Statistical Science
1999, Vol. 14, No. 2, 214–222

On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares

Anders Hald

Abstract. It is shown that the method of maximum likelihood occurs in rudimentary forms before Fisher [Messenger of Mathematics **41** (1912) 155–160], but not under this name. Some of the estimates called "most probable" would today have been called "most likely." Gauss [*Z. Astronom. Verwandte Wiss.* **1** (1816) 185–196] used invariance under parameter transformation when deriving his estimate of the standard deviation in the normal case. Hagen [*Grundzüge der Wahrscheinlichkeits-Rechnung*, Dümmler, Berlin (1837)] used the maximum likelihood argument for deriving the frequentist version of the method of least squares for the linear normal model. Edgeworth [*J. Roy. Statist. Soc.* **72** (1909) 81–90] proved the asymptotic normality and optimality of the maximum likelihood estimate for a restricted class of distributions. Fisher had two aversions: noninvariance and unbiasedness. Replacing the posterior mode by the maximum likelihood estimate he achieved invariance, and using a two-stage method of maximum likelihood he avoided appealing to unbiasedness for the linear normal model.

Key words and phrases: Chauvenet, confidence limits, credible limits, Edgeworth, Encke, Fisher, Gauss, Gosset, Hagen, invariance, inverse probability, Laplace, least squares, likelihood limits, linear normal model, maximum likelihood, Merriman, posterior mode, reparameterization, *t*-distribution, two-stage maximum likelihood method, unbiasedness.

1. INTRODUCTION

The modern version of the method of maximum likelihood was created single-handedly by R. A. Fisher between 1912 and 1922; see the recent discussions by Edwards (1997), Aldrich (1997) and Hald (1998, Sections 28.4–28.5). Here we present some further information on the history of the method, as a supplement to Edwards (1974). The method of maximum likelihood occurs in various rudimentary forms before Fisher, but not under this name. Some of the estimates called "the most probable values of the unknowns" are maximum likelihood estimates. The terminological confusion was not cleared up until Fisher in 1921 introduced the term "likelihood" and in 1922 the "maximum likelihood estimate."

It is well known (see Hald, 1998) that the method of maximum likelihood was proposed independently by Lambert and Daniel Bernoulli, but with no practical effect because the maximum likelihood equation for the error distribution considered was intractable.

Gauss (1809) combined the Lambert–Bernoulli idea with Laplace's principle of inverse probability, which led him to maximize the posterior density of the location parameter in the error distribution, assuming that the prior distribution is uniform. Requiring that the posterior mode equals the arithmetic mean, Gauss derived the normal distribution and thus gave a probabilistic justification for the method of least squares.

The methods considered lead to the same estimates of the location parameters (regression coefficients) in the linear normal model, so when looking for differences we have to study the estimation of

Anders Hald is Professor Emeritus, Department of Theoretical Statistics, University of Copenhagen. His address is Furesøvej 87 A, DK-2830 Virum, Denmark.

the variance. The method of least squares does not specify a rule for estimating the variance. However, as noted by Laplace (1812, II, Section 20) in connection with his large-sample theory of this method, it is natural, by analogy, to use the empirical second moment $\hat{\sigma}^2 = \sum_1^n (x_i - \bar{x})^2/n$ an estimate of σ^2. He used this estimate to obtain large-sample confidence limits for the location parameter in the form $\bar{x} \pm u\hat{\sigma}/\sqrt{n}$, where u denotes the standardized normal deviate for the confidence level chosen.

In his frequentist version of the method of least squares, Gauss (1823) proved that $s^2 = \sum_1^n (x_i - \bar{x})^2/(n-1)$ is an unbiased estimate of σ^2, and this became the estimate ordinarily used by astronomers and surveyors.

However, Edgeworth's genuine inverse method and Fisher's method of maximum likelihood both lead to the estimate $\hat{\sigma}^2$, so they had to modify their methods to obtain the Gaussian estimate. Edgeworth (1908) showed that the estimate s may be obtained as the mode in the marginal posterior distribution, and Fisher (1922b) derived s by a two-stage maximum likelihood method. They both avoided appealing to unbiasedness, a concept foreign to their methods.

We shall in particular discuss three contributions that imply the method of maximum likelihood: namely, the contributions by Gauss (1816), Hagen (1837) and Edgeworth (1909); see Sections 2–4. Fisher did not know these results when he wrote his first papers on maximum likelihood.

In Section 5 we discuss some of Fisher's results in relation to noninvariance, unbiasedness and confidence intervals.

2. GAUSS'S ESTIMATES OF THE STANDARD DEVIATION

Gauss (1809) wrote the normal distribution in the form

$$f(x|\theta, h) = \pi^{-1/2} h \exp[-h^2(x-\theta)^2], \quad -\infty < x < \infty,\ -\infty < \theta < \infty,\ 0 < h < \infty,$$

the precision constant being $h = 1/\sigma\sqrt{2}$. He considered the estimation of θ and h as separate one-parameter problems; in 1809 he estimated θ, assuming that h is known, and in 1816 he estimated h, assuming that θ is known; in both cases he used the posterior mode as estimate.

For a sample of n observations $\mathbf{x} = (x_1, \cdots, x_n)$, we shall introduce the errors $\varepsilon_i = x_i - \theta$ and the residuals $e_i = x_i - \bar{x}$, $i = 1, \cdots, n$. Moreover, we use the Gaussian summation notation $[\varepsilon\varepsilon] = \sum_1^n \varepsilon_i^2$, so that

$$[\varepsilon\varepsilon] = [ee] + n(\bar{x} - \theta)^2. \tag{1}$$

The probability density of the sample is

$$p(\mathbf{x}|\theta, h) = \pi^{-n/2} h^n \exp(-h^2[ee]) \times \exp[-h^2 n(\bar{x} - \theta)^2]. \tag{2}$$

Assuming that θ and h are uniformly distributed, the posterior distribution $p(\theta, h|\mathbf{x})$, say, is proportional to $p(\mathbf{x}|\theta, h)$.

Applying Laplace's principle of inverse probability in its (likelihood) ratio form, Gauss (1816, Art. 3) stated that

$$\frac{p(h_1|\boldsymbol{\varepsilon})}{p(h_2|\boldsymbol{\varepsilon})} = \frac{p(\boldsymbol{\varepsilon}|h_1)}{p(\boldsymbol{\varepsilon}|h_2)}, \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n), \tag{3}$$

for all values of h_1 and h_2 on the positive real line. He concluded that

$$p(h|\boldsymbol{\varepsilon}) \propto p(\boldsymbol{\varepsilon}|h) \propto h^n \exp(-h^2[\varepsilon\varepsilon]). \tag{4}$$

The most probable value of the true value of h is thus $\hat{h} = \sqrt{n/2[\varepsilon\varepsilon]}$.

However, Gauss's goal was to find probability limits for θ, and he therefore needed an estimate of σ. He writes (in our notation):

> The most probable value of σ is consequently $1/\hat{h}\sqrt{2}$. This result holds generally, whether n be large or small.

Hence, Gauss transformed $\hat{h}$ to $\hat{\sigma} = \sqrt{[\varepsilon\varepsilon]/n}$ as if the estimates were parameters. We shall call this rule the Gaussian rule of invariance. Gauss did not give any argument for this rule, perhaps because it seemed obvious to him for the following reason. Reparameterizing the error distribution from h to σ, and denoting the corresponding densities by p^*, it follows that

$$p^*(\sigma|\boldsymbol{\varepsilon}) \propto p^*(\boldsymbol{\varepsilon}|\sigma) \propto p(\boldsymbol{\varepsilon}|h) \quad \text{for } h = 1/\sigma\sqrt{2},$$

which demonstrates the fact that $\hat{\sigma} = 1/\hat{h}\sqrt{2}$. It is clear that the Gaussian rule holds for any one-to-one transformation of the parameters.

Gauss used the probable error $r = 0.6744897\sigma$ as parameter; we have rewritten his formulas in terms of σ.

Expressed in terms of prior distributions, it can be said that Gauss used a modified version of the principle of inverse probability, namely: to estimate h he assumed the prior distribution of h to be uniform, and to estimate a one-to-one transformation of h, $\sigma = \sigma(h)$ say, he assumed the prior distribution of σ to be uniform. However, this is equivalent to using the method of maximum likelihood.

Gauss did not mention the fact that if he had used the normalized version of Laplace's principle,

he would have found that

$$p(\sigma|\boldsymbol{\varepsilon}) = p(h|\boldsymbol{\varepsilon})|dh/d\sigma|,$$

which leads to a contradiction of his rule of invariance.

Gauss's rule was accepted by most statisticians, although several writers pointed out that a uniform prior distribution of h implies a prior distribution for σ proportional to σ^{-2}.

Expanding $\ln p(h|\boldsymbol{\varepsilon})$ in Taylor's series about $h = \hat{h}$, and neglecting terms of smaller order of magnitude than the first, Gauss obtained

$$(5) \qquad p(h|\boldsymbol{\varepsilon}) = p(\hat{h}|\boldsymbol{\varepsilon})\exp\left[-n(h-\hat{h})^2/\hat{h}^2\right],$$

so h is asymptotically normal with mean $\hat{h}$ and variance $\hat{h}^2/2n$. He concluded that the large-sample limits for the true value of h are $\hat{h}(1 \pm u/\sqrt{2n})$. Making the substitutions $h = 1/\sigma\sqrt{2}$ and $\hat{h} = 1/\hat{\sigma}\sqrt{2}$, Gauss found the corresponding limits for σ as $\hat{\sigma}(1 \pm u/\sqrt{2n})$; that is, σ is asymptotically normal with mean $\hat{\sigma}$ and variance $\hat{\sigma}^2/2n$.

Referring to Laplace's central limit theorem, Gauss noted that the sampling distribution of $[\varepsilon\varepsilon]$ for large n is normal with mean $n\sigma^2$ and variance $2n\sigma^4$. It follows that the probability limits for $\hat{\sigma}$ are $\sigma(1 \pm u/\sqrt{2n})$, and solving for σ Gauss obtained the limits $\hat{\sigma}(1 \pm u/\sqrt{2n})$. He remarked that these limits are identical to those found above by inverse probability.

Gauss had only the single word "probability" at his disposal for characterizing the two sets of limits that today are called credible and confidence limits, respectively. However, there is no ambiguity in Gauss's description of the different probabilistic backgrounds. Gauss (1816, Art. 4) also distinguished between parameter and estimate by using lowercase letters for parameters and capitals for estimates.

Gauss left three problems unsolved: (1) Why did he not maximize the joint posterior distribution $p(\theta, h|\mathbf{x})$ with respect to both parameters? (2) Why did he not use the marginal distribution of θ for finding credible limits? (3) Why did he not use the marginal distribution of h to estimate the true value of h? As we shall see in Section 4, these problems were discussed by Edgeworth.

We have used $\hat{\sigma}^2$ as notation for $[ee]/n$ as well as $[\varepsilon\varepsilon]/n$, because both are maximum likelihood estimates of σ^2 depending upon assumptions made about θ.

Consider now the linear normal model with $m < n$ parameters in the usual matrix notation $y = X\beta + \varepsilon$. Gauss (1809) proved that β_r, the rth component of β, is normally distributed with mean b_r, the least squares estimate, and precision $h/\sqrt{q_{rr}}$, where q_{rr} is the rth diagonal element of $Q = (X'X)^{-1}$, $r = 1, \cdots, m$. His main tool in the proof was the decomposition

$$(6) \qquad \varepsilon'\varepsilon = e'e + v'v, \quad e = y - Xb,$$

where $v = U(\beta - b)$, U being upper triangular, and the elements of v are independently distributed as $N(0, \sigma^2)$, like the elements of ε.

Gauss (1823) gave up the method of inverse probability, because he considered it as metaphysical, and he also abandoned the assumption of normality as too narrow. Instead he justified the method of least squares as the method leading to minimum variance estimates within the class of linear unbiased estimates. He called these estimates the most plausible values of the parameters. Moreover, he proved that

$$s^2 = (y - Xb)'(y - Xb)/(n - m)$$

is an unbiased estimate of σ^2, and remarked that the sum of the n squared residuals under normality is distributed in the same way as the sum of $n - m$ squared true errors; presumably he based this remark on relation (6). He used s as an estimate of σ, that is, he did not require unbiasedness.

3. THE FREQUENTIST VERSION OF THE METHOD OF LEAST SQUARES FOR THE LINEAR NORMAL MODEL

The task of rewriting Gauss's first proof (Gauss, 1809) in frequentist terms was carried out by astronomers and geodesists writing elementary textbooks on the method of least squares. They found Gauss's second proof (Gauss, 1823) too cumbersome for their readers and did not need the generalization involved, because the measurement errors encountered in their fields were in most cases nearly normally distributed. They realized that the method of maximizing the posterior density could be replaced by the method of maximizing the density of the observations, because $p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)$. Hence, the roles of parameters and estimates in Gauss's formulas were interchanged.

They did not relate their method to the Lambert–Bernoulli idea, which presumably was unknown to them. Only Todhunter (1865, pages 236–237) noted that the method of maximizing $p(\mathbf{x}|\theta)$ with respect to θ for normally distributed observations was an application of Bernoulli's method of maximum likelihood.

These elementary textbooks authors wrote the probability of the observed system of errors as

$$f(\varepsilon_1) \cdots f(\varepsilon_n)\, d\varepsilon_1 \cdots d\varepsilon_n = P\, d\varepsilon_1 \cdots d\varepsilon_n,$$

calling P the probability of the system of errors, the term "probability density" being of a later date.

The estimates obtained by maximizing P were called "the most probable values of the unknowns"; they thus adopted Gauss's expression although they did not use inverse probability. They did not grasp the importance of Gauss's distinction between "most probable" and "most plausible."

They followed Gauss by treating the maximization of the density

$$p(\mathbf{x}|\theta, h) = \pi^{-n/2} h^n \exp\left\{-h^2 \sum (x_i - \theta)^2\right\}$$

as two successive one-parameter problems. Maximization with respect to θ leads to $\hat{\theta} = \bar{x}$, and for the linear normal model to $\hat{\beta}_r = b_r$, $r = 1, \cdots, m$. Maximization with respect to h leads to $\hat{h} = \sqrt{n/2[\varepsilon\varepsilon]}$, and maximization with respect to σ gives $\hat{\sigma} = \sqrt{[\varepsilon\varepsilon]/n}$. It follows that $\hat{\sigma} = 1/\hat{h}\sqrt{2}$, in accordance with Gauss's rule.

We shall consider some early books and papers on this topic to demonstrate how the method evolved.

Encke (1832–1834) wrote a comprehensive survey of Gauss's work on the method of least squares, covering both of Gauss's proofs and adding some modifications of his own. He (1832, page 276) maximizes $p(\mathbf{x}|\theta, h)$ with respect to θ, and afterwards he notes that the same estimate is obtained by maximizing $p(\theta, h|\mathbf{x})$. He reproduces Gauss's derivation of $\hat{\sigma}^2 = [\varepsilon\varepsilon]/n$, and using equation (1) he (1832, pages 283–284) remarks that $[ee]$ is smaller than $[\varepsilon\varepsilon]$ by the quantity $n(\bar{x} - \theta)^2$, which on the average equals σ^2. He concludes that the estimate of σ^2 should be $s^2 = [ee]/(n-1)$ when θ is unknown. He (1833, page 320) extends this proof to the linear normal model. Taking expectations of both sides of equation (6) he gets $E(e'e) = (n - m)\sigma^2$, so that $s^2 = e'e/(n-m)$ is an unbiased estimate of σ^2. This simple proof became standard in textbooks on the method of least squares.

In his textbook for civil engineers, Hagen (1837) begins by deriving the normal distribution of errors by a simplification of Laplace's central limit theorem. His hypothesis of elementary errors (Hagen, 1837, page 34) says that

> the error in the result of a measurement is the algebraic sum of an infinitely large number of elementary errors which are all equally large, and each of which can be positive or negative with equal ease.

This means that the distribution of the sum of n elementary errors is the symmetric binomial, which converges to the normal for $n \to \infty$. To avoid the complicated proofs of de Moivre and Laplace, Hagen finds the relative slope of the binomial frequency curve, which for $n \to \infty$ leads to a differential equation with the normal distribution as solution. Because of its simplicity this proof was adopted by many textbook writers.

Hagen (1837, page 75) emphasizes that, by using this derivation of the normal distribution as starting point, one avoids the circularity involved in Gauss's (1809) proof of the arithmetic mean as the best estimate of θ.

Assuming that the observational errors are normally distributed, and setting $d\varepsilon_i = d\varepsilon$ for all i, Hagen (1837, page 67) obtains (in our notation)

$$(7)\quad p(\boldsymbol{\varepsilon})(d\varepsilon)^n = (d\varepsilon/\sqrt{\pi})^n h^n \exp(-h^2[\varepsilon\varepsilon]).$$

He remarks that

> the first factor of this expression will be unchanged if we attach another hypothesis [regarding the true value] to the observations and the individual errors therefore take on other values; the second factor will however be changed. Among all hypotheses of this kind, which can be attached to the observations, the most probable is consequently the one which makes $Y[p(\varepsilon)(d\varepsilon)^n]$ a maximum, which means that the exponent of e should be a minimum, that is, the sum of the squares of the resulting errors should be as small as possible.

Hagen's second factor is thus the likelihood function for θ, which he maximizes to find the most likely hypothesis.

For the linear normal model Hagen gets

$$\varepsilon'\varepsilon = e'e + (b - \beta)' X'X (b - \beta),$$

which inserted into $p(\boldsymbol{\varepsilon})$ gives the likelihood function for β. To find the likelihood for β_1, he (Hagen, 1837, page 80) maximizes $p(\boldsymbol{\varepsilon})$ with respect to the other elements of β and finds

$$(8)\quad \max_{(\beta_2, \cdots, \beta_m)} p(\boldsymbol{\varepsilon}) \propto \exp\left(\frac{-h^2(b_1 - \beta_1)^2}{q_{11}}\right),$$

$$Q = (X'X)^{-1}.$$

He concludes that b_r is normally distributed with mean β_r and variance $\sigma^2 q_{rr}$, $r = 1, \cdots, m$. This is the likelihood version of Gauss's 1809 proof.

In his textbook on astronomy, Chauvenet (1863) wrote an appendix on the method of least squares that essentially is an abridged English version of Encke's 1832 paper, except that it leaves out all material on inverse probability. He thus proved Gauss's basic results for normally distributed observations by operating on the likelihood function instead of the posterior distribution.

A more consistent exposition of the likelihood theory is due to Merriman (1884) in *The Method of*

Least Squares, written for "civil engineers who have not had the benefit of extended mathematical training." Merriman had an extraordinarily good background for writing this book because in 1877 he had provided a "List of writings relating to the method of least squares, with historical and critical notes" (Merriman, 1877), containing his comments on 408 books and papers published between 1722 and 1876. It should be noted, however, that all his comments are based on the principle of maximizing the probability of the sample; he does not even mention inverse probability.

Merriman (1884, Art. 13) defines "most probable" as follows: "The most probable event among several events is that which has the greatest mathematical probability," and in Art. 17 he writes that "The probability of an assigned accidental error in a set of measurements is the ratio of the number of errors of that magnitude to the total number of errors."

In Art. 41 he formulates the method of estimation:

> The most probable system of errors will be that for which P is a maximum (Art. 13) and the most probable values of the unknowns will correspond to the most probable system of errors.

Merriman's postulate above obviously leads to the maximum likelihood estimate, which for the location parameter in the linear normal model equals the least squares estimate.

To estimate h he maximizes the right-hand side of formula (4), leading to the maximum likelihood estimate $\hat{h}$. He remarks that because $\sigma h\sqrt{2} = 1$, the estimate of σ becomes $\hat{\sigma} = \sqrt{[\varepsilon\varepsilon]/n}$, and using Encke's argument he obtains the estimate $s = \sqrt{[ee]/(n-1)}$.

To find the uncertainty of $\hat{h}$, Merriman (1884, Art. 164) expands $p(\boldsymbol{\varepsilon}|h)$ around $\hat{h}$ and, omitting terms of smaller order of magnitude, he obtains

$$(9) \qquad p(\boldsymbol{\varepsilon}|h) = p(\boldsymbol{\varepsilon}|\hat{h})\exp\left[-n(h-\hat{h})^2/\hat{h}^2\right].$$

He concludes that the sampling distribution of $\hat{h}$ in large samples is normal with mean h and standard error $\hat{h}/\sqrt{2n}$. This is the likelihood version of Gauss's 1816 proof; see (5).

It will be seen that Hagen and Merriman derived the likelihood functions for the parameters involved; see (8) and (9). Because the likelihood function is proportional to the probability density, they interpreted these results as the sampling distributions of the estimates. In this way they came to the same frequentist results as Laplace and Gauss, but with much simpler proofs because they considered only the linear normal model.

Merriman's book was well known among British statisticians; both Pearson and Edgeworth refer to it.

4. EDGEWORTH'S GENUINE INVERSE METHOD AND THE ASYMPTOTIC SAMPLING DISTRIBUTION AND OPTIMALITY OF THE POSTERIOR MODE

Edgeworth (1883) maximizes the density $p(\theta, \sigma|\mathbf{x})$ with respect to both parameters and thus obtains the estimates $\bar{x}$ and $\sqrt{[ee]/n}$. He emphasizes that the solution holds whether the number of observations be finite or infinite. (Edgeworth uses the modulus $c = \sigma\sqrt{2}$ as parameter.) It worries him that Gauss and his followers use $n-1$ instead of n as denominator, and he therefore provides a further argument for using n.

If σ is known, then the credible limits for θ are found by means of the standard error $\sigma/\sqrt{n}$, but how are we to find these limits when σ is unknown? Edgeworth derives the marginal distribution of θ for a uniform distribution of h, which gives

$$p(\theta|\mathbf{x}) = \int p(\theta, h|\mathbf{x})\, dh$$

$$\propto \{[ee] + n(\bar{x}-\theta)^2\}^{-(n+1)/2}.$$

This means that

$$t = \frac{(\theta - \bar{x})\sqrt{n}}{\sqrt{[ee]/n}}$$

is distributed as Student's t with n degrees of freedom. Edgeworth concludes that the probable limits (50% credibility) for θ,

$$\bar{x} \pm un^{-1/2}\sigma \cong \bar{x} \pm un^{-1/2}\sqrt{[ee]/n},$$

$$\Phi(u) - \Phi(-u) = 0.5,$$

should be replaced by the limits

$$(10) \qquad \bar{x} \pm tn^{-1/2}\sqrt{[ee]/n}, \quad P(t) - P(-t) = 0.5,$$

where $P(t)$ denotes the cumulative distribution function of the t-distribution. Edgeworth takes this result as further evidence for using n as denominator in the estimate of σ.

Returning to this problem in 1908, Edgeworth says (Edgeworth, 1908, pages 393–394) that the probable limits (10) appear to be of no great significance. Presumably he did not realize the importance of the t-distribution because he considered only the probable limits where the difference between u and t is rather small.

He remarks that the estimate $\sqrt{(n-1)/2[ee]}$ can be found as the mode in the posterior marginal

distribution of h because

$$p(h|\mathbf{x}) = \int p(\theta, h|\mathbf{x})\, d\theta \propto h^{n-1} \exp(-h^2[ee]). \tag{11}$$

After the invention of Karl Pearson's four-parameter system of frequency curves it became clear that new methods of estimation were needed. Pearson used the method of moments, whereas Edgeworth developed a generalized version of Gauss's (1809, 1816) method and applied it to nonnormal distributions.

Edgeworth (1908, pages 392, 396) points out that the posterior mode is noninvariant to parameter transformation. However, limiting himself to large-sample theory, he remarks that this fact is of no importance, because ordinary transformations are nearly linear in a neighborhood of $\hat{\theta}$ of order $n^{-1/2}$.

For large n he introduces the "genuine inverse method," which may be summarized as follows:

1. Use a uniform distribution for the parameters in the model, whatever parameterization has been chosen.
2. Maximize the joint posterior distribution to find the estimates.
3. The parameters are asymptotically multivariate normal with the posterior mode as mean and the inverse of the observed information matrix as dispersion matrix.
4. Interchanging the roles of parameters and estimates, it follows that the estimates are asymptotically multivariate normal with the parameters as means and the inverse of the expected information matrix as dispersion matrix.
5. The posterior mode minimizes the posterior expected squared error.

Edgeworth next takes the remarkable step of investigating the sampling distribution of $\hat{\theta}$ to show the optimality of $\hat{\theta}$ within the class of consistent and asymptotically normal statistics $t = t(\mathbf{x})$ that are symmetric functions of the observations. Assuming that $f(x - \theta)$ is symmetric about zero, he proves by means of Schwarz's inequality that $\text{var}(\hat{\theta}|\theta) \leqq \text{var}(t|\theta)$ for t equal to the arithmetic mean and the median, respectively. For a skew distribution, Pearson's Type III, he proves that $\hat{\theta}$ is at least as good as the estimate found by the method of moments. However, he does not find a general proof of $\hat{\theta}$'s optimality by this method.

Returning to the problem in 1909, Edgeworth limits the investigation to location densities and to the class of estimates (today called M-estimates) satisfying an equation of the form $\Sigma g(x_i - \theta) = 0$, where $E\{g(y)\} = 0$ and $g'(0) \neq 0$, $y = x - \theta$. The error of estimation $u = t - \theta$ of an estimate t is thus found from the equation $\Sigma g(y_i - u) = 0$, and using Taylor's expansion he finds to a first approximation that

$$u = \frac{\Sigma g(y_i)}{\Sigma g'(y_i)}.$$

From the central limit theorem it follows that u is asymptotically normal with zero mean and that

$$n\ \text{var}(u) = E\{g^2(y)\}/[E\{g'(y)\}]^2.$$

Using the calculus of variations, Edgeworth proves that the function minimizing $\text{var}(u)$ is proportional to $f'(y)/f(y)$, which means that $\hat{\theta}$ has minimum variance within the class of estimates considered.

He gives a similar analysis of the estimation problem for the scale parameter model and indicates that analogous results hold for the location–scale model.

More detailed discussions of Edgeworth's proofs are given by Pratt (1976) and Hald (1998, Section 28.3).

It is an astounding fact that Edgeworth's papers were unknown to Fisher when he wrote his paper on maximum likelihood estimation in 1912.

5. ON THE DEVELOPMENT OF FISHER'S METHOD OF MAXIMUM LIKELIHOOD

Let $f(x|\boldsymbol{\theta})\, dx$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$, be the chance of an observation falling within the range dx, and set

$$P' = f(x_1|\theta) \cdots f(x_n|\theta)\, dx_1 \cdots dx_n.$$

Fisher (1912) introduces the method of maximum likelihood by the following argument:

> The factors $\delta x[dx_1, \cdots, dx_n]$ are independent of the theoretical curve, so the probability of any particular set of θ's is proportional to P, where
>
> $$\log P = \sum_1^n \log f.$$
>
> The most probable set of values for the θ's will make P a maximum.

Fisher's argument is thus the same as Hagen's (1837), and his terminology and notation are those used by Chauvenet and Merriman. However, Fisher was not aware of the distinction between the frequentist and the inverse probability versions of the method of least squares, so confusingly he calls $p(\mathbf{x}|\theta, h)$ "the inverse probability system." At the end of the paper he explains that P, after all, is not an inverse probability but "must be considered as

the relative probability of the set of values $\theta_1, \cdots, \theta_m$."

Like Edgeworth, Fisher notes that the posterior mode depends on the parameterization of the model. While Edgeworth dismisses this problem, because it is of no importance in large samples, Fisher makes the crucial remark that "the relative values of P would be unchanged by such a transformation," whereas the (inverse) probability that the true value lies within a given region would change unless the Jacobian of the transformation equals unity. Hence, Fisher rejects inverse probability because of its noninvariance.

For normally distributed observations $P = p(\mathbf{x}|\theta, h)$; hence, Fisher finds the estimates $\hat{\theta} = \bar{x}$ and $\hat{h} = \sqrt{n/2[ee]}$, and thus $\hat{\sigma}^2 = [ee]/n$. To defend this result, as opposed to the ordinary estimate s^2, Fisher criticizes Chauvenet's proof because it depends on unbiasedness, leading to noninvariant estimates. Instead he suggests maximizing $p(\hat{\sigma}^2|h)$ with respect to h to find the most probable value of h. This is rather bewildering for the reader; Fisher had just explained why he preferred $\hat{h}$ to the classical estimate, and now he proposes to replace $\hat{h}$ by another (more probable?) estimate. If he had known the papers by Helmert (1876) or "Student" (1908), in which $p(\hat{\sigma}^2|h)$ is derived, he would have found that the proposed procedure leads to the estimate $s^2 = [ee]/(n-1)$ that he had just criticized.

As pointed out by E. S. Pearson (1968), Fisher came into correspondence with Gosset ("Student") in September 1912. Fisher's letters to Gosset are lost, but Gosset's correspondence with K. Pearson reveals that Gosset's arguments for $n-1$ as denominator made Fisher reconsider the problem. In his second letter to Gosset, Fisher gave an exact proof of Gosset's ("Student," 1908) joint distribution of $\bar{x}$ and $\hat{\sigma}$ and of the distribution of $z = (\bar{x} - \theta)/\hat{\sigma}$, which was Gosset's ("Student," 1908) original form of the t-distribution.

Fisher's proof was not published until 1915; he used (Fisher, 1915, page 509) a geometrical argument, which may be expressed in algebraic terms as

$$\begin{aligned} p(\mathbf{x}|\theta, \sigma)d(\mathbf{x}) &\propto \sigma^{-n} \exp\left\{-\tfrac{1}{2}\sigma^{-2}(n(\bar{x}-\theta)^2 + n\hat{\sigma}^2)\right\} d(\mathbf{x}) \\ &\propto \left\{\sigma^{-1} \exp\left(-n(\bar{x}-\theta)^2/2\sigma^2\right) d\bar{x}\right\} \\ &\quad \cdot\{\sigma^{-(n-1)}\hat{\sigma}^{n-2} \exp(-n\hat{\sigma}^2/2\sigma^2)\, d\hat{\sigma}\} \\ &\propto p(\bar{x}|\theta, \sigma)\, d\bar{x}\, p(\hat{\sigma}|\sigma)\, d\hat{\sigma}. \end{aligned}$$

This result led Fisher to introduce a two-stage maximum likelihood method. At the first stage, the ordinary maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is found by maximizing $L(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})$; at the second stage the sampling distribution of $\hat{\theta}_1$, say, is derived, and if this distribution depends on θ_1 only, then a new estimate $\hat{\hat{\theta}}_1$, say, is found by maximizing $L(\theta_1|\hat{\theta}_1) \propto p(\hat{\theta}_1|\theta_1)$. Aldrich (1997) calls this "the second criterion"; like Savage (1976, page 455), we suggest that it is better considered as the second step in an extended maximum likelihood procedure.

Estimating σ by this method, we first find $\hat{\sigma}^2 = [ee]/n$, and next $\hat{\hat{\sigma}}^2 = [ee]/(n-1)$ by maximizing $L(\sigma|\hat{\sigma}) \propto p(\hat{\sigma}|\sigma)$. Transforming from σ to h, it will be seen that $L(h|\hat{h})$ is proportional to Edgeworth's $p(h|\mathbf{x})$. Hence, the two-stage method leads to the same result as the posterior marginal distribution, and as the frequentist method with its requirement of unbiasedness.

Fisher's first example of the two-stage method is given in his paper on the distribution of the correlation coefficient $p(r|\rho)$. The first-stage estimate is $\hat{\rho} = r$, and maximizing $L(\rho|r) \propto p(r|\rho)$ Fisher (1915, page 521) finds, to a first approximation, that

$$r = \hat{\hat{\rho}}\left(1 + \frac{1-r^2}{2n}\right).$$

Both r and $\hat{\hat{\rho}}$ are biased estimates of ρ.

Suppose that $\hat{\theta}$ is normally distributed with mean $\theta + b(\theta)/n$ and variance ω^2/n, where ω^2 is independent of θ. It follows that $\hat{\hat{\theta}} = \hat{\theta} - b(\hat{\theta})/n$, to a first approximation, so that the bias at the first stage is removed at the second stage.

This is the idea underlying Fisher's second example, in which he considers the transformed correlation coefficient

$$\zeta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}.$$

Fisher (1921) proves that the distribution of the first-stage estimate, which is obtained by replacing ρ by r, is nearly normal with mean $\zeta + \rho/2(n-1)$ and variance $1/(n-3)$, so that

$$\hat{\hat{\zeta}} = \frac{1}{2}\ln\frac{1+r}{1-r} - \frac{r}{2(n-1)}$$

is a nearly unbiased estimate of ζ. Fisher remarks that the correction to the first-stage estimate is immaterial for a single sample, because it is of higher order than the standard deviation, but for pooling the information in several samples the correction should be taken into account to avoid a systematic error of estimation.

When giving his final definition of the likelihood of $\boldsymbol{\theta}$, Fisher (1922a, page 310) mentions only the first stage. Nevertheless, he (Fisher, 1922b, page

600) uses the two-stage method for estimating the within-group variance in a one-way layout and the variance about the regression line. Under the usual assumptions for the analysis of variance, the first-stage estimate becomes

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^m n_i \hat{\sigma}_i^2}{\sum_{i=1}^m n_i}, \quad n_i \hat{\sigma}_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2.$$

Noting that the within-group variances are independent, Fisher gets

$$p(\hat{\sigma}^2 | \sigma^2) \propto \prod_{i=1}^m \sigma^{-(n_i-1)} (\hat{\sigma}_i^2)^{(n_i-3)/2} \times \exp(-n_i \hat{\sigma}_i^2 / 2\sigma^2),$$

and, maximizing with respect to σ^2, he obtains

$$\hat{\hat{\sigma}}^2 = \frac{\sum_{i=1}^m n_i \hat{\sigma}_i^2}{\sum_{i=1}^m (n_i - 1)}.$$

In this way he provides a maximum likelihood argument for using the number of degrees of freedom as denominator; he has thus obtained an unbiased estimate of σ^2 without invoking the criterion of unbiasedness.

Suppose that the m samples are of the same size, n say, so that $\hat{\sigma}^2 = \sum_1^m \hat{\sigma}_i^2 / m$. For a fixed value of n and $m \to \infty$, $\hat{\sigma}^2$ tends in probability to $\sigma^2(n - 1)/n$. This inconsistency of the maximum likelihood estimate was pointed out by Neyman and Scott (1948); they have, however, overlooked that in this problem Fisher would have used the two-stage method; see Savage (1976, page 455).

Fisher considered the likelihood function as measuring the support, given by the observations, to the various possible hypotheses within the model. He suggested (Fisher, 1921) that the ratio of the likelihood function to its maximum may be used to find "likelihood intervals" (our term) for the parameter, presumably to replace the two other types of probability intervals in ordinary use. He remarked that if "the sampling curves are normal and equivariant," then the normed likelihood function will be proportional to $\exp\{-(\theta - \hat{\theta})^2 / 2\omega^2\}$, so that intervals can be found within which the normed likelihood exceeds any chosen value.

Ten years after postulating the superiority of the absolute criterion, and after having used it only for normally distributed observations, Fisher (1922a) realized the need to support his statement by an investigation of the sampling distribution of the maximum likelihood estimate for the parameters of a distribution of arbitrary form, satisfying only some regularity conditions, which he left the reader to explore. It is well known (see Hald, 1998, Section 28.5) that he proved the asymptotic normality and optimality of $\hat{\theta}$ within the class of statistics that are asymptotically normal with mean θ and variance of order n^{-1}, and that he linked the maximum likelihood estimate to sufficiency. His proofs are simpler and more general than Edgeworth's.

ACKNOWLEDGMENTS

I am grateful to the Editor and two referees for their remarks, which made me rewrite the original version of this paper.

REFERENCES

ALDRICH, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statist. Sci.* **12** 162–176.

CHAUVENET, W. (1863). On the method of least squares. In *A Manual of Spherical and Practical Astronomy* **2** 469–566. Lippincott, Philadelphia. (An appendix.)

EDGEWORTH, F. Y. (1883). The method of least squares. *Philos. Mag.* (*5*) **16** 360–375.

EDGEWORTH, F. Y. (1908). On the probable error of frequency constants. *J. Roy. Statist. Soc.* **71** 381–397, 499–512, 651–678.

EDGEWORTH, F. Y. (1909). Addendum on "Probable errors of frequency constants." *J. Roy. Statist. Soc.* **72** 81–90.

EDWARDS, A. W. F. (1974). The history of likelihood. *Internat. Statist. Rev.* **42** 9–15.

EDWARDS, A. W. F. (1997). What did Fisher mean by "inverse probability" in 1912–1922? *Statist. Sci.* **12** 177–184.

ENCKE, J. F. (1832–1834). Über die Methode der kleinsten Quadrate. In *Berliner Astronomisches Jahrbuch für 1834* 249–312; *für 1835* 253–320; *für 1836* 253–308.

FISHER, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41** 155–160. [Reprinted in *Statist. Sci.* **12** (1997) 39–41.]

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.

FISHER, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

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

FISHER, R. A. (1922b). The goodness of fit of regression formulæ, and the distribution of regression coefficients. *J. Roy. Statist. Soc.* **85** 597–612.

GAUSS, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.* Perthes et Besser, Hamburg.

GAUSS, C. F. (1816). Bestimmung der Genauigkeit der Beobachtungen. *Z. Astronom. Verwandte Wiss.* **1** 185–196.

GAUSS, C. F. (1823). Theoria combinationis observationum erroribus minimis obnoxiae. *Comm. Soc. Reg. Gottingensis Rec.* **5** 33–62, 63–90.

HAGEN, G. (1837). Grundzüge der Wahrscheinlichkeits-Rechnung. Dümmler, Berlin.

HALD, A. (1998). *A History of Mathematical Statistics from 1750 to 1930.* Wiley, New York.

HELMERT, F. R. (1876). Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehler direchter Beobachtungen gleicher Genauigkeit. *Astronom. Nachr.* **88** 113–132.

LAPLACE, P. S. DE (1812). *Théorie Analytique des Probabilités.* Courcier, Paris.

MERRIMAN, M. (1877). A list of writings relating to the method of least squares, with historical and critical notes. *Trans. Conn. Acad. Arts Sci.* **4** 151–232.

MERRIMAN, M. (1884). *A Text-Book on the Method of Least Squares*. Wiley, New York. [References are to the 8th ed. (1915).]

NEYMAN, J. AND SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

PEARSON, E. S. (1968). Some early correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments. *Biometrika* **55** 445–457.

PRATT, J. W. (1976). F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation. *Ann. Statist.* **4** 501–514.

SAVAGE, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.* **4** 441–483.

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

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace.* Macmillan, London.