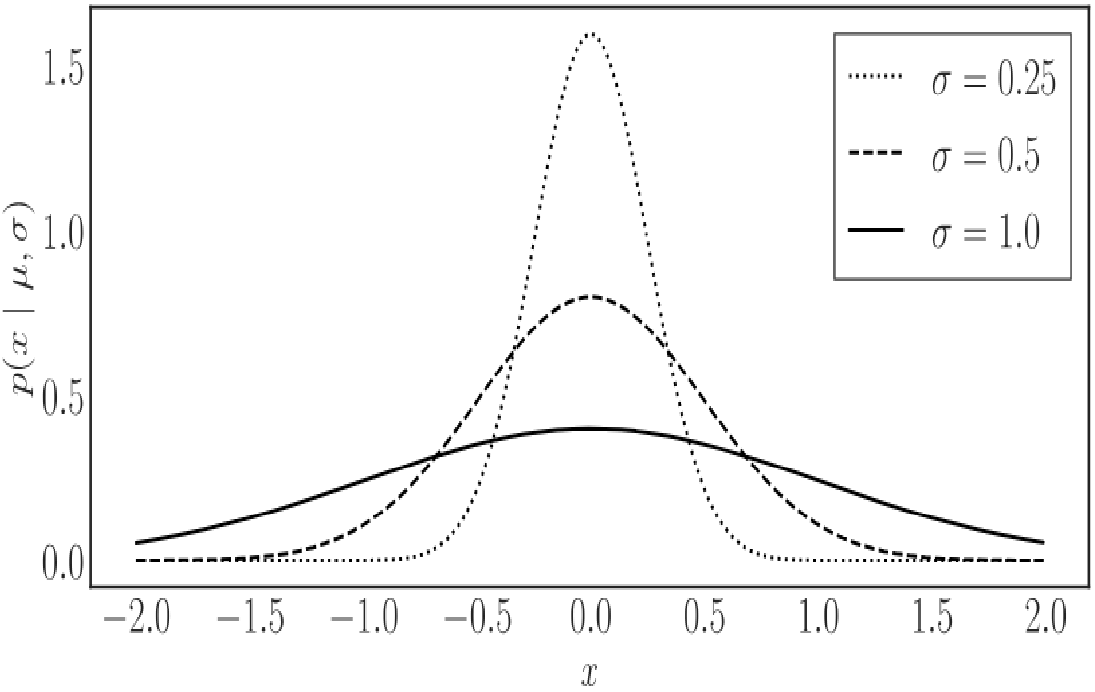
DATA SCIENCE		
MSDSDC201T		**Statistical Methods**
Unit 2.1 : Computational Method for finding Estimators - Fisher Information		

WHAT'S IN A PARAMETER

How much information does a data-point contain about a parameter? This seems an abstract idea, however, it is actually a measurable quantity.

Let's say we have a sample from a Normal distribution with a mean μ and variance σ^2. The variance σ^2 is known, and the goal is to estimate the mean, μ. To understand how difficult this is, we would like to know how much information we can expect the sample to contain about μ. Consider for instance,three possible scenarios with three different variances, as shown here:



We expect that the mean is easiest to estimate for the Normal distribution with the smallest variance. In this case, the samples would cluster quite close to the mean value, hence all of them would be similar and close in value to each other. A random sample is more likely to be close to the mean when the variance is small than when the variance is large. This implies that the information content should grow as σ shrinks.

Let's use $\mathbf{I}_x(\mu)$ to represent the information content of a sample **x** at the mean μ. In the case of the Normal distribution, we might expect the information content of the sample to be inversely proportional to the variance, $\mathbf{I}_x(\mu) \propto 1/\sigma^2$. In general, the information content will be a function of μ, the parameter we want to estimate. Different values of the parameter could be easier or harder to estimate. However, for the Normal distribution, μ only shifts the mode of the distribution, so the information content only depends on σ and not on μ.

Another important point is that **x** is a random sample. We don't want to specify a value for **x**. Instead, we'd like the information content of **x** to consider all possible values for x and their corresponding probabilities. A value for **x** which might tell us a lot about the parameter but is exceedingly unlikely shouldn't contribute much to the expected information content of the sample. Taking an expectation over **x** is a natural way to account for this.

The ***Fisher Information*** attempts to quantify the sensitivity of the random variable **x** to the value of the parameter θ. If small changes in θ result in large changes in the likely values of **x**, then the samples we observe tell us a lot about θ. In this case the Fisher information should be high. Continuing the Normal distribution example, a small variance means we will see large changes in the observed **x** with small changes in the mean. In this case the Fisher information of **x** about the mean μ is large.

We will formalize this idea by first introducing the idea of Score

SCORE

Score is a derivative of log-likelihood function with respect to the parameter. It is usually evaluated at a particular value of the parameter. We define it as:

$$s(\theta) = \frac{\partial\, log\, L(\theta)}{\partial\theta}$$

Intuitively, we want to see the sensitivity of the log-Likelihood function to small changes in θ.

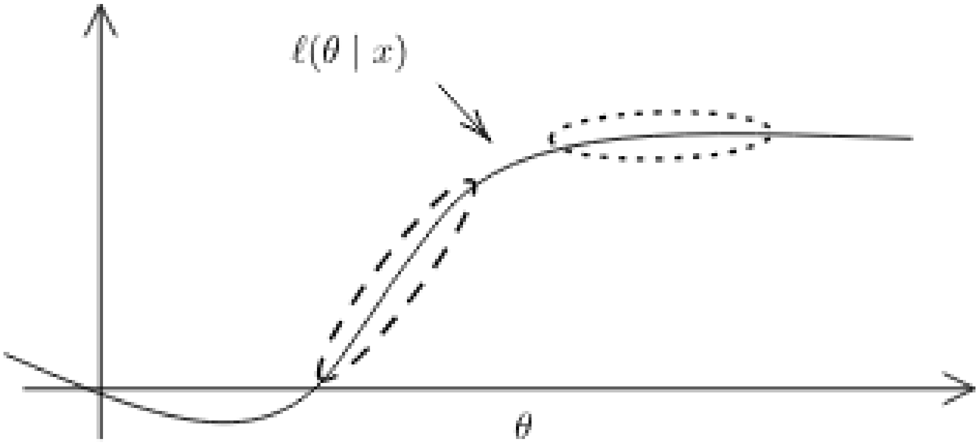


FIG : An example log-likelihood function ℓ (θ ∣ x). The section of the log-likelihood surrounded by the dashed line changes rapidly as a function of θ. The section of the log-likelihood surrounded by the dotted line changes very slowly as a function of θ

What exactly is the score function telling us?
The Likelihood function measures how likely a particular value of a parameter **θ** is, given the value of the data-point. In other words, it shows us the dependence of the parameter on the data, i.e we obtain information about the parameter from the data.

Let us recall that the log-Likelihood function is a proxy for the likelihood function, except that it is generally easier to deal with. However, what would be the effect of a small change in the value of **θ** on the Likelihood function or indirectly, the log-Likelihood function? This is the question that the score function helps us deal with. We measure that by measuring the gradient of the log-Likelihood function. This is what the score function is telling us.

WORKING EXAMPLE : SCORE OF A BINOMIAL VARIABLE

Consider the Likelihood function for a Binomial variable Bin(n,p).

Let us assume the data-point x we are interested in x has r-successes.

Then the Likelihood function for this particular data-point is:

L (**p | x**) = P(**x | p**)
= $p^r(1-p)^{n-r}$

Then let us consider the log-likelihood l(**p | x**) = r log(p) + (n-r) log(1-p)
Then score is defined as l'(**p | x**) = r/ p - (n-r)/(1-p) = (r - np) / p(1-p)

One of the most important properties of the score function is that it has zero expectation.
This is because, considering a pdf function has a total probability of one:

$$\int p(x;\theta)\,\mathrm{d}x = 1$$

Take derivatives on both sides

$$\frac{\partial}{\partial\theta}\int p(x;\theta)\,\mathrm{d}x = 0$$

However,

$$\begin{aligned}\frac{\partial}{\partial\theta}\int p(x;\theta)\,\mathrm{d}x &= \int \frac{\partial p(x;\theta)}{\partial\theta}\mathrm{d}x \\ &= \int \frac{\frac{\partial p(x;\theta)}{\partial\theta}}{p(x;\theta)}p(x;\theta)\,\mathrm{d}x \\ &= \int \frac{\partial \log p(x;\theta)}{\partial\theta}p(x;\theta)\,\mathrm{d}x \\ &= E\left[\frac{\partial \ell(\theta;x)}{\partial\theta}\right]\end{aligned}$$

So, we can conclude that **E**[s(θ)] = 0

FISHER INFORMATION

Fisher information is a way to measure how much information is contained in a known random variable X about the unknown population parameter **θ** that is supposed to model X. It is calculated as the **variance** of the **score**.

Imagine you want to estimate how good an estimate is given all our knowledge about it. Fisher's information describes the probability f(x), given a known value of **θ**. If f(x) is sharply peaked at a value of X, it indicates we have a good estimate of the true X. If f(x) is more evenly spread, we know little about the true value. Consequently we would need many more samples to accurately know what the value should be. Theoretically, we would need to know the entire population.

Intuitively, if an event has small probability, then the occurrence of this event brings us much information. For a random variable X ~ f(x|**θ**), if **θ** were the true value of the parameter, the likelihood function should take a big value, or equivalently, the derivative log-likelihood function should be close to zero, and this is the basic principle of maximum likelihood estimation. if **l'**(X|**θ**) is close to zero, then it is expected, thus the random variable does not provide much information about **θ**; on the other hand, if |**l'**(X | **θ**)| or [**l'**(X| **θ**)]^2 is large, the random variable provides much information about **θ**. Thus, we can use [**l'**(X | **θ**)]^2 to measure the amount of information provided by X. However, since X is a random variable, we should consider the average case.

This intuitive understanding tells us that there would have to be some way to know how much information we have or how much information we need. This indicates that we should be looking at the measure of variance of the score with respect to **θ**. We will see that this is equivalent to the [**l'**(X | **θ**)]^2 based definition outlined above.

Fisher's information is defined as the variance of the score, λ'(θ).

We note here that the Expectation of the Score function E[Score] is 0, hence using the identity Var(**X**) = **E[X^2] - {E[X]}^{2}**, we must have that Var (Score) = E (Score^2). This allows us to define the Fisher Information as:

$$I(\theta) = \mathsf{E}_\theta\left[\lambda'(X \mid \theta)^2\right]$$

where, λ' refers to the score function.

We will now derive an alternative version of the Fisher Information: Note that Expected Score is zero, so

$$E\left[\frac{\partial \ell\left(\theta; x\right)}{\partial \theta}\right] = 0$$

$$\int \frac{\partial \ell\left(\theta; x\right)}{\partial \theta} p\left(x; \theta\right) dx = 0$$

Differentiating both sides of the equation, we get:

$$\frac{\partial}{\partial \theta} \int \frac{\partial \ell(\theta; x)}{\partial \theta} p(x; \theta)\, dx = 0$$

$$\int \frac{\partial^2 \ell(\theta; x)}{\partial \theta^2} p(x; \theta)\, dx + \int \frac{\partial \ell(\theta; x)}{\partial \theta} \frac{\partial p(x; \theta)}{\partial \theta} dx = 0$$

The second term here is:

$$\begin{aligned}
\int \frac{\partial \ell(\theta; x)}{\partial \theta} \frac{\partial p(x; \theta)}{\partial \theta} dx &= \int \frac{\partial \log p(x; \theta)}{\partial \theta} \frac{\partial p(x; \theta)}{\partial \theta} dx \\
&= \int \frac{\partial \log p(x; \theta)}{\partial \theta} \frac{\frac{\partial p(x;\theta)}{\partial \theta}}{p(x; \theta)} p(x; \theta)\, dx \\
&= \int \left(\frac{\partial \log p(x; \theta)}{\partial \theta} \right)^2 p(x; \theta)\, dx \\
&= V\left[\frac{\partial \ell(\theta; x)}{\partial \theta} \right]
\end{aligned}$$

<u>This gives us an alternate definition of the Fisher information as follows:</u>

$$\begin{aligned}
V\left[\frac{\partial \ell(\theta; x)}{\partial \theta} \right] &= -\int \frac{\partial^2 \ell(\theta; x)}{\partial \theta^2} p(x; \theta)\, dx \\
&= -E\left[\frac{\partial^2 \ell(\theta; x)}{\partial \theta^2} \right]
\end{aligned}$$

<u>Thus we now have three definitions of the Fisher Information:</u>

- $I(\theta) = Var\,[\,Score(\theta)\,] = Var\,[\frac{\delta l\,(\theta\,;\,x)}{\delta\theta}]$
- $I(\theta) = E_{\theta}[\,Score\,(\theta)^2] = E_{\theta}[\{\frac{\delta l\,(\theta\,;\,x)}{\delta\theta}\}^2]$
- $I(\theta) = -\,E_{\theta}[\frac{\delta^2 l\,(\theta\,;\,x)}{\delta\theta^2}]$

We have seen that these three forms are equivalent. In practice, some forms are more easy to use depending on the problem at hand. It is expected that we are familiar with all three and use these definitions as convenience demands.

Fisher information helps us measure the “information content” about a parameter. If the information content is low, then obviously it would take many samples before a conclusion can be made. On the other hand, a large amount of information content would help us out, as few samples would be sufficient. This makes Fisher information crucial to study and evaluate.

WORKING EXAMPLE : FISHER INFORMATION - BERNOULLI DISTRIBUTION

Let's use the Bernoulli distribution as an example. The Bernoulli distribution is that of a biased coin which has probability θ of turning up heads (or 1) and probability (1−θ) of turning up tails (or 0). We should expect that the more biased the coin, the easier it is to identify the bias from an observation of the coin toss. As an extreme example, if θ is 1 or 0, then a single coin toss will tell us the value of θ. The Fisher information of the sample x (the result of the coin toss) will be higher the closer θ is to either 1 or 0.

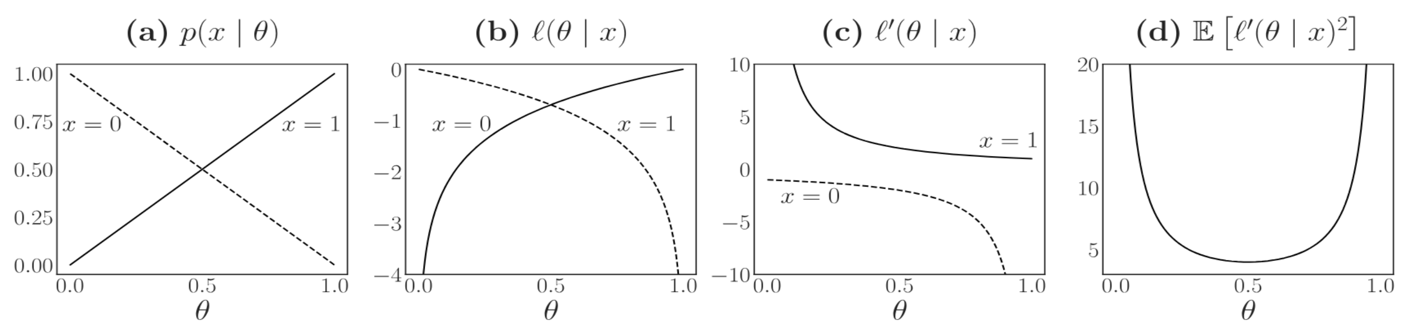


FIG : The sequence of terms needed to compute the Fisher information for a Bernoulli distribution at the parameter θ. All of the terms are plotted as a function of the parameter θ for both values of x, namely 0 and 1

The Bernoulli distribution p(**x**|θ) is plotted as a function of the parameter θ in figure (a) above. The two values of the distribution as a function of θ are p(**x**=1|θ) = θ and p(**x** = 0| θ) = 1−θ. The log-likelihood for each value of x is plotted as a function of θ in figure (b) above, and the score function is plotted in figure (c). The derivatives are:

$$\frac{d}{d\theta}\log p(x=1 \mid \theta) = \frac{1}{\theta} \quad \text{and} \quad \frac{d}{d\theta}\log p(x=0 \mid \theta) = \frac{1}{\theta - 1}.$$

To get the Fisher information, shown in figure (d), we take the expectation over **x** of the squared derivatives:

$$\mathcal{I}_x(\theta) = \theta\frac{1}{\theta^2} + (1-\theta)\frac{1}{(\theta-1)^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

The Fisher information in figure (d) has the shape we expect. As θ approaches 0 or 1, the Fisher information grows rapidly. Clearly, the Fisher information is inversely proportional to the variance of the Bernoulli distribution which is Var(**x**) = θ(1 − θ). The smaller the variance, the more we expect the sample of **x** to tell us about the parameter θ and hence the higher the Fisher information.

WORKING EXAMPLE : FISHER INFORMATION - EXPONENTIAL DISTRIBUTION

Consider the score function λ for the Exponential Distribution i.e. X~Exp(θ)
We omit the details of the calculation of the score function.

We just need to remember that we calculate the score by first estimating the Likelihood function, then the log-Likelihood, and then differentiate it to obtain the score function

$$\begin{aligned}\lambda(x \mid \theta) &= \log\theta - x\theta \\ \lambda'(x \mid \theta) &= \frac{1}{\theta} - x \\ \lambda''(x \mid \theta) &= -\frac{1}{\theta^2} \\ I(\theta) &= \mathsf{E}_\theta\left[\frac{(1 - X\theta)^2}{\theta^2}\right] = \frac{1}{\theta^2}.\end{aligned}$$

Hence, we can see that the Information is inversely proportional to the square of the parameter.

WORKING EXAMPLE : FISHER INFORMATION - POISSON DISTRIBUTION

Consider the score function λ for the Poisson Distribution i.e. X~Po(θ)

$$\begin{aligned}\lambda(x \mid \theta) &= x\log\theta - \log x! - \theta \\ \lambda'(x \mid \theta) &= \frac{x-\theta}{\theta} \\ \lambda''(x \mid \theta) &= -\frac{x}{\theta^2} \\ I(\theta) &= \mathsf{E}_\theta\left[\frac{(X - \theta)^2}{\theta^2}\right] = \frac{1}{\theta}.\end{aligned}$$

WORKING EXAMPLE : FISHER INFORMATION - NORMAL DISTRIBUTION with fixed variance

Normal: For the $\mathsf{No}(\theta, \sigma^2)$ distribution with fixed $\sigma^2 > 0$,

$$\begin{aligned}\lambda(x \mid \theta) &= -\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\theta)^2 \\ \lambda'(x \mid \theta) &= \frac{1}{\sigma^2}(x-\theta) \\ \lambda''(x \mid \theta) &= -\frac{1}{\sigma^2} \\ I(\theta) &= \mathsf{E}_\theta\left[\frac{(X - \theta)^2}{\sigma^4}\right] = \frac{1}{\sigma^2},\end{aligned}$$

This makes intuitive sense to us.

As we observed in the introduction, the Fisher information is inversely proportional to the variance in the sample.

In fact, the Fisher information is exactly the reciprocal of the variance.

PROPERTIES OF THE FISHER INFORMATION

- **ADDITIVE** : The value of the datapoint **X** can represent a single sample drawn from a single distribution or can represent a collection of samples drawn from a collection of distributions. If there are n samples and the corresponding n distributions are statistically independent then the Fisher information will necessarily be the sum of the single-sample Fisher information values, one for each single sample from its distribution. In particular, if the n distributions are independent and identically distributed then the Fisher information will necessarily be n times the Fisher information of a single sample from the common distribution.

  If X_n are i.i.d., then

$$\begin{aligned}\mathcal{I}_N(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\sum_{n=1}^{N}\log f_\theta(X_n)\right] \\ &= -\sum_{n=1}^{N}\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_n)\right] \\ &= -N\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_n)\right].\end{aligned}$$

- **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:

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

  In particular, if X and Y are independent, we have:

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

- **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:

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

  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.

- **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. In other words, the precision to which we can estimate θ is fundamentally limited by the Fisher information of the likelihood function. We can express this as:

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

MULTIVARIATE FISHER INFORMATION

If the family of distributions we are interested in has more than one parameter, then we can create a multivariate version of the Fisher Information. This is called the **Fisher Information Matrix**.

Consider the gradient operator defined as follows:

$$u(\theta) = \nabla_\theta \log p(x|\theta)$$

In practice there are multiple parameters, so this actually results in a vector:
For a distribution X with p parameters:

$$\underset{\sim}{X} \sim f(\underset{\sim}{x} \mid \theta),\ \theta = (\theta_1, \theta_2, \ldots, \theta_p) \text{ and}$$

$$\frac{\partial}{\partial\theta} = \begin{pmatrix} \frac{\partial}{\partial\theta_1} \\ \vdots \\ \frac{\partial}{\partial\theta_p} \end{pmatrix}$$

We obtain:

$$\frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_1} \log f(\underset{\sim}{X} \mid \theta) \\ \vdots \\ \frac{\partial}{\partial\theta_p} \log f(\underset{\sim}{X} \mid \theta) \end{pmatrix}$$

I.e the gradient of the score in this case is a p x 1 column vector, consisting of the partial derivatives of the log-Likelihood function with respect to each of the parameters.

DEFINITION : We define the **Fisher Information Matrix (FIM)** $I_X(\theta) = \nabla_\theta\, log\, p(\bar{x} \mid \theta)$ as:

$$(p \times p) \text{ matrix} \quad I_{\underset{\sim}{X}}(\theta) = E(S_{p\times 1} S'_{1\times p})$$

where S’ refers to the transpose of the matrix S.

$$\begin{aligned} \mathrm{I}(\theta) &= \mathrm{K}_{s(\theta)} \\ &= \mathbb{E}[(s(\theta) - 0)(s(\theta) - 0)^\top] \\ &= \mathbb{E}[s(\theta)s(\theta)^\top] \end{aligned}$$

The variance-covariance matrix is simply a matrix that contains information on the covariance of multiple random variables in a neat, compact matrix form. In our case, if we consider the variance-covariance matrix as follows:

$$\mathrm{K} = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top]$$

and we look at this matrix for our case, we can also consider the fact that, the expectation of the score is zero:

$$\begin{aligned} \mathrm{I}(\theta) &= \mathrm{K}_{s(\theta)} \\ &= \mathbb{E}[(s(\theta) - 0)(s(\theta) - 0)^\top] \\ &= \mathbb{E}[s(\theta)s(\theta)^\top] \end{aligned}$$

Hence, another way of thinking about the FIM is as the variance-covariance matrix of the score function.

Intuitively, Fisher's information gives us an estimate of how certain we are about the estimate of the parameter θ.

We also have a third representation of the FIM, the negative expected value of the second derivative of the log likelihood. In our multivariate context where θ is a vector, the second derivative is effectively the Hessian :

$$\begin{aligned} \mathrm{I}(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} s(\theta)\right] \\ &= -\mathbb{E}[\mathrm{H}_{\log p(x|\theta)}] \end{aligned}$$

WORKING EXAMPLE : FISHER INFORMATION - Multivariate NORMAL

Consider a Normal distribution with unknown mean and variance. We know the pdf function:

$$f(x \mid \mu, \xi) = \frac{1}{\sqrt{2\pi\xi}} e^{-(x-\mu)^2/(2\xi)}$$

Which also gives us the likelihood function. Note:

$$l = \log f = -\frac{1}{2}\log(2\pi\xi) - \frac{(x-\mu)^2}{2\xi}$$

$$\frac{\partial}{\partial \theta} \log f(X \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial \mu} \log f \\ \frac{\partial}{\partial \xi} \log f \end{pmatrix} = \begin{pmatrix} \frac{x-\mu}{\xi} \\ -\frac{1}{2\xi} + \frac{(x-\mu)^2}{2\xi^2} \end{pmatrix}$$

Hence, the Fisher information matrix is:

$$I(\theta) = -E\begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \xi} \\ \frac{\partial^2 l}{\partial \xi \partial \mu} & \frac{\partial^2 l}{\partial \xi^2} \end{pmatrix} = -E\begin{pmatrix} \frac{-1}{\xi} & \frac{-(X-\mu)}{\xi^2} \\ \frac{-(X-\mu)}{\xi^2} & \frac{1}{2\xi^2} - \frac{(X-\mu)^2}{\xi^3} \end{pmatrix} = \begin{pmatrix} \frac{1}{\xi} & 0 \\ 0 & \frac{1}{2\xi^2} \end{pmatrix}$$

As the parameters are independent, their covariance is 0.

WORKING EXAMPLE : FISHER INFORMATION - GAMMA DISTRIBUTION

Let us recall the Gamma distribution, with its pdf given by:

$$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

Then the log-likelihood is:

$$l = \log f = -\log \Gamma(\alpha) - \alpha \log \beta + (\alpha - 1)\log x - x/\beta.$$

Hence:

$$\frac{\partial}{\partial \theta} \log f(X \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial \alpha} \log f \\ \frac{\partial}{\partial \beta} \log f \end{pmatrix} = \begin{pmatrix} -\psi(\alpha) - \log \beta + \log X \\ -\frac{\alpha}{\beta} + \frac{X}{\beta^2} \end{pmatrix}$$

Which gives:

$$I(\theta) = -E\begin{pmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\ \frac{\partial^2 l}{\partial \beta \partial \alpha} & \frac{\partial^2 l}{\partial \beta^2} \end{pmatrix} = -E\begin{pmatrix} -\psi'(\alpha) & \frac{-1}{\beta} \\ \frac{-1}{\beta} & \frac{\alpha}{\beta^2} - \frac{2X}{\beta^3} \end{pmatrix} = \begin{pmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}$$

This is the required FIM.

.

EXERCISES

- Find the score function of a Poisson distribution i.e X~Po(λ) when 3 independent data points are $X_1 = 3$, $X_2 = 5$, $X_3 = 7$, at parameter value λ = 0.1, 1 , 10, 100. Where is this most sensitive?
- Calculate the Fisher information for a Geometric distribution
- What is the Fisher information for a uniform distribution?
- How could we calculate the Fisher information for a discrete distribution?
- Calculate the FIM for a Weibull distribution