

Likelihoods

The distribution of a random variable Y with a discrete sample space (e.g. a finite sample space or the integers) can be characterized by its probability mass function (pmf):

$$P(Y = y) = f(y).$$

For example, suppose Y has a geometric distribution on $1, 2, \dots$ with parameter p . Then the pmf is

$$P(Y = y) = (1 - p)^{y-1}p.$$

You can check that $\sum_{y=1}^{\infty} P(Y = y) = 1$ using properties of geometric series.

The distribution of a random variable Y with a continuous sample space (e.g. $(0, 1)$ or $(-\infty, \infty)$) can be characterized by its probability density function (pdf) $f(y)$, which tells us the probability that Y lies in an interval (a, b) :

$$P(a \leq Y \leq b) = \int_a^b f(y) dy.$$

For example, if Y is standard normal, then the density is

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

where μ and σ^2 are parameters.

Suppose we observe independent data Y_1, \dots, Y_n , where the probability density or mass function of Y_j is f_j .

The joint density of Y_1, \dots, Y_n is

$$\prod_{i=1}^n f_j(Y_j).$$

Suppose the f_j are identical functions, and include a parameters θ , so we can write $f_j(Y) = f(Y; \theta)$. Taking the logarithm of the joint density yields the log-likelihood function

$$\sum_i \log f(Y_j; \theta).$$

The log-likelihood function

$$\sum_i \log f(Y_j; \theta).$$

is a function of θ , since we observe the Y_j and can substitute their numerical values into the expression above.

For example, suppose the Y_j follow exponential distributions

$$f_j(y, \lambda) = \lambda^{-1} \exp(-y/\lambda).$$

If we want to estimate λ from the data, one principle is to maximize the log-likelihood function. This estimate $\hat{\lambda}$ is called the maximum likelihood estimate (MLE).

As an example, consider the exponential distribution. The log-likelihood is

$$\begin{aligned} L(\lambda) &= -n \log \lambda - \sum_j Y_j / \lambda \\ &= -n \log \lambda - n \bar{Y} / \lambda \end{aligned}$$

To maximize this as function of λ , we calculate its derivative

$$L'(\lambda) = -n/\lambda + n\bar{Y}/\lambda^2$$

and solve $L'(\lambda) = 0$ yielding the MLE

$$\hat{\lambda} = \sum Y_j / n = \bar{Y}.$$

To be rigorous, we should check that this is a local maximum ($L''(\hat{\lambda}) < 0$) rather than a local minimum ($L''(\hat{\lambda}) > 0$).

$$L''(\lambda) = n\lambda^{-2} - 2n\bar{Y}/\lambda^3.$$

$$L''(\hat{\lambda}) = n/\bar{Y}^2 - 2n/\bar{Y}^3 = -n/\bar{Y}^2 \leq 0.$$

The inequality is strict as long as not all Y_j are zero.

The curvature of $L(\theta)$ around $\hat{\theta}$ tells us how precisely we are able to estimate θ using $\hat{\theta}$.

This program simulates exponential data and plots the likelihood function $L(\lambda)$. The λ region such that $L(\lambda)$ is within one unit of the peak value are shown in red.

```
## Sample sizes.
N = c(5,10,20,40,80)

## A grid of lambda values.
G = seq(0.1, 5, by=0.05)

for (k in 1:length(N)) {

  ## The sample size.
  n = N[k]

  ## Replications.
  for (j in 1:10) {
    X = rexp(n)
```

```
L = -n*log(G) - sum(X)/G
```

```
plot(G, L, t='l', xlab='Lambda', ylab='log-likelihood', main='')  
title(main=sprintf('Sample size %d', n))
```

```
ii = which(L >= (max(L)-1))  
lines(G[ii], L[ii], col='red')
```

```
scan()
```

```
}
```

```
}
```

The curvature in $L(\theta)$ is measured by the Fisher information $L''(\theta)$.

As a general principal, the sampling variance of the MLE $\hat{\theta}$ is approximately the negative inverse of the Fisher information:

$$-1/L''(\hat{\theta})$$

For the exponential example, we would get

$$\text{var}\hat{\lambda} \approx \bar{Y}^2/n.$$

Since the mean of the exponential distribution is λ and its variance is λ^2 , we expect $\bar{Y}^2 \approx \hat{\sigma}_Y^2$, so this will give similar results to the usual $\text{var}\bar{Y} = \sigma^2/n$ formula.

As another example, suppose we observe data following a logistic distribution with an unknown mean μ . The density is

$$\frac{\exp(-(y - \mu))}{(1 + \exp(-(y - \mu)))^2}.$$

The log-likelihood from an independent sample Y_1, \dots, Y_n is

$$L(\mu) = -Y_{\cdot} + n\mu - 2 \sum_i \log(1 + \exp(-(Y_i - \mu))),$$

where $Y_{\cdot} = \sum_i Y_i$.

The derivative of the log-likelihood is

$$L'(\mu) = n - 2 \sum_i \frac{\exp(-(Y_i - \mu))}{1 + \exp(-(Y_i - \mu))}$$

In the logistic example, it will not be possible to solve $L'(\mu) = 0$ symbolically. We can still use the computer to find the MLE numerically.

One approach is Newton's method. Begin with a reasonable estimate of μ , say $\mu_0 = \bar{Y}$. Then construct a linear approximation to $L'(\mu)$ at μ_0 :

$$L'(\mu) \approx L'(\mu_0) + (\mu - \mu_0)L''(\mu_0).$$

This linear expression can be equated to zero and solved for μ , yielding

$$\mu_1 \approx \mu_0 - L'(\mu_0)/L''(\mu_0).$$

Then μ_1 can be used to produce μ_2 , and so on. This process converges rapidly to a point that is usually a local maximum of L .

In the logistic example

$$L''(\mu) = -2 \sum_i \frac{\exp(-(Y_i - \mu))}{(1 + \exp(-(Y_i - \mu)))^2}$$

The following program simulates samples from a logistic distribution, and uses Newton's method to calculate the MLE. The approximate standard errors $\sqrt{-1/L''(\hat{\mu})}$ are also stored, and the proportion of the time that the true value is within two standard errors of the estimate (which should be 0.95) is recorded.

```
## The population mean.  
mu = 1  
  
## The sample size.  
n = 10  
  
## The number of simulation replications.  
nrep = 1e3  
  
Q = array(0, c(nrep,3))  
  
for (k in 1:1e3) {  
  
  ## Generate the logistic data.  
  U = runif(n)  
  Y = mu + log(U/(1-U))  
  
  ## Starting value.  
  m = mean(Y)
```

```
## Newton's method.
while (TRUE) {

  V = exp(-(Y-m))

  ## The first derivative of the log-likelihood.
  D1 = n - 2*sum(V/(1+V))

  ## Check for convergence.
  if (abs(D1) < 1e-10) { break }

  ## The second derivative of the log-likelihood.
  D2 = -2*sum(V/(1+V)^2)

  ## Newton's step.
  m = m - D1/D2
}

## The approximate standard error.
SE = sqrt(-1/D2)
```

```
Q[k,] = c(m, SE, 1*(abs(m-mu) < 2*SE))  
}
```