Maximum Likelihood Estimation

Guy Lebanon

February 19, 2011

Maximum likelihood estimation is the most popular general purpose method for obtaining estimating a distribution from a finite sample. It was proposed by Fisher about 100 years ago and has been widely used since.

Definition 1. Let $X^{(1)}, \ldots, X^{(n)}$ be sampled iid[1] from a distribution with a parameter θ that lies in a set Θ. The maximum likelihood estimator (MLE) is the $\theta \in \Theta$ that maximizes the likelihood function

$$L(\theta) = L(\theta | X^{(1)}, \ldots, X^{(n)}) = p_\theta(X^{(1)}, \ldots, X^{(n)}) = \prod_{i=1}^{n} p_\theta(X^{(i)})$$

where p above is the density function if X is continuous and the mass function if X is discrete. The MLE is denoted $\hat{\theta}$ or $\hat{\theta}_n$ if we wish to emphasize the sample size.

Above, we suppress the dependency of L on $X_1, \ldots, X^{(n)}$ to emphasize that we are treating the likelihood as a function of θ. Note that both $X^{(i)}$ and θ may be scalars or vectors (not necessarily of the same dimension) and that L may be discrete or continuous in either X and θ or both or neither.

1. Strictly monotonic increasing functions g preserve order in the sense that the maximizer of $L(\theta)$ is the same as the maximizer of $g(L(\theta))$. As a consequence, we can find the MLE by obtaining the maximizer of $\log L(\theta)$ rather than the likelihood itself which is helpful since it transforms the multiplicative likelihood into a sum (sums are easier to differentiate than products). A common notation for the log of the likelihood is $\ell(\theta) = \log L(\theta)$.

2. Any additive and multiplicative terms in $\ell(\theta)$ that are not a function of θ may be ignored since it dropping them will not change the maximizer.

3. If $L(\theta)$ is differentiable in θ, we can try to find the MLE by solving the equation $d\ell(\theta)/d\theta = 0$ for scalar θ or the system of equations $\nabla \ell(\theta) = 0$ i.e., $\partial \ell(\theta)/\partial \theta_j = 0$, $j = 1, \ldots, d$ for vector θ. The obtained solutions are necessarily critical points (maximum, minimum or inflection) of the log-likelihood. To actually prove that the solution is a maximum we need to show in the scalar case $d^2\ell(\theta)/d\theta^2 < 0$ or if θ is a vector that the Hessian matrix $H(\theta)$ defined by $[H(\theta)]_{ij} = \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}$ is negative definite[2] (at the solution of $\nabla \ell(\theta) = 0$).

4. If the above method does not work (we can't solve $\nabla \ell(\theta) = 0$) we can find the MLE by iteratively following in the direction of the gradient: initialize θ randomly and iterate $\theta \leftarrow \theta + \alpha \nabla \ell(\theta)$ (where α is a sufficiently small step size) until convergence e.g., $\|\nabla \ell(\theta)\| \leq \epsilon$ or $\|\theta^{(t+1)} - \theta^{(t)}\| < \epsilon$. Since the gradient vector points in the direction of steepest ascent this will bring us to a maximum point.

5. The MLE is invariant in the sense that for all 1-1 functions h: denoting the MLE for θ by $\hat{\theta}$ we have that the MLE for $h(\theta)$ is $h(\hat{\theta})$. The key property is that 1-1 functions have an inverse. If we use the parametrization $h(\theta)$ rather than θ, the likelihood function is $L \circ h^{-1}$ rather than L. This can be shown by noting that for 1-1 $\eta = h(\theta)$, $p_\theta(X) = p_{h^{-1}(\eta)}(X)$ and thus the likelihood function of η is $K(\eta) = L(h^{-1}(\eta))$. We conclude that if $\hat{\theta}$ is the MLE of θ then

$$L(h^{-1}(h(\hat{\theta}))) = L(\hat{\theta}) \geq L(\theta) = L(h^{-1}(h(\theta)))$$

and $h(\hat{\theta})$ is the MLE for $h(\theta)$.

Potential problems are:

- When ℓ and L are not differentiable we can not solve $\nabla \ell(\theta) = 0$ nor can we follow the gradient to convergence. Solution: if θ is low dimensional vector use a grid search, otherwise use non-smooth optimization techniques.

[1] iid stands for independent identically distributed, that is the samples are independent draws from the same distribution.

[2] A negative definite matrix H is one that satisfies $v^\top H v < 0$ for all vectors v. A condition that is equivalent to negative definiteness in symmetric matrices (H is symmetric) and that is often easier to verify is that all eigenvalues are negative.

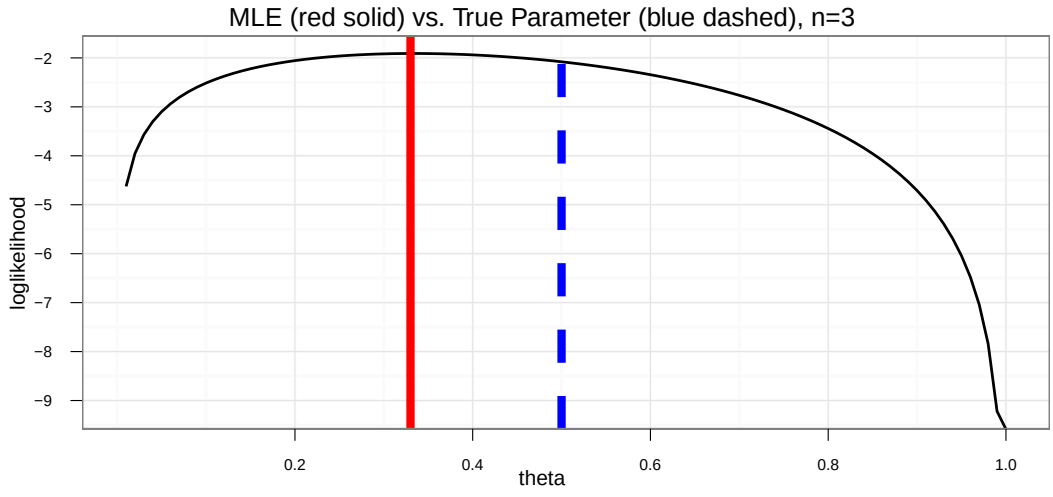
- $\ell(\theta)$ has multiple local maxima. Solution: Restart multiple gradient ascent procedures with different starting positions and pick the solution that is the highest of all the maximizers.
- There are multiple global maxima. Solution: pick one or pick multiple ones and report all as potential estimation values.
- $\ell(\theta)$ has no maximum point in which case the iterative algorithm may never converge. Solution: Use regularization to penalize high values of θ i.e., $\ell(\theta) + \lambda\|\theta\|^2$ or use early stopping.

Example 1: Let $X^{(1)}, \ldots, X^{(n)} \sim \text{Ber}(\theta)$ (Bernoulli distribution means $X = 1$ with probability θ and 0 with probability $1 - \theta$ (and 0 probability for all other values); It is identical to binomial with parameters $N = 1$. The likelihood is $L(\theta) = \prod \theta^{X^{(i)}}(1-\theta)^{1-X^{(i)}} = \theta^{\sum X^{(i)}}(1-\theta)^{n-\sum X^{(i)}}$ and the log-likelihood is $\ell(\theta) = (\sum X^{(i)}) \log\theta + (n - \sum X^{(i)}) \log(1-\theta)$. Setting the loglikelihood derivative to zero yields $0 = \sum X^{(i)}/\theta - (n - \sum X^{(i)})/(1-\theta)$ or $0 = (1-\theta)\sum X^{(i)} - (n - \sum X^{(i)})\theta = \sum X^{(i)} - n\theta$ which implies $\hat{\theta} = \frac{1}{n}\sum X^{(i)}$. Note that the probability p and the loglikelihood are neither continuous nor smooth in X but are both continuous and smooth in θ. For example the graph below are the loglikelihood of a sample from a Bernouli distribution with $\theta = 0.5$ and $n = 3$ which is maximized at $\hat{\theta}$ being the empirical average in accordance with the math solution above.

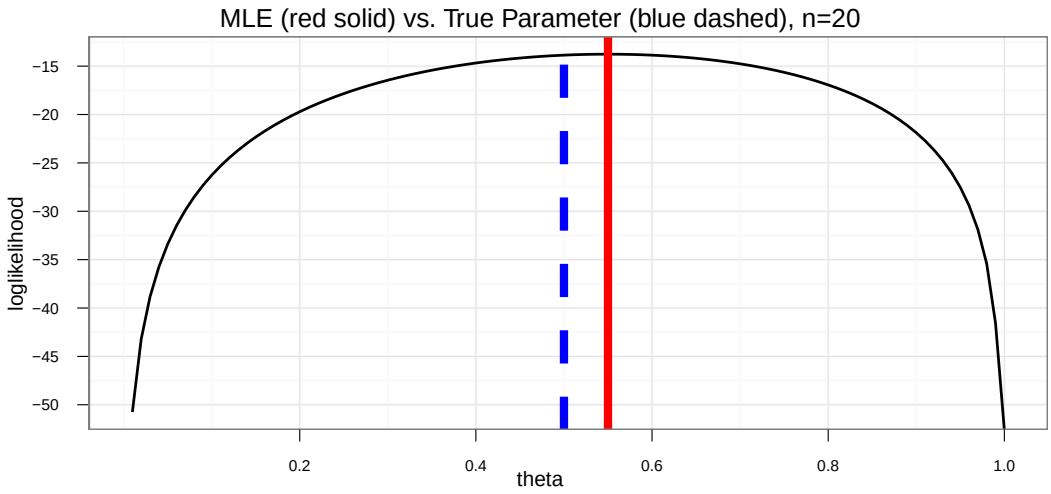
```
> theme_set(theme_bw(base_size=8));set.seed(0); # for reproducable experiment
> n=3;theta=0.5;samples=rbinom(n,1,theta);samples
```

```
[1] 1 0 0
```

```
> D=data.frame(theta=seq(0.01,1,length=100));
> D$loglikelihood=sum(samples) * log(D$theta) + (n-sum(samples)) * log (1-D$theta);
> mle=which.max(D$loglikelihood);
> p=ggplot(D,aes(theta,loglikelihood)) + geom_line()
> print(qplot(theta,loglikelihood,geom='line',data=D,
+             main='MLE (red solid) vs. True Parameter (blue dashed), n=3')+
+       geom_vline(aes(xintercept=theta[mle]),color=I('red'),size=1.5)+
+       geom_vline(aes(xintercept=0.5),color=I('blue'),lty=2,size=1.5))
```



Contrast the above graph with the one below, which corresponds to $n = 20$ to see the improvement of $\hat{\theta}_{20}$ over $\hat{\theta}_3$.



Example 2: Let $X^{(1)}, \ldots, X^{(n)} \sim N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$. The log-likelihood is

$$\ell(\theta) = \log \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} e^{-(X^{(i)}-\mu)^2/(2\sigma^2)} = c - \frac{n}{2}\log\sigma^2 + \log e^{-\sum_{i=1}^{n}(X^{(i)}-\mu)^2/(2\sigma^2)} = c - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(X^{(i)}-\mu)^2}{2\sigma^2}$$

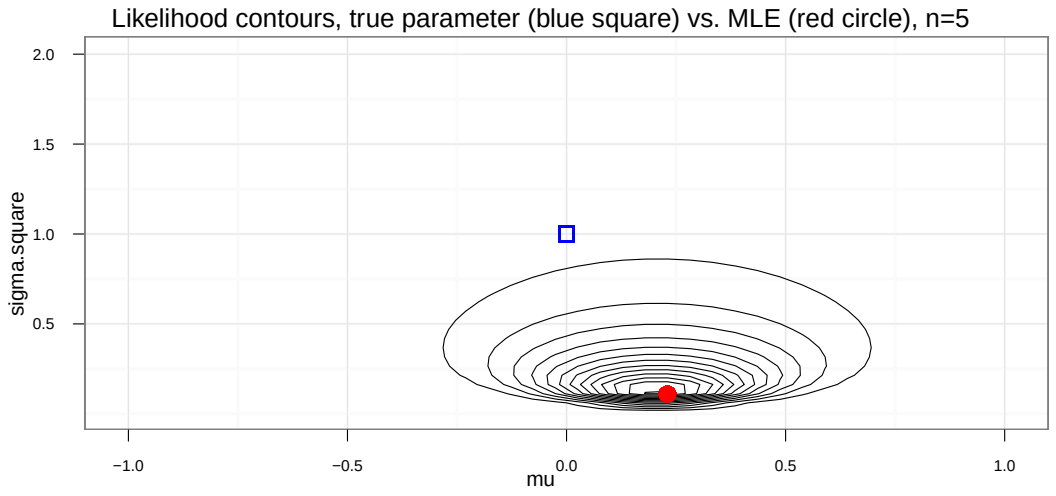
where c is an inconsequential additive constant. Setting the partial derivative with respect to μ to zero gives

$$\frac{\partial\ell(\theta)}{\partial\mu} = \sum \frac{X^{(i)}-\mu}{\sigma^2} = 0 \quad \Rightarrow \quad \hat{\mu} = \bar{X} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X^{(i)}.$$

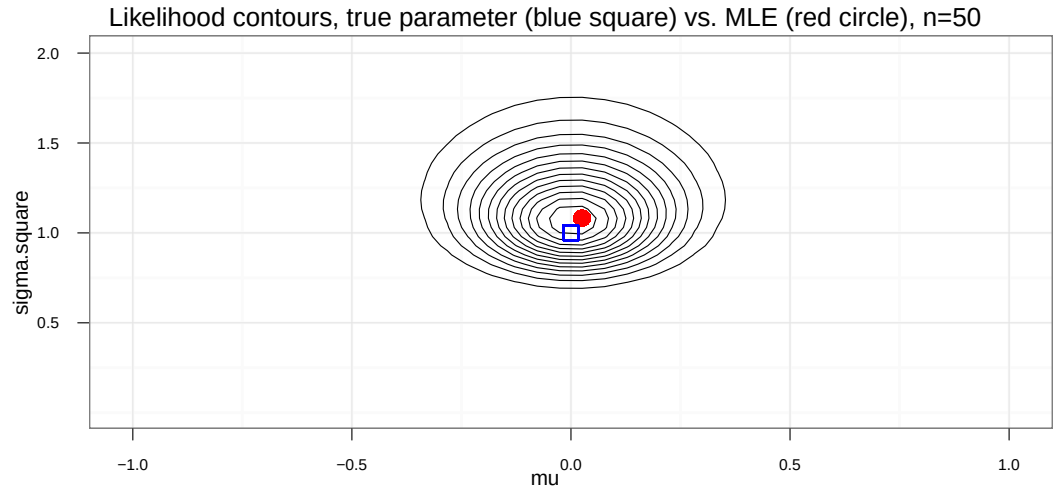
Substituting this in the equation resulting from setting the partial derivative with respect to σ^2 to 0: $0 = \frac{\partial\ell(\theta)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \sum \frac{(X^{(i)}-\mu)^2}{2\sigma^4}$ or $0 = -\frac{n}{2\sigma^2} + \sum \frac{(X^{(i)}-\bar{x})^2}{2\sigma^4}$ or $\sigma^2 n + \sum (X^{(i)} - \bar{x})^2 = 0$ which implies $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \bar{X})^2$. By property 4 above, the MLE for the standard deviation is $\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \bar{X})^2}$.

In this case θ is a two dimensional vector and so the likelihood function forms a surface in three dimensions. It is convenient to visualize using a heat-map or contour plot which plots the value of $\ell(\theta)$ in terms of colors or equal value contours. For example compare the two graphs below showing the likelihood as a function of $\theta = (\mu, \sigma^2)$ for $n = 5$ and $n = 50$. The MLE is clearly more accurate in the latter case.

```
> n=5;samples=rnorm(n,0,1); # 5 samples from N(0,1)
> mu=seq(-1,1,length=40);v=seq(0.01,2,length=40);
> D=expand.grid(mu=mu,sigma.square=v); # create all combinations of the two parameters
> nloglik=function(theta,samples) { # loglikelihood function
+     l=-log(theta[,2])*n/2;  # works for multiple theta values arranged in data frame
+     for (s in samples) l=l-(s-theta[,1])^2/(2*theta[,2]);
+     return(l);
+ }
> D$loglikelihood=nloglik(D,samples);
> p=ggplot(D,aes(mu,sigma.square,z=exp(loglikelihood)))+stat_contour(size=0.2);
> mle=which.max(D$loglikelihood);
> print(p+geom_point(aes(mu[mle],sigma.square[mle]),color=I('red'),size=3)+
+       geom_point(aes(0,1),color=I('blue'),size=3,shape=22)+opts(title=
+    'Likelihood contours, true parameter (blue square) vs. MLE (red circle), n=5'))
```




```
> n=50;samples=rnorm(n,0,1); # 50 samples from N(0,1)
> mu=seq(-1,1,length=40);v=seq(0.01,2,length=40);
> D=expand.grid(mu=mu,sigma.square=v);
> D$loglikelihood=exp(nloglik(D,samples));
> p=ggplot(D,aes(mu,sigma.square,z=exp(loglikelihood)))+stat_contour(size=0.2);
> mle=which.max(D$loglikelihood);
> print(p+geom_point(aes(mu[mle],sigma.square[mle]),color=I('red'),size=3)+
+       geom_point(aes(0,1),color=I('blue'),size=3,shape=22)+opts(title=
+   'Likelihood␣contours,␣true␣parameter␣(blue␣square)␣vs.␣MLE␣(red␣circle),␣n=50'))
```



R can also be used to find the MLE using numeric iterative methods such as gradient descent. This can be very useful when setting the loglikelihood gradient to zero does not result in closed form solution. Below is an example of using R's `optim()` function to numerically optimize the Gaussian likelihood (even though we don't really need to in this case since there is a closed form solution). The `optim()` function takes three arguments: an initial parameter value to start the iterative optimization, the objective function to minimize (minus the logikelihood in our case), and the parameter to pass to the objective function. Comparing the MLE obtained using the iterative numeric algorithm with the grid search MLE obtained in the previous example we see that the former is more accurate than latter. As the grid resolution is increased that difference will converge to 0.

```
> # numerical MLE for Gaussian parameters using 50 samples from the previous example
> nnloglik=function(theta,samples) { # loglikelihood function
+     l=-log(theta[2])*n/2;  # works for multiple theta values arranged in data frame
+     for (s in samples) l=l-(s-theta[1])^2/(2*theta[2]);
+     return(-l);
```

```
+ }
> iterative.mle=optim(c(0.5,0.5),nnloglik,samples=samples);
> iterative.mle$par # mle using numeric gradient descent
```

```
[1] 0.004178593 1.065951133
```

```
> c(D$mu[mle],D$sigma.square[mle]) # mle using grid search (above example)
```

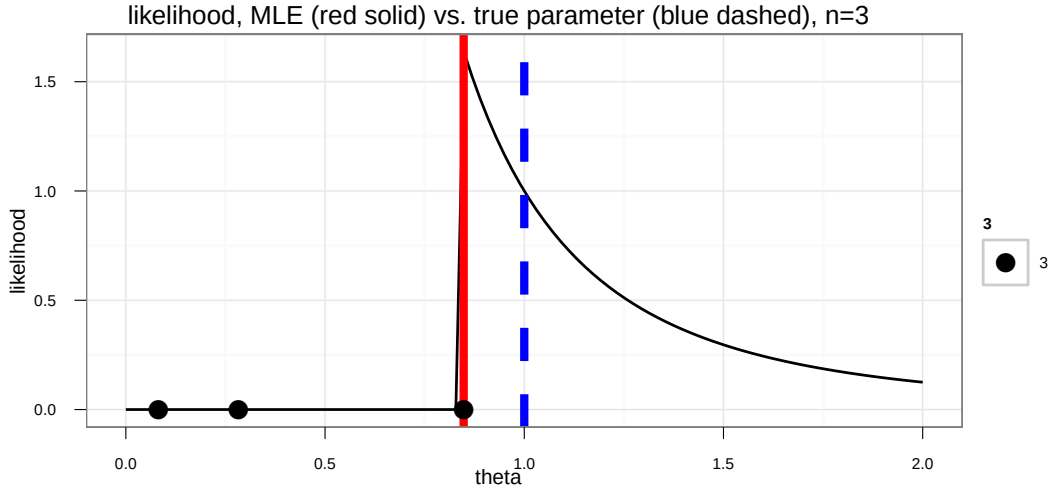
```
[1] 0.02564103 1.08153846
```

Example 3: $X^{(1)},\ldots,X^{(n)} \sim U[0,\theta]$. In this case $p_\theta(X^{(i)}) = \theta^{-1}$ if $X^{(i)} \in [0,\theta]$ and 0 otherwise. The likelihood is $L(\theta) = \theta^{-n}$ if $0 \leq X^{(1)},\ldots,X^{(n)} \leq \theta$ and 0 otherwise. We need to exercise care in this situation since the definition of the likelihood branches to two option depending on the value of the parameter θ. To treat only one case and not two we write the likelihood as $L(\theta) = \theta^{-n} 1_{\{0\leq X^{(1)},\ldots,X^{(n)}\leq\theta\}}$ where $1_{\{A\}}$ is the indicator function which equals 1 if A is true and 0 otherwise. We can't at this point proceed as before since the likelihood is not a differentiable function of θ (and neither will the log-likelihood be). We therefore do not take derivatives and simply examine the function $L(\theta)$: it will be zero for $\theta < \max(X^{(1)},\ldots,X^{(n)})$ and non-zero for $\theta \geq \max(X^{(1)},\ldots,X^{(n)})$ in which case the likelihood function will be monotonically decreasing in θ. It follows then the $\hat{\theta} = \max(X^{(1)},\ldots,X^{(n)})$. The graph below plots the likelihood function for $\theta = 1$ overlayed on the samples ($n = 3$) with the MLE and true parameter indicated by vertical lines.

```
> D=data.frame(theta=seq(0,2,length=100));
> n=3;samples=runif(n,0,1); samples; # 5 samples from U[0,theta], theta=1
```

```
[1] 0.72540527 0.48614910 0.06380247
```

```
> D$likelihood[D$theta<max(samples)]=0;
> D$likelihood[D$theta≥max(samples)]=D$theta[D$theta≥max(samples)]^(-n);
> p=ggplot(D,aes(theta,likelihood)) + geom_line()
> print(p+opts(title='likelihood, MLE (red solid) vs. true parameter (blue dashed), n=3')+
+       geom_vline(aes(xintercept=max(samples)),color=I('red'),size=1.5)+
+       geom_vline(aes(xintercept=1),color=I('blue'),lty=2,size=1.5)+
+       geom_point(aes(x=samples,y=0,size=3)))
```



A variation of this situation has $X^{(1)},\ldots,X^{(n)} \sim U(0,\theta)$ (as before but this time the interval is open and not closed). We start as before, but the likelihood at $\max(X^{(1)},\ldots,X^{(n)})$ is zero. Since the likelihood increases as we get θ closer to $\max(X^{(1)},\ldots,X^{(n)})$ (from the right), and at $\max(X^{(1)},\ldots,X^{(n)})$ it is zero. That is there is no MLE! For any specific value $\hat{\theta}$, we can always come up with $\hat{\theta}'$ that will result in a higher likelihood. Thus there is no value of θ that maximizes the likelihood.

Another variation has $X^{(1)},\ldots,X^{(n)} \sim U[\theta,\theta+1]$. In this case the likelihood is $L(\theta) = 1_{\{\theta\leq X^{(1)},\ldots,X^{(n)}\leq\theta+1\}}$. The likelihood is thus either zero or 1: it is 1 for many possible values, there are multiple maximizers or MLEs (all $\hat{\theta}$ for which $\hat{\theta} \leq X^{(1)},\ldots,X^{(n)} \leq \hat{\theta}+1$)) rather than a unique one.

Theoretical Properties

The MLE may be motivated on practical grounds as the most popular estimation technique in statistics. It also has some nice theoretical properties that motivate it. Below is a brief non-formal description of thes properties.

Consistency The MLE $\hat{\theta}_n$ is a function of $X^{(1)}, \ldots, X^{(n)}$ and is therefore a random variable. It is sometimes more accurate and sometime less accurate depending on the samples. It is generally true, however, that as n increases the value of the random variable $\hat{\theta}_n$ converges to the true parameter value with probability 1

$$P\left(\lim_{n\to\infty} \hat{\theta}_n = \theta\right) = 1.$$

The main condition for this is identifiability, defined as $\theta' \neq \theta'' \Rightarrow p_{\theta'} \not\equiv p_{\theta''}$ (that is for two distinct parameter values we get two distinct probability functions).

Asymptotic Variance The random variable $\hat{\theta}_n$ for large n is approximately normally distributed $N(\theta, I^{-1}(\theta)/n)$, with expectation equal to the true parameter and variance decaying linearly with n.

Asymptotic Efficiency Among all other unbiased estimators the MLE has the smallest asymptotic variance.