DATA SCIENCE		
MSDSDC201T		**Statistical Methods**
Unit 2.1 : Computational Method for finding Estimators - Likelihood and MLE		

LIKELIHOOD

If we flip a coin and find that it gives 60 heads and 40 tails, what would we guess to be the probability of getting heads? It is pretty reasonable to think that the probability of getting heads would be 60/100 = 0.6. However, let us examine this in some more detail. What exactly have we done here? We wanted to estimate the value of the parameter **p**(probability of getting heads), so to do that, we created an estimator (average number of heads), and then used this estimator to measure the value of **p**. In other words, we have used the data from the sample (60 heads and 40 tails) to estimate the value of **p**. However, multiple values of **p** are possible, for example, **p=0.1** could also generate this sample, and **p=0.9** could also generate a sample with 60 heads and 40 tails. In that case, how do we choose the correct value of **p**? In other words, given information in the form of a sample, how do we then infer the value of the parameter from this sample? The general procedure for this asks us to consider the probabilities of each value of **p**, given the sample information, and then choose the value of **p** which maximizes the probability function. This is called the method of ***maximum likelihood estimation***. Before we do that we need to understand Likelihood function.

Let us take another example to understand this process:
I have a bag that contains 3 balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it θ, might be 0, 1, 2, or 3. I am allowed to choose 4 balls at random from the bag with replacement. We define the random variables X_1, X_2, X_3, and X_4 as follows

$$X_i = \begin{cases} 1 & i-th\ ball\ chosen\ is\ blue \\ 0 & i-th\ ball\ chosen\ is\ red \end{cases}$$

Note that X_i's are i.i.d. and X_i~Bernoulli(θ/3). After doing my experiment, I observe the following values for X_i's : $x_1=1, x_2=0, x_3=1, x_4=1$.
Thus, I observe 3 blue balls and 1 red ball. I need to estimate θ from this. Let me consider the probability of getting each value of θ, given this sample by noting that
$P(X_1=x_1, X_2=x_2, X_3=x_3, X_4=x_4 \mid \theta) = P(X_1=x_1 \mid \theta) * P(X_2=x_2 \mid \theta) * P(X_3=x_3 \mid \theta) * P(X_4=x_4 \mid \theta)$
And since each X_i ~ Bernoulli (θ/3)

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3} & for\ x = 1 \\ 1-\frac{\theta}{3} & for\ x = 0 \end{cases}$$

We have $P(1,0,1,1) = \frac{\theta}{3} \cdot (1 - \frac{\theta}{3}) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} = \theta^3(1 - \frac{\theta}{3})$

We can calculate this in a table

θ	**P(θ) =** $\theta^3(1-\frac{\theta}{3})$
0	0
1	0.0247
2	0.0988
3	0

What would now be a good estimate of **θ?** It would make sense to choose **θ = 2** as this is what seems to have the highest probability. In short, we found the value of **θ** that maximizes the probability of occurring. This is the ***MAXIMUM LIKELIHOOD ESTIMATOR***

Let us recall that the goal of using estimators is to *estimate* the value of a population parameter i.e given samples, we wish to estimate the parameter from this data. In order to do that, we create a function called the Likelihood Function, which basically measures the probability of the estimating the parameter θ with the value $\hat{\theta}$.

The Likelihood function gives us an idea of how well the data summarizes these parameters. The "parameters" here are the parameters for a probability distribution function (PDF). In other words, they are the building blocks for a PDF, or what you need for parameterization. The likelihood function (often simply called the likelihood) is the joint probability (or probability density) of observed data viewed as a function of the parameters of a statistical model.

Suppose the joint probability density function of your sample **X** = (X_1,…X_n) is f(**x** | θ), where θ is a parameter,and X = **x** is an observed sample point. Then the function of θ defined as

$$\mathbf{L}(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta)$$

is called the **LIKELIHOOD FUNCTION**.

It certainly looks like we're just taking our PDF and cleverly relabeling it as a likelihood function. The reality, though, is actually quite different. For your PDF, you thought of θ as a constant and focused on an ever changing **x**. In the likelihood function, you let a sample point **x** be a constant and imagine θ to be varying over the whole range of possible parameter values.

If we compare two points on our probability density function, we'll be looking at two different values of **x** and examining which one has more probability of occurring. But for the likelihood function, we compare two different parameter points. For example, if we find that $L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x})$, we know that our observed point **x** is more likely to have been observed under parameter conditions $\theta = \theta_1$ rather than $\theta = \theta_2$.

Unlike probability density functions, likelihoods aren't normalized. The area under their curves does not have to add up to 1.

WORKING EXAMPLE : BAGS OF CANDY

Consider the candy M&M's sold in the USA and Canada. The bags sold in the United States have 50% red candies compared to 30% in those sold in Canada. In an experimental study, a sample of 5 candies were drawn from an unlabelled bag and 2 red candies were observed. Is it more plausible that this bag was from the United States or from Canada?

To understand this let us consider the Likelihood function for the system.
Consider the case that a red candy can be drawn out of the bag with probability **p**. This **p** serves as our parameter. Now if each draw is independent and drawing a red candy is labeled as "success", clearly this is a Bin(5,**p**) distribution. In this case what is the probability of drawing 2 reds? P(X=2) = $C_2^5 \cdot p^2 \cdot (1-p)^3$ i.e given that we know probability of getting red is **p**, then the Likelihood function is just $10 \cdot p^2 \cdot (1-p)^3$

The likelihood function is: $\mathbf{L}(p|\mathbf{x}) \propto p^2(1-p)^3$, and we know that it is either p=0.3 or 0.5.

$L(0.3|\mathbf{x}) = 0.03087 < 0.03125 = L(0.5|\mathbf{x})$, suggesting that it is more plausible that the bag used in the experiment was from the United States

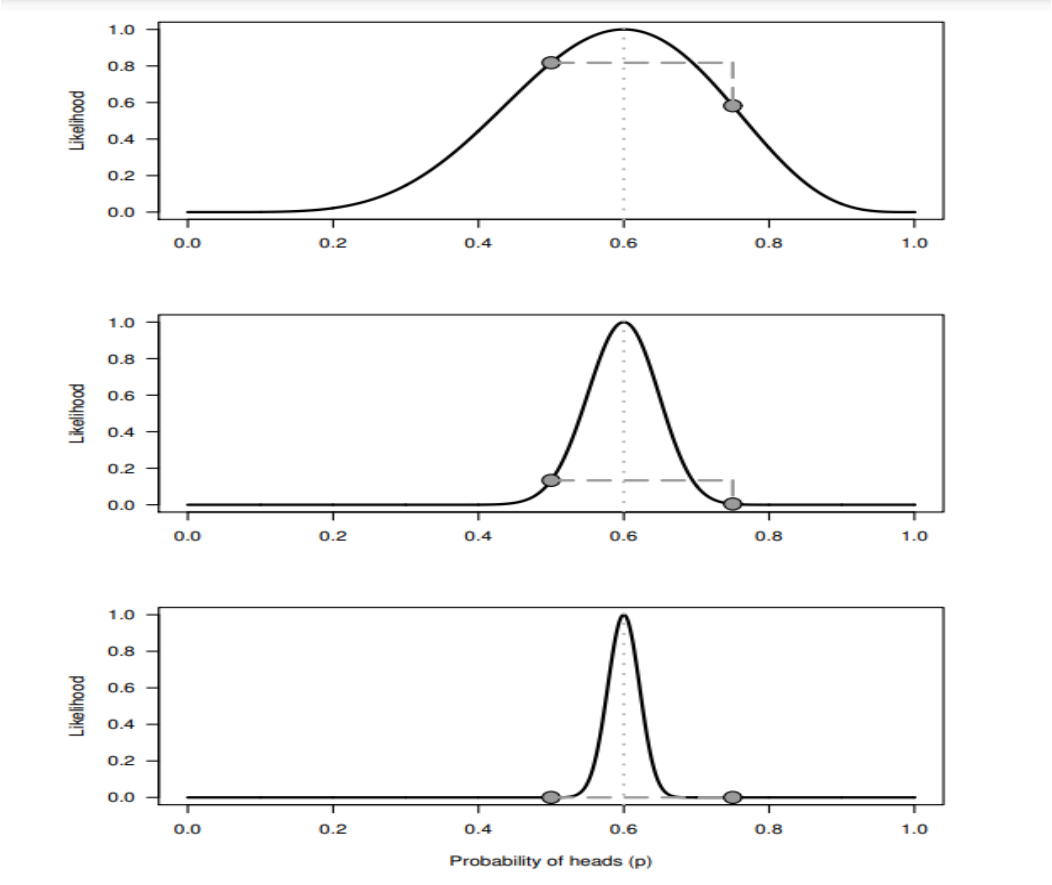


Figure 1. **Top.** The likelihood function for the case of 6 heads in 10 flips. **Middle.** The likelihood function for 60 heads in 100 flips. **Bottom.** The likelihood function for 300 heads in 500 flips.

LOG-LIKELIHOOD FUNCTION

A likelihood method is a measure of how well a particular model fits the data; They explain how well a parameter (θ) explains the observed data. The logarithms of likelihood, the log likelihood function, does the same job and is usually preferred for a few reasons:

- The log likelihood function in maximum likelihood estimations is usually computationally simpler.
- Likelihoods are often tiny numbers (or large products) which makes them difficult to graph. Taking the natural (base e) logarithm results in a better graph with large sums instead of products.
- The log likelihood function is usually (not always!) easier to optimize.

For example, let's say you had a set of iid observations x_1, $x_2 \ldots x_n$ with individual probability density function $f_X(x)$. Their joint density function is:

$f_{X1, X2 \ldots Xn}(x_1, x_2 \ldots x_n) = f_X(x_1) * f_X(x_2) * \ldots * f_X(x_n) =$

$$\prod_{i=1}^{n} f_X(x_i)$$

The log of a product is the sum of the logs of the multiplied terms, so we can rewrite the above equation with summation instead of products:

ln $[f_X(x_1) * f_X(x_2) * \ldots * f_X(x_n)] \quad =$

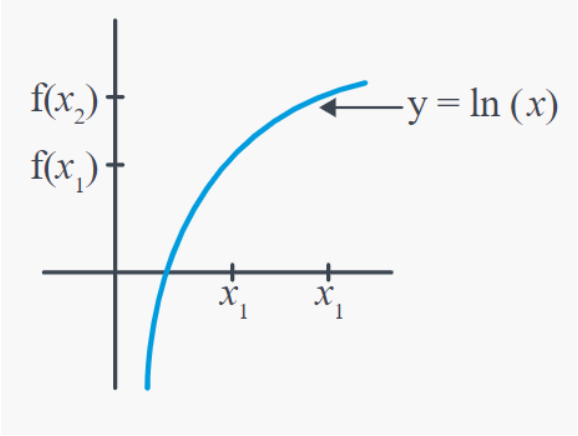
$$\sum_{i=1}^{n} \ln[\, f_X(x_i)\,]$$

<u>The above relationship leads directly to the log likelihood function:</u>

$$\mathbf{l(\Theta) = ln\ [\ L(\Theta)\ ]}$$

Where **L(Θ)** is the likelihood function and **l(Θ)** is called the ***Log-Likelihood function.***

We note here that log is an increasing function so any maximum of the function is also a maximum of the log of a function. This is because as we can see from the graph alongside, $x_1 < x_2$ means $\log(x_1) < \log(x_2)$



It turns out that finding the maximum of the log likelihood function is generally easier to do than finding the maximum of the likelihood function, this is why it is very important.

WORKING EXAMPLE : BINOMIAL DISTRIBUTION and LOG-LIKELIHOOD

Consider the Binomial Distribution with n-trials and probability of success being **p**.

Let us study what happens in case n=3.

No	Event	Probability
1	**000**	$(1-p)^3$
2	**001**	$p^1(1-p)^2$
3	**010**	$p^1(1-p)^2$
4	**011**	$p^2(1-p)^1$
5	**100**	$p^1(1-p)^2$
6	**101**	$p^2(1-p)^1$
7	**110**	$p^2(1-p)^1$
8	**111**	p^3

We can generalize this idea. Consider a sequence of n-trials with outcomes 11001…1 (say)

Then L[p|θ] = $f_X(x_1) * f_X(x_2) * \ldots * f_X(1) = f_X(1) * f_X(0) * \ldots * f_X(1) = p * p * (1-p) * \ldots * p$
= $p^r(1-p)^{n-r}$ where r is the number of heads in the sample i.e $r = X_1 + X_2 + \ldots X_n$.

Correspondingly, the Log Likelihood function is given by:
l(Θ) = $\ln[p^r(1-p)^{n-r}] = ln[p^r] + ln(1-p)^{n-r} = r * ln[p] + (n-r)ln(1-p)$ where r is the number of heads in the sample i.e $r = X_1 + X_2 + \ldots X_n$.

We want to choose the **p** which maximizes the likelihood function, and thus also, the log-likelihood function. To do this we use our knowledge of Calculus, and hence, we need to find the derivative of this log likelihood function:

$\frac{d}{dp}l(\Theta) = \frac{r}{p} - \frac{(n-r)}{1-p} = \frac{r(1-p)}{p(1-p)} - \frac{p(n-r)}{p(1-p)} = \frac{r-pn}{p(1-p)}$

Hence, we see that to minimize the log-likelihood, we must have $r - pn = 0$ i.e $p = \frac{r}{n}$.

This makes intuitive sense to us. If we flipped a coin 5 times and got 4 heads, it makes sense that the most likely value of **p** is 0.8. It turns out this is actually the most likely answer, as confirmed by Calculus.

MAXIMUM LIKELIHOOD ESTIMATION

The examples above give us the essence of what we are aiming to do: We want to estimate the (hidden) **parameter** from the available data. To do this we use estimators, as we have seen before. However, we need to calculate the value of the parameter, which we do by first calculating the *likelihood* function. The value of the parameter is then the most likely value of the parameter as suggested by the likelihood function. In other words, we find the value of the parameter with the ***maximum probability*** for the given data, and conclude that this is indeed the value of the parameter. Often, it is easier to compute the log-likelihood function to apply Calculus ideas of maxima/minima so we use that as an intermediate step.

This process is what we call **MAXIMUM LIKELIHOOD ESTIMATION** and the value of the parameter obtained is called the maximum likelihood estimate.

Maximum likelihood is easy to understand, intuitive and has important properties.

Let us study it in some detail, but first, let us formalize our understanding of the ideas.

We model a set of observations as a random sample from an unknown joint probability distribution which is expressed in terms of a set of parameters. The goal of maximum likelihood estimation is to determine the parameters for which the observed data have the highest joint probability. We write the parameters governing the joint distribution as a vector $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^\mathsf{T}$ so that this distribution falls within a parametric family $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$. Evaluating the joint density at the observed data sample, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ we get a function $\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta; \mathbf{y}) = f_n(\mathbf{y}; \theta)$,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:

$$f_n(\mathbf{y}; \theta) = \prod_{k=1}^{n} f_k^{\text{univar}}(y_k; \theta) \ .$$

The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space, i.e.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta\,; \mathbf{y}) \ .$$

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 $\mathcal{L}_n$ is called the **maximum likelihood estimate**.

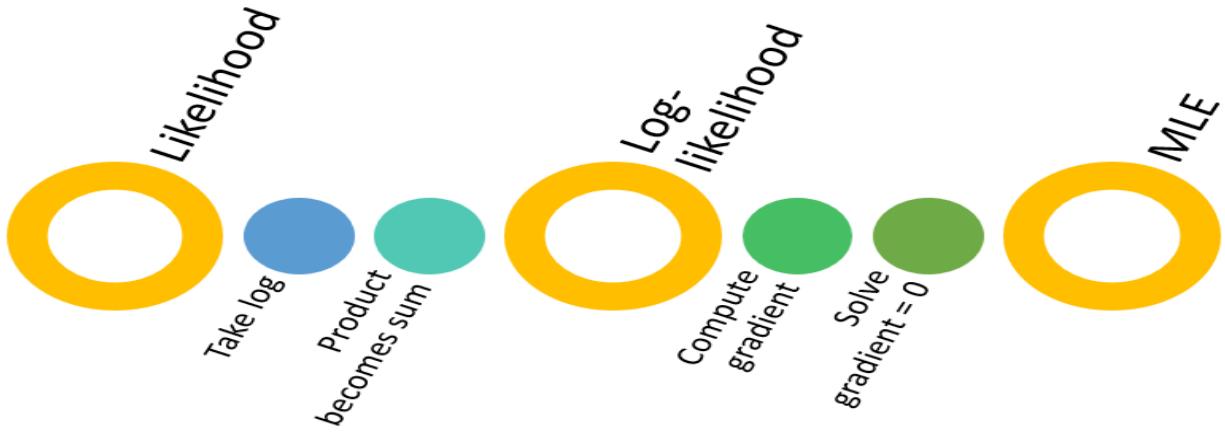
In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log-likelihood:

$$\ell(\theta\,;\mathbf{y}) = \ln \mathcal{L}_n(\theta\,;\mathbf{y})\ .$$

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.



WORKING EXAMPLE : BINOMIAL COIN PROBLEM

Let us say that we have a bag with three coins with known probability of getting heads. Coin 1(C1) has a probability ⅓ of getting heads, Coin 2(C2) has probability ½ of getting heads, and Coin 3 (C3) has probability ⅔ of getting heads. The coins look identical and cannot be distinguished from each other by physical means. How could we identify which coin is which?

One method would be for us to flip the coins and use the data obtained to identify the coins. For example, let us say that we grabbed a coin at random from the bag, and we flipped the coin 80 times. We obtain 49 heads and 31 tails. What can we say from the data? To do that, let us estimate the likelihood as follows: assuming that we know the coin is C1, what is the probability of getting 49H, 31T and so on. Using formulas for Bin(n,**p**) we get:

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

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

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

Hence we see that p=⅔ is the most likely outcome, which would allow us to conclude C3 is the most probable. Hence **p**=⅔ is the ***maximum likelihood estimate.***

PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

MLEs have many nice properties. Let θ_0 denote the true unknown value of the parameter θ. Under certain regularity conditions, we have the following:

- **CONSISTENCY** - The consistency means that if the data were generated by $f(\,.\,,\theta_0)$ 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 i.e. $\widehat{\theta} \rightarrow \theta_0$ in probability. In other words, the probability of an MLE deviating from the truth by any given fixed amount goes to 0 as the sample size increases. Another way of stating this is that, as the number of observations increase, the distribution of the maximum likelihood estimator becomes more and more concentrated about the true state of nature.

- **FUNCTIONAL INVARIANCE** - If $\widehat{\theta}$ is the maximum likelihood estimator for θ, and if g is any transformation of θ, then the maximum likelihood estimator for $\alpha = g(\theta)$ is just $\widehat{\alpha} = g(\widehat{\theta})$. This property is less commonly known as *functional equivariance*. The invariance property holds for arbitrary transformation g, although the proof simplifies if *g* is restricted to one-to-one transformations. This is an amazing property and it basically says that $\widehat{g(\theta)} = g(\widehat{\theta})$.

- **ASYMPTOTICALLY NORMAL** - Regardless of what the probability density f($\overline{X}$ | θ) of the sample is, we have that as the sample size increases, the distribution of the MLE approaches the normal density:

  $$\widehat{\theta} \sim \mathbf{N}(\theta_0, \mathbf{I}(\theta_0)^{-1})$$

  where $\mathbf{I}(\theta_0)$ is the Fisher information matrix, defined as the negative expectation of the second derivative of the log likelihood. The expectation we refer to represents the average over all repeated samples of the same size, generated by the same model (with the same true parameter θ_0) as the sample at hand. This tells us that the estimator converges to a normal distribution eventually.

- **EFFICIENCY** - If we consider a large enough sample, we will find that

  $$\mathrm{Var}_{\theta_0}(\hat{\theta}_n(X)) \approx \frac{1}{nI(\theta_0)}$$

  the lowest variance possible under the Cramer-Rao lower bound. We can write this in terms of the z-score. Let :

  $$Z_n = \frac{\hat{\theta}(X) - \theta_0}{1/\sqrt{nI(\theta_0)}}$$

  Then, as with the central limit theorem, Z_n converges in distribution to a standard normal random variable i.e achieves this lowest possible bound. This property is called **asymptotic efficiency**.

EXACT ESTIMATION OF MAXIMUM LIKELIHOOD ESTIMATORS

Now that we are clear about what Maximum Likelihood Estimation is and the properties of maximum likelihood estimators, the question becomes, how do we find them?

Generally the procedures for finding these estimators can be classified as:

- Exact methods (using Calculus)
- Approximate methods (numerical estimation)

While exact estimation is desirable, it is not always possible. Hence, we also need to have numerical estimation methods to help us out.

The general procedure of finding the exact Maximum Likelihood Estimator is to do the following:

1. Find the Likelihood function L(θ | **X**) i.e the probability of θ given the sample **X.**

2. Find the Log-Likelihood function ***l***(θ | **X**) of the Likelihood function L(θ | **X**). This is generally easier to compute and use.

3. Compute the maxima of ***l***(θ | **X**) by using ideas from Calculus: differentiate ***l***(θ | **X**) with respect to **X** (partial derivatives). Equate these to zero. Find the points of θ from this computation.

4. Verify this is a maxima by checking the second derivative (it should be negative).

WORKING EXAMPLE : MLE FOR POISSON DISTRIBUTION

We observe n independent draws from a Poisson distribution.
In other words, there are n independent Poisson random variables $X_1, X_2, \ldots X_n$ and we observe their realizations $x_1, x_2, x_3, \ldots x_n$.

The probability mass function of a single draw X_j is

$$p_X(x_j) = \begin{cases} \exp(-\lambda_0)\frac{1}{x_j!}\lambda_0^{x_j} & \text{if } x_j \in R_X \\ 0 & \text{if } x_j \notin R_X \end{cases}$$

where:

- λ_0 is the parameter of interest (for which we want to derive the MLE);
- the support of the distribution is the set of non-negative integer numbers
- $x_j!$ is the factorial of x_j.

The n observations are independent. As a consequence, the likelihood function is equal to the product of their probability mass functions. Hence, the likelihood function is

$$L(\lambda; x_1, \ldots, x_n) = \prod_{j=1}^{n} \exp(-\lambda)\frac{1}{x_j!}\lambda^{x_j}$$

From this, we can calculate the log-Likelihood function:

$$l(\lambda; x_1, \ldots, x_n) = -n\lambda - \sum_{j=1}^{n} \ln(x_j!) + \ln(\lambda) \sum_{j=1}^{n} x_j$$

From this we can now proceed to estimate the Maximum Likelihood Estimator:

$$\begin{aligned} l(\lambda; x_1, \ldots, x_n) &= \ln\left(\prod_{j=1}^{n} \exp(-\lambda) \frac{1}{x_j!} \lambda^{x_j} \right) \\ &= \sum_{j=1}^{n} \ln\left(\exp(-\lambda) \frac{1}{x_j!} \lambda^{x_j} \right) \\ &= \sum_{j=1}^{n} [\ln(\exp(-\lambda)) - \ln(x_j!) + \ln(\lambda^{x_j})] \\ &= \sum_{j=1}^{n} [-\lambda - \ln(x_j!) + x_j \ln(\lambda)] \\ &= -n\lambda - \sum_{j=1}^{n} \ln(x_j!) + \ln(\lambda) \sum_{j=1}^{n} x_j \end{aligned}$$

Differentiating wrt λ and setting it to zero we obtain:

$$\widehat{\lambda}_n = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Therefore, the estimator $\widehat{\lambda}$ is just the sample mean of the n observations in the sample.

This makes intuitive sense because the expected value of a Poisson random variable is equal to its parameter λ_0, and the sample mean is an unbiased estimator of the expected value.

WORKING EXAMPLE : MLE FOR NORMAL DISTRIBUTION

Our sample is made up of the first n terms of an IID sequence $\{X_n\}$ of normal random variables having mean μ_0 and variance σ_0^2. The probability density function of a generic term of the sequence is

$$f_X(x_j) = \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left(-\frac{1}{2} \frac{(x_j - \mu_0)^2}{\sigma_0^2} \right)$$

The mean μ_0 and the variance σ_0^2 are the two parameters that need to be estimated.
Given the assumption that the observations from the sample are IID, the likelihood function can be written as

$$L(\mu,\sigma^2;x_1,\ldots,x_n) = \prod_{j=1}^{n} f_X(x_j;\mu,\sigma^2)$$
$$= \prod_{j=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\frac{(x_j-\mu)^2}{\sigma^2}\right)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

Thus, the likelihood function is

$$L(\mu,\sigma^2;x_1,\ldots,x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

We can find the log-Likelihood function by taking the logarithm:

$$l(\mu,\sigma^2;x_1,\ldots,x_n) = \ln\left(L(\mu,\sigma^2;x_1,\ldots,x_n)\right)$$
$$= \ln\left((2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)\right)$$
$$= \ln\left((2\pi\sigma^2)^{-n/2}\right) + \ln\left(\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)\right)$$
$$= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$
$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

Hence, the log-likelihood function is

$$l(\mu,\sigma^2;x_1,\ldots,x_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

To find the MLE, we need to solve the following maximization problem: $\max_{\mu,\sigma^2} l(\mu,\sigma^2;x_1,\ldots,x_n)$

To do this, we need to calculate:

$$\frac{\partial}{\partial\mu} l(\mu,\sigma^2;x_1,\ldots,x_n) = 0$$
$$\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2;x_1,\ldots,x_n) = 0$$

The partial derivative of the log-likelihood with respect to the mean is:

$$\frac{\partial}{\partial\mu} l(\mu,\sigma^2;x_1,\ldots,x_n)$$

$$= \frac{\partial}{\partial\mu}\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

$$= \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j-\mu)$$

$$= \frac{1}{\sigma^2}\left(\sum_{j=1}^{n}x_j - n\mu\right)$$

which is equal to zero only if

$$\sum_{j=1}^{n}x_j - n\mu = 0 \quad \text{thus,} \quad \mu = \frac{1}{n}\sum_{j=1}^{n}x_j$$

The partial derivative of the log-likelihood with respect to the variance is

$$\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2;x_1,\ldots,x_n)$$

$$= \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

$$= -\frac{n}{2\sigma^2} - \left[\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2\right]\frac{d}{d\sigma^2}\left(\frac{1}{\sigma^2}\right)$$

$$= -\frac{n}{2\sigma^2} - \left[\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2\right]\left(-\frac{1}{(\sigma^2)^2}\right)$$

$$= -\frac{n}{2\sigma^2} + \left[\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2\right]\frac{1}{(\sigma^2)^2}$$

$$= \frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2 - n\right]$$

Ruling out $\sigma^2 = 0$ this tells $\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)^2$

Thus, the MLE estimators are just the sample mean and the sample variance.

$$\hat{\mu}_n = \frac{1}{n}\sum_{j=1}^{n}x_j$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\hat{\mu})^2$$

EXERCISES:

- Bag A contains 9 green balls and 4 red balls. Ball B contains 10 red balls and 3 green balls. Mr. Hu is blindfolded and given a bag, from which he draws first a green ball, then a red ball, then a green ball, each time replacing the ball in the bag. Construct a Likelihood function for Bag A and Bag B for this sample. Which bag is more likely?

- There are two types of trains that stop at Station X. Local trains which have an average frequency of 3 every hour, and out-station trains which have an average frequency of 1 every 3 hours. Assuming these trains follow a Poisson distribution with the corresponding average rates, construct a Likelihood function for each type of train. One Sunday, Mr. Y calls the station and is informed that only one train has stopped in the past hour. Which type of train is it more likely to have been?

- Dinesh runs two chaat stalls serving three types of chaat: Bhelpuri and Sevpuri and Panipuri. The percentage of customers ordering these is given in table:

	Bhelpuri(B)	**Sevpuri(S)**	**Panipuri(P)**
Stall A	33%	50%	17%
Stall B	20%	40%	40%

  One day, he finds a bill with an order of two bhelpuris, and one panipuri. Which Stall is more likely to have been where the order was placed?

- Suppose we have a random sample$X_1,X_2,X_3,..,X_n$ where:
  - $X_i=0$, if a randomly selected student does not own a sports car, and
  - $X_i=1$, if a randomly selected student does own a sports car.

  Assuming that the X_is are independent Bernoulli random variables with unknown parameter **p**, find the maximum likelihood estimator of **p**, the proportion of students who own a sports car.

- A bag contains N cards. What would be the MLE for N, if the sampling procedure we use is to just pull out a card?