21) Maximum Likelihood

The maximum likelihood gives us a rationale on how to estimate a value for a model parameter.

The <u>maximum likelihood principle</u> can be stated prosaically as:

> "<u>Given a dataset</u>, choose the value of the model parameters in such a way that the data is most likely."

<u>Example: 21.1</u> You have a coin that has probability p of landing on heads but you do not know this probability

So you flip coin 5 times and you observe the sequence

$$H \quad H \quad T \quad H \quad H$$

Let us first assume that you know that there are <u>only two possible</u> values for p, namely

$$p = 3/4 \quad \text{and} \quad p = 2/3$$

If you had to guess which of the two would you choose, what would you guess?

You start by calculating the probability of observing the data for each of the two possibilities:

$$P\left(\text{data} \,\middle|\, p=\frac{3}{4}\right) = \left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right) = \frac{81}{1024} \approx 0.079$$

$$P\left(\text{data} \,\middle|\, p=\frac{2}{3}\right) = \left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right) = \frac{16}{243} \approx 0.066$$

On the basis that the data is more likely to arise given $p=3/4$, you might guess that on the balance of probabilities $p=3/4$

Next let us assume you have no information about p.

Then you can still calculate

$$P(\text{data} \mid p) = p^4(1-p)$$
$$= p^4 - p^5$$

and there is one value of p that makes the probability of the observed data the largest.

The way to find this probability, we use that the slope of function is 0 at maximum.
In this case we have:

$$\frac{d}{dp}\left(p^4 - p^5\right) = p^3(4-5p) = 0$$

$$\Rightarrow 4 - 5p = 0$$

$$\Rightarrow p = 4/5$$

$p = 4/5$ is the maximum likelihood estimate for p.

Defn 21.2: Let $x_1, x_2, \ldots, x_n$ be modelled by random variables $X_1, X_2, \ldots, X_n$ with a joint parameter θ. Then the likelihood $L(\theta)$ is the probability (density) of observing data.

If X is discrete then

$$\boxed{L(\theta) = P(X = x_1, \ldots X = x_n) = p_{X_1, \ldots, X_n}(x_1, \ldots, x_n)}$$

If X is continuous then

$$\boxed{L(\theta) = f_{X_1, \ldots, X_n}(x_1, \ldots x_n)}$$

The maximum likelihood estimate for θ is a value $\hat{\theta}$ that maximises $L(\theta)$.

If we write $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ then

$$\boxed{\hat{\Theta} = h(X_1, X_2, \ldots, X_n)}$$

is the maximum likelihood estimator for θ.

Example: 21.3 A dataset $x_1, x_2, \ldots, x_n$ is modelled as an iid sample $X_1, \ldots, X_n$ from an $Exp(\lambda)$ distribution, i.e.

$$f_{X_i}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

We are interested in the value of λ. The

The likelelihood is given by

$$L(\lambda) = f_{X_1, X_2 \ldots, X_n}(x_1, x_2, \ldots, x_n)$$

$$= f_{X_1}(x_1) \cdot f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

[by iid, they are independant]

$$= \lambda e^{-\lambda x_1} \ldots \lambda e^{-\lambda x_n}$$

Here we used that the variables are iid sample and therefore independant. Hence joint density function factorises into product of densities of the individual variables.

The likelihood function $L(\lambda)$ is a continuous function and we can find its extrema by looking for 0's of the derivative.

$$\frac{d}{d\lambda} L(\lambda) = n \lambda^{n-1} e^{-\lambda(x_1 + \cdots + x_n)} + \lambda(x_1 + \cdots + x_n) e^{-\lambda(x_1 + \cdots + x_n)}$$

(by product rule)

$$= n \lambda^{n-1} e^{-\lambda(x_1 + \cdots + x_n)} \left(1 - \lambda \underbrace{\frac{(x_1 + \cdots + x_n)}{n}}_{\bar{x}_n} \right)$$

$$= n \lambda^{n-1} e^{-\lambda(x_1 + \cdots + x_n)} \left(1 - \lambda \bar{x}_n \right)$$

We know that $L(\lambda)$ has an <u>extremum at $\lambda = \hat{\lambda}$</u> if and only if

$$\frac{d}{d\lambda} L(\hat{\lambda}) = 0$$

$$\frac{d}{d\lambda} L(\hat{\lambda}) = 0$$

$$\Leftrightarrow$$

$$n\hat{\lambda} e^{-\lambda(x_1+\cdots+x_n)} \left(1 - \hat{\lambda}\bar{x}_n\right) = 0$$

$$\Leftrightarrow$$

$$\hat{\lambda} = 0 \quad \text{or} \quad 1 - \hat{\lambda}\bar{x}_n = 0$$

$$\Leftrightarrow$$

$$\hat{\lambda} = 0 \quad \text{or} \quad \hat{\lambda} = \frac{1}{\bar{x}_n}$$

We are <u>not</u> <u>interested</u> in the <u>extremum at $\hat{\lambda} = 0$</u>.

Therefore this implies that

$$\hat{\lambda} = \frac{1}{\bar{x}_n}$$

So see that we have a maximum and not a minimum, we can make a qualitative sketch or observe that

$$L(0) = 0,$$

$$L(\hat{\lambda}) = x^{-n} e^{-n} > 0$$

$$L(\infty) = 0 \quad (L(\lambda) \to 0 \text{ as } \lambda \to \infty)$$

So we have found the maximum likelihood estimate $\hat{\lambda}$ for λ.

Thus the maximum likelihood estimator for λ is

$$\hat{\Lambda} = \frac{1}{\overline{X}_n}$$

If as in this example, the likelihood consists of many factors, it is usually easier to work with log likelihood.

$$\boxed{\ell(\theta) = \log(L(\theta))}$$

because where the likelihood is a product of many factors requiring us to use the product rule to differentiate, the log likelihood is just a sum of many terms.

The log likelihood takes its maximum at the same location as the likelihood itself.
This is because

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\log(L(\theta)) = \frac{1}{L(\theta)} \cdot \frac{d}{d\theta}L(\theta)$$

and thus

$$\frac{d}{d\theta}\ell(\hat{\theta}) = 0 \iff \frac{dL(\hat{\theta})}{d\theta} = 0$$

Furthermore because logarithm is a strictly increasing function, a maximum of $\ell(\theta)$ is also a maximum of $L(\theta)$

The log likelihood in this example is

$$\ell(\lambda) = \log(L(\lambda))$$

$$= \log(\lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n})$$

$$= \log(\lambda e^{-\lambda x_1}) + \cdots + \log(\lambda e^{-\lambda x_n})$$

$$= [\log(\lambda) + -\lambda x_1 \log e] + \cdots + [\log(\lambda) + -\lambda x_n \log e]$$

$$= (\log(\lambda) - \lambda x_1) + \cdots + (\log(\lambda) - \lambda x_n)$$

$$= n \log \lambda - \lambda(x_1 + \cdots + x_n)$$

$\Rightarrow$

$$\ell(\lambda) = n \log \lambda - \lambda(x_1 + \cdots + x_n)$$

Differentiating $\ell(\lambda)$ with respect to λ,

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - (x_1 + \cdots + x_n)$$

$$= n\left(\frac{1}{\lambda} - \frac{(x_1 + \cdots + x_n)}{n}\right)$$

$$= n\left(\frac{1}{\lambda} - \bar{x}_n\right)$$

$$\frac{d}{d\lambda}\ell(\hat{\lambda}) = 0 \iff n\left(\frac{1}{\hat{\lambda}} - \bar{x}_n\right) = 0$$

$$n \in \mathbb{N} \neq 0 \Rightarrow \frac{1}{\hat{\lambda}} - \bar{x}_n = 0$$

$$\Rightarrow \hat{\lambda} = \frac{1}{\bar{x}_n}$$

It is now easy to check that this extremum is indeed a maximum by calculating the second derivative and observing it is negative

$$\frac{d^2 \ell(\hat{\lambda})}{d\lambda^2} = -\frac{n}{\hat{\lambda}^2} < 0$$

Thus the maximum likelihood estimator is

$$\hat{\Lambda} = \frac{1}{\overline{X}_n}$$

Example: 21.4 (Example 19.3 continued):
In this example dataset will be modelled by random variables N_1, N_2, N_3, N_4 are jointly multinomially distributed

$$P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) =$$

$$p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} \cdot \frac{n!}{n_1! \, n_2! \, n_3! \, n_4!}$$

where $p_1 = \frac{\theta + 2}{4}$, $p_2 = p_3 = \frac{(1-\theta)}{4}$, $p_4 = \frac{\theta}{4}$

We are interested in the parameter θ. Now we want to derive the maximum likelihood estimator for θ.

The likelihood is the probability to obtain the data, i.e.

$$L(\theta) = P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4)$$

$$= p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} \frac{n!}{n_1! n_2! n_3! n_4!}$$

$$= (\theta+2)^{n_1} (1-\theta)^{n_2+n_3} \theta^{n_4} \frac{1}{4^n} \frac{n!}{n_1! n_2! n_3! n_4!}$$

Again, it is easier to work with <u>log likelihood</u>

$$\ell(\theta) = \log\left((\theta+2)^{n_1} (1-\theta)^{n_2+n_3} \theta^{n_4} \frac{1}{4^n} \frac{n!}{n_1! n_2! n_3! n_4!}\right)$$

$$= n_1 \log(\theta+2) + (n_2+n_3)\log(1-\theta) + n_4 \log(\theta)$$

$$+ \log\left(\frac{1}{4^n} \frac{n!}{n_1! n_2! n_3! n_4!}\right)$$

To maximise we take derivative with respect to θ

$$\frac{d}{d\theta} \ell(\theta) = \frac{n_1}{\theta+2} - \frac{n_2+n_3}{1-\theta} + \frac{n_4}{\theta}$$

The maximum is at the value $\theta = \hat{\theta}$ where this derivative vanishes. Hence

$$\frac{d}{d\theta} \ell(\hat{\theta}) = 0$$

$$\Rightarrow \frac{n_1}{\theta+2} - \frac{n_2+n_3}{1-\theta} + \frac{n_4}{\theta} = 0$$

$\Rightarrow$

$$n(1-\theta)\theta - (n_2+n_3)(\theta+2)(\theta) + n_4(\theta+2)(1-\theta) = 0$$

$\Rightarrow$

$$-n\theta^2 + (n_1 - 2(n_2+n_3) - n_4)\,\theta + 2n_4$$

Let $m = n_1 - 2(n_1+n_3) - n_4$

Therefore

$$-n\theta^2 + m\theta + 2n_4 = 0$$

Using quadratic formula and taking positwe square root:

$$\hat{\theta} = \frac{m + \sqrt{m^2 + 8n_1 n_4}}{2n}$$

Putting numerical values from example $n_1 = 1997$, $n_2 = 906$, $n_3 = 904$, $n_4 = 32$, $n = 3839$

$$\hat{\theta} \approx 0.0357$$

The estimator for parameter θ is

$$\hat{\Theta} = \frac{M + \sqrt{M^2 + 8nM_4}}{2n}$$

where $M = N_1 - 2(N_2 + N_3) - N_4$

Example 21.5: Observations of number of earthquakes in UK in 3 different years:

$$n_1 = 16, \quad n_2 = 12, \quad n_3 = 20$$

Model these as realisations of <u>3 independant</u> <u>poisson distributed</u> random variable

$$N_i \sim Pos(\lambda) \qquad \text{for } i = 1, 2, 3$$

This means that

$$P(N_i = n_i) = \begin{cases} \frac{\lambda}{n_i!} e^{-\lambda} & \text{if } n_i \in \{0, 1, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

We want to find maximum likelihood estimator for the parameter λ. The likelihood is

$$L(\lambda) = p_{N_1, N_2, N_3}(n_1, n_2, n_3)$$

$$= p_{N_1}(n_1) \cdot p_{N_2}(n_2) \cdot p_{N_3}(n_3)$$

$$= \frac{\lambda^{n_1}}{n_1!} e^{-\lambda} \frac{\lambda^{n_2}}{n_2!} e^{-\lambda} \frac{\lambda^{n_3}}{n_3!} e^{-\lambda}$$

$$= \frac{\lambda^{n_1+n_2+n_3}}{n_1!\, n_2!\, n_3!} e^{-3\lambda}$$

Again, nicer to work with log likelihood.

$$\ell(\lambda) = \log(L(\lambda))$$

$$= (n_1+n_2+n_3)\log(\lambda) - 3\lambda - \log(n_1!\, n_2!\, n_3!)$$

At max, derivative is 0 at $\lambda = \hat{\lambda}$

$$\frac{d}{d\lambda}\ell(\hat{\lambda}) = \frac{n_1+n_2+n_3}{\hat{\lambda}} - 3 = 0$$

$$\Rightarrow \hat{\lambda} = \frac{n_1+n_2+n_3}{3}$$

Substituting data gives us

$$\hat{\lambda} = \frac{n_1 + n_2 + n_3}{3} = 16$$

The corresponding maximum likelihood estimator is

$$\hat{\Lambda} = \frac{N_1 + N_2 + N_3}{3} = \overline{N}_3$$

This estimator is <u>unbiased</u> because

$$E[\hat{\Lambda}] = E\left[\frac{N_1 + N_2 + N_3}{3}\right]$$

$$= \frac{1}{3}\left(E[N_1] + E[N_2] + E[N_3]\right)$$

$$= \lambda$$

The <u>mean squared</u> error MSE is

$$MSE(\hat{\Lambda}) = Var(\hat{\Lambda}) = Var\left[\frac{N_1 + N_2 + N_3}{3}\right]$$

$$= \frac{1}{9}\left(Var(N_1) + Var(N_2) + Var(N_3) \right)$$

$$= \frac{\lambda}{3}$$

$\Rightarrow$

$$MSE(\hat{\Lambda}) = \frac{\lambda}{3}$$

<u>Example</u>: Let dataset $x_1, \ldots, x_n > 0$ be modelled by iid
<u>21.6</u> sample from uniform distribution $X_i \sim U(0, \theta)$
We want to find the maximum likelihood estimator for θ.

The likelihood function is

$$L(\theta) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) = \begin{cases} 0 & \text{if } \theta < \max\{x_1 \ldots x_n\} \\ (1/\theta)^n & \text{otherwise} \end{cases}$$

because if any observed data is larger than θ, then the corresponding density function is 0 and if it is below θ, then the density is equal to $1/\theta$

The maximum likelihood estimate $\hat{\theta}$ for θ is the value for θ at which $L(\theta)$ takes its maximal value and thus

$$\hat{\theta} = \max\{x_1, \dots, x_n\}$$

The maximum likelihood estimator is

$$\hat{\Theta} = \max\{X_1, \dots, X_n\}$$

Next example, we <u>model 2 parameters</u>

<u>Example: 21.7</u> Let a dataset $x_1, \dots, x_n$ be modelled as an iid sample $X_1, \dots, X_n$ from a normal distribution $N(\mu, \sigma^2)$

We want to <u>determine maximum likelihood estimator</u> for

$$\mu \text{ and } \sigma^2$$

So in this case, model parameters is

$$\theta = (\mu, \sigma^2)$$

This just means that the likelihood will be a function of μ and σ^2 and we have to maximise it with respect to both.

The likelihood is

$$L(\mu, \sigma^2) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

where the probability density for the $N(\mu, \sigma^2)$ distributed random variable is

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

We will work with log likelihood so we need to calculate

$$\log f_X(x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$$

This is gives for the log likelihood:

$$\ell(\mu,\sigma^2) = \log L(\mu,\sigma^2)$$

$$= \log f_{X_1}(x_1) + \cdots + \log f_{X_n}(x_n)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left((x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right)$$

We want to find out about the extrema of this function of two variable.
Done by using partial derivatives.

$$\frac{\partial \ell}{\partial \mu}(\mu,\sigma^2) = \frac{-1}{2\sigma^2}\left(-2(x_1-\mu)^2 - \cdots - (x_n-\mu)^2\right)$$

$$= \frac{-1}{\sigma^2}(x_1 + \cdots + x_n - n\mu)$$

$$= \frac{n}{\sigma^2}(\bar{x}_n - \mu)$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^2}\left((x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right)$$

$$= \frac{-n}{2\sigma^4}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

There is an extremum at $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$ if and only if

$$\frac{\partial}{\partial l} l(\hat{\mu},\hat{\sigma}^2) = 0 = \frac{\partial}{\partial \sigma^2} l(\hat{\mu},\hat{\sigma}^2)$$

From the first condition

$$\frac{\partial}{\partial l} l(\hat{\mu},\hat{\sigma}^2) = 0 \Rightarrow \frac{n}{\hat{\sigma}^2}(\bar{x}_n - \hat{\mu}) = 0$$

$$\Rightarrow \hat{\mu} = \bar{x}_n$$

From second condition

$$\frac{\partial \ell(\hat{\mu}, \hat{\sigma}^2)}{\partial \sigma^2} = 0 \Rightarrow \frac{-n}{2\hat{\sigma}^2}\left(\hat{\sigma}^2 - \frac{1}{n}\sum (x_i - \bar{x}_n)^2\right) = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x}_n)^2$$

This extremum is infact a maximum. (To verify look at Thomas Calculus).

We have thus found the maximum likelihood estimates for the mean and standard deviation. Correspondingly, the maximum likelihood estimators for the mean is

$$\hat{M}_n = \bar{X}_n$$

and maximum likelihood estimator for the variance is

$$\hat{\Sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Note that this estimator has a different normalisation from the unbiased estimator S_n^2 from Thm 19.2

This means that $\hat{\Sigma}_n^2$ is not unbiased. Instead

$$E[\hat{\Sigma}_n^2] = \frac{n+1}{n} E[S_n^2] = \frac{n+1}{n} \sigma^2$$

However the bias gets smaller as sample size n increases and goes away in limit $n \to \infty$

$$\lim_{n \to \infty} E[\hat{\Sigma}_n^2] = 0$$

We say that the estimator is asymptotically unbiased