WIKIPEDIA

Fisher information

In mathematical statistics, the **Fisher information** (sometimes simply called **information**[1]) is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X. Formally, it is the variance of the score, or the expected value of the observed information. In Bayesian statistics, the asymptotic distribution of the posterior mode depends on the Fisher information and not on the prior (according to the Bernstein–von Mises theorem, which was anticipated by Laplace for exponential families).[2] The role of the Fisher information in the asymptotic theory of maximum-likelihood estimation was emphasized by the statistician Ronald Fisher (following some initial results by Francis Ysidro Edgeworth). The Fisher information is also used in the calculation of the Jeffreys prior, which is used in Bayesian statistics.

The Fisher-information matrix is used to calculate the covariance matrices associated with maximum-likelihood estimates. It can also be used in the formulation of test statistics, such as the Wald test.

Statistical systems of a scientific nature (physical, biological, etc.) whose likelihood functions obey shift invariance have been shown to obey maximum Fisher information.[3] The level of the maximum depends upon the nature of the system constraints.

Contents

Definition

The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends. Let $f(X;\ \theta)$ be the probability density function (or probability mass function) for X conditional on the value of θ. It describes the probability that we observe a given outcome of X, *given* a known value of θ. If f is sharply peaked with respect to changes in θ, it is easy to indicate the "correct" value of θ from the data, or equivalently, that the data X provides a lot of information about the parameter θ. If the likelihood f is flat and spread-out, then it would take many samples of X to estimate the actual "true" value of θ that *would* be obtained using the entire population being sampled. This suggests studying some kind of variance with respect to θ.

Formally, the partial derivative with respect to θ of the natural logarithm of the likelihood function is called the "*score*". Under certain regularity conditions, if θ is the true parameter (i.e. X is actually distributed as $f(X;\ \theta)$), it can be shown that the expected value (the first moment) of the score is 0:[4]

$$\begin{aligned}\mathbf{E}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\middle|\theta\right] &= \int \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx \\ &= \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx \\ &= \frac{\partial}{\partial\theta}1 = 0.\end{aligned}$$

The variance of the score is defined to be the **Fisher information**:

$$\mathcal{I}(\theta) = \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\middle|\theta\right] = \int\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx,$$

Note that $0 \leq \mathcal{I}(\theta)$. A random variable carrying high Fisher information implies that the absolute value of the score is often high. The Fisher information is not a function of a particular observation, as the random variable X has been averaged out.

If $\log f(x;\ \theta)$ is twice differentiable with respect to θ, and under certain regularity conditions,[4] then the Fisher information may also be written as[5]

$$\mathcal{I}(\theta) = -\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\middle|\theta\right],$$

since

$$\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right)^2 = \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2$$

and

$$\mathbf{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\middle|\theta\right] = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,dx = 0.$$

Thus, the Fisher information may be seen as the curvature of the support curve (the graph of the log-likelihood). Near the maximum likelihood estimate, low Fisher information therefore indicates that the maximum appears "blunt", that is, the maximum is shallow and there are many nearby values with a similar log-likelihood. Conversely, high Fisher information indicates that the maximum is sharp.

Discrepancy in definition

There exist two versions of the definition of Fisher information. Some books and notes define

$$\mathcal{I}(\theta) := \mathrm{E}\left[-\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]$$

where $\log f(X|\theta)$ is the log-likelihood for one observation, whereas others define

$\mathcal{I}(\theta) := \mathrm{E}\left[-\frac{\partial^2}{\partial\theta^2}\ell(X \mid \theta)\right]$ where ℓ is the log-likelihood function for all observations. Some textbooks may even use the same symbol $\mathcal{I}(\theta)$ to denote both versions under different topics (for example, a book that defines $\mathcal{I}(\theta)$ to be the all observation version when discussing Cramer–Rao lower bound and may still let the same symbol refer to the one observation version when presenting the asymptotic normal distribution of maximum likelihood estimator). One should be careful with the meaning of $\mathcal{I}(\theta)$ in a specific context.

Informal derivation of the Cramér–Rao bound

The Cramér–Rao bound states that the inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ. H.L. Van Trees (1968) and B. Roy Frieden (2004) provide the following method of deriving the Cramér–Rao bound, a result which describes use of the Fisher information.

Informally, we begin by considering an unbiased estimator $\hat{\theta}(X)$. Mathematically, "unbiased" means that

$$\mathrm{E}\left[\hat{\theta}(X) - \theta \middle| \theta\right] = \int \left(\hat{\theta}(x) - \theta\right) f(x;\theta)\,dx = 0.$$

This expression is zero independent of θ, so its partial derivative with respect to θ must also be zero. By the product rule, this partial derivative is also equal to

$$0 = \frac{\partial}{\partial\theta}\int \left(\hat{\theta}(x) - \theta\right) f(x;\theta)\,dx = \int \left(\hat{\theta}(x) - \theta\right)\frac{\partial f}{\partial\theta}\,dx - \int f\,dx.$$

For each θ, the likelihood function is a probability density function, and therefore $\int f\,dx = 1$. A basic computation implies that

$$\frac{\partial f}{\partial\theta} = f\,\frac{\partial \log f}{\partial\theta}.$$

Using these two facts in the above, we get

$$\int \left(\hat{\theta} - \theta\right) f\,\frac{\partial \log f}{\partial\theta}\,dx = 1.$$

Factoring the integrand gives

$$\int \left(\left(\hat{\theta}-\theta\right)\sqrt{f}\right)\left(\sqrt{f}\,\frac{\partial \log f}{\partial \theta}\right)\,dx = 1.$$

Squaring the expression in the integral, the Cauchy–Schwarz inequality yields

$$1 = \left(\int \left[\left(\hat{\theta}-\theta\right)\sqrt{f}\right]\cdot\left[\sqrt{f}\,\frac{\partial \log f}{\partial \theta}\right]\,dx\right)^2 \leq \left[\int \left(\hat{\theta}-\theta\right)^2 f\,dx\right]\cdot\left[\int \left(\frac{\partial \log f}{\partial \theta}\right)^2 f\,dx\right].$$

The second bracketed factor is defined to be the Fisher Information, while the first bracketed factor is the expected mean-squared error of the estimator $\hat{\theta}$. By rearranging, the inequality tells us that

$$\operatorname{Var}\left(\hat{\theta}\right) \geq \frac{1}{\mathcal{I}(\theta)}.$$

In other words, the precision to which we can estimate θ is fundamentally limited by the Fisher information of the likelihood function.

Single-parameter Bernoulli experiment

A Bernoulli trial is a random variable with two possible outcomes, "success" and "failure", with success having a probability of θ. The outcome can be thought of as determined by a coin toss, with the probability of heads being θ and the probability of tails being $1-\theta$.

Let X be a Bernoulli trial. The Fisher information contained in X may be calculated to be

$$\begin{aligned}
\mathcal{I}(\theta) &= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\middle|\theta\right] \\
&= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\left(X\log\theta+(1-X)\log(1-\theta)\right)\middle|\theta\right] \\
&= \operatorname{E}\left[\frac{X}{\theta^2}+\frac{1-X}{(1-\theta)^2}\middle|\theta\right] \\
&= \frac{\theta}{\theta^2}+\frac{1-\theta}{(1-\theta)^2} \\
&= \frac{1}{\theta(1-\theta)}.
\end{aligned}$$

Because Fisher information is additive, the Fisher information contained in n independent Bernoulli trials is therefore

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}.$$

This is the reciprocal of the variance of the mean number of successes in n Bernoulli trials, so in this case, the Cramér–Rao bound is an equality.

Matrix form

When there are N parameters, so that θ is an $N \times 1$ vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_N]^{\mathrm{T}}$, then the Fisher information takes the form of an $N \times N$ matrix. This matrix is called the **Fisher information matrix** (FIM) and has typical element

$$\left[\mathcal{I}(\theta)\right]_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\middle|\,\theta\right].$$

The FIM is a $N \times N$ positive semidefinite matrix. If it is positive definite, then it defines a Riemannian metric on the N-dimensional parameter space. The topic information geometry uses this to connect Fisher information to differential geometry, and in that context, this metric is known as the Fisher information metric.

Under certain regularity conditions, the Fisher information matrix may also be written as

$$\left[\mathcal{I}(\theta)\right]_{i,j} = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(X;\theta)\middle|\,\theta\right].$$

The result is interesting in several ways:

- It can be derived as the Hessian of the relative entropy.
- It can be understood as a metric induced from the Euclidean metric, after appropriate change of variable.
- In its complex-valued form, it is the Fubini–Study metric.
- It is the key part of the proof of Wilks' theorem, which allows confidence region estimates for maximum likelihood estimation (for those conditions for which it applies) without needing the Likelihood Principle.
- In cases where the analytical calculations of the FIM above are difficult, it is possible to form an average of easy Monte Carlo estimates of the Hessian of the negative log-likelihood function as an estimate of the FIM.[6][7][8] The estimates may be based on values of the negative log-likelihood function or the gradient of the negative log-likelihood function; no analytical calculation of the Hessian of the negative log-likelihood function is needed.

Orthogonal parameters

We say that two parameters θ_i and θ_j are orthogonal if the element of the ith row and jth column of the Fisher information matrix is zero. Orthogonal parameters are easy to deal with in the sense that their maximum likelihood estimates are independent and can be calculated separately. When dealing with research problems, it is very common for the researcher to invest some time searching for an orthogonal parametrization of the densities involved in the problem.

Singular statistical model

If the Fisher information matrix is positive definite for all θ, then the corresponding statistical model is said to be *regular*; otherwise, the statistical model is said to be *singular*.[9] Examples of singular statistical models include the following: normal mixtures, binomial mixtures, multinomial mixtures, Bayesian networks, neural networks, radial basis functions, hidden Markov models, stochastic context-free grammars, reduced rank regressions, Boltzmann machines.

In machine learning, if a statistical model is devised so that it extracts hidden structure from a random phenomenon, then it naturally becomes singular.[10]

Multivariate normal distribution

The FIM for a N-variate multivariate normal distribution, $X \sim N(\mu(\theta), \Sigma(\theta))$ has a special form. Let the K-dimensional vector of parameters be $\theta = [\theta_1, \dots, \theta_K]^{\mathrm{T}}$ and the vector of random normal variables be $X = [X_1, \dots, X_N]^{\mathrm{T}}$. Assume that the mean values of these random variables are $\mu(\theta) = [\mu_1(\theta), \dots, \mu_N(\theta)]^{\mathrm{T}}$, and let $\Sigma(\theta)$ be the covariance matrix. Then, for $1 \le m, n \le K$, the (m, n) entry of the FIM is:[11]

$$\mathcal{I}_{m,n} = \frac{\partial \mu^{\mathrm{T}}}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n} + \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_m} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_n} \right),$$

where $(\cdot)^{\mathrm{T}}$ denotes the transpose of a vector, $\operatorname{tr}(\cdot)$ denotes the trace of a square matrix, and:

- $\frac{\partial \mu}{\partial \theta_m} = \begin{bmatrix} \frac{\partial \mu_1}{\partial \theta_m} & \frac{\partial \mu_2}{\partial \theta_m} & \cdots & \frac{\partial \mu_N}{\partial \theta_m} \end{bmatrix}^{\mathrm{T}}$;
- $\frac{\partial \Sigma}{\partial \theta_m} = \begin{bmatrix} \frac{\partial \Sigma_{1,1}}{\partial \theta_m} & \frac{\partial \Sigma_{1,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{1,N}}{\partial \theta_m} \\ \frac{\partial \Sigma_{2,1}}{\partial \theta_m} & \frac{\partial \Sigma_{2,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{2,N}}{\partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \Sigma_{N,1}}{\partial \theta_m} & \frac{\partial \Sigma_{N,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{N,N}}{\partial \theta_m} \end{bmatrix}$.

Note that a special, but very common, case is the one where $\Sigma(\theta) = \Sigma$, a constant. Then

$$\mathcal{I}_{m,n} = \frac{\partial \mu^{\mathrm{T}}}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n}.$$

In this case the Fisher information matrix may be identified with the coefficient matrix of the normal equations of least squares estimation theory.

Another special case occurs when the mean and covariance depend on two different vector parameters, say, β and θ. This is especially popular in the analysis of spatial data, which often uses a linear model with correlated residuals. In this case,[12]

$$\mathcal{I}(\beta, \theta) = \operatorname{diag}(\mathcal{I}(\beta), \mathcal{I}(\theta))$$

where

$$\mathcal{I}(\beta)_{m,n} = \frac{\partial \mu^{\mathrm{T}}}{\partial \beta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \beta_n},$$

$$\mathcal{I}(\theta)_{m,n} = \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_m} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_n} \right)$$

Properties

Chain rule

Similar to the entropy or mutual information, the Fisher information also possesses a **chain rule** decomposition. In particular, if X and Y are jointly distributed random variables, it follows that:[13]

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_{Y|X}(\theta),$$

where $\mathcal{I}_{Y|X}(\theta)$ is the Fisher information of Y relative to $\boldsymbol{\theta}$ calculated with respect to the conditional density of Y given a specific value $X = x$.

As a special case, if the two random variables are independent, the information yielded by the two random variables is the sum of the information from each random variable separately:

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta).$$

Consequently, the information in a random sample of n independent and identically distributed observations is n times the information in a sample of size 1.

Sufficient statistic

The information provided by a sufficient statistic is the same as that of the sample X. This may be seen by using Neyman's factorization criterion for a sufficient statistic. If $T(X)$ is sufficient for θ, then

$$f(X;\theta) = g(T(X),\theta)h(X)$$

for some functions g and h. The independence of $h(X)$ from θ implies

$$\frac{\partial}{\partial\theta}\log[f(X;\theta)] = \frac{\partial}{\partial\theta}\log[g(T(X);\theta)],$$

and the equality of information then follows from the definition of Fisher information. More generally, if $T = t(X)$ is a statistic, then

$$\mathcal{I}_T(\theta) \leq \mathcal{I}_X(\theta)$$

with equality if and only if T is a sufficient statistic.

Reparametrization

The Fisher information depends on the parametrization of the problem. If θ and η are two scalar parametrizations of an estimation problem, and θ is a continuously differentiable function of η, then

$$\mathcal{I}_\eta(\eta) = \mathcal{I}_\theta(\theta(\eta))\left(\frac{d\theta}{d\eta}\right)^2$$

where $\mathcal{I}_\eta$ and $\mathcal{I}_\theta$ are the Fisher information measures of η and θ, respectively.[14]

In the vector case, suppose $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are k-vectors which parametrize an estimation problem, and suppose that $\boldsymbol{\theta}$ is a continuously differentiable function of $\boldsymbol{\eta}$, then,[15]

$$\mathcal{I}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \boldsymbol{J}^{\mathrm{T}}\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{J}$$

where the (i, j)th element of the $k \times k$ Jacobian matrix $\boldsymbol{J}$ is defined by

$$J_{ij} = \frac{\partial\theta_i}{\partial\eta_j},$$

and where $\boldsymbol{J}^{\mathrm{T}}$ is the matrix transpose of $\boldsymbol{J}$.

In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrization. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher–Rao metric) for the manifold of thermodynamic states, and can be used as an information-geometric complexity measure for a classification of phase transitions, e.g., the scalar curvature of the thermodynamic metric tensor diverges at (and only at) a phase transition point.[16]

In the thermodynamic context, the Fisher information matrix is directly related to the rate of change in the corresponding order parameters.[17] In particular, such relations identify second-order phase transitions via divergences of individual elements of the Fisher information matrix.

Applications

Optimal design of experiments

Fisher information is widely used in optimal experimental design. Because of the reciprocity of estimator-variance and Fisher information, *minimizing* the *variance* corresponds to *maximizing* the *information*.

When the linear (or linearized) statistical model has several parameters, the mean of the parameter estimator is a vector and its variance is a matrix. The inverse of the variance matrix is called the "information matrix". Because the variance of the estimator of a parameter vector is a matrix, the problem of "minimizing the variance" is complicated. Using statistical theory, statisticians compress the information-matrix using real-valued summary statistics; being real-valued functions, these "information criteria" can be maximized.

Traditionally, statisticians have evaluated estimators and designs by considering some summary statistic of the covariance matrix (of an unbiased estimator), usually with positive real values (like the determinant or matrix trace). Working with positive real numbers brings several advantages: If the estimator of a single parameter has a positive variance, then the variance and the Fisher information are both positive real numbers; hence they are members of the convex cone of nonnegative real numbers (whose nonzero members have reciprocals in this same cone). For several parameters, the covariance matrices and information matrices are elements of the convex cone of nonnegative-definite symmetric matrices in a partially ordered vector space, under the Loewner (Löwner) order. This cone is closed under matrix addition and inversion, as well as under the multiplication of positive real numbers and matrices. An exposition of matrix theory and Loewner order appears in Pukelsheim.[18]

The traditional optimality criteria are the information matrix's invariants, in the sense of invariant theory; algebraically, the traditional optimality criteria are functionals of the eigenvalues of the (Fisher) information matrix (see optimal design).

Jeffreys prior in Bayesian statistics

In Bayesian statistics, the Fisher information is used to calculate the Jeffreys prior, which is a standard, non-informative prior for continuous distribution parameters.[19]

Computational neuroscience

The Fisher information has been used to find bounds on the accuracy of neural codes. In that case, X is typically the joint responses of many neurons representing a low dimensional variable θ (such as a stimulus parameter). In particular the role of correlations in the noise of the neural responses has been studied.[20]

Derivation of physical laws

Fisher information plays a central role in a controversial principle put forward by Frieden as the basis of physical laws, a claim that has been disputed.[21]

Machine learning

The Fisher information is used in machine learning techniques such as elastic weight consolidation,[22] which reduces catastrophic forgetting in artificial neural networks.

Relation to relative entropy

Fisher information is related to relative entropy.[23] Consider a family of probability distributions $f(x;\theta)$ where θ is a parameter which lies in a range of values. Then the relative entropy, or Kullback–Leibler divergence, between two distributions in the family can be written as

$$D(\theta \parallel \theta') = \int f(x;\theta) \log \frac{f(x;\theta)}{f(x;\theta')} dx = \int f(x;\theta) \left(\log f(x;\theta) - \log f(x;\theta')\right) dx,$$

while the Fisher information matrix is:

$$[\mathcal{I}(\theta)]_{ij} = \left(\frac{\partial^2}{\partial \theta'_i \partial \theta'_j} D(\theta \parallel \theta') \right)_{\theta'=\theta} = -\int f(x;\theta) \frac{\partial^2 \log f(x;\theta)}{\partial \theta_i \partial \theta_j} dx.$$

If θ is fixed, then the relative entropy between two distributions of the same family is minimized at $\theta' = \theta$. For θ' close to θ, one may expand the previous expression in a series up to second order:

$$D(\theta \parallel \theta') = \frac{1}{2} (\theta' - \theta)^\top \underbrace{\left(\frac{\partial^2}{\partial \theta'_i \, \partial \theta'_j} D(\theta \parallel \theta') \right)_{\theta'=\theta}}_{\textbf{Fisher information}} (\theta' - \theta) + \cdots$$

Thus the Fisher information represents the curvature of the relative entropy.

Schervish (1995: §2.3) says the following.

> One advantage Kullback-Leibler information has over Fisher information is that it is not affected by changes in parameterization. Another advantage is that Kullback-Leibler information can be used even if the distributions under consideration are not all members of a parametric family.
>
> ...
> Another advantage to Kullback-Leibler information is that no smoothness conditions on the densities ... are needed.

History

The Fisher information was discussed by several early statisticians, notably F. Y. Edgeworth.[24] For example, Savage[25] says: "In it [Fisher information], he [Fisher] was to some extent anticipated (Edgeworth 1908–9 esp. 502, 507–8, 662, 677–8, 82–5 and references he [Edgeworth] cites including Pearson and Filon 1898 [. . .])." There are a number of early historical sources[26] and a number of reviews of this early work.[27][28][29]

See also

- Efficiency (statistics)
- Observed information
- Fisher information metric
- Formation matrix
- Information geometry
- Jeffreys prior

- Cramér–Rao bound

Other measures employed in information theory:

- Entropy (information theory)
- Kullback–Leibler divergence
- Self-information

Notes

1. Lehmann & Casella, p. 115
2. Lucien Le Cam (1986) *Asymptotic Methods in Statistical Decision Theory*: Pages 336 and 618–621 (von Mises and Bernstein).
3. Frieden & Gatenby (2013)
4. Suba Rao. "Lectures on statistical inference" (http://www.stat.tamu.edu/~suhasini/teaching613/inference.pdf) (PDF).
5. Lehmann & Casella, eq. (2.5.16), Lemma 5.3, p.116.
6. Spall, J. C. (2005), "Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings," *Journal of Computational and Graphical Statistics*, vol. 14(4), pp. 889–909. https://doi.org/10.1198/106186005X78800
7. Spall, J. C. (2008), "Improved Methods for Monte Carlo Estimation of the Fisher Information Matrix," *Proceedings of the American Control Conference*, Seattle, WA, 11–13 June 2008, pp. 2395–2400. https://doi.org/10.1109/ACC.2008.4586850
8. Das, S., Spall, J. C., and Ghanem, R. (2010), "Efficient Monte Carlo Computation of Fisher Information Matrix Using Prior Information," *Computational Statistics and Data Analysis*, vol. 54(2), pp. 272–289. https://doi.org/10.1016/j.csda.2009.09.018
9. Watanabe, S. (2008), Accardi, L.; Freudenberg, W.; Ohya, M. (eds.), "Algebraic geometrical method in singular statistical estimation", *Quantum Bio-Informatics*, World Scientific, pp. 325–336.
10. Watanabe, S (2013). "A Widely Applicable Bayesian Information Criterion". *Journal of Machine Learning Research*. **14**: 867–897.
11. Malagò, Luigi; Pistone, Giovanni (2015). *Information geometry of the Gaussian distribution in view of stochastic optimization*. *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*. pp. 150–162. doi:10.1145/2725494.2725510 (https://doi.org/10.1145%2F2725494.2725510). ISBN 9781450334341.
12. Mardia, K. V.; Marshall, R. J. (1984). "Maximum likelihood estimation of models for residual covariance in spatial regression". *Biometrika*. **71** (1): 135–46. doi:10.1093/biomet/71.1.135 (https://doi.org/10.1093%2Fbiomet%2F71.1.135).
13. Zamir, R. (1998). "A proof of the Fisher information inequality via a data processing argument". *IEEE Transactions on Information Theory*. **44** (3): 1246–1250. CiteSeerX 10.1.1.49.6628 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.6628). doi:10.1109/18.669301 (https://doi.org/10.1109%2F18.669301).
14. Lehmann & Casella, eq. (2.5.11).
15. Lehmann & Casella, eq. (2.6.16)
16. Janke, W.; Johnston, D. A.; Kenna, R. (2004). "Information Geometry and Phase Transitions". *Physica A*. **336** (1–2): 181. arXiv:cond-mat/0401092 (https://arxiv.org/abs/cond-mat/0401092). Bibcode:2004PhyA..336..181J (https://ui.adsabs.harvard.edu/abs/2004PhyA..336..181J). doi:10.1016/j.physa.2004.01.023 (https://doi.org/10.1016%2Fj.physa.2004.01.023).
17. Prokopenko, M.; Lizier, Joseph T.; Lizier, J. T.; Obst, O.; Wang, X. R. (2011). "Relating Fisher information to order parameters". *Physical Review E*. **84** (4): 041116. Bibcode:2011PhRvE..84d1116P (https://ui.adsabs.harvard.edu/abs/2011PhRvE..84d1116P). doi:10.1103/PhysRevE.84.041116 (https://doi.org/10.1103%2FPhysRevE.84.041116).
18. Pukelsheim, Friedrick (1993). *Optimal Design of Experiments*. New York: Wiley. ISBN 978-0-471-61971-0.
19. Bernardo, Jose M.; Smith, Adrian F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons. ISBN 978-0-471-92416-6.
20. Abbott, Larry F., and Peter Dayan. "The effect of correlated variability on the accuracy of a population code." Neural computation 11.1 (1999): 91-101.
21. Streater, R. F. (2007). *Lost Causes in and beyond Physics*. Springer. p. 69. ISBN 978-3-540-36581-5.
22. Kirkpatrick, James; Pascanu, Razvan; Rabinowitz, Neil; Veness, Joel; Desjardins, Guillaume; Rusu, Andrei A.; Milan, Kieran; Quan, John; Ramalho, Tiago (2017-03-28). "Overcoming catastrophic forgetting in neural networks" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5380101). *Proceedings of the National Academy of Sciences*. **114** (13): 3521–3526. doi:10.1073/pnas.1611835114 (https://doi.org/10.1073%2Fpnas.1611835114). ISSN 0027-8424 (https://www.worldcat.org/issn/0027-8424). PMC 5380101 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5380101). PMID 28292907 (https://www.ncbi.nlm.nih.gov/pubmed/28292907).
23. Gourieroux & Montfort (1995), page 87 (https://books.google.com/books?id=gqI-pAP2JZ8C&pg=PA87)
24. Savage (1976)
25. Savage(1976), page 156
26. Edgeworth (September 1908, December 1908)
27. Pratt (1976)
28. Stigler (1978, 1986, 1999)
29. Hald (1998, 1999)

References

- Edgeworth, F. Y. (Jun 1908). "On the Probable Errors of Frequency-Constants" (https://zenodo.org/record/1449470). *Journal of the Royal Statistical Society*. **71** (2): 381–397. doi:10.2307/2339461 (https://doi.org/10.2307%2F2339461). JSTOR 2339461 (https://www.jstor.org/stable/2339461).

- Edgeworth, F. Y. (Sep 1908). "On the Probable Errors of Frequency-Constants (Contd.)" (https://zenodo.org/record/1449468). *Journal of the Royal Statistical Society*. **71** (3): 499–512. doi:10.2307/2339293 (https://doi.org/10.2307%2F2339293). JSTOR 2339293 (https://www.jstor.org/stable/2339293).
- Edgeworth, F. Y. (Dec 1908). "On the Probable Errors of Frequency-Constants (Contd.)" (https://zenodo.org/record/1449468). *Journal of the Royal Statistical Society*. **71** (4): 651–678. doi:10.2307/2339378 (https://doi.org/10.2307%2F2339378). JSTOR 2339378 (https://www.jstor.org/stable/2339378).
- Frieden, B. R. (2004) *Science from Fisher Information: A Unification*. Cambridge Univ. Press. ISBN 0-521-00911-1.
- Frieden, B. Roy; Gatenby, Robert A. (2013). "Principle of maximum Fisher information from Hardy's axioms applied to statistical systems". *Physical Review E*. **88** (4): 042144. arXiv:1405.0007 (https://arxiv.org/abs/1405.0007). Bibcode:2013PhRvE..88d2144F (https://ui.adsabs.harvard.edu/abs/2013PhRvE..88d2144F). doi:10.1103/PhysRevE.88.042144 (https://doi.org/10.1103%2FPhysRevE.88.042144). PMID 24229152 (https://www.ncbi.nlm.nih.gov/pubmed/24229152).
- Hald, A. (May 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 (https://doi.org/10.1214%2Fss%2F1009212248). JSTOR 2676741 (https://www.jstor.org/stable/2676741).
- Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley. ISBN 978-0-471-17912-2.
- Lehmann, E. L.; Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer. ISBN 978-0-387-98502-2.
- Le Cam, Lucien (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag. ISBN 978-0-387-96307-5.
- Pratt, John W. (May 1976). "F. Y. Edgeworth and R. A. Fisher on the Efficiency of Maximum Likelihood Estimation". *Annals of Statistics*. **4** (3): 501–514. doi:10.1214/aos/1176343457 (https://doi.org/10.1214%2Faos%2F1176343457). JSTOR 2958222 (https://www.jstor.org/stable/2958222).
- Savage, L. J. (May 1976). "On Rereading R. A. Fisher". *Annals of Statistics*. **4** (3): 441–500. doi:10.1214/aos/1176343456 (https://doi.org/10.1214%2Faos%2F1176343456). JSTOR 2958221 (https://www.jstor.org/stable/2958221).
- Schervish, Mark J. (1995). *Theory of Statistics*. New York: Springer. ISBN 978-0-387-94546-0.
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900* (https://archive.org/details/historyofstatist00stig). Harvard University Press. ISBN 978-0-674-40340-6.
- Stigler, S. M. (1978). "Francis Ysidro Edgeworth, Statistician". *Journal of the Royal Statistical Society, Series A*. **141** (3): 287–322. doi:10.2307/2344804 (https://doi.org/10.2307%2F2344804). JSTOR 2344804 (https://www.jstor.org/stable/2344804).
- Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press. ISBN 978-0-674-83601-3.
- Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley. ISBN 978-0-471-09517-0.
