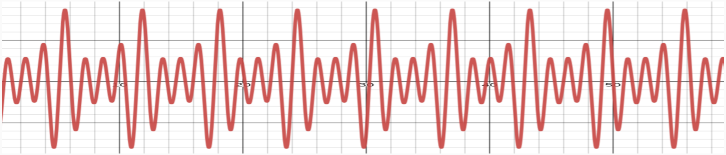
Likelihood Functions



Prof. Fabrício Olivetti de França

Federal University of ABC

05 February, 2024

Likelihood Functions

Likelihood Function

Let us recall how we describe the data distribution using a mass function, for discrete events, or density function, for continuous:

$$\sum_x f(x) = 1$$

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

Likelihood Function

If the shape of the distribution depends on parameters we write $f(x;\theta)$ to say the probability distribution of a random variable x when parameterized by θ.

Likelihood Function

The likelihood function is defined as $\mathcal{L}(\theta; x)$ also written as $\mathcal{L}(\theta \mid x)$ and has the same form as $f(x; \theta)$.

The main difference is that $f(x; \theta)$ is a function of x given a fixed θ and $\mathcal{L}(\theta; x)$ is a function of θ with a fixed x.

Interpretation

> The likelihood function should be interpreted as the probability of observing x if the true value of the parameter is θ.

This is not the probability of θ given x!!!

Likelihood Function

Assuming coin flipping events[1] where the probability distribution is parameterized by the probability of observing heads (p_H).

Assuming a fair coin, we have $p_H = 0.5$ and the probability of observing two heads is:

$$\boxed{P(HH; p_H = 0.5) = 0.25}$$

[1] https://en.m.wikipedia.org/wiki/Likelihood_function#Example

Likelihood Function

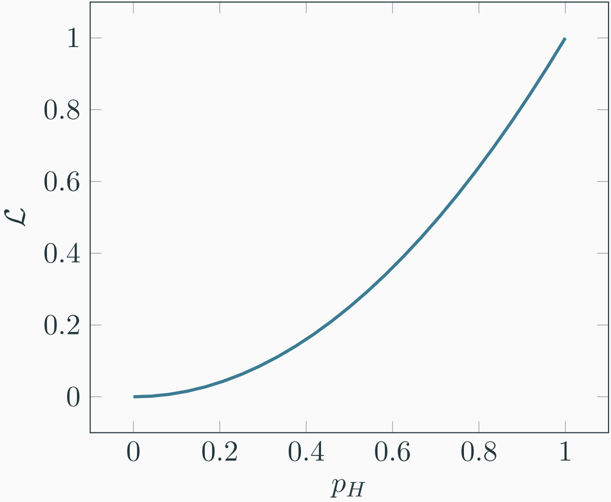
Now, the likelihood of $p_H = 0.5$ is given by:

$$\boxed{\mathcal{L}(p_H = 0.5; HH) = 0.25}$$

That is the probability of observing HH if $p_H = 0.5$.

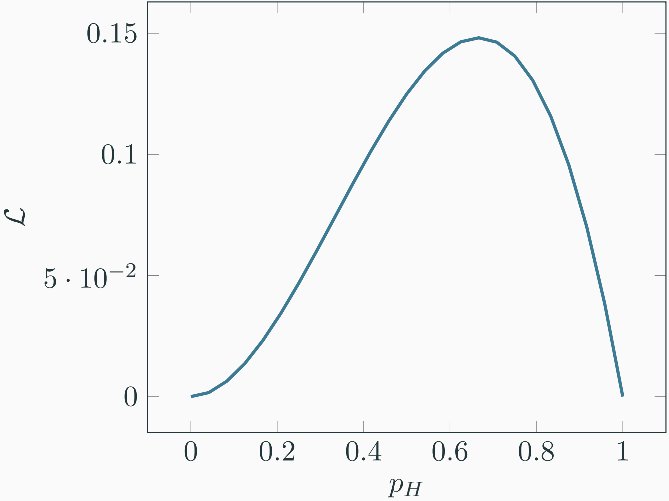
Likelihood Function

Let us see a graphical representation of the likelihood function.



Likelihood Function

Similarly for HHT we have that the likelihood is $p_h^2(1 - p_H)$.



Likelihood Function

For a continuous distribution we have that the probability density funcion describes the probability that a random variable x is within a range of value.

If we define this range as $[x^{(i)}, x^{(i)} + h]$ for $h > 0$, then the likelihood functions is

$$\mathcal{L}(\theta; x \in [x^{(i)}, x^{(i)} + h])$$

Likelihood Function

Since we want to find θ that maximizes the likelihood, we have:

$$\operatorname*{argmax}_{\theta} \mathcal{L}(\theta; x \in [x^{(i)}, x^{(i)} + h]) = \operatorname*{argmax}_{\theta} \frac{1}{h} \mathcal{L}(\theta; x \in [x^{(i)}, x^{(i)} + h])$$

by multiplying the likelihood function with a positive constant does not change the optima.

Likelihood Function

Since the probability density function and the likelihood has the same form, it follows:

$$\begin{aligned}\operatorname*{argmax}_{\theta} \frac{1}{h}\mathcal{L}(\theta; x \in [x^{(i)}, x^{(i)} + h]) &= \operatorname*{argmax}_{\theta} \frac{1}{h} P(x \in [x^{(i)}, x^{(i)} + h]; \theta) \\ &= \operatorname*{argmax}_{\theta} \frac{1}{h} \int_{x^{(i)}}^{x^{(i)}+h} f(x;\theta) dx\end{aligned}$$

Likelihood Function

Taking the limit of $h \rightarrow 0^+$ we can apply the fundamental theorem of calculus.

$$\lim_{h \rightarrow 0^+} \frac{1}{h} \int_{x^{(i)}}^{x^{(i)}+h} f(x;\theta)dx = f(x;\theta)$$

Likelihood Function

Which leads us to

$$\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta; x) = \underset{\theta}{\operatorname{argmax}}\, f(x; \theta)$$

Regularity conditions

The likelihood function is assumed to obey certain conditions called **regularity conditions**.

These conditions ensure that, asymptotically, the likelihood can be approximatted by the likelihood of a normal distribution.

This will help us to extend the calculation of confidence intervals for different distributions.

Regularity conditions

The three regularity conditions are:

1. The random variable is independently and identically distributed (we already assume that for our application).
2. There exists an open set $\Theta^* \subset \Theta \subset \mathbb{R}^P$ containing $\hat{\theta}$.
3. For all x, $f(x;\theta)$ is continuously differentiable w.r.t. θ up to third order derivative on Θ^*.

Regularity conditions

An **open set** generalizes open intervals.

For example, the points (x, y) such that $x^2 + y^2 < r^2$ represents an open set with the points $x^2 + y^2 = r^2$ the boundary set and the union of both sets, a **closed set**.

Regularity conditions

There are some conditions necessary to ensure the third condition, but one that we will use is the fact that the **Fisher Information matrix** is positive definite.

Fisher Information

Fisher Information measures how much information a random variable carries about the model parameters.

Fisher Information

The main intuition about the Fisher Information is that it measures the variance (second moment) of the probability distribution $f(x;\theta)$ w.r.t. θ.

If a change in θ propagates into a significant change in f, it means that the current data is enough to determine a good estimate of the true θ. Otherwise, if f is flat, it requires a large amount of data to find the true parameter.

The *flatter* the surface, the more data is required up until the whole population.

Fisher Information

The **score** is defined as the partial derivatives of $log f$ w.r.t. θ. Assuming the regularity conditions, if θ is the true parameter, the first moment is 0:

$$\begin{aligned}
E\left[\frac{\partial}{\partial\theta} log f(x;\theta)\Big|_{\theta=\hat{\theta}}\right] &= \int \frac{\partial}{\partial\theta} log f(x;\theta) dx \\
&= \int \frac{\partial}{\partial\theta} f(x;\theta) \frac{f(x;\theta)}{f(x;\theta)} dx \\
&= \int \frac{\partial}{\partial\theta} f(x;\theta) dx \\
&= \frac{\partial}{\partial\theta} \int f(x;\theta) dx \\
&= \frac{\partial}{\partial\theta} 1 \\
&= 0
\end{aligned}$$

Fisher Information

The Fisher Information is the second moment (variance) of the score:

$$\mathcal{F}(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f(x;\theta) \right)^2 \Bigg|_{\theta=\hat{\theta}} \right]$$

Fisher Information

By the regularity condition, $f(x;\theta)$ is twice differentiable, and we have that:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) &= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right)^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2
\end{aligned}$$

Fisher Information

And the expected value is:

$$\begin{aligned}
E\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right] &= E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] \\
&= E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)}\right] - E\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] \\
&= -E\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] \\
&= -\mathcal{F}(\theta) \\
\mathcal{F}(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]
\end{aligned}$$

So the Fisher information can be calculated as the expected value of the second order partial derivatives of the logarithm of $p(x;\theta)$

Fisher Information

The fact that the Fisher Information is positive definite means that the negative log-likelihood has a minimum value (is concave up) within Θ^*.

Negative Log-Likelihood

It is common to frame the optimization problem to fit the model to the data as the minimization of the **negative log-liklihood** (nll).

Fisher Information

Let us exemplify the Fisher Information for a Bernoulli distribution. The likelihood of this distribution is given by:

$$\mathcal{L}(\theta; x) = \theta^x (1 - \theta)^{1-x}$$

wih $x \in 0, 1$.

Fisher Information

The Fisher Information is then:

$$
\begin{aligned}
\mathcal{F}(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta;x)\right] \\
&= -E\left[\frac{\partial^2}{\partial\theta^2}\log(\theta^x(1-\theta)^{1-x})\right] \\
&= -E\left[\frac{\partial^2}{\partial\theta^2}(x\log\theta+(1-x)\log(1-\theta))\right] \\
&= E\left[\frac{x}{\theta^2}+\frac{1-x}{(1-\theta)^2}\right]
\end{aligned}
$$

Fisher Information

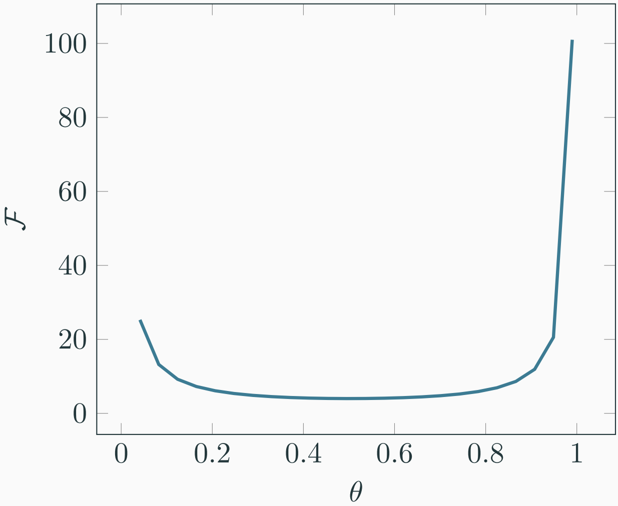
Since in Bernoulli distribuion $E[x] = \theta$

$$\begin{aligned}\mathcal{F}(\theta) &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)}\end{aligned}$$

For a fair coin where $\theta = 0.5$, the Fisher Information is 4.

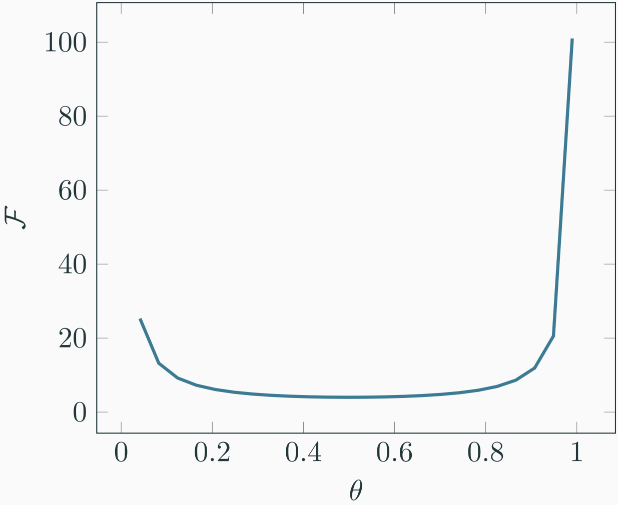
Fisher Information

The Fisher Information is the reciprocal of the variance!



Fisher Information

As $\theta \to 0^-$ or $\theta \to 1^+$, the Fisher Information grows toward infinite. This means that just a few data points are enough to find the correct value of θ.



Fisher Information Matrix

When we have P parameters, the Fisher Information is a $P \times P$ matrix called **Fisher Information Matrix**.

Laplace Approximation

One important consequence of the Fisher Information and its properties, is that we can approximate any likelihood function as a multivariate normal distribution:

$$f(x;\theta)\,\mathcal{N}(\theta, \mathcal{F}^{-1})$$

This can be used to approximate the confidence intervals for any likelihood function with a similar procedure we used for the linear regression models.

Likelihood Ratio

We can calculate the Likelihood Ratio as:

$$\Lambda(\theta_1 : \theta_2; x) = \frac{\mathcal{L}(\theta_1; x)}{\mathcal{L}(\theta_2; x)}$$

And it can be interpreted as how much the data support one parameter versus the other. It can also be used for statistical test comparing two hypotheses.

Relative Likelihood Ratio

Once you find the $th\hat{e}ta$ that maximizes the likelihood function (also called **maximum likelihood estimate**), we can calculate the relative likelihood ratio for any value of θ as:

$$R(\theta) = \frac{\mathcal{L}(\theta; x)}{\mathcal{L}(\hat{\theta}; x)}$$

This standardizes the likelihood to a maximum value of 1.

Likelihod Intervals

The relative likelihood ratio can be used to define a likelihood interval by finding the set of values θ such that:

$$\left\{\theta : R(\theta) \geq \frac{p}{100}\right\}$$

This is interpreted as the region in which the likelihood ratio is within a certain percentage. Not to be confused with the coverage Interpretation of confidence intervals.

Wilk's theorem says that two times the difference of the log-likelihood is approximately a chi-squared distribution. This will be explored later!

Likelihood of Common Regression Distributions

Likelihood of Common Regression Distributions

In regression analysis, the **Generalized Linear Models** (GLM) extends the linear model to support different distribution by means of a **link function** (g).

$$E[y \mid x; \beta] = g(\mu) = f(x; \beta) = (x\beta)$$

The linear model can be transformed with the inverse link function g^{-1} that maps the linear relationship to the mean of the desired distribution.

Likelihood of Common Regression Distributions

Likewise, given a nonlinear model $f(x;\theta)$ we can apply the same transformation to generalize to different distributions.

In the following slides we will see the link function, negative log-likelihood (nll), first and second order derivatives for commons distributions we will use throughout the remainder of the course.

The first oder derivative will be used during the parameter optimization process. The second-order will be also used for the optimization process and the calculation of the Fisher Information Matrix.

Likelihood of Common Regression Distributions

The following slides will depict the likelihood $\mathcal{L}(\mu; y)$ as the probability of observing the targer value y when the true mean of the distribution is μ.

We will use μ to represent $g^{-1}(f(x;\theta))$, the inverse of the link function applied to the nonlinear model.

Gaussian Likelihood

$$\mathcal{L}(\mu; y) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y^{(i)} - \mu^{(i)})^2/(2\sigma^2)}$$

$$\text{nll}(\mu; y) = -\sum_i \log((2\pi\sigma^2)^{-0.5} e^{-(y^{(i)} - \mu^{(i)})^2/(2\sigma^2)})$$

$$\text{nll}(\mu; y) = -\sum_i -0.5 \log(2\pi\sigma^2) - 0.5(y^{(i)} - \mu^{(i)})^2/\sigma^2$$

$$\text{nll}(\mu; y) = 0.5 \sum_i \left(\log(2\pi\sigma^2) + (y^{(i)} - \mu^{(i)})^2/\sigma^2\right)$$

Gaussian Likelihood

$$\begin{aligned}
\mu &= f(x;\theta) \\
\mathrm{nll}(\mu;y) &= 0.5\sum_i \Big(\log(2\pi\sigma^2) + (y^{(i)} - \mu^{(i)})^2/\sigma^2\Big) \\
\frac{\partial}{\partial\theta_j}\,\mathrm{nll}(\mu;y) &= \frac{1}{\sigma^2}\sum_i (\mu^{(i)} - y^{(i)}\frac{\partial}{\partial\theta_j} f(x^{(i)};\theta) \\
\frac{\partial}{\partial\theta_j\theta_k}\,\mathrm{nll}(\mu;y) &= \frac{1}{\sigma^2}\sum_i (\mu^{(i)} - y^{(i)})\frac{\partial}{\partial\theta_j\theta_k} f(x^{(i)};\theta) \\
&\quad - \frac{\partial}{\partial\theta_j} f(x^{(i)};\theta)\frac{\partial}{\partial\theta_k} f(x^{(i)};\theta)
\end{aligned}$$

Bernoulli Likelihood

$$\mathcal{L}(\mu; y) = \prod_i \mu^{y^{(i)}} (1-\mu)^{1-y^{(i)}}$$

$$\text{nll}(\mu; y) = -\sum_i \log(\mu^{y^{(i)}} (1-\mu)^{1-y^{(i)}})$$

$$\text{nll}(\mu; y) = -\sum_i (y^{(i)} \log(\mu) + (1-y^{(i)}) \log(1-\mu))$$

Bernoulli Likelihood

$$
\begin{aligned}
\mu &= \frac{1}{1 + e^{-f(x;\theta)}} \\
\mathrm{nll}(\mu; y) &= -\sum_i \left(y^{(i)} \log(\mu) + (1 - y^{(i)}) \log(1 - \mu)\right) \\
\frac{\partial}{\partial\theta_j} \mathrm{nll}(\mu; y) &= \sum_i (\mu^{(i)} - y^{(i)}) \frac{\partial}{\partial\theta_k} f(x^{(i)}; \theta) \\
\frac{\partial}{\partial\theta_j\theta_k} \mathrm{nll}(\mu; y) &= \sum_i \mu^{(i)}(1 - \mu^{(i)}) \frac{\partial}{\partial\theta_j} f(x^{(i)}; \theta) \frac{\partial}{\partial\theta_k} f(x^{(i)}; \theta) \\
&\quad + (\mu^{(i)} - y^{(i)}) \frac{\partial}{\partial\theta_j\theta_k} f(x^{(i)}; \theta)
\end{aligned}
$$

Poisson Likelihood

$$\mathcal{L}(\mu; y) = \prod_i \frac{(\mu^{(i)})^{y^{(i)}}}{y^{(i)}!} e^{-\mu^{(i)}}$$

$$\text{nll}(\mu; y) = -\sum_i \log(\frac{(\mu^{(i)})^{y^{(i)}}}{y^{(i)}!} e^{-\mu^{(i)}})$$

$$\text{nll}(\mu; y) = -\sum_i y^{(i)} \log \mu^{(i)} - \mu^{(i)} - y^{(i)} \log y^{(i)}$$

$$\text{nll}(\mu; y) = \sum_i \mu^{(i)} + y^{(i)} \log y^{(i)} - y^{(i)} \log \mu^{(i)}$$

Poisson Likelihood

$$
\begin{aligned}
\mu &= e^{f(x;\theta)} \\
\mathrm{nll}(\mu; y) &= \sum_i \mu^{(i)} + y^{(i)} \log y^{(i)} - y^{(i)} \log \mu^{(i)} \\
\frac{\partial}{\partial \theta_j} \mathrm{nll}(\mu; y) &= \sum_i (\mu^{(i)} - y^{(i)}) \frac{\partial}{\partial \theta_j} f(x^{(i)}; \theta) \\
\frac{\partial}{\partial \theta_j \theta_k} \mathrm{nll}(\mu; y) &= \sum_i \mu^{(i)} \frac{\partial}{\partial \theta_j \theta_k} f(x^{(i)}; \theta) + (\mu^{(i)} \\
&\qquad - y^{(i)}) \frac{\partial}{\partial \theta_j} f(x^{(i)}; \theta) \frac{\partial}{\partial \theta_k} f(x^{(i)}; \theta)
\end{aligned}
$$

Terminology learned today

- **Fisher Information:** measures how much information a random variable carries about the model parameters.
- **Laplace Approximation:** approximates a distribution as a multinomial normal distribution with covariance equals to the inverse of the Fisher Information Matrix.
- **Likelihood Ratio:** the ratio of likelihood for two different parameters.
- **Relative Likelihod Ratio:** the likelihood standardized by the maximum likelihood estimate.
- **Maximum Likelihood Estimate:** the value $\hat{\theta}$ that maximizes the likelihood.
- **Likelihod Interval:** the interval of θ that keeps a relatve likelihood ratio to a percentage.

Next lecture

- Nonlinear Optimization

Acknowledgments