The History of Robust Estimation at the Turn of the 19th and 20th Century

H. Striteska

Masaryk University, Faculty of Science, Brno, Czech Republic.

Abstract. This article reviews some of the history of robust estimation in the end of the 19th century and the beginning of the 20th century. Special attention is given to mixtures of normal densities, linear functions of order statistics and M-estimators.

Introduction

In the 18th century the word "robust" was used to refer for someone who was strong, but crude and vulgar. In 1953 G. E. P. Box first gave the word its statistical meaning and the word lost his negative connotation. Today there are various definitions for this term, but in general, it is used in the sense "insensitive to small departures from the idealized assumptions for which the estimator is optimized". [3]

In the 19th century, scientists have been concerned with "robustness" in the sense of insensitivity of procedures to departures from assumptions, especially the assumption of normality. For example, A. M. Legendre (1752-1833) explicitly used the rejection of outliers in his first published work on least squares in 1805.

Among the longest used robust estimations belong the median and the interquartile range. In 1818 P. S. Laplace published his mathematical work on robust estimation on the distribution of the median. Laplace (1749-1827) compares the large-sample densities and he shows that the median is superior to the mean. The interquartile range appears in works of men such as C. F. Gauss (1816) or the Belgian "father of biometry" A. Quetelet (1846). Direct consideration of the interquartile range and the interdecile range began with Galton (1882). He arranged the observations in increasing order of magnitude and then took out the required fraction from each end. He used interpolation to obtain the exact cut-off points.

Situation in the second half of the 19th century

One of the statistical problems connected with robust estimation was the rejection of outliers. In 1852 B. Peirce (1809-1880) published the first proposal of the determination of outliers. Peirce`s study (and most others on this subject) were not really about robust estimation. They did not concern themselves with the properties of the final estimators. They implicitly assumed, that after rejection of outliers the estimation could be done regardless of what information might be lost. Soon British astronomer G. B. Airy (1801-1892) criticized this narrowness in his paper (1856). A lively debate followed with participants such as J. W. L. Glaisher or Peirce`s son.

The year 1885 can be taken as the start of very innovative period in the history of mathematical statistics, due to such men as K. Pearson (1857-1936), F. Y. Edgeworth (1845-1926) [6].

Before 1885 also other techniques than simply "reject outliers and then count sample mean" were employed. For example weighted means were used. In 1763 J. Short (American astronomer and famous manufacturer of telescopes, 1710-1768) estimated the sun's parallax by observations of the transit of Venus in 1761. He averaged three means: the sample mean, the mean of all observations with residuals less than one second, and the mean of observations with residuals less than half a second.

In the second half of the 19th century weighted least squares were the standard topic in the theory of errors. It was a common routine to weight astronomical observations differently. It depended on the astronomer's estimate of the probable error of the observation.

In 1847 De Morgan (1806-1871) proposed a scheme for discounting more extreme observations. This was further developed by Glaisher (1873). The procedure began with the sample mean. Then it assigned to the observations different probable errors, based on the value of the likelihood function at these observations, and iterated this process.

Around this period F. Galton (1822-1911) published a various use of median (1875). His motivation was rather than a mistrust of normal distribution, an easiness of calculation and easiness of interpretation of median. A lot of similar features we can find in the independent work of G. T. Fechner (1878, 1801-1887).

Mixtures of normal densities

Simon Newcomb seems to be the first man, who has introduced a mixture of normal densities as a model for a heavy-tailed distribution. He has employed this model to get an estimator of location which was more robust than the sample mean. At this time also Francis Galton and Karl Pearson have used normal mixtures, but for entirely different reasons – for a demonstration how a single population can be splitted into components.

Newcomb (1835-1901) is the best known American astronomer of the 19th century. He has determined many of astronomical constants, which are still accepted. He was also talented applied mathematician and co-founder of American Journal of Mathematics. Due to his book Principles of Political Economy he has become a major American economic theorist.

Newcomb has also frequently used weighted means in his estimation of astronomical constants. He usually thought the weights in terms of "probable errors". Newcomb determined weights subjectively on the basis of his own assessment of the relative correctness of the process of observation. Newcomb also rejected outliers if necessary, but usually only in cases of enormous deviations or on the basis of an external evidence.

During the observations of the transits of Mercury (6.5.1878) he recognized that a set of 684 residuals had much heavier tails than the corresponding normal distribution. He knew that he could not distinguish the observations with large probable errors from those with small probable errors. He wrote:

"It is also evident that in such a case the arithmetical mean does not necessarily give the most probable result. … 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."

In 1886 Newcomb published a significant paper in his own journal American Journal of Mathematics, where he used this model to reach more robust estimator of location than the sample mean. In this paper and in his later work Newcomb made an early use of a simple version of Tukey's sensitivity function.

L-estimators

L-estimators are based on the ordered observations (order statistics). Two members of this class, the median and the midrange, indeed have a long history.

P. S. Laplace seems to be the first who published (1818) the first extensive mathematical analysis involving order statistics. Laplace compared the large-sample behaviour of L-estimators of type I and II. Among others he considered as a special case the estimation of the centre of a symmetric distribution by the median and by the mean. He showed that the median is superior to the mean.

F. Galton (1875) and F. Y. Edgeworth (1885) proposed using the median in the cases where heavier tails than normal could be expected.

In 1889 Galton suggested a more complicated linear estimator of the mean and of the standard deviation of a normal distribution in the form:

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

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

where p, q are arbitrary, but fixed ($0 < p < q < 1$), ξ_p a ξ_q are p a q percentiles of a standard normal distribution and $X^{(np)}$ a $X^{(nq)}$ are $100p$ and $100q$ sample percentiles. In 1899 Sheppard published the proof of the joint asymptotic normality of Galton's estimators for normal population.

In 1920 P. J. Daniell (1889-1946) published in American Journal of Mathematics the article "Observations Weighted According to Order". But this paper was absolutely ignored. It took thirty years until his results were rediscovered. In his article Daniell studied the asymptotic estimation of both location and scale parameters by general linear functions of the order statistics.

This study was inspired by H. Poincare and his Calcul des Probabilités (1912). Especially concerning the put-off extreme observations before calculation of the mean. Daniell wrote to it [1]: "Besides such a discard-average we 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 minimum), and the quartile deviation can all be regarded as calculated by a process in 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 seems to be the first who introduced the probability integral transformation and used it to find the moments of a function of order statistics. In this study he among others derived the optimal weighted function that minimizes S^2 for estimating both scale and location parameters.

Daniell wrote this article in Rice Institute in Houston, Texas. Daniell's most important works are about theory of integration (Daniell integral). Maybe his isolation from the active statistical research and the fact, that his study from 1920 seems to be only related work to statistics, are responsible for an oversight of the article.

M-estimators

A class of M-estimator introduced P. J. Huber in his work "Robust estimation of a location parameter" in 1964. "M" means maximum likelihood type.

The first appearance of these estimators seems to be in the work of H. Jeffreys in 1932. But much earlier R. L. Ellis has claimed the consistency of M-estimator (without proof or regularity conditions) [4]. He even suggested a stability test, which may be useful for judging to what degree an estimated value depends on the choice of estimator. Ellis came to this at examination the various "proofs" of the method of least squares (1844).

In 1888 a professor of engineering R. H. Smith sent a letter to Nature with a proposal, which includes weight function very similar to Tukey biweight function. Among others A. E. Beaton & J. W. Tukey (1974) a F. Mosteller & J. W. Tukey (1977) considered combining independent measurements $X_1, ..., X_n$ into a single estimate T of a location parameter by solving for T in the relationship

$$(1)\ T = \frac{\sum_i X_i w\left(\left(X_i - T\right)/s\right)}{\sum_i w\left(\left(X_i - T\right)/s\right)}$$

where the summations are over $i = 1, \ldots, n$, s is an estimate of scale based upon $X_1, ..., X_n$ and $w(u)$ is a weight function:

$$(2)\ w(u) = \begin{cases} \left(1-u^2\right)^2 & |u| \le 1 \\ 0 & |u| > 1 \end{cases}$$

They called this function "biweight". This function is very similar to the weight function in Smith's proposal. Smith assumed that data passed through the preliminary screening. In this screening, an investigator fixed lower limit and upper limit. Then all measurements outside these limits were rejected. But Smith didn't specify an exact process of determination these limits.

In modern notation, that used S. M. Stigler in [8], we can write Smith's weight function in the form:

$$(3)\ w(u) = \begin{cases} 1-u^2 & |u| \le 1 \\ 0 & |u| > 1 \end{cases}$$

Smith's weight function is the square root of Tukey biweight function. Smith proposed a closed form noniterative solution for his estimate T:

(4) $T = \frac{3}{4}\frac{\sum x^2}{\sum x}\left\{1+\sqrt{1-\frac{8}{9}\frac{\sum x \sum x^3}{\left(\sum x^2\right)^2}}\right\}$

Smith didn't provide any analysis of properties of this proposal. He probably assumed, that the population was symmetrical.

Much earlier in 1785 D. Bernoulli occupied himself with a similar estimate. In his article he gave a geometric description of an estimate much like Smith's, containing a weight function $w(u) = (1-u^2)^{1/2}$. Bernoulli suggested an iterative solution. In addition he was more inexplicit than Smith in his specification of the scale used.

Conclusion

We can say that till 1885 a traditional approach was a provident use of the sample mean – sometimes weighted mean, sometimes after put-off outliers, but still the sample mean. Major events from the next period that I deal with in my article are as follows:

In 1886 S. Newcomb came up with modern approach to robust estimation. As the first he used mixtures of normal densities for heavy-tailed distributions. P. Daniell (1920) seems to be the first, who analyzed the class of estimators which are linear functions of order statistics. He also derived the weighted functions for estimating scale and location parameters. M-estimators were introduced as late as 1964, but back in 1888 R. H. Smith suggested a weight function very similar to Tukey biweight function and noniterative evaluation of the estimate.

References

[1] Daniell, P. J., Observations Weighted According to Order, American Journal of Mathematics, 42, 222-36, 1920.
[2] David, H. A., Early Sample Measures of Variability, Statistical Science, Vol. 13, No. 4, 368-377, 1998.
[3] Huber, P. J., Robust Estimation of a Location Parameter, Annals of Mathematical Statistics, 35, No. 1, 73-101, 1964.
[4] Huber, P. J., Robust Statistics, New York, Wiley, 1981.
[5] Jurečková, J., Picek, J., Robust Statistical Methods with R, Chapman & Hall/CRC, Boca Raton, 2006.
[6] Pearson, E. S., Studies in the History of Probability and Statistics XVII: Some Reflections on Continuity in the Development of Mathematical Statistics, 1885-1920, Biometrika, 54, No. 2, 341-355, 1967.
[7] Stigler, S., Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920, Journal of the American Statistical Association, Vol. 68, No. 344, 872-879, 1973.
[8] Stigler, S., Studies in the History of Probability and Statistics XXXVIII, R. H. Smith, a Victorian interested in robustness, Biometrika, 67, 1980.
[9] Tukey, J. W., Beaton, A. E., The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, Technometrics 16, 147-85, 1974.
[10] Tukey, J. W., Mosteller, F., Data Analysis and Regression. Reading, Mass: Addison-Wesley, 1977.