Biometrika (1973), **60**, 3, *p.* 439
Printed in Great Britain

Studies in the History of Probability and Statistics. XXXII

Laplace, Fisher, and the discovery of the concept of sufficiency

BY STEPHEN M. STIGLER
University of Wisconsin, Madison

SUMMARY

R. A. Fisher's 1920 discovery of sufficiency while studying competing estimators of standard deviations is discussed. It is shown how Laplace completed a similar investigation a century earlier attempting to improve upon least squares, yet did not abstract the concept.

Some key words: Laplace; Fisher; Sufficiency; History of statistics; History of estimation; Least squares.

1. INTRODUCTION

Because Fisher's concept of sufficiency depends so strongly on the assumed form of the population distribution, its importance to applied statistics has been questioned in recent years. There can be no question, however, as to its importance to theoretical statistics. Sufficiency served as a cornerstone to Fisher's theory of estimation, and has helped illuminate topics as diverse as the optimal design of experiments and Bayesian inference. That the concept did not appear before Fisher's time is not surprising. The very definition of sufficiency requires the notion of the joint distribution of a statistic and the sample, and many statisticians before Fisher were not mathematically able to handle the necessary complicated operations in multiple dimensions, much less abstract the concept. Accordingly, it may come as a surprise to some just how close Laplace came to discovering sufficiency in 1818, following a chain of reasoning remarkably similar to that used by Fisher in 1920.

The aim of this present paper is to examine the relevant works of Fisher and Laplace and show how they set out on parallel routes: Laplace essentially comparing the sample mean and median as estimators of the centre of a symmetrical population, Fisher comparing the sample standard deviation and mean deviation as estimators of the standard deviation of a normal population. Yet we shall see how Fisher continued slightly beyond the point Laplace stopped, and how he was then led to the modern concept of sufficiency.

2. FISHER'S DISCOVERY OF SUFFICIENCY

R. A. Fisher first hit upon the concept of sufficiency in 1920, although he did not use the name 'sufficiency' until the following year. The paper, 'A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error' (Fisher, 1920), was written in answer to a statement of the astronomer A. S. Eddington in his book *Stellar Movements*. Eddington had written:

> ...in calculating the mean error of a series of observations it is preferable to use the simple mean residual irrespective of sign rather than the mean square residual,

to which Eddington added a footnote:

> This is contrary to the advice of most text-books; but it can be shown to be true [Eddington, 1914, p. 147].

In his brief paper, Fisher examined Eddington's claim. We shall outline the steps of his argument.

Fisher began by assuming that each of n measurements is normally distributed with mean m and variance σ^2, and that the object was to estimate σ. He considered in particular the estimators σ_1, the mean deviation from the sample mean, suitably scaled, and σ_2, the sample standard deviation. He derived the standard deviations of both σ_1 and σ_2, and observed that, contrary to Eddington's statement, the standard deviation of σ_1 is 14% greater than that of σ_2 for large n. At this point he permitted Eddington to add a footnote, which read in part:

> I think it accords with the general experience of astronomers that, for the errors commonly occurring in practice, the mean error is a safer criterion of accuracy than the mean square error, especially if any doubtful observations have been rejected; but I was wrong in claiming a theoretical advantage for the mean error in the case of a truly Gaussian distribution.

Fisher went on to show that of all estimators based on the sum of the pth powers of the residuals, the estimator with $p = 2$, i.e. σ_2, has the smallest variance for large samples.

Although both Fisher and Eddington were unaware of it at the time, the paper up to this point contained nothing new, save Fisher's elegant geometrical approach to the distribution theory. In fact, Gauss had published essentially the same investigation in 1816! Luckily, Fisher did not stop here. Rather, he went on to a more detailed study of the behaviour of σ_1 and σ_2. He observed that for large n the variances of σ_1 and σ_2 might tell the story, as both are then approximately normally distributed, but that for small n the situation is more complicated. To illustrate their behaviour, he derived the joint distribution of σ_1 and σ_2 for the special case $n = 4$. After calculating the first few moments of σ_1 and σ_2 for this case, a reconsideration of their joint distribution produced the following revelation:

> So far the variables have been compared only in respect of the quantitative characters of their frequency distributions. There exists also in the form of the frequency surface a qualitative distinction, which reveals the unique character of σ_2.
>
> From the manner in which the frequency surface has been derived...it is evident that:
>
> *For a given value of σ_2, the distribution of σ_1 is independent of σ.* [Fisher's italics.] On the other hand, it is clear... that for a given value of σ_1 the distribution of σ_2 does involve σ. In other words, if, in seeking information as to the value of σ, we first determine σ_1, then we can still further improve our estimate by determining σ_2; but if we had first determined σ_2, the frequency curve for σ_1 being entirely independent of σ, the actual value of σ_1 can give us no further information as to the value of σ. The whole of the information to be obtained from σ_1 is included in that supplied by a knowledge of σ_2.
>
> This remarkable property of σ_2, as the methods which we have used to determine the frequency surface demonstrate, follows from the distribution of frequency density in concentric spheres over each of which σ_2 is constant. It therefore holds equally if σ_3 or any other derivate be substituted for σ_1. If this is so, then it must be admitted that:
>
> *The whole of the information respecting σ, which a sample provides, is summed up in the value of σ_2.* [Fisher's italics.]

Fisher went on to observe that this property of σ_2 is quite dependent on the assumption that the population is normal, and showed that indeed σ_1 is preferable to σ_2, at least in large samples, for estimating the scale parameter of the double exponential distribution,

providing both estimators are appropriately rescaled. He closed the paper by proposing that in actual situations the sample measure of kurtosis, β_2, be calculated, and

> If this is near 3 the Mean Square Error will be required; if, on the other hand, it approaches 6, its value for the double exponential curve, it may be that σ_1 is a more suitable measure of dispersion.

By 1922, Fisher had named sufficiency, related it to maximum likelihood, stated the factorization theorem, and applied the concept in a variety of situations. The theory of estimation had taken a giant leap forward; the concept of sufficiency, one of Fisher's most original contributions, had been born.

3. Laplace and the method of situation

In the last part of the Second Supplement (1818) to his monumental *Théorie Analytique des Probabilités*, Laplace presented an investigation which is strikingly similar to Fisher's work of 1920. Laplace considered the problem we would now refer to as linear regression through the origin. In his notation, we have a system of n equations,

$$p_i y - a_i + x_i = 0 \quad (i = 1, \ldots, n),$$

where the p_i's and a_i's are known, and y and the x_i's are unknown, the x_i's being the errors of observation. The problem is to estimate y. Laplace had earlier in the volume discussed the method of least squares, for which he here used the name 'most advantageous method'; the purpose of this section of the Second Supplement was to discuss an alternative method of estimation introduced by Boscovich in 1757, which Laplace called the 'method of situation'.

Laplace began by repeating the definition of the 'method of situation' and a computational algorithm which had appeared in his *Mécanique Céleste* in 1799. Briefly, the method is: estimate y by that value which minimizes

$$\sum_{i=1}^{n} |p_i y - a_i|. \tag{1}$$

If all the p_i are positive and the observations are indexed in such a way that

$$\frac{a_1}{p_1} \geqslant \frac{a_2}{p_2} \geqslant \ldots \geqslant \frac{a_n}{p_n},$$

and if the integer r is defined to be such that

$$p_1 + \ldots + p_{r-1} < p_r + \ldots + p_n, \quad p_1 + \ldots + p_r > p_{r+1} + \ldots + p_n,$$

then the value of y which minimizes (1) is given by

$$y = a_r / p_r. \tag{2}$$

Of course, in the special case where all $p_i \equiv 1$, this simply gives the median of the a_i's. Edgeworth (1923) has called the solution (2) a weighted median; we use the symbol y_{MS}, MS for 'method of situation', for the expression (2) in order to distinguish it from the least squares or 'most advantageous' estimator,

$$y_{\mathrm{LS}} = \frac{\Sigma p_i a_i}{\Sigma p_i^2}.$$

After proving that y_{MS} given by (2) does in fact minimize (1), Laplace proceeded to investigate the distribution of y_{MS}. He assumed that the observational errors x_i all obeyed

the same probability density $\phi(x)$, but he assumed nothing more about ϕ than that it was an even function, $\phi(x) = \phi(-x)$. His later analysis assumes implicitly that ϕ is twice differentiable at zero and $\phi(0) > 0$. Laplace then correctly derived the density of $y_{\text{MS}} - y$, and showed that as n increases, this density approaches the normal density with mean zero and variance $\{4\phi^2(0)\Sigma p_i^2\}^{-1}$. In the special case where all p_i's are 1 this agrees with the now standard results for the sample median; indeed, it seems likely that Laplace was the first to derive the asymptotic distribution of a single order statistic.

Earlier in the volume Laplace had derived the asymptotic distribution of the least squares estimator y_{LS}; he now compared y_{MS} and y_{LS} as estimators of y on the basis of the variances of their asymptotic distributions, and concluded that the 'method of situation' was preferable to the 'most advantageous method' if and only if

$$\{2\phi(0)\}^2 > \left\{\int_{-\infty}^{\infty} x^2\phi(x)\,dx\right\}^{-1}.$$

He noted that y_{LS} was to be preferred if ϕ is a normal density.

If Laplace had halted his investigation at this point and gone on to other matters, it would still have been a notable work. For the first time, two reasonable methods of estimation had been compared for general populations and the precise conditions under which one would be preferable to another spelled out. In addition, he had presented the first large sample theory for a single order statistic, using an argument that would still be considered modern today, an argument which can be easily extended to nonsymmetrical error distributions and sample percentiles other than the median. But Laplace did not stop here.

As Fisher did in a similar situation one hundred years later, Laplace, having examined the distributions of y_{MS} and y_{LS} separately, went on to consider their joint distribution. His object in doing this was to show how the two estimators could be combined to provide a new estimator which would be better than either:

> In combining the results of these two methods, one can obtain a result whose probability law of error will be more rapidly decreasing.

He accomplished this as follows. First, he presented an expression for the joint density of $y_{\text{MS}} - y$ and $y_{\text{LS}} - y$ in terms of the inversion of a characteristic function. Next, he developed the integrand as a series. Finally, he was able to perform the required integration by taking n large, showing that the joint asymptotic distribution of $\zeta = y_{\text{MS}} - y$ and $\zeta' = y_{\text{LS}} - y$ had a density proportional to

$$\exp\left[-\frac{k}{2k''}\zeta'^2\Sigma p_i^2 - \frac{k''}{k}\frac{\left\{\zeta\dfrac{\phi(0)}{k} - \zeta'\dfrac{k'}{k''}\right\}^2}{2\left(\dfrac{k''}{k} - \dfrac{k'^2}{k^2}\right)}\Sigma p_i^2\right], \tag{3}$$

where

$$k = \tfrac{1}{2}, \quad k' = \int_0^{\infty} x\phi(x)\,dx, \quad k'' = \int_0^{\infty} x^2\phi(x)\,dx.$$

We would now recognize this as the density of the bivariate normal distribution with mean $(0, 0)'$ and covariance matrix $(\Sigma p_i^2)^{-1}\{\sigma_{jk}\}$, where

$$\sigma_{jk} = \{2\phi(0)\}^{-(4-j-k)}\int_{-\infty}^{\infty} |x|^{j+k-2}\phi(x)\,dx.$$

Laplace then used this result to find, in terms of ϕ, the value of C for which the asymptotic variance of $y_{\mathrm{LS}} - (y_{\mathrm{LS}} - y_{\mathrm{MS}})\,C$ is smallest. The appropriate value of C is, again in Laplace's notation,

$$C = \frac{\frac{\phi(0)}{k}\left\{\frac{\phi(0)}{k} - \frac{k'}{k''}\right\}}{\frac{k}{k''} - \frac{k'^2}{k''^2} + \left\{\frac{\phi(0)}{k} - \frac{k'}{k''}\right\}^2}.$$

Laplace wrote, in closing the Second Supplement,

> the result of the most advantageous method [i.e. y_{LS}] must therefore be diminished by the quantity $[(y_{\mathrm{LS}} - y_{\mathrm{MS}})C]$; and the probability of an error u using this corrected result will be proportional to the preceding exponential. [The correct expression had been given.] The importance of this new result will be increased if $(\phi(0)/k) - (k'/k'')$ is not zero; there is accordingly an advantage to correcting the most advantageous method in this way. When one does not know the distribution of the errors of observation this correction is not feasible; but it is remarkable that in the case where this probability is proportional to e^{-hx^2}; that is, when $\phi(x) \propto e^{-hx^2}$, the quantity
>
> $$\frac{\phi(0)}{k} - \frac{k'}{k''}$$
>
> will be zero. Then the result of the most advantageous method will receive no correction from the result of the method of situation, and the probability law of an error is unchanged.

Thus we see that the consideration of the joint distribution of y_{LS} and y_{MS} led Laplace to the 'remarkable' conclusion that not only is y_{LS} a better estimator of y than y_{MS} when the errors are normally distributed, but y_{LS} is better than any other linear combination of the two. Earlier in the Second Supplement, Laplace had performed a similar but less complicated calculation to also show that y_{LS} could not be improved by linear combination with any linear unbiased estimator of y. He thus clearly realized that, in this limited sense, neither y_{MS} nor any linear estimator could add information about y to y_{LS}.

4. Conclusion

While it is true that Laplace had clearly come across and described one aspect of sufficiency, at least as it related to his problem, it must be admitted that he did not go so far as Fisher, and he did not isolate the concept. Both Laplace and Fisher took the important step of considering the joint distribution of competing estimators, rather than merely looking at their distributions separately. Both were led by this step to realize that for normally distributed data, one of their estimators could add nothing to the other. But Laplace stopped where Fisher continued, and did not abstract that element of the problem which was responsible for this state of affairs.

Ironically, whereas Fisher considered only normally distributed data for the greater part of his paper, Laplace considered a more general class of distributions, and this greater generality apparently kept him from realizing the special nature of the density (3) when $\phi(x) \propto \exp(-hx^2)$. It is tempting to speculate that if Laplace had lived in an age where the normal distribution was considered as 'normal' as it was by Fisher, he might have looked at the density (3) when $\phi(0)/k = k'/k''$, realize that the conditional distribution of y_{MS} given y_{LS} did not change with y, and hit upon the concept of sufficiency, as Fisher did a century later. However, such speculation is worse than idle. For it not only serves to mask the fact that it is the generality of Laplace's work which both motivated his achievement and made

it important, but also serves to diminish unfairly the magnitude of Fisher's great accomplishment, leaping from a particular fact to the general concept of sufficiency with such speed and force that the momentum carried him on to develop a whole new theory of estimation.

Actually, the great difference between the historical contexts of these works makes a real comparison impossible. Laplace's work took place in the infancy of mathematical statistics; the method of least squares had only been published 13 years before and the problem of point estimation was fuzzily understood at best. By Fisher's time, the issues had become much clearer, through the work of Karl Pearson and others. Yet, the similarities between the two men's work, despite their different circumstances, sheds some light on the creative processes of great minds.

BIBLIOGRAPHICAL NOTES

Laplace's *Deuxième Supplément a la Théorie Analytique des Probabilités* was published separately in 1818 dated 'Février 1818' and appeared as an appendix to the *Théorie Analytique* with the publication of the third edition in 1820. The first part of this Supplement, which is not discussed here, originally appeared in *Connaissance des Temps pour l'an 1820* (published 1818), pp. 422–43, bearing the legend 'Lu à l'Academie des Sciences, le 4 août 1817'. This would apparently date the work discussed here some time in the latter part of 1817.

Boscovich's description of the method of situation and the relation between his work and Laplace's have been discussed by Eisenhart (1961). While the method itself and a geometrical version of the algorithm presented here are due to Boscovich, the analysis of the probabilistic properties of the estimator seems to have been original with Laplace.

Boscovich, and Laplace in *Mécanique Céleste* (Première Partie, Livre III, No. 40), had considered the more general model $p_i y + z - a_i + x_i = 0$, and determined the estimate of the parameter z from that of y by the requirement that the sum of the residuals be zero.

Fisher does not appear to have been aware of Laplace's Second Supplement. He nowhere refers to it, and the only explicit reference to *Théorie Analytique* in his collected papers is to the second edition (1814), which did not contain the Second Supplement. He does refer to the third edition in his books, in connexion with Laplace's use of Bayes's theorem.

While the first paper of Fisher's to use the name sufficiency was his 1922 paper 'On the mathematical foundations of theoretical statistics', the name did appear in *Nature* for 24 November 1921, in the abstract of Fisher's 17 November presentation to the Royal Society.

This research was initiated in the Department of Statistics, University of Wisconsin, under the partial support of the Office of Naval Research and completed while the author was on leave in the Department of Statistics, University of Chicago, under the partial support of the National Science Foundation.

REFERENCES

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

EDGEWORTH, F. Y. (1923). On the use of medians for reducing observations relating to several quantities. *Phil. Magazine* **46** (6th series), 1074–88.

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 the Fordham University Press, New York.

FISHER, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices R. Astronomical Soc.* **80**, 758–70. Reprinted (1950) in Fisher's *Contributions to Mathematical Statistics*. New York: Wiley.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond.* A **222**, 309–68. Reprinted (1950) in Fisher's *Contributions to Mathematical Statistics*. New York: Wiley. Abstract in *Nature, Lond.* **108** (1921), 421.

GAUSS, C. F. (1816). Bestimmung der Genauigkeit der Beobachtungen. In *Carl Friedrich Gauss Werke* Band 4, pp. 109–17, Göttingen: Königlichen Gesellschaft der Wissenschaften (1880).

LAPLACE, P. S. DE (1818). *Deuxième Supplément a la Théorie Analytique des Probabilités*. Paris: Courcier. Reprinted (1847) in *Oeuvres de Laplace* **7**, pp. 569–623. Paris: Imprimerie Royale; (1886) in *Oeuvres Complètes de Laplace* **7**, pp. 531–80. Paris: Gauthier-Villars. The first Part of this Supplement, which is not discussed here, originally appeared (1818) in *Connaissance des Temps pour l'an 1820*, pp. 422–43.

[*Received December* 1972. *Revised February* 1973]