Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikiped

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
العربية
Deutsch
Español
Français
□□□
Português
Русский
اردو
□□
文A 24 more
Edit links

Maximum likelihood estimation

From Wikipedia, the free encyclopedia

This article is about the statistical techniques. For computer data storage, see Partial response maximum likelihood.

In statistics, **maximum likelihood estimation** (**MLE**) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.[1] The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.[2][3][4]

If the likelihood function is differentiable, the derivative test for determining maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved explicitly; for instance, the ordinary least squares estimator maximizes the likelihood of the linear regression model.[5] Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function.

From the vantage point of Bayesian inference, MLE is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters. In frequentist inference, MLE is a special case of an extremum estimator, with the objective function being the likelihood.

Contents [hide]

Principles [edit]

From a statistical standpoint, a given set of observations are a random sample from an unknown population. The goal of maximum likelihood estimation is to make inferences about the population that is most likely to have generated the sample,[6] specifically the joint probability distribution of the random variables $\{y_1, y_2, \ldots\}$, not necessarily independent and identically distributed.

Associated with each probability distribution is a unique vector $\theta = \left[\theta_1,\, \theta_2,\, \ldots,\, \theta_k\right]^{\mathsf T}$ of parameters that index the probability distribution within a parametric family $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$, where Θ is called the parameter space, a finite-dimensional subset of Euclidean space. Evaluating the joint density at the observed data sample $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ gives a real-valued function,

$$L_n(\theta) = L_n(\theta; \mathbf{y}) = f_n(\mathbf{y}; \theta)$$

which is called the likelihood function. For independent and identically distributed random variables, $f_n(\mathbf{y}; \theta)$ will be the product of univariate density functions.

The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space,[6] that is

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{arg\;max}}\ \widehat{L}_n(\theta\,;\mathbf{y})$$

Intuitively, this selects the parameter values that make the observed data most probable. The specific value $\hat{\theta} = \hat{\theta}_n(\mathbf{y}) \in \Theta$ that maximizes the likelihood function L_n is called the maximum likelihood estimate. Further, if the function $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ so defined is measurable, then it is called the maximum likelihood estimator. It is generally a function defined over the sample space, i.e. taking a given sample as its argument. A sufficient but not necessary condition for its existence is for the likelihood function to be continuous over a parameter space Θ that is compact.[7] For an open Θ the likelihood function may increase without ever reaching a supremum value.

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log-likelihood:

$$\ell(\theta\,;\mathbf{y}) = \ln L_n(\theta\,;\mathbf{y}).$$

Since the logarithm is a monotonic function, the maximum of $\ell(\theta\,;\mathbf{y})$ occurs at the same value of θ as does the maximum of L_n.[8] If $\ell(\theta\,;\mathbf{y})$ is differentiable in θ, the necessary conditions for the occurrence of a maximum (or a minimum) are

$$\frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell}{\partial \theta_k} = 0,$$

known as the likelihood equations. For some models, these equations can be explicitly solved for $\widehat{\theta}$, but in general no closed-form solution to the maximization problem is known or available, and an MLE can only be found via numerical optimization. Another problem is that in finite samples, there may exist multiple roots for the likelihood equations.[9] Whether the identified root $\widehat{\theta}$ of the likelihood equations is indeed a (local) maximum depends on whether the matrix of second-order partial and cross-partial derivatives,

$$\mathbf{H}\left(\widehat{\theta}\right) = \begin{bmatrix} \left.\frac{\partial^2 \ell}{\partial \theta_1^2}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \dots & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_2^2}\right|_{\theta=\widehat{\theta}} & \dots & \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \dots & \left.\frac{\partial^2 \ell}{\partial \theta_k^2}\right|_{\theta=\widehat{\theta}} \end{bmatrix},$$

known as the Hessian matrix is negative semi-definite at $\widehat{\theta}$, which indicates local concavity. Conveniently, most common probability distributions—in particular the exponential family—are logarithmically concave.[10][11]

Restricted parameter space [edit]

Not to be confused with restricted maximum likelihood.

While the domain of the likelihood function—the parameter space—is generally a finite-dimensional subset of Euclidean space, additional restrictions sometimes need to be incorporated into the estimation process. The parameter space can be expressed as

$$\Theta = \left\{ \theta : \theta \in \mathbb{R}^k,\; h(\theta) = 0 \right\},$$

where $h(\theta) = \left[h_1(\theta), h_2(\theta), \ldots, h_r(\theta) \right]$ is a vector-valued function mapping $\mathbb{R}^k$ into $\mathbb{R}^r$. Estimating the true parameter θ belonging to Θ then, as a practical matter, means to find the maximum of the likelihood function subject to the constraint $h(\theta) = 0$.

Theoretically, the most natural approach to this constrained optimization problem is the method of substitution, that is "filling out" the restrictions $h_1, h_2, \ldots, h_r$ to a set $h_1, h_2, \ldots, h_r, h_{r+1}, \ldots, h_k$ in such a way that $h^\ast = \left[h_1, h_2, \ldots, h_k \right]$ is a one-to-one function from $\mathbb{R}^k$ to itself, and reparameterize the likelihood function by setting $\phi_i = h_i(\theta_1, \theta_2, \ldots, \theta_k)$.[12] Because of the invariance of the maximum likelihood estimator, the properties of the MLE apply to the restricted estimates also.[13] For instance, in a multivariate normal distribution the covariance matrix Σ must be positive-definite; this restriction can be imposed by replacing $\Sigma = \Gamma^{\mathsf T} \Gamma$, where Γ is a real upper triangular matrix and $\Gamma^{\mathsf T}$ is its transpose.[14]

In practice, restrictions are usually imposed using the method of Lagrange which, given the constraints as defined above, leads to the restricted likelihood equations

$$\frac{\partial \ell}{\partial \theta} - \frac{\partial h(\theta)^\mathsf{T}}{\partial \theta}\lambda = 0 \text{ and } h(\theta) = 0,$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$ is a column-vector of Lagrange multipliers and $\frac{\partial h(\theta)^\mathsf{T}}{\partial \theta}$ is the *k* × *r* Jacobian matrix of partial derivatives.[12] Naturally, if the constraints are nonbinding at the maximum, the Lagrange multipliers should be zero.[15] This in turn allows for a statistical test of the "validity" of the constraint, known as the Lagrange multiplier test.

Properties [edit]

A maximum likelihood estimator is an extremum estimator obtained by maximizing, as a function of *θ*, the objective function $\widehat{\ell\,}(\theta\,;x)$. If the data are independent and identically distributed, then we have

$$\widehat{\ell\,}(\theta\,;x) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i \mid \theta),$$

this being the sample analogue of the expected log-likelihood $\ell(\theta) = \operatorname{\mathbb E}[\,\ln f(x_i \mid \theta)\,]$, where this expectation is taken with respect to the true density.

Maximum-likelihood estimators have no optimum properties for finite samples, in the sense that (when evaluated on finite samples) other estimators may have greater concentration around the true parameter-value.[16] However, like other estimation methods, maximum likelihood estimation possesses a number of attractive limiting properties: As the sample size increases to infinity, sequences of maximum likelihood estimators have these properties:

- Consistency: the sequence of MLEs converges in probability to the value being estimated.
- Functional Invariance: If $\hat{\theta}$ is the maximum likelihood estimator for θ, and if $g(\theta)$ is any transformation of θ, then the maximum likelihood estimator for $\alpha = g(\theta)$ is $\hat{\alpha} = g(\hat{\theta})$.
- Efficiency, i.e. it achieves the Cramér–Rao lower bound when the sample size tends to infinity. This means that no consistent estimator has lower asymptotic mean squared error than the MLE (or other estimators attaining this bound), which also means that MLE has asymptotic normality.
- Second-order efficiency after correction for bias.

Consistency [edit]

Under the conditions outlined below, the maximum likelihood estimator is consistent. The consistency means that if the data were generated by $f(\cdot\,;\theta_0)$ and we have a sufficiently large number of observations *n*, then it is possible to find the value of θ_0 with arbitrary precision. In mathematical terms this means that as *n* goes to infinity the estimator $\widehat{\theta\,}$ converges in probability to its true value:

$$\widehat{\theta\,}_\mathrm{mle} \xrightarrow{\text{p}} \theta_0.$$

Under slightly stronger conditions, the estimator converges almost surely (or *strongly*):

$$\widehat{\theta\,}_\mathrm{mle} \xrightarrow{\text{a.s.}} \theta_0.$$

In practical applications, data is never generated by $f(\cdot\,;\theta_0)$. Rather, $f(\cdot\,;\theta_0)$ is a model, often in idealized form, of the process that generated by the data. It is a common aphorism in statistics that *all models are wrong*. Thus, true consistency does not occur in practical applications. Nevertheless, consistency is often considered to be a desirable property for an estimator to have.

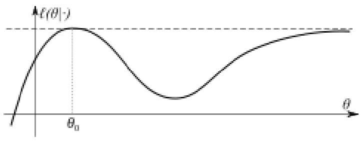
To establish consistency, the following conditions are sufficient.[17]

1. Identification of the model:

   $$\theta \neq \theta_0 \quad \Leftrightarrow \quad f(\cdot\mid\theta) \neq f(\cdot\mid\theta_0).$$

   In other words, different parameter values *θ* correspond to different distributions within the model. If this condition did not hold, there would be some value θ_1 such that θ_0 and θ_1 generate an identical distribution of the observable data. Then we would not be able to distinguish between these two parameters even with an infinite amount of data—these parameters would have been observationally equivalent.

   The identification condition is absolutely necessary for the ML estimator to be consistent. When this condition holds, the limiting likelihood function *ℓ*(*θ*|·) has unique global maximum at θ_0.
2. Compactness: the parameter space Θ of the model is compact.



   The identification condition establishes that the log-likelihood has a unique global maximum. Compactness implies that the likelihood cannot approach the maximum value arbitrarily close at some other point (as demonstrated for example in the picture on the right).

Compactness is only a sufficient condition and not a necessary condition. Compactness can be replaced by some other conditions, such as:

- both concavity of the log-likelihood function and compactness of some (nonempty) upper level sets of the log-likelihood function, or
- existence of a compact neighborhood *N* of θ_0 such that outside of *N* the log-likelihood function is less than the maximum by at least some $\varepsilon > 0$.

3. Continuity: the function ln $f(x \mid \theta)$ is continuous in θ for almost all values of x:

$$\mathrm{P}\Big[\;\ln f(x\mid\theta)\;\in\;C^0(\Theta)\;\Big]=1.$$

   The continuity here can be replaced with a slightly weaker condition of upper semi-continuity.

4. Dominance: there exists $D(x)$ integrable with respect to the distribution $f(x \mid \theta_0)$ such that

$$\Big|\ln f(x\mid\theta)\Big|<D(x)\quad\text{ for all }\theta\in\Theta.$$

   By the uniform law of large numbers, the dominance condition together with continuity establish the uniform convergence in probability of the log-likelihood:

$$\sup_{\theta\in\Theta}\left|\widehat{\ell\,}(\theta\mid x)-\ell(\theta)\,\right|\ \xrightarrow{\text{p}}\ 0.$$

The dominance condition can be employed in the case of i.i.d. observations. In the non-i.i.d. case, the uniform convergence in probability can be checked by showing that the sequence $\widehat{\ell\,}(\theta\mid x)$ is stochastically equicontinuous. If one wants to demonstrate that the ML estimator $\widehat{\theta\,}$ converges to θ_0 almost surely, then a stronger condition of uniform convergence almost surely has to be imposed:

$$\sup_{\theta\in\Theta}\left\|\;\widehat{\ell\,}(\theta\mid x)-\ell(\theta)\;\right\|\ \xrightarrow{\text{a.s.}}\ 0.$$

Additionally, if (as assumed above) the data were generated by $f(\cdot\,;\theta_0)$, then under certain conditions, it can also be shown that the maximum likelihood estimator converges in distribution to a normal distribution. Specifically,[18]

$$\sqrt{n}\left(\widehat{\theta}_\mathrm{mle}-\theta_0\right)\ \xrightarrow{d}\ \mathcal{N}\left(0,\,I^{-1}\right)$$

where I is the Fisher information matrix.

Functional invariance [edit]

The maximum likelihood estimator selects the parameter value which gives the observed data the largest possible probability (or probability density, in the continuous case). If the parameter consists of a number of components, then we define their separate maximum likelihood estimators, as the corresponding component of the MLE of the complete parameter. Consistent with this, if $\widehat{\theta\,}$ is the MLE for θ, and if $g(\theta)$ is any transformation of θ, then the MLE for $\alpha = g(\theta)$ is by definition[19]

$$\widehat{\alpha}=g(\,\widehat{\theta\,}\,).\,$$

It maximizes the so-called profile likelihood:

$$\bar{L}(\alpha)=\sup_{\theta:\alpha=g(\theta)}L(\theta).\,$$

The MLE is also invariant with respect to certain transformations of the data. If $y = g(x)$ where g is one to one and does not depend on the parameters to be estimated, then the density functions satisfy

$$f_Y(y)=\frac{f_X(x)}{|g'(x)|}$$

and hence the likelihood functions for X and Y differ only by a factor that does not depend on the model parameters.

For example, the MLE parameters of the log-normal distribution are the same as those of the normal distribution fitted to the logarithm of the data.

Efficiency [edit]

As assumed above, the data were generated by $f(\cdot\,;\theta_0)$, then under certain conditions, it can also be shown that the maximum likelihood estimator converges in distribution to a normal distribution. It is $\sqrt{n}$-consistent and asymptotically efficient, meaning that it reaches the Cramér–Rao bound. Specifically,[18]

$$\sqrt{n}(\widehat{\theta}_\text{mle}-\theta_0)\ \xrightarrow{d}\ \mathcal{N}(0,\ I^{-1}),$$

where I is the Fisher information matrix:

$$I_{jk} = \operatorname{E}\left[\, -\frac{\partial^2 \ln f_{\theta_0}(X_t)}{\partial\theta_j\,\partial\theta_k} \,\right].$$

In particular, it means that the bias of the maximum likelihood estimator is equal to zero up to the order $1/\sqrt{n}$.

Second-order efficiency after correction for bias [edit]

However, when we consider the higher-order terms in the expansion of the distribution of this estimator, it turns out that θ_{mle} has bias of order $1/n$. This bias is equal to (componentwise)[20]

$$b_h \equiv \operatorname{E}\left[\, (\hat\theta_{\text{mle}} - \theta_0)_h \,\right] = \frac{1}{n} \sum_{i,j,k=1}^{m} I^{hi} I^{jk} \left(\frac{1}{2} K_{ijk} + J_{j,ik} \right)$$

where I^{jk} denotes the (*j,k*)-th component of the *inverse* Fisher information matrix I^{-1}, and

$$\tfrac{1}{2} K_{ijk} + J_{j,ik} = \operatorname{E}\left[\, \frac{1}{2} \frac{\partial^3 \ln f_{\theta_0}(X_t)}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k} + \frac{\partial \ln f_{\theta_0}(X_t)}{\partial\theta_j} \frac{\partial^2 \ln f_{\theta_0}(X_t)}{\partial\theta_i\,\partial\theta_k} \,\right].$$

Using these formulae it is possible to estimate the second-order bias of the maximum likelihood estimator, and *correct* for that bias by subtracting it:

$$\widehat\theta^{*}_{\text{mle}} = \widehat\theta_{\text{mle}} - \widehat b\,.$$

This estimator is unbiased up to the terms of order $1/n$, and is called the bias-corrected maximum likelihood estimator.

This bias-corrected estimator is *second-order efficient* (at least within the curved exponential family), meaning that it has minimal mean squared error among all second-order bias-corrected estimators, up to the terms of the order $1/n^2$. It is possible to continue this process, that is to derive the third-order bias-correction term, and so on. However the maximum likelihood estimator is *not* third-order efficient.[21]

Relation to Bayesian inference [edit]

A maximum likelihood estimator coincides with the most probable Bayesian estimator given a uniform prior distribution on the parameters. Indeed, the maximum a posteriori estimate is the parameter θ that maximizes the probability of θ given the data, given by Bayes' theorem:

$$\operatorname{P}(\theta \mid x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n \mid \theta)\operatorname{P}(\theta)}{\operatorname{P}(x_1, x_2, \ldots, x_n)}$$

where $P(\theta)$ is the prior distribution for the parameter θ and where $\operatorname{P}(x_1, x_2, \ldots, x_n)$ is the probability of the data averaged over all parameters. Since the denominator is independent of θ, the Bayesian estimator is obtained by maximizing $f(x_1, x_2, \ldots, x_n \mid \theta)\operatorname{P}(\theta)$ with respect to θ. If we further assume that the prior $P(\theta)$ is a uniform distribution, the Bayesian estimator is obtained by maximizing the likelihood function $f(x_1, x_2, \ldots, x_n \mid \theta)$. Thus the Bayesian estimator coincides with the maximum likelihood estimator for a uniform prior distribution $\operatorname{P}(\theta)$.

Application of maximum-likelihood estimation in Bayes decision theory [edit]

In many practical applications in machine learning, maximum-likelihood estimation is used as the model for parameter estimation.

The Bayesian Decision theory is about designing a classifier that minimizes total expected risk, especially, when the costs (the loss function) associated with different decisions are equal, the classifier is minimizing the error over the whole distribution.[22]

Thus, the Bayes Decision Rule is stated as "decide w_1 if $P(w_1|x) > P(w_2|x)$; otherwise w_2", where w_1, w_2 are predictions of different classes. From a perspective of minimizing error, it can also be stated as

$w = \arg\min_w \int_{-\infty}^{\infty} P(\text{error} \mid x) P(x)\, dx$, where $P(\text{error} \mid x) = P(w_1 \mid x)$ if we decide w_2 and $P(\text{error} \mid x) = P(w_2|x)$ if we decide w_1.

By applying Bayes' theorem : $P(w_i \mid x) = \frac{P(x \mid w_i) P(w_i)}{P(x)}$, and if we further assume the zero/one loss function, which is a same loss for all errors, the Bayes Decision rule can be reformulated as:

$h_{\text{Bayes}} = \arg\max_w P(x \mid w) P(w)$, where h_{Bayes} is the prediction and $P(w)$ is the priori probability.

Relation to minimizing Kullback–Leibler divergence and cross entropy [edit]

Finding $\hat\theta$ that maximizes the likelihood is asymptotically equivalent to finding the $\hat\theta$ that defines a probability distribution ($Q_{\hat\theta}$) that has a minimal distance, in terms of Kullback–Leibler divergence, to the real probability distribution from which our data was generated (i.e., generated by P_{θ_0}).[23] In an ideal world, P and Q are the same (and the only thing unknown is θ that defines P), but even if they are not and the model we use is misspecified, still the MLE will give us the "closest" distribution (within the restriction of a model Q that depends on $\hat{\theta}$) to the real distribution P_{θ_0}.[24]

Proof. [show]

Since cross entropy is just Shannon's Entropy plus KL divergence, and since the Entropy of P_{θ_0} is constant, then the MLE is also asymptotically minimizing cross entropy.[25]

Examples [edit]

Discrete uniform distribution [edit]

Main article: German tank problem

Consider a case where *n* tickets numbered from 1 to *n* are placed in a box and one is selected at random (*see uniform distribution*); thus, the sample size is 1. If *n* is unknown, then the maximum likelihood estimator $\widehat{n}$ of *n* is the number *m* on the drawn ticket. (The likelihood is 0 for *n* < *m*, $1/n$ for *n* ≥ *m*, and this is greatest when *n* = *m*. Note that the maximum likelihood estimate of *n* occurs at the lower extreme of possible values {*m*, *m* + 1, ...}, rather than somewhere in the "middle" of the range of possible values, which would result in less bias.) The expected value of the number *m* on the drawn ticket, and therefore the expected value of $\widehat{n}$, is (*n* + 1)/2. As a result, with a sample size of 1, the maximum likelihood estimator for *n* will systematically underestimate *n* by (*n* − 1)/2.

Discrete distribution, finite parameter space [edit]

Suppose one wishes to determine just how biased an unfair coin is. Call the probability of tossing a 'head' *p*. The goal then becomes to determine *p*.

Suppose the coin is tossed 80 times: i.e. the sample might be something like x_1 = H, x_2 = T, ..., x_{80} = T, and the count of the number of heads "H" is observed.

The probability of tossing tails is 1 − *p* (so here *p* is *θ* above). Suppose the outcome is 49 heads and 31 tails, and suppose the coin was taken from a box containing three coins: one which gives heads with probability *p* = $1/3$, one which gives heads with probability *p* = $1/2$ and another which gives heads with probability *p* = $2/3$. The coins have lost their labels, so which one it was is unknown. Using maximum likelihood estimation, the coin that has the largest likelihood can be found, given the data that were observed. By using the probability mass function of the binomial distribution with sample size equal to 80, number successes equal to 49 but for different values of *p* (the "probability of success"), the likelihood function (defined below) takes one of three values:

$$\operatorname{P}\left[\mathrm{H}=49 \mid p=\tfrac{1}{3}\right] = \binom{80}{49}\left(\tfrac{1}{3}\right)^{49}\left(1-\tfrac{1}{3}\right)^{31} \approx 0.000,$$

$$\operatorname{P}\left[\mathrm{H}=49 \mid p=\tfrac{1}{2}\right] = \binom{80}{49}\left(\tfrac{1}{2}\right)^{49}\left(1-\tfrac{1}{2}\right)^{31} \approx 0.012,$$

$$\operatorname{P}\left[\mathrm{H}=49 \mid p=\tfrac{2}{3}\right] = \binom{80}{49}\left(\tfrac{2}{3}\right)^{49}\left(1-\tfrac{2}{3}\right)^{31} \approx 0.054.$$

The likelihood is maximized when *p* = $2/3$, and so this is the *maximum likelihood estimate* for *p*.

Discrete distribution, continuous parameter space [edit]

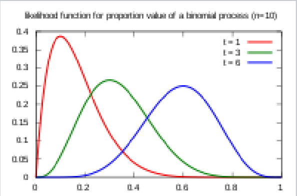
Now suppose that there was only one coin but its *p* could have been any value 0 ≤ *p* ≤ 1. The likelihood function to be maximised is

$$L(p) = f_D(\mathrm{H}=49 \mid p) = \binom{80}{49} p^{49}(1-p)^{31},$$

and the maximisation is over all possible values 0 ≤ *p* ≤ 1.

One way to maximize this function is by differentiating with respect to *p* and setting to zero:

$$\begin{aligned}
0 &= \frac{\partial}{\partial p}\left(\binom{80}{49} p^{49}(1-p)^{31}\right), \\
0 &= 49p^{48}(1-p)^{31} - 31p^{49}(1-p)^{30} \\
&= p^{48}(1-p)^{30}\left[49(1-p) - 31p\right] \\
&= p^{48}(1-p)^{30}\left[49 - 80p\right].
\end{aligned}$$



likelihood function for proportion value of a binomial process (*n* = 10)

This is a product of three terms. The first term is 0 when *p* = 0. The second is 0 when *p* = 1. The third is zero when *p* = 49⁄80. The solution that maximizes the likelihood is clearly *p* = 49⁄80 (since *p* = 0 and *p* = 1 result in a likelihood of 0). Thus the *maximum likelihood estimator* for *p* is 49⁄80.

This result is easily generalized by substituting a letter such as *s* in the place of 49 to represent the observed number of 'successes' of our Bernoulli trials, and a letter such as *n* in the place of 80 to represent the number of Bernoulli trials. Exactly the same calculation yields *s*⁄*n* which is the maximum likelihood estimator for any sequence of *n* Bernoulli trials resulting in *s* 'successes'.

Continuous distribution, continuous parameter space [edit]

For the normal distribution $\mathcal{N}(\mu,\sigma^2)$ which has probability density function

$$f(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

the corresponding probability density function for a sample of n independent identically distributed normal random variables (the likelihood) is

$$f(x_1,\ldots,x_n\mid\mu,\sigma^2)=\prod_{i=1}^{n}f(x_i\mid\mu,\sigma^2)=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

This family of distributions has two parameters: $\theta=(\mu,\sigma)$; so we maximize the likelihood, $\mathcal{L}(\mu,\sigma)=f(x_1,\ldots,x_n\mid\mu,\sigma)$, over both parameters simultaneously, or if possible, individually.

Since the logarithm function itself is a continuous strictly increasing function over the range of the likelihood, the values which maximize the likelihood will also maximize its logarithm (the log-likelihood itself is not necessarily strictly increasing). The log-likelihood can be written as follows:

$$\log\Bigl(\mathcal{L}(\mu,\sigma)\Bigr)=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\,x_i-\mu\,)^2$$

(Note: the log-likelihood is closely related to information entropy and Fisher information.)

We now compute the derivatives of this log-likelihood as follows.

$$0=\frac{\partial}{\partial\mu}\log\Bigl(\mathcal{L}(\mu,\sigma)\Bigr)=0-\frac{-2n(\bar{x}-\mu)}{2\sigma^2}.$$

where $\bar{x}$ is the sample mean. This is solved by

$$\widehat{\mu}=\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n}.$$

This is indeed the maximum of the function, since it is the only turning point in μ and the second derivative is strictly less than zero. Its expected value is equal to the parameter μ of the given distribution,

$$\operatorname{E}\bigl[\,\widehat{\mu}\,\bigr]=\mu,$$

which means that the maximum likelihood estimator $\widehat{\mu}$ is unbiased.

Similarly we differentiate the log-likelihood with respect to σ and equate to zero:

$$0=\frac{\partial}{\partial\sigma}\log\Bigl(\mathcal{L}(\mu,\sigma)\Bigr)=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(\,x_i-\mu\,)^2.$$

which is solved by

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Inserting the estimate $\mu=\widehat{\mu}$ we obtain

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j.$$

To calculate its expected value, it is convenient to rewrite the expression in terms of zero-mean random variables (statistical error) $\delta_i\equiv\mu-x_i$. Expressing the estimate in these variables yields

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(\mu-\delta_i)^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mu-\delta_i)(\mu-\delta_j).$$

Simplifying the expression above, utilizing the facts that $\operatorname{E}\left[\;\delta_i\;\right] = 0$ and $\operatorname{E}\left[\;\delta_i^2\;\right] = \sigma^2$, allows us to obtain

$$\operatorname{E}\left[\;\widehat\sigma^2\;\right] = \frac{n-1}{n}\sigma^2.$$

This means that the estimator $\widehat\sigma$ is biased. However, $\widehat\sigma$ is consistent.

Formally we say that the *maximum likelihood estimator* for $\theta = (\mu, \sigma^2)$ is

$$\widehat{\theta} = \left(\widehat{\mu}, \widehat{\sigma}^2\right).$$

In this case the MLEs could be obtained individually. In general this may not be the case, and the MLEs would have to be obtained simultaneously.

The normal log-likelihood at its maximum takes a particularly simple form:

$$\log\Bigl(\mathcal{L}(\widehat\mu,\widehat\sigma)\Bigr) = \frac{\,-n\;\;}{2} \bigl(\,\log(2\pi\widehat\sigma^2) + 1\,\bigr)$$

This maximum log-likelihood can be shown to be the same for more general least squares, even for non-linear least squares. This is often used in determining likelihood-based approximate confidence intervals and confidence regions, which are generally more accurate than those using the asymptotic normality discussed above.

Non-independent variables [edit]

It may be the case that variables are correlated, that is, not independent. Two random variables y_1 and y_2 are independent only if their joint probability density function is the product of the individual probability density functions, i.e.

$$f(y_1,y_2)=f(y_1)f(y_2)$$

Suppose one constructs an order-*n* Gaussian vector out of random variables $(y_1,\ldots,y_n)$, where each variable has means given by $(\mu_1, \ldots, \mu_n)$. Furthermore, let the covariance matrix be denoted by $\mathit\Sigma$. The joint probability density function of these *n* random variables is then follows a multivariate normal distribution given by:

$$f(y_1,\ldots,y_n)=\frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathit\Sigma)}} \exp\left(-\frac{1}{2} \left[y_1-\mu_1,\ldots,y_n-\mu_n\right]\mathit\Sigma^{-1} \left[y_1-\mu_1,\ldots,y_n-\mu_n\right]^\mathrm{T} \right)$$

In the bivariate case, the joint probability density function is given by:

$$f(y_1,y_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{(y_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} \right) \right]$$

In this and other cases where a joint density function exists, the likelihood function is defined as above, in the section "principles," using this density.

Example [edit]

$X_1,\ X_2,\ldots,\ X_m$ are counts in cells / boxes 1 up to m; each box has a different probability (think of the boxes being bigger or smaller) and we fix the number of balls that fall to be n:$x_1+x_2+\cdots+x_m=n$. The probability of each box is p_i, with a constraint: $p_1+p_2+\cdots+p_m=1$. This is a case in which the X_i *s* are not independent, the joint probability of a vector $x_1,\ x_2,\ldots,x_m$ is called the multinomial and has the form:

$$f(x_1,x_2,\ldots,x_m\mid p_1,p_2,\ldots,p_m)=\frac{n!}{\Pi x_i!}\Pi p_i^{x_i}= \binom{n}{x_1,x_2,\ldots,x_m} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

Each box taken separately against all the other boxes is a binomial and this is an extension thereof.

The log-likelihood of this is:

$$\ell(p_1,p_2,\ldots,p_m)=\log n!-\sum_{i=1}^m \log x_i!+\sum_{i=1}^m x_i\log p_i$$

The constraint has to be taken into account and use the Lagrange multipliers:

$$L(p_1,p_2,\ldots,p_m,\lambda)=\ell(p_1,p_2,\ldots,p_m)+\lambda\left(1-\sum_{i=1}^m p_i\right)$$

By posing all the derivatives to be 0, the most natural estimate is derived

$$\hat{p}_i=\frac{x_i}{n}$$

Maximizing log likelihood, with and without constraints, can be an unsolvable problem in closed form, then we have to use iterative procedures.

Iterative procedures [edit]

Except for special cases, the likelihood equations

$$\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta} = 0$$

cannot be solved explicitly for an estimator $\hat{\theta} = \hat{\theta}(\mathbf{y})$. Instead, they need to be solved iteratively: starting from an initial guess of θ (say $\hat{\theta}_1$), one seeks to obtain a convergent sequence $\left\{\hat{\theta}_r\right\}$. Many methods for this kind of optimization problem are available,[26][27] but the most commonly used ones are algorithms based on an updating formula of the form

$$\hat{\theta}_{r+1} = \hat{\theta}_r + \eta_r \mathbf{d}_r\left(\hat{\theta}\right)$$

where the vector $\mathbf{d}_r\left(\hat{\theta}\right)$ indicates the descent direction of the *r*th "step," and the scalar η_r captures the "step length,"[28][29] also known as the learning rate.[30]

Gradient descent method [edit]

(Note: here it is a maximization problem, so the sign before gradient is flipped)

$\eta_r \in \mathbb{R}^+$ that is small enough for convergence and $\mathbf{d}_r\left(\hat{\theta}\right) = \nabla \ell\left(\hat{\theta}_r;\mathbf{y}\right)$

Gradient descent method requires to calculate the gradient at the rth iteration, but no need to calculate the inverse of second-order derivative, i.e., the Hessian matrix. Therefore, it is computationally faster than Newton-Raphson method.

Newton–Raphson method [edit]

$\eta_r = 1$ and $\mathbf{d}_r\left(\hat{\theta}\right) = -\mathbf{H}_r^{-1}\left(\hat{\theta}\right)\mathbf{s}_r\left(\hat{\theta}\right)$

where $\mathbf{s}_r(\hat{\theta})$ is the score and $\mathbf{H}_r^{-1}\left(\hat{\theta}\right)$ is the inverse of the Hessian matrix of the log-likelihood function, both evaluated the *r*th iteration.[31][32] But because the calculation of the Hessian matrix is computationally costly, numerous alternatives have been proposed. The popular Berndt–Hall–Hall–Hausman algorithm approximates the Hessian with the outer product of the expected gradient, such that

$$\mathbf{d}_r\left(\hat{\theta}\right) = -\left[\frac{1}{n}\sum_{t=1}^{n}\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta}\left(\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta}\right)^{\mathsf{T}}\right]^{-1}\mathbf{s}_r\left(\hat{\theta}\right)$$

Quasi-Newton methods [edit]

Other quasi-Newton methods use more elaborate secant updates to give approximation of Hessian matrix.

Davidon–Fletcher–Powell formula [edit]

DFP formula finds a solution that is symmetric, positive-definite and closest to the current approximate value of second-order derivative:

$$\mathbf{H}_{k+1} = \left(I - \gamma_k y_k s_k^{\mathsf{T}}\right)\mathbf{H}_k\left(I - \gamma_k s_k y_k^{\mathsf{T}}\right) + \gamma_k y_k y_k^{\mathsf{T}},$$

where

$$y_k = \nabla \ell(x_k + s_k) - \nabla \ell(x_k),$$
$$\gamma_k = \frac{1}{y_k^T s_k},$$
$$s_k = x_{k+1} - x_k.$$

Broyden–Fletcher–Goldfarb–Shanno algorithm [edit]

BFGS also gives a solution that is symmetric and positive-definite:

$$B_{k+1} = B_k + \frac{y_k y_k^{\mathsf{T}}}{y_k^{\mathsf{T}} s_k} - \frac{B_k s_k s_k^{\mathsf{T}} B_k^{\mathsf{T}}}{s_k^{\mathsf{T}} B_k s_k},$$

where

$$y_k = \nabla \ell(x_k + s_k) - \nabla \ell(x_k),$$
$$s_k = x_{k+1} - x_k.$$

BFGS method is not guaranteed to converge unless the function has a quadratic Taylor expansion near an optimum. However, BFGS can have acceptable performance even for non-smooth optimization instances

Fisher's scoring [edit]

Another popular method is to replace the Hessian with the Fisher information matrix, $\mathcal{I}(\theta) = \mathbf{E}\left[\mathbf{H}_r\left(\hat{\theta}\right)\right]$, giving us the Fisher scoring algorithm. This procedure is standard in the estimation of many methods, such as generalized linear models.

Although popular, quasi-Newton methods may converge to a stationary point that is not necessarily a local or global maximum,[33] but rather a local minimum or a saddle point. Therefore, it is important to assess the validity of the obtained solution to the likelihood equations, by verifying that the Hessian, evaluated at the solution, is both negative definite and well-conditioned.[34]

History [edit]

Early users of maximum likelihood were Carl Friedrich Gauss, Pierre-Simon Laplace, Thorvald N. Thiele, and Francis Ysidro Edgeworth.[35][36] However, its widespread use rose between 1912 and 1922 when Ronald Fisher recommended, widely popularized, and carefully analyzed maximum-likelihood estimation (with fruitless attempts at proofs).[37]

Maximum-likelihood estimation finally transcended heuristic justification in a proof published by Samuel S. Wilks in 1938, now called Wilks' theorem.[38] The theorem shows that the error in the logarithm of likelihood values for estimates from multiple independent observations is asymptotically χ^2-distributed, which enables convenient determination of a confidence region around any estimate of the parameters. The only difficult part of Wilks' proof depends on the expected value of the Fisher information matrix, which is provided by a theorem proven by Fisher.[39] Wilks continued to improve on the generality of the theorem throughout his life, with his most general proof published in 1962.[40]

Reviews of the development of maximum likelihood estimation have been provided by a number of authors.[41][42][43][44][45][46][47][48]

Ronald Fisher in 1913

See also [edit]

Other estimation methods [edit]

Mathematics portal

- Generalized method of moments are methods related to the likelihood equation in maximum likelihood estimation
- M-estimator, an approach used in robust statistics
- Maximum a posteriori (MAP) estimator, for a contrast in the way to calculate estimators when prior knowledge is postulated
- Maximum spacing estimation, a related method that is more robust in many situations
- Maximum entropy estimation
- Method of moments (statistics), another popular method for finding parameters of distributions
- Method of support, a variation of the maximum likelihood technique
- Minimum distance estimation
- Partial likelihood methods for panel data
- Quasi-maximum likelihood estimator, an MLE estimator that is misspecified, but still consistent
- Restricted maximum likelihood, a variation using a likelihood function calculated from a transformed set of data

Related concepts [edit]

- Akaike information criterion, a criterion to compare statistical models, based on MLE
- Extremum estimator, a more general class of estimators to which MLE belongs
- Fisher information, information matrix, its relationship to covariance matrix of ML estimates
- Mean squared error, a measure of how 'good' an estimator of a distributional parameter is (be it the maximum likelihood estimator or some other estimator)
- RANSAC, a method to estimate parameters of a mathematical model given data that contains outliers
- Rao–Blackwell theorem, which yields a process for finding the best possible unbiased estimator (in the sense of having minimal mean squared error); the MLE is often a good starting place for the process
- Wilks' theorem provides a means of estimating the size and shape of the region of roughly equally-probable estimates for the population's parameter values, using the information from a single sample, using a chi-squared distribution

References [edit]

1. ^ Rossi, Richard J. (2018). *Mathematical Statistics : An Introduction to Likelihood Based Inference*. New York: John

Wiley & Sons. p. 227. ISBN 978-1-118-77104-4.

2. ^ Hendry, David F.; Nielsen, Bent (2007). *Econometric Modeling: A Likelihood Approach*. Princeton: Princeton University Press. ISBN 978-0-691-13128-3.
3. ^ Chambers, Raymond L.; Steel, David G.; Wang, Suojin; Welsh, Alan (2012). *Maximum Likelihood Estimation for Sample Surveys*. Boca Raton: CRC Press. ISBN 978-1-58488-632-7.
4. ^ Ward, Michael Don; Ahlquist, John S. (2018). *Maximum Likelihood for Social Science : Strategies for Analysis*. New York: Cambridge University Press. ISBN 978-1-107-18582-1.
5. ^ Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. (1992). "Least Squares as a Maximum Likelihood Estimator". *Numerical Recipes in FORTRAN: The Art of Scientific Computing* (2nd ed.). Cambridge: Cambridge University Press. pp. 651–655. ISBN 0-521-43064-X.
6. ^ *a* *b* Myung, I. J. (2003). "Tutorial on Maximum Likelihood Estimation". *Journal of Mathematical Psychology*. **47** (1): 90–100. doi:10.1016/S0022-2496(02)00028-7.
7. ^ Gourieroux, Christian; Monfort, Alain (1995). *Statistics and Econometrics Models*. Cambridge University Press. p. 161. ISBN 0-521-40551-3.
8. ^ Kane, Edward J. (1968). *Economic Statistics and Econometrics*. New York: Harper & Row. p. 179.
9. ^ Small, Christoper G.; Wang, Jinfang (2003). "Working with Roots". *Numerical Methods for Nonlinear Estimating Equations*. Oxford University Press. pp. 74–124. ISBN 0-19-850688-0.
10. ^ Kass, Robert E.; Vos, Paul W. (1997). *Geometrical Foundations of Asymptotic Inference*. New York: John Wiley & Sons. p. 14. ISBN 0-471-82668-5.
11. ^ Papadopoulos, Alecos (September 25, 2013). "Why we always put log() before the joint pdf when we use MLE (Maximum likelihood Estimation)?". *Stack Exchange*.
12. ^ *a* *b* Silvey, S. D. (1975). *Statistical Inference*. London: Chapman and Hall. p. 79. ISBN 0-412-13820-4.
13. ^ Olive, David (2004). "Does the MLE Maximize the Likelihood?" (PDF).
14. ^ Schwallie, Daniel P. (1985). "Positive Definite Maximum Likelihood Covariance Estimators". *Economics Letters*. **17** (1–2): 115–117. doi:10.1016/0165-1765(85)90139-9.
15. ^ Magnus, Jan R. (2017). *Introduction to the Theory of Econometrics*. Amsterdam: VU University Press. pp. 64–65. ISBN 978-90-8659-766-6.
16. ^ Pfanzagl (1994, p. 206)
17. ^ By Theorem 2.5 in Newey, Whitney K.; McFadden, Daniel (1994). "Chapter 36: Large sample estimation and hypothesis testing". In Engle, Robert; McFadden, Dan (eds.). *Handbook of Econometrics, Vol.4*. Elsevier Science. pp. 2111–2245. ISBN 978-0-444-88766-5.
18. ^ *a* *b* By Theorem 3.3 in Newey, Whitney K.; McFadden, Daniel (1994). "Chapter 36: Large sample estimation and hypothesis testing". In Engle, Robert; McFadden, Dan (eds.). *Handbook of Econometrics, Vol.4*. Elsevier Science. pp. 2111–2245. ISBN 978-0-444-88766-5.
19. ^ Zacks, Shelemyahu (1971). *The Theory of Statistical Inference*. New York: John Wiley & Sons. p. 223. ISBN 0-471-98103-6.
20. ^ See formula 20 in Cox, David R.; Snell, E. Joyce (1968). "A general definition of residuals". *Journal of the Royal Statistical Society, Series B*. **30** (2): 248–275. JSTOR 2984505.
21. ^ Kano, Yutaka (1996). "Third-order efficiency implies fourth-order efficiency". *Journal of the Japan Statistical Society*. **26**: 101–117. doi:10.14490/jjss1995.26.101.
22. ^ Christensen, Henrik I., *Bayesian Decision Theory - CS 7616 - Pattern Recognition* (PDF) (presentation)
23. ^ cmplx96 (https://stats.stackexchange.com/users/177679/cmplx96), Kullback–Leibler divergence, URL (version: 2017-11-18): https://stats.stackexchange.com/q/314472 (at the youtube video, look at minutes 13 to 25)
24. ^ Introduction to Statistical Inference | Stanford (Lecture 16 — MLE under model misspecification)
25. ^ Sycorax says Reinstate Monica (https://stats.stackexchange.com/users/22311/sycorax-says-reinstate-monica), the relationship between maximizing the likelihood and minimizing the cross-entropy, URL (version: 2019-11-06): https://stats.stackexchange.com/q/364237
26. ^ Fletcher, R. (1987). *Practical Methods of Optimization* (Second ed.). New York: John Wiley & Sons. ISBN 0-471-91547-5.
27. ^ Nocedal, Jorge; Wright, Stephen J. (2006). *Numerical Optimization* (Second ed.). New York: Springer. ISBN 0-387-30303-0.
28. ^ Daganzo, Carlos (1979). *Multinomial Probit : The Theory and its Application to Demand Forecasting*. New York: Academic Press. pp. 61–78. ISBN 0-12-201150-3.
29. ^ Gould, William; Pitblado, Jeffrey; Poi, Brian (2010). *Maximum Likelihood Estimation with Stata* (Fourth ed.). College Station: Stata Press. pp. 13–20. ISBN 978-1-59718-078-8.
30. ^ Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press. p. 247. ISBN 978-0-262-01802-9.
31. ^ Amemiya, Takeshi (1985). *Advanced Econometrics*. Cambridge: Harvard University Press. pp. 137–138. ISBN 0-674-00560-0.
32. ^ Sargan, Denis (1988). "Methods of Numerical Optimization". *Lecture Notes on Advanced Econometric Theory*. Oxford: Basil Blackwell. pp. 161–169. ISBN 0-631-14956-2.
33. ^ See theorem 10.1 in Avriel, Mordecai (1976). *Nonlinear Programming: Analysis and Methods*. Englewood Cliffs: Prentice-Hall. pp. 293–294. ISBN 9780486432274.
34. ^ Gill, Philip E.; Murray, Walter; Wright, Margaret H. (1981). *Practical Optimization*. London: Academic Press. pp. 312–313. ISBN 0-12-283950-1.
35. ^ Edgeworth, Francis Y. (Sep 1908). "On the probable errors of frequency-constants". *Journal of the Royal Statistical Society*. **71** (3): 499–512. doi:10.2307/2339293. JSTOR 2339293.
36. ^ Edgeworth, Francis Y. (Dec 1908). "On the probable errors of frequency-constants". *Journal of the Royal Statistical Society*. **71** (4): 651–678. doi:10.2307/2339378. JSTOR 2339378.
37. ^ Pfanzagl, Johann, with the assistance of R. Hamböker (1994). *Parametric Statistical Theory*. Walter de Gruyter. pp. 207–208. ISBN 978-3-11-013863-4.
38. ^ Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". *Annals of Mathematical Statistics*. **9**: 60–62. doi:10.1214/aoms/1177732360.
39. ^ Owen, Art B. (2001). *Empirical Likelihood*. London: Chapman & Hall/Boca Raton, FL: CRC Press. ISBN 978-1584880714.
40. ^ Wilks, Samuel S. (1962), *Mathematical Statistics*, New York: John Wiley & Sons. ISBN 978-0471946502.
41. ^ Savage, Leonard J. (1976). "On rereading R. A. Fisher". *The Annals of Statistics*. **4** (3): 441–500. doi:10.1214/aos/1176343456. JSTOR 2958221.
42. ^ Pratt, John W. (1976). "F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation". *The Annals of Statistics*. **4** (3): 501–514. doi:10.1214/aos/1176343457. JSTOR 2958222.

43. ^ Stigler, Stephen M. (1978). "Francis Ysidro Edgeworth, statistician". *Journal of the Royal Statistical Society, Series A*. **141** (3): 287–322. doi:10.2307/2344804. JSTOR 2344804.
44. ^ Stigler, Stephen M. (1986). *The history of statistics: the measurement of uncertainty before 1900*. Harvard University Press. ISBN 978-0-674-40340-6.
45. ^ Stigler, Stephen M. (1999). *Statistics on the table: the history of statistical concepts and methods*. Harvard University Press. ISBN 978-0-674-83601-3.
46. ^ Hald, Anders (1998). *A history of mathematical statistics from 1750 to 1930*. New York, NY: Wiley. ISBN 978-0-471-17912-2.
47. ^ Hald, Anders (1999). "On the history of maximum likelihood in relation to inverse probability and least squares". *Statistical Science*. **14** (2): 214–222. doi:10.1214/ss/1009212248. JSTOR 2676741.
48. ^ Aldrich, John (1997). "R. A. Fisher and the making of maximum likelihood 1912–1922". *Statistical Science*. **12** (3): 162–176. doi:10.1214/ss/1030037906. MR 1617519.

Further reading [edit]

- Cramer, J. S. (1986). *Econometric Applications of Maximum Likelihood Methods*. New York: Cambridge University Press. ISBN 0-521-25317-9.
- Eliason, Scott R. (1993). *Maximum Likelihood Estimation : Logic and Practice*. Newbury Park: Sage. ISBN 0-8039-4107-2.
- King, Gary (1989). *Unifying Political Methodology : the Likehood Theory of Statistical Inference*. Cambridge University Press. ISBN 0-521-36697-6.
- Le Cam, Lucien (1990). "Maximum likelihood : An Introduction". *ISI Review*. **58** (2): 153–171. JSTOR 1403464.
- Magnus, Jan R. (2017). "Maximum Likelihood". *Introduction to the Theory of Econometrics*. Amsterdam: VU University Press. pp. 53–68. ISBN 978-90-8659-766-6.
- Millar, Russell B. (2011). *Maximum Likelihood Estimation and Inference*. Hoboken: Wiley. ISBN 978-0-470-09482-2.
- Pickles, Andrew (1986). *An Introduction to Likelihood Analysis*. Norwich: W. H. Hutchins & Sons. ISBN 0-86094-190-6.
- Severini, Thomas A. (2000). *Likelihood Methods in Statistics*. New York: Oxford University Press. ISBN 0-19-850650-3.
- Ward, Michael D.; Ahlquist, John S. (2018). *Maximum Likelihood for Social Science : Strategies for Analysis*. Cambridge University Press. ISBN 978-1-316-63682-4.

External links [edit]

- "Maximum-likelihood method", *Encyclopedia of Mathematics*, EMS Press, 2001 [1994]
- Purcell, S. "Maximum Likelihood Estimation".
- Sargent, Thomas; Stachurski, John. "Maximum Likelihood Estimation". *Quantitative Economics with Python*.
- Toomet, Ott; Henningsen, Arne (2019-05-19). "maxLik: A package for maximum likelihood estimation in R".

V · T · E **Statistics** [hide]

Outline · Index

Descriptive statistics [show]

Data collection [show]

Statistical inference [show]

Correlation · Regression analysis [show]

Categorical / Multivariate / Time-series / Survival analysis [show]

Applications [show]

Category · **Mathematics portal** · **Commons** · **WikiProject**

Categories: Maximum likelihood estimation | M-estimators | Probability distribution fitting