Likelihood and Information

David C Muller

Imperial College London

Outline

Likelihood

Maximum Likelihood

Information

Summary

CAVEAT: This is necessarily a rather superficial overview, intended to give you a conceptual familiarity with these important principles of statistical inference.

Likelihood

Likelihood

Much of Epidemiology and Public Health is concerned with estimating values for population parameters:

- The proportion of a population infected with HPV
- The rate of incidence of diabetes
- The rate ratio of lung cancer among smokers compared with never smokers

To do so, we posit statistical models, and estimate the parameters of these models.
Likelihood is a key concept that helps us to derive estimators of these parameters.

Likelihood

Definition (phrased as a question)
What is the probability of observing the data I have, given a specific value of a parameter?

So we can consider the likelihood associated with a parameter value $\theta = \theta_0$, given an observed sample value $Y = y$.

If the observed data provide more support for one value of the parameter than for another value, then the likelihood is higher for the former parameter value.

Binomial example

We wish to estimate the population prevalence of HPV, and take a sample of 1000 people. We find 121 people have evidence of HPV infection. A natural estimate of the population prevalence is the sample prevalence: $\hat{\theta} = 121/1000 = 0.121$ or 12.1%.

What is the likelihood of this estimate compared with other values? Assuming a binomial distribution for the outcome:

$$
\begin{aligned}
P(Y = y|\theta = \hat{\theta}) &= \frac{N!}{(N-y)!y!} p^y (1-p)^{N-y} \\
P(Y = 121|\theta = 0.121) &= f_Y(121; 0.121) \\
&= \frac{1000!}{879!121!} 0.121^{121} (1-0.121)^{879} \\
&= 0.0387
\end{aligned}
$$

Binomial example

0.0387 may seem small, but given 1001 possible outcomes, perhaps not. In fact, likelihood is mostly useful as a relative concept: e.g., the likelihood of 12.1% greater than that of 15%? Setting $\hat{\theta} = 0.15$:

$$\begin{aligned} P(Y = 121 | \theta = 0.15) &= f_Y(121; 0.15) \\ &= \frac{1000!}{879!121!} 0.15^{121} (1 - 0.15)^{879} \\ &= 0.002.\text{‘} \end{aligned}$$

So the value of 0.15 is less supported by the data than 0.121. We can calculate the relative magnitude of the likelihoods by their ratio: $0.0387/0.0012 = 32.25$.

Likelihood function

We have so far considered the likelihood of specific parameter values. We can consider the likelihood across all possible values of the parameter: the likelihood function.

$$L(\theta), \text{or } L(\theta; y)$$

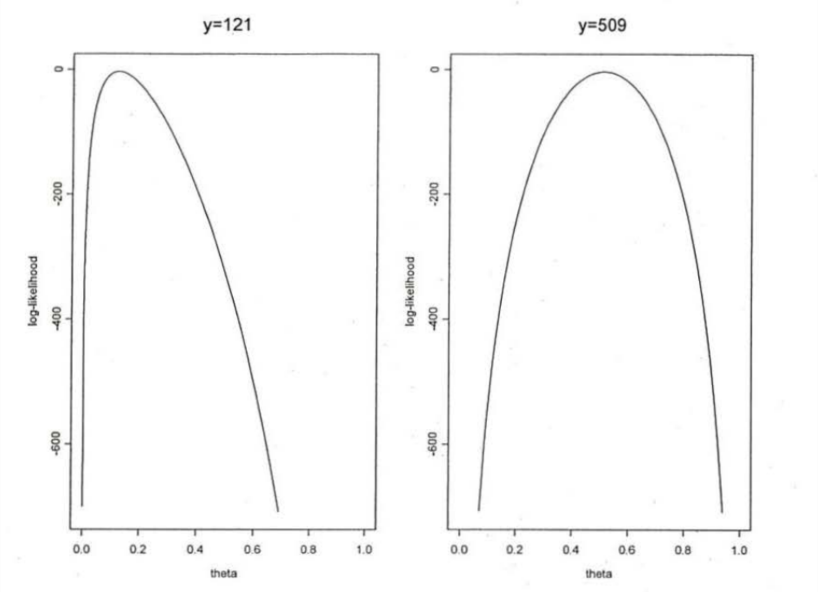
Often it is easier to work with the log-likelihood function

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

As in the case of independent observations, the observation level likelihood is of the form

$$L(\theta) = \prod_{i=1}^{n} f_i(y_i;\theta) \text{ , so; } \ell(\theta) = \sum_{i=1}^{n} \log f_i(y_i;\theta)$$

Likelihood function



Maximum Likelihood

Maximum likelihood

- The likelihood associated with a particular parameter value θ_0 is the probability (density) of obtaining the sample data, assuming that the true value of the parameter is θ_0
- This measures the support that the data has for that particular parameter value
- The most supported parameter value $\hat{\theta}$ will be the value for which $L(\hat{\theta}) > L(\theta_0)$ for all $\theta_0 \neq \hat{\theta}$

The parameter value that maximises the likelihood function is thus that which is most supported by the data. It is referred as the Maximum Likelihood Estimate (MLE).

Maximum likelihood

- General approach
- In principle can be applied to any situation where we can write down a likelihood function

Some details to be aware of:

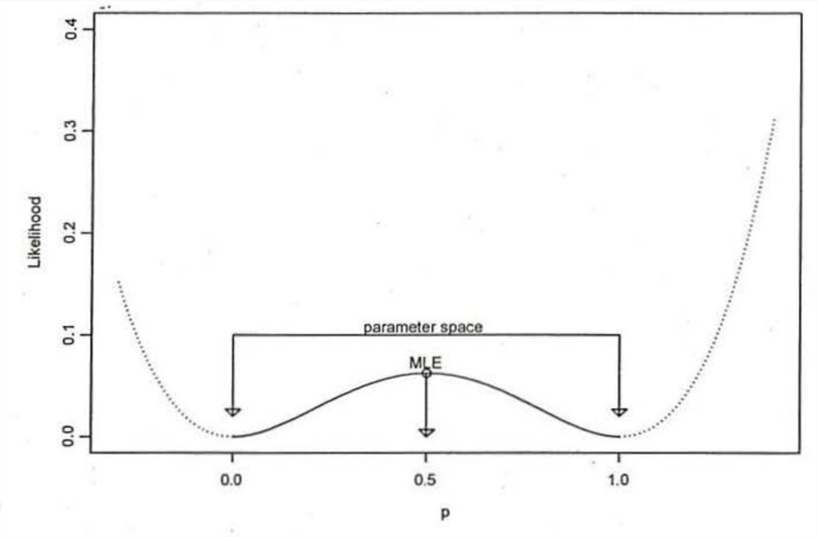
- Assumes existence of a unique maximum
- Choice of parameter space – only consider valid values of a parameter (e.g., the interval [0, 1] for a proportion)

Maximum likelihood

Population prevalence example:

$N = 4$
$y = 2$
$L(p) \propto p^y (1 - p)^{N-y}$
$L(p) \propto p^2 (1 - p)^2$

Maximum likelihood



Derivation of MLEs

How do we derive and calculate MLEs?

Derivation of MLEs

How do we derive and calculate MLEs?

Calculus!

When the MLE occurs at a stationary point, we can find it by solving an equation where we set the derivative of the log-likelihood function to 0. This is called a score equation,

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

Derivation of MLEs: binomial example

Likelihood function: $L(p) \propto p^y(1-p)^{N-y}$

$$\ell(p) = \log(L(p)) = y\log(p) + (N-y)\log(1-p)$$
$$\frac{d}{dp}\ell(p) = \frac{y}{p} - \frac{N-y}{1-p}$$

Setting to 0 and solving for p, we have

$$\hat{p} = \frac{y}{N}$$

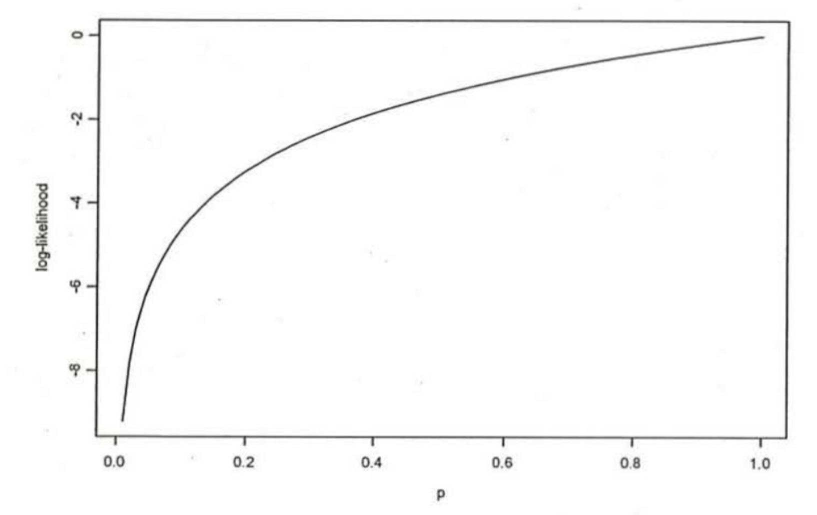
Derivation of MLEs: binomial example

What if by chance in our sample, $y = N$?

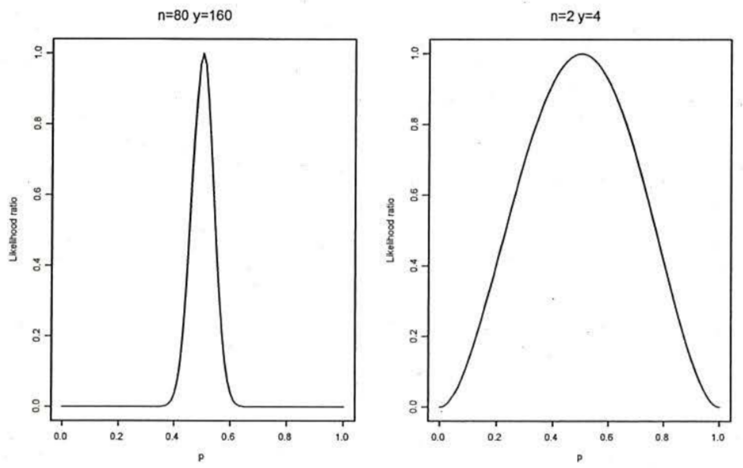
Derivation of MLEs: binomial example

What if by chance in our sample, $y = N$?

The log likelihood reduces to $y \log(p)$



Derivation of MLEs: binomial example



Maximum likelihood summary

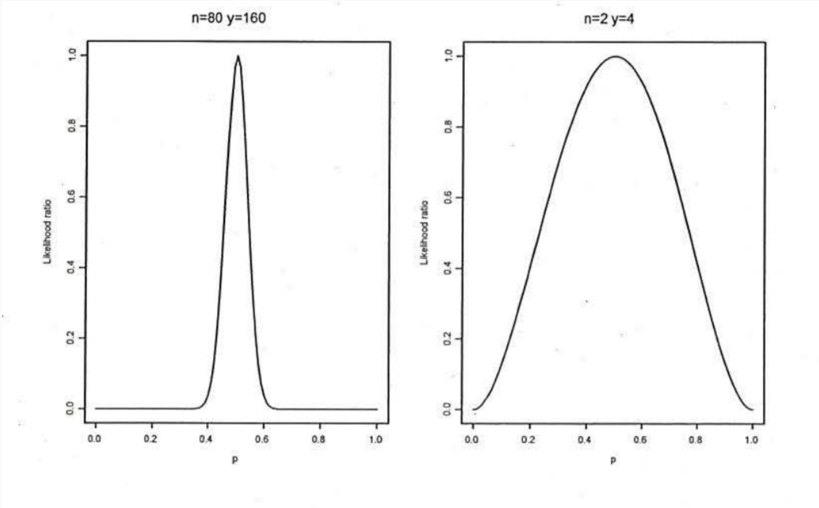
- Powerful general method for deriving estimators
- We have seen simple and "explicit" examples that can be solved by analytically maximising the likelihood function, can be used in much more complex situations using numerical solvers (Newton-Raphson algorithm and related root finders).
- Care needs to be taken with the parameter space, and that the maximum of the likelihood is at a stationary point.

Information

Beyond point estimation

- We have our nice maximum likelihood estimates of parameters of interest
- We would like an estimate of the uncertainty in our estimates, such as confidence intervals
- we will see that the curvature of the (log) likelihood function can be used to calculate standard errors for MLEs.

Binomial example redux



Second derivative of the likelihood function

Intuitively, the sharper the peak of the likelihood function, the more confident we will be in our MLE. The second derivative of the (log) likelihood function formalises this. Recall:

- First derivative is the gradient of the function
- Second derivative tells us how rapidly the gradient is changing

If the gradient is becoming smaller at a slow rate, the second derivative will be only slightly less than 0, indicating a relatively flat likelihood function around the maximum.

If the gradient is changing rapidly, then the second derivative will be substantially less than 0, indicating a more sharply pointed likelihood.

Information

Define the information or observed information to be the negative second derivative of the log-likelihood.

$$I_o(\theta) = -\frac{d^2}{d\theta^2}\ell(\theta)$$

It is called the observed information because it depends on the observed sample $I_o(\hat{\theta}) = I_o(\hat{\theta}; y)$.

Information

We can take the expectation with respect to y to obtain the expected information, also called Fisher information

$$I(\theta) = \mathbb{E}\{I_o(\theta; Y)\}.$$

This is related to the asymptotic variance of the MLE:

$$\begin{aligned} \text{Var}(\hat{\Theta}) &\approx I(\hat{\theta})^{-1} \quad \text{when N is large, and} \\ \text{SE}(\hat{\Theta}) &= \sqrt{I(\hat{\theta})^{-1}}. \end{aligned}$$

Important properties of the MLE

- Not unbiased (though in specific circumstances it may be)
- Consistent (asymptotically, as N approaches infinity, the MLE converges in probability to the true value).
- Its sampling distribution is asymptotically (multivariate) normal, so in large samples

$$\hat{\Theta} \overset{\mathrm{d}}{\approx} \mathrm{N}(\theta, I(\theta)^{-1})$$

So we can calculate confidence intervals using percentiles of the standard normal distribution.

Example: deriving the variance of the binomial proportion

We can obtain the observed information from the second derivative of the binomial likelihood:

$$
\begin{aligned}
I_o(p; y) = -\frac{d^2}{dp^2}\ell(\theta) &= -\left\{-\frac{y}{p^2} - \frac{N-y}{(1-p)^2}\right\} \\
&= \frac{y}{p^2} + \frac{N-y}{(1-p)^2}
\end{aligned}
$$

The Fisher information is the obtained by taking the expectation with respect to *y*. Using the fact that the expectation of a binomial variate is *Np*, we have

$$
\begin{aligned}
I(p) = \mathbb{E}\{I_o(p; y)\} &= \frac{Np}{p^2} + \frac{N-Np}{(1-p)^2} = \frac{N}{p(1-p)}, \\
\mathrm{Var}(\hat{\Theta}) \approx I(\hat{\theta})^{-1} &= \frac{p(1-p)}{N}.
\end{aligned}
$$

Summary

Multiple parameters

We have focussed on examples with a single parameter for illustrative purposes. Maximum likelihood estimation can be applied to multiple parameter situations with obvious extensions:

- One score equation for each parameter, which are partial derivatives of the likelihood with respect to that parameter;
- The information becomes matrix valued, and the inverse of this matrix approximates the variance-covariance matrix of the MLE.

Summary

- Statistical likelihood quantifies the level of support that the data in hand provide to a given parameter value
- Likelihood functions can be used to compare support for different parameter values
- Maximum likelihood estimation is a general and powerful approach to deriving estimates for parameters in statistical models
- Information is crucial, as it relates to how informative a sample is in terms of estimating parameters of interest, and is related to the variance of MLEs.