally involved with, and upon the number of observers, general opinion of the reporting observatory, and "number and force of the doubtful circumstances "(Newcomb, 1891b, p. 383), in cases involving combination of other's measurements. He was apparently aware of criticism of the subjective nature of these assignments, but he maintained that

> "Opinions may doubtless differ as to whether a judicious system of weights has always been applied, but it is not likely that any unbiased reassignment would materially affect the result". (Newcomb, 1898, p. 211)

Newcomb also rejected outliers when necessary, but usually only based on external evidence or really huge deviations.

With this experience in dealing with observations made with differing degrees of precision, it is not surprising that, when faced with a collection of non-normal observations for which there was no satisfactory way to weight them individually, he should consider a mixture of normal densities with different variances as a model. For, having observed that a collection of 684 residuals based on observations of the transits of Mercury had much heavier tails than the corresponding normal distribution (even with excessive deviations ignored), he wrote (Newcomb, 1882, p. 382):

> "It is evident that if we have a collection of observations of different degrees of probable error, in which, however, there is no way of distinguishing those of great probable error from those of small probable error, the law of the errors will not be that usually adopted, but there will be a comparative excess of large residuals. It is also evident that in such a case the arithmetical mean does not necessarily give the most probable result. For, in the case of an observation of large residual, there is evidently a preponderance of probability that it belongs to a class with large probable error,

> and therefore should be assigned least weight. ... That any general collection of observations of transits of Mercury must be a mixture of observations with different probable errors was made evident to the writer by his observations of the transit of May 6, 1878, which may be here described as an illustration of the subject."

Four years after writing this, Newcomb published a remarkable paper in his own journal, the American Journal of Mathematics, in which he used this model to arrive at a more robust estimator of location than the sample mean. In this paper (Newcomb, 1886)[*], after criticizing the overuse of outlier criteria and presenting his mixture model, he proceeded to develop an estimator upon the principles of Bayesian decision theory that gave "less weight to the more discordant observations". Adopting squared error as a loss function (Newcomb's word for loss was "evil"), he demonstrated that in general the posterior mean minimizes the expected mean square error, and he suggested the following procedure. 1) Calculate the residuals based on the sample mean, and, using trial and error, fit a mixture of a finite number of normal densities with zero means to these residuals. 2) Take this fitted mixture and, considering the location family it generates, estimate the desired mean by the posterior mean with respect to a uniform prior given the original observations. Newcomb realized that this procedure presented practical difficulties and gave a number of simplifying approximations to arrive at a usable estimator. He illustrated its use with the data on the transits of Mercury.[**]

[*] Some of his arguments also appear in Newcomb (1895), p. 81-86.

[**] Ogorodnikoff (1928) provided a different simplification of Newcomb's estimator based on a Charlier expansion of the posterior distribution. The relationship between Newcomb's simplified estimator and the maximum likelihood estimator was discussed by Hulme and Symms (1939).

As an interesting sidelight, we note that in this paper and in a later work, Newcomb made an early use of a simple version of Tukey's sensitivity function (see Andrews et. al., 1972, p. 96). In Newcomb (1912, p. 212), discussing the unsatisfactory nature of outlier criteria, he wrote that if all observations with large residuals are rejected (and the mean estimated from the remaining observations), then the final result

> "becomes a discontinuous function of the residual of the rejected observation, the continuity being broken at the point regarded as the limit of normal error. A simple example will make the case clear. If we have three observed results a,b,c of which the mean is to be taken, and if c be the result which may be abnormal, then so long as c is retained we shall have
>
> $$\text{mean} = \frac{1}{3}(a + b + c);$$
>
> the mean will then continuously increase with c. When c passes the normal limit, the mean changes per saltum to
>
> $$\frac{1}{2}(a + b)".$$

In the same posthumous paper (Newcomb, 1912, p. 214), he also proposed a very simple estimator in the spirit of his 1886 paper: weight the observation X_i by $w_i = c/\max(|X_i - \bar{X}|, c)$, where c is a constant to be specified.

3. Laplace, Sheppard, Daniell, and linear functions of order statistics

With few exceptions, statisticians were quite late in coming to consider any but the simplest linear functions of order statistics as estimators of means. By a linear function of order statistics we shall mean any weighted linear combination of observations where the weights depend *only* on their order, not on their magnitudes or the size of their residuals. The median and the midrange, two members of this class, evidently have a long history (Eisenhart (1964), (1971)), but perhaps the first extensive mathematical analysis to be published involving order statistics was by Laplace. In the second supplement (1818) to his monumental _Théorie Analytique des Probabilités_, Laplace considered the problem we would now call linear regression through the origin: $a_i = p_i y + x_i$, a_i, p_i known, y to be estimated, where the errors x_i were assumed to have an arbitrary continuous, symmetric distribution. By looking for that estimator which minimized the sum of the absolute values of the residuals, he was led to consider an estimator of y which reduces to the median of the a_i's in the case $p_i \equiv 1$. Laplace derived the density of this estimator, showed that this density approaches the normal density as the sample size increases, and gave the necessary and sufficient condition on the error distribution that the median have a smaller asymptotic variance than the sample mean.* Laplace's proof is easily adopted to any sample percentile and asymmetrical populations, as was in fact later noted by Edgeworth (1885, 1886). In addition, Laplace derived the joint asymptotic density of the sample mean and median, and used it to find which linear combination of these estimators has the smallest asymptotic variance. (As the weights depend upon the unknown error distribution, he termed this result "impracticable", but noted that if the error distribution were normal, the best linear combination was the sample

*Laplace actually carried through his entire investigation in the more general regression situation, comparing the general estimator with the least squares estimator for this situation. For other views of Laplace's work and its historical context, see Eisenhart (1961) and Stigler (1972).

mean alone.)* Two years before Laplace's investigation, Gauss (1816), considering the problem of estimating the probable error of a normal distribution, had suggested the use of the median of the absolute values of the residuals, and stated (without proof) the asymptotic probable error of the median for this special case. Gauss apparently never published or circulated a proof, for 18 years later Encke (1834), who had corresponded with Gauss, felt it necessary to provide one, attributing it to Dirichlet. It seems likely that Dirichlet's proof for this special case was simply an adaptation of Laplace's, as Dirichlet was quite familiar with Laplace's work, the second supplement in particular (see Dirichlet (1836)).

Later in the nineteenth century, Galton (1875) and particularly Edgeworth (1885, 1887b, 1888), touted the use of the median in situations where heavier tails than the normal could be expected. Specifically, Edgeworth (1888) used Laplace's results to conclude that the median may well be better than the mean when the population distribution is one of Newcomb's mixtures of normal distributions. Also, Edgeworth (1886) seems to have been the first to realize that the median may possess an advantage over the sample mean for serially correlated data.

More complicated linear estimators began to appear in 1889, when Galton (in a footnote on p. 61-62 of Natural Inheritance) suggested estimating the mean and standard deviation of a normal distribution by what amounts to taking

*The possibility of a linear function of two estimators outperforming both has been more fully exploited in the recent Princeton robustness study (see Andrews et. al (1972) p. 132).

$$\hat{\mu} = \frac{\xi_p X^{(nq)} - \xi_q X^{(np)}}{\xi_p - \xi_q} ,$$

$$\hat{\sigma} = \frac{X^{(np)} - X^{(nq)}}{\xi_p - \xi_q} ,$$

where ξ_p and ξ_q are the p and q percentiles of the standard normal distribution, $X^{(np)}$ and $X^{(nq)}$ are the sample p and q percentiles, and p and q are arbitrary but fixed $(0 < p < q < 1)$. In 1899 in a long paper on the multivariate normal distribution and its applications, Sheppard proved the joint asymptotic normality of Galton's estimators when the population is normal. He also showed the joint asymptotic normality of $X^{(np)}$ and $X^{(nq)}$, and gave analogues to $\hat{\mu}$ and $\hat{\sigma}$ based on any finite number of sample percentiles (Sheppard, 1899, p. 131-132). Sheppard's (sketchy) proof, which is based on an implicit use of the probability integral transformation, can be easily adapted to any regular distribution.*

Sheppard's paper also represented the first attempt since Laplace to optimize performance within a class of linear functions of order statistics. He both showed how the best choice (for normal populations) of p and q can be made (1899, p. 135) and found which linear combination of the three quartiles has the smallest asymptotic variance (again for normal populations) (1899, footnote, p. 134). Such functions

*Twenty years later, Karl Pearson (1920) presented part of Sheppard's proof in more detail, made the obvious step to more general distributions than the normal, and much more fully examined the consequences of the result.

of the three quartiles had been considered earlier by Edgeworth (1893), who neglected the quartiles' correlation and erroneously claimed the estimator with weights in proportions 5:7:5 to be superior to the sample mean for normal populations. Recent work, however, seems to bear out Edgeworth's claim that such an estimator is to be recommended on grounds of robustness. (See Gastwirth (1966) and Andrews et. al. (1972), for example.)

The next mathematical work to appear on order statistics was Karl Pearson's (1902) examination of the Galton difference problem. In this paper, which was inspired by an inquiry of Galton's (1902) as to the most suitable proportion between the values of first and second prizes, Pearson gave the joint density of any two consecutive order statistics and found their expected difference. He remarked in a footnote that

> "I propose on another occasion to consider the application of Galton's problem to a new theory for the rejection of outlying individuals".

This proposal was later carried out by J. O. Irwin (1925).

In 1920, a remarkable paper appeared in the American Journal of Mathematics (the journal Simon Newcomb co-founded) by the English mathematician P. J. Daniell. This paper, "Observations weighted according to order", has been all but totally overlooked since it's publication. It could in fact be claimed that Daniell was at least thirty years ahead of his time, for it took that long for his major results to be rediscovered. While his paper itself is a model of clarity and rigor, its relevance to modern work is such that it merits a short summary, in his own notation.

The work was apparently inspired by a reading of Poincaré's Calcul des Probabilités (1912). After remarking how Poincaré had suggested discarding extreme observations (when normality is suspect) before taking the mean, Daniell wrote:

> "Besides such a discard-average [ie. the trimmed mean] we might invent others in which weights might be assigned to the measures according to their order. In fact the ordinary average or mean, the median, the discard-average, the numerical deviation (from the median, which makes it a minimum), and the quartile deviation can all be regarded as calculated by a process in which the measures are multiplied by factors which are functions of order. It is the general purpose of this paper to obtain a formula for the mean square deviation of any such expression. This formula may then be used to measure the relative accuracies of all such expressions".

Daniell's analysis proceeded as follows: First he explicitly introduced the probability integral transformation (apparently the first time this was done*) and explained how it can be used to find the moments of any function of order statistics. Then, he assumed the population density $p(t)$ was regular (and indefinitely differentiable), and he expanded the inverse of the distribution function in a Taylor series to derive asymptotic expressions for the mean of an order statistic t_r and the mean product of any two. He thus duplicated some of Sheppard's (1899) results, but in a much more rigorous manner.

Daniell then considered statistics of the form $\bar{t} = \sum_{r=1}^{n} f_r t_r$, where he assumed that the weight f_r associated with the r^{th} order statistic t_r was given by

$$f_r = \frac{1}{n} f(\frac{r}{n+1}),$$

*The next being Karl Pearson (1931).

and put things together to obtain the (now standard*) expression for the asymptotic variance of $\bar{t}$,

$$S^2 = \int_{-\infty}^{\infty} \phi^2(t)\, p(t)dt,$$

where $\phi(t)$ is the indefinite integral of $f(x(t))$, $x(t) = \int_{-\infty}^{t} p(u)du$. If he was less than specific as to why the remainder terms are uniformly negligible, his standard of rigor was nonetheless far above that of the statistical literature of the time.

In the third section of the paper, Daniell gave the conditions on f under which the asymptotic mean of $\bar{t}$ is the population mean or standard deviation, and defined the "accuracy" of $\bar{t}$ as the ratio of the asymptotic variance of the sample mean (or sample standard deviation, as the case may be) to that of $\bar{t}$. (He also derived the asymptotic variance of the sample standard deviation here.) In the fourth section, Daniell gave the optimal weight function f - that which minimizes S^2 - for both the location and scale cases, using standard results from the calculus of variations, and noted that the optimal estimate of σ for the normal case is as accurate as the sample standard deviation in this case. These results were not to appear in print again until Jung (1955), although they are in Bennett's (1952) unpublished thesis.

The final two sections were concerned with applications. Daniell gave special attention to the "discard-average" (the trimmed mean),

*See Chernoff, Gastwirth, and Johns (1967).

presenting the (now standard) expression for its asymptotic variance and evaluating its performance for various Pearson densities, including Student's t. He also gave conditions under which the quartile-discard average is superior to the sample mean. The paper ended with a number of applications to other estimators of location and scale*, with numerical results. Daniell did not derive the asymptotic normality of $\bar{t}$, nor did he try to state minimal regularity conditions (indeed, some of his regularity conditions were implied rather than stated). However, taken altogether it is a thoroughly modern paper which almost appears to have been gleaned from the literature of the 1950's and 1960's.

How could such a paper have gone unnoticed for all these years?** To see why, we need to learn something of Daniell's life. Percy John Daniell (1889-1946) received a B.A. degree at Cambridge in 1910 (and an M.A. in 1914), where his honors included Senior Wrangler in Mathematics (1909), First Class Physics Tripos (1910), and the Raleigh Prize (1912). His stay at Cambridge would have overlapped R. A. Fisher's, but they were at different colleges and may not have met. After graduation (and brief stays at Göttingen and Liverpool), Daniell went to Rice Institute in Houston, Texas in 1912 as a travelling fellow. He remained at Rice until 1923, becoming a full professor in 1920. It was at Rice he did his most important work, principally on the theory of integration (including the development of what is now known as the Daniell integral.) In 1924 he returned to

*Including the "discard-deviation", where the <u>inner</u> quartiles are discarded.

**A fairly complete review of the literature reveals only two published citations, Dodd (1922) and Greenberg (1968), and the descriptions there are superficial and misleading. Daniell's paper came to my attention as the result of a systematic inspection of the <u>American Journal of Mathematics</u>.

England to the University of Sheffield, where he remained until his death at the age of 57. In the latter part of his life he published occasional papers on applied mathematics, on such topics as flame motion, potentials, and quadrature formulae.

The paper, Daniell (1920), written at Rice, seems to have been his only related work in statistics. This fact, together with his isolation from active statistical research (both at Rice and Sheffield), was largely responsible for the obscurity of the paper. Daniell's death before his results were rediscovered and widely discussed, and Wilks' overlooking his work in the survey paper of 1948 also served to delay recognition of his priority. As a further irony, these circumstances have helped relegate to obscurity another important paper of Daniell's, "Integral products and probability" (1921), in which he presents one of the earliest mathematical treatments of continuous time Markov processes, including the Chapman-Kolmogorov equation (ten years before Kolmogorov) and a short treatment of the Wiener process (two years before Wiener).

4. M-estimators

Recently, much attention has been given to a class of robust estimators which Huber calls "M-estimators", M for maximum-likelihood type. (See Huber, 1972). T is said to be an M-estimator corresponding to a function ϕ if T is a solution to $\sum\phi(X_i - T) = 0$. Each choice of ϕ determines an estimator; if $\phi = p'/p$, T is the maximum likelihood estimator for the location parameter of the population with density $p(t - \theta)$.

As the first appearance of these estimators in the context of robustness seems to be in the work of Jeffreys after 1920 (see Jeffreys (1932) and (1939) in particular), and as this work is outside the scope of this study, we shall not dwell on this subject. However, we cannot resist calling attention to an early reference in which the class of M-estimators is introduced and their consistency claimed.

In a paper examining the various "proofs" of the method of least squares, Ellis (1844) began with Gauss's first proof. Letting x_i's denote observed values, a the quantity to be estimated, and $e_i = x_i - a$, Ellis questions Gauss's a priori designation of the arithmetic mean (the solution to $\sum(x_i - a) = 0$) as the most probable value.

> "It [the arithmetic mean] is not the only rule to which these considerations might lead us. For not only is $\sum e = 0$ ultimately, but $\sum fe = 0$, where fe is any function such that $fe = -f(-e)$; and therefore we should have
>
> $$\sum f(x-a) = 0,$$
>
> as an equation which ultimately would give the true value of x when the number of observations increases sine limite, and which therefore for a finite number of observations may be looked on in precisely the same way as the equation which expresses the rule of the arithmetic mean. There is no discrepancy between these two results. At the limit they coincide: short of the limit both are approximations to the truth. Indeed we might form some idea how far the action of fortuitous causes had disappeared from a given series of observations by assigning different forms of f, and comparing the different values thus found for a.
>
> "No satisfactory reason can be assigned why, setting aside mere convenience, the rule of the arithmetic mean should be singled out from other rules which are included in the general equation $\sum f(x-a) = 0$".

Thus Ellis has claimed (without proof or regularity conditions) the consistency of M-estimators*, and even suggested the class may be useful for judging to what degree an estimated value depends on the choice of estimator, a stability test. Of course Ellis was not really concerned with robustness, only with illuminating the arbitrary nature of Gauss's proof, but his comments are of interest nonetheless.

A Note on the References

In addition to those works cited below, many other works were consulted for references and general information. The information on the life of Simon Newcomb came principally from the _Encyclopedia Britannica_, the _International Encyclopedia of the Social Sciences_, and Newcomb's autobiography (1903). The information on the life of Percy Daniell came from various editions of _Who's Who_ and _American Men of Science_, and Stewart (1947). Merriman's (1872) bibliography on least squares was quite useful for the period prior to 1877. I would also like to thank William Kruskal, Churchill Eisenhart, and Oscar B. Sheynin for a number of references and helpful comments. A good bibliography of work since 1920 can be found in H. A. David's _Order Statistics_ (1970).

[*] Huber proved this in Huber (1964).

References

[1] Airy, G. B. (1856). Letter from Professor Airy, Astronomer Royal, to the editor, Astronomical Journal 4, 137-138.

[2] 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: Princeton University Press.

[3] Anscombe, F. J. (1960). Rejection of outliers. Technometrics 2, 123-147.

[4] Bennett, C. A. (1952). Asymptotic properties of ideal linear estimators. Unpublished dissertation, University of Michigan.

[5] Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika 40, 318-335.

[6] Chernoff, H., Gastwirth, J., and Johns, M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. Annals of Mathematical Statistics 38, 52-72.

[7] Daniell, P. J. (1920). Observations weighted according to order. American Journal of Mathematics 42, 222-236.

[8] Daniell, P. J. (1921). Integral products and probability. American Journal of Mathematics 43, 143-162.

[9] David, H. A. (1970). Order Statistics. New York: Wiley.

[10] De Morgan, A. (1847). Theory of Probabilities. Encyclopedia of Pure Mathematics (Part of Encyclopedia Metropolitana), 393-490.

[11] Dirichlet, G. L. (1836). Ueber die Methode der kleinsten Quadrate. In G. Lejeune Dirichlet's Werke, Vol. I, 281-282. Berlin: Reimer (1889).

[12] Dodd, E. L. (1922). Functions of measurements under general laws of error. Skandinavisk Aktuarietidskrift 5, 133-158.

[13] Edgeworth, F. Y. (1883). The method of least squares. Philosophical Magazine 16 (Fifth Series), 360-375.

[14] Edgeworth, F. Y. (1885). Observations and statistics. An essay on the theory of errors of observation and the first principles of statistics. Transactions of the Cambridge Philosophical Society 14, 138-169.

[15] Edgeworth, F. Y. (1886). Problems in probabilities. Philosophical Magazine 22 (Fifth Series), 371-384.

[16] Edgeworth, F. Y. (1887a). On discordant observations. Philosophical Magazine 23 (Fifth Series), 364-375.

[17] Edgeworth, F. Y. (1887b). The choice of means. Philosophical Magazine 24 (Fifth Series), 268-271.

[18] Edgeworth, F. Y. (1888). On a new method of reducing observations relating to several quantities. Philosophical Magazine 25 (Fifth Series), 184-191.

[19] Edgeworth, F. Y. (1893). Exercises in the calculation of errors. Philosophical Magazine 36 (Fifth Series), 98-111.

[20] Eisenhart, C. (1961). Boscovich and the combination of observations. Chapter 7 in R. J. Boscovich Studies of His Life and Work, (ed. L. L. Whyte). London: Allen and Unwin. (reissued 1963 by Fordam University Press, New York)

[21] Eisenhart, C. (1964). The meaning of "least" in least squares. Journal of the Washington Academy of Sciences 54, 24-33.

[22] Eisenhart, C. (1971). The development of the concept of the best mean of a set of measurements from antiquity to the present day. 1971 A.S.A. Presidential Address.

[23] Ellis, R. L. (1844). On the method of least squares. Transactions of the Cambridge Philosophical Society 8, 204-219.

[24] Encke, J. F. (1834). On the method of least squares. Translated from the German in Scientific Memoirs, Selected from the Transactions of Foreign Academies of Science and Learned Societies and from Foreign Journals (Ed. R. Taylor). Vol. II (1841).

[25] Galton, F. (1875). Statistics by intercomparison, with remarks on the law of frequency of error. Philosophical Magazine 49 (Fourth Series), 33-46.

[26] Galton, F. (1889). Natural Inheritance. London: Macmillan.

[27] Galton, F. (1902). The most suitable proportion between the values of first and second prizes. Biometrika 1, 385-390.

[28] Gastwirth, J. (1966). On robust procedures. Journal of the American Statistical Association 61, 929-948.

[29] Gauss, C. F. (1816). Bestimmung der Genauigkeit der Beobachtungen. In Carl Friedrich Gauss Werke, Band 4, 109-117, Göttingen: Königlichen Gesellschaft der Wissenschaften (1880).

[30] Glaisher, J. W. L. (1872). On the law of facility of errors of observations, and on the method of least squares. *Memoirs of the Royal Astronomical Society* **39** (Part II), 75-124.

[31] Glaisher, J. W. L. (1873). On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society* **33**, 391-402.

[32] Glaisher, J. W. L. (1874). Note on a paper by Mr. Stone, "On the rejection of discordant observations". *Monthly Notices of the Royal Astronomical Society* **34**, 251.

[33] Greenberg, B. G. (1968). Nonparametric Statistics: Order Statistics. Article in the *International Encyclopedia of the Social Sciences*, New York: The Macmillan Company and the Free Press.

[34] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73-101.

[35] Huber, P. J. (1972). Robust statistics: a review. *Annals of Mathematical Statistics* **43**, 1041-1067.

[36] Hulme, H. R. and Symms, L. S. T. (1939). The law of error and the combination of observations. *Monthly Notices of the Royal Astronomical Society* **99**, 642-649.

[37] Irwin, J. O. (1925). On a criterion for the rejection of outlying observations. *Biometrika* **17**, 238-250.

[38] Jeffreys, H. (1932). An alternative to the rejection of observations. *Proceedings of the Royal Society*, Series A, **137**, 78-87.

[39] Jeffreys, H. (1939). The law of error and the combination of observations. *Philosophical Transactions of the Royal Society of London*, Series A, **237**, 231-271.

[40] Jevons, W. S. (1874). *The Principles of Science* **I**. London: Macmillan.

[41] Jung, J. (1955). On linear estimates defined by a continuous weight function. *Arkiv For Matematik* Band 3 nr 15, 199-209.

[42] Laplace, P. S. de (1818). *Deuxieme Supplement a la Théorie Analytique des Probabilités*. Paris: Courcier. (Pp. 569-623 in *Oeuvres de Laplace* **7**, Paris: Imprimerie Royale (1847); pp. 531-580 in *Oeuvres Complétes de Laplace* **7**, Paris: Gauthier-Villars (1886).)

[43] Legendre, A. M. (1805). On the method of least squares. In A Source Book in Mathematics, p. 576-579. New York: Dover.

[44] Merriman, M. (1872). A list of writings relating to the method of least squares, with historical and critical notes. Transactions of the Connecticut Academy of Arts and Sciences 4, 151-232.

[45] Newcomb, S. (1878). Researches on the motion of the moon, I. Washington Observations for 1875 - Appendix II (published by the U.S. Naval Observatory, Washington.)

[46] Newcomb, S. (1882). Discussion and results of observations on transits of Mercury from 1677 to 1881. Astronomical Papers 1, 363-487.

[47] Newcomb, S. (1885). Principles of Political Economy. New York: Harper and Brothers.

[48] Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics 8, 343-366.

[49] Newcomb, S. (1891a). Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880 to 1882. Astronomical Papers 2, 107-230.

[50] Newcomb, S. (1891b). Discussion of observations of the transits of Venus in 1761 and 1769. Astronomical Papers 2, 259-405.

[51] Newcomb, S. (1895). The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy. (Supplement to the American Ephemeris and Nautical Almanac for 1897.) Washington: Government Printing Office.

[52] Newcomb, S. (1898). Catalogue of fundamental stars for the epochs 1875 and 1900 reduced to an absolute system. Astronomical Papers 8, 77-403.

[53] Newcomb, S. (1903). The Reminiscences of an Astronomer. Boston: Houghton Mifflin.

[54] Newcomb, S. (1912). Researches on the motion of the moon, II. Astronomical Papers 9, 1-249.

[55] Og[o]rodnikoff, K. (1928). On the occurrence of discordant observations and a new method of treating them. Monthly Notices of the Royal Astronomical Society 88, 523-532.

[56] Pearson, E. S. (1967). Studies in the history of probability and statistics XVII. Some reflexions on continuity in the development of mathematical statistics, 1885-1920. Biometrika 54, 341-355.

[57] Pearson, K. (1902). Note on Francis Galton's Problem. Biometrika 1, 390-399.

[58] Pearson, K. (1920). On the probable errors of frequency constants, III. Biometrika 13, 113-132.

[59] Pearson, K., with Pearson, M. V. (1931). On the mean character and variance of a ranked individual, and on the mean and variance of the intervals, between ranked individuals, I: Symmetrical distributions (normal and rectangular). Biometrika 23, 364-397.

[60] Peirce, B. (1852). Criterion for the rejection of doubtful observations. Astronomical Journal 2, 161-163.

[61] Peirce, C. S. (1873). On the theory of errors of observations. Report of the Superintendent of the United States Coast Survey (for 1870), 200-224.

[62] Poincaré, H. (1912). Calcul des Probabilités, Paris: Gauthier-Villars.

[63] Rider, P. R. (1933). Criterion for rejection of observations. Washington University Studies - New Series, Science and Technology, No. 8.

[64] Sheppard, W. F. (1899). On the application of the theory of error to cases of normal distribution and normal correlation. Philosophical Transactions of the Royal Society of London (Series A) 192, 101-167.

[65] Short, J. (1763). Second paper concerning the parallax of the sun determined from the observations of the late transit of Venus; in which this subject is treated of more at length, and the quantity of the parallax more fully ascertained. Philosophical Transactions of the Royal Society of London 53, 300-345.

[66] Stewart, C. A. (1947). P. J. Daniell. Journal of the London Mathematical Society 22, 75-80.

[67] Stigler, S. M. (1972). Laplace, Fisher, and the discovery of the concept of sufficiency.

[68] Stone, E. J. (1873). On the rejection of discordant observations. Monthly Notices of the Royal Astronomical Society 34, 9-15.

[69] Wilks, S. S. (1943). Order statistics. Bulletin of the American Mathematical Society 5, 6-50.