Maximum Likelihood Estimation

The likelihood and log-likelihood functions are the basis for deriving estimators for parameters, given data. While the shapes of these two functions are different, they have their maximum point at the same value. In fact, the value of p that corresponds to this maximum point is defined as the Maximum Likelihood Estimate (MLE) and that value is denoted as $\hat{p}$. This is the value that is "mostly likely" relative to the other values. This is a simple, compelling concept and it has a host of good statistical properties.

Thus, in general, we seek $\hat{\theta}$, such that this value maximizes the log-likelihood function. In the binomial model, the $\log_e(\mathcal{L})$ is a function of only one variable, so it is easy to plot and visualize.

The maximum likelihood estimate of p (the unknown parameter in the model) is that value that maximizes the log-likelihood, given the data. We denote this as $\hat{p}$. In the binomial model, there is an analytical form (termed "closed form") of the MLE, thus maximization of the log-likelihood is not required. In this simple case,

$$\hat{p} = y/n = 7/11 = 0.6363636363...$$

of course, if the observed data were different, $\hat{p}$ would differ.

The log-likelihood links the data, unknown model parameters and assumptions and allows rigorous, statistical inferences.

Real world problems have more that one variable or parameter (e.g., p, in the example). Computers can find the maximum of the multi-dimensional log-likelihood function, the biologist need not to be terribly concerned with these details.

The actual numerical value of the log-likelihood at its maximum point is of substantial importance. In the binomial coin flipping example with $n = 11$ and $y = 7$, $\max(\log \mathcal{L}) = -1.411$ (see graph).

The log-likelihood function is of fundamental importance in the theory of inference and in all of statistics. It is the basis for the methods explored in FW-663. Students should make every effort to get comfortable with this function in the simple cases. Then, extending the concepts to more complex cases will come easy.

Likelihood Theory -- What Good Is It?

1. The basis for deriving estimators or estimates of model parameters (e.g., survival probabilities). These are termed "maximum likelihood estimates," MLEs.

2. Estimates of the precision (or repeatability). This is usually the conditional (on the model) sampling variance
-covariance matrix (to be discussed).

3. Profile likelihood intervals (asymmetric confidence intervals).

4. A basis for testing hypotheses:
Tests between nested models (so-called likelihood ratio tests)
Goodness of fit tests for a given model

5. Model selection criterion, based on Kullback-Leibler information.

Numbers 1-3 (above) require a model to be "given." Number 4, statistical hypothesis testing, has become less useful in many respects in the past two decades and we do not stress this approach as much as others might. Likelihood theory is also important in Bayesian statistics.

Properties of Maximum Likelihood Estimators

For "large" samples ("asymptotically"), MLEs are optimal.

1. MLEs are asymptotically normally distributed.

2. MLEs are asymptotically "minimum variance."

3. MLEs are asymptotically unbiased (MLEs are often biased, but the bias $\rightarrow 0$ as $n \rightarrow \infty$).

One to one transformations are also MLEs. For example, mean life span $\bar{L}$ is defined as

$-1/\log_e(S)$. Thus, an estimator of $\bar{L} = -1/\log_e(\hat{S})$ and then $\hat{\bar{L}}$ is also an MLE.

Maximum likelihood estimation represents the backbone of statistical estimation. It is based on deep theory, originally developed by R. A. Fisher (his first paper on this theory was published in 1912 when he was 22 years old!). While beginning classes often focus on least squares estimation ("regression"); likelihood theory is the omnibus approach across the sciences, engineering and medicine.

The **Likelihood Principle** states that all the relevant information in the sample is contained in the likelihood function. The likelihood function is also the basis for Bayesian statistics. See Royall (1997) and Azzalini (1996) for more information on likelihood theory.

Maximum Likelihood Estimates

Generally, the calculus is used to find the maximum point of the log-likelihood function and obtain MLEs is closed form. This is tedious for biologists and often not useful in real problems (where a closed form estimator may often not even exist).

The log-likelihood functions we will see have a single mode or maximum point and no local optima. These conditions make the use of numerical methods appealing and efficient.

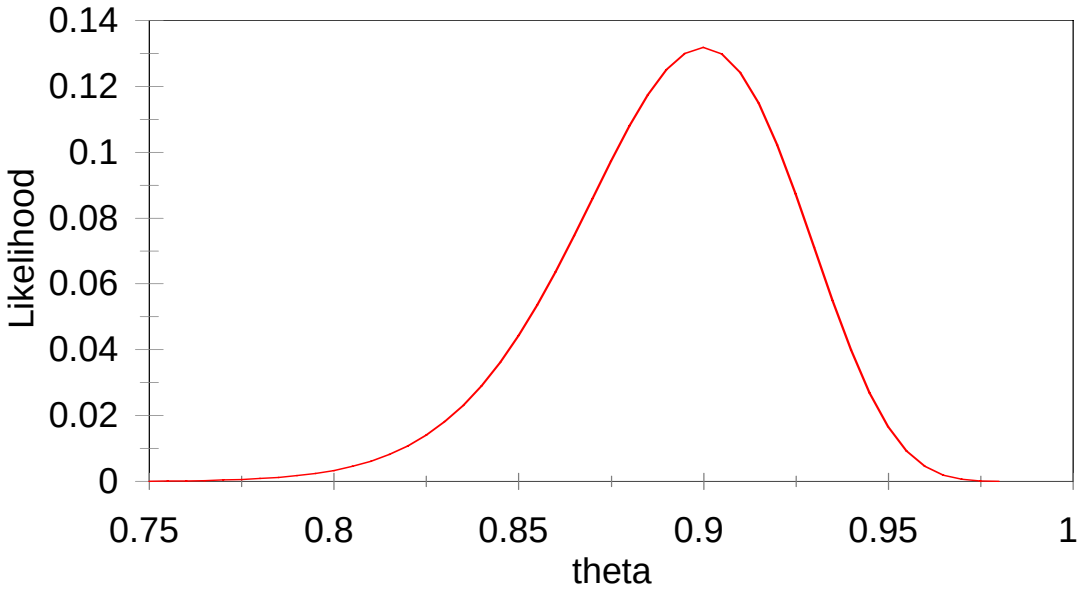
Consider, first, the binomial model with a single unknown parameter, p. Using calculus one could take the first partial derivative of the log-likelihood function with respect to the p, set it to zero and solve for p. This solution will give $\hat{p}$, the MLE. This value of $\hat{p}$ is the one the maximizes the log-likelihood function. It is the value of the parameter that is *most likely*, given the data.

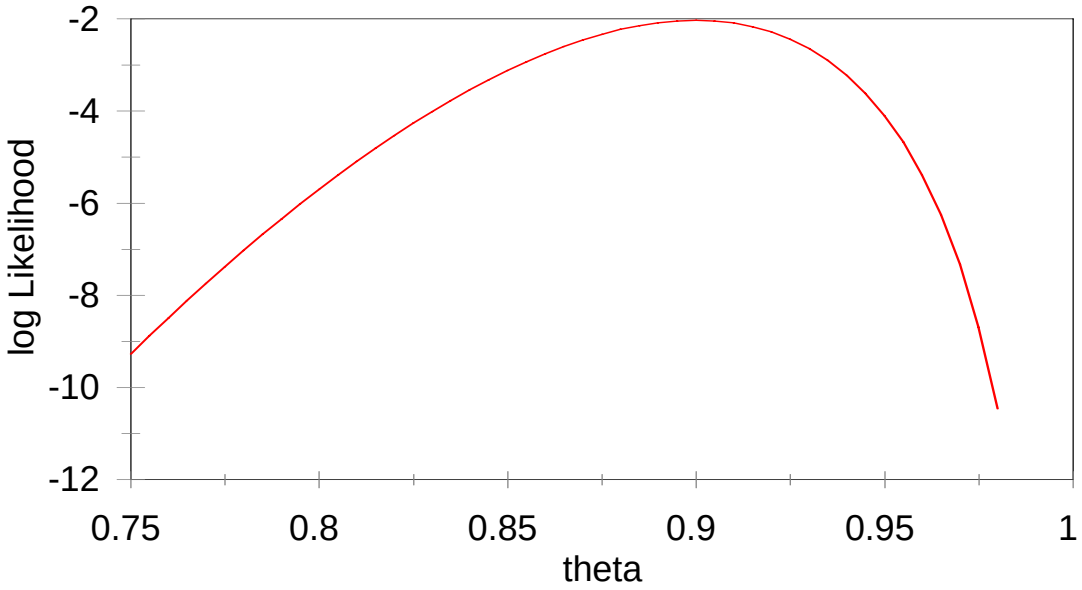
> The likelihood function provides information on the *relative likelihood* of various parameter values, given the data and the model (here, a binomial). Think of 10 of your friends, 9 of which have one raffle ticket, while the 10^{th} has 4 tickets. The person with 4 tickets has a higher likelihood of winning, relative to the other 9. If you were to try to select the mostly likely winner of the raffle, which person would you pick? Most would select the person with 4 tickets (would you?). Would you feel strongly that this person would win? Why? or Why not?
>
> Now, what if 8 people had a single ticket, one had 4 tickets, but the last had 80 tickets. Surely, the person with 80 tickets is most likely to win (but not with certainty). In this simple example you have a feeling about the "strength of evidence" about the likely winner. In the first case, one person has an edge, but not much more. In the second case, the person with 80 tickets is relatively very likely to win.

The **shape** of the log-likelihood function is important in a conceptual way to the raffle ticket example. If the log-likelihood function is relatively flat, one can make the interpretation that several (perhaps many) values of p are nearly equally likely. They are relatively alike; this is quantified as the sampling variance or standard error.

If the log-likelihood function is fairly flat, this implies considerable uncertainty and this is reflected in large sampling variances and standard errors, and wide confidence intervals. On the other hand, if the log-likelihood function is fairly peaked near its maximum point, this indicates some values of p are relatively very likely compared to others (like the person with 80 raffle tickets). There is some considerable degree of certainty implied and this is reflected in small sampling variances and standard errors, and narrow confidence intervals. So, the log-likelihood function at its maximum point is important as well as the shape of the function near this maximum point.

The shape of the likelihood function near the maximum point can be measured by the analytical second partial derivatives and these can be closely approximated numerically by a computer. Such numerical derivatives are important in complicated problems where the log-likelihood exists in 20-60 dimensions (i.e., has 20-60 unknown parameters).





The standard, analytical method of finding the MLEs is to take the first partial derivatives of the log-likelihood with respect to each parameter in the model. For example:

$$\frac{\partial \ell n(\mathcal{L}(p))}{\partial p} = \frac{11}{p} - \frac{5}{1-p} \qquad (n = 16)$$

Set to zero:

$$\frac{\partial \ell n(\mathcal{L}(p))}{\partial p} = 0$$

$$\frac{11}{p} - \frac{5}{1-p} = 0$$

and solve to get $\hat{p} = 11/16$, the MLE.

For most models we have more than one parameter. In general, let there be K parameters, $\theta_1, \theta_2, \ldots, \theta_K$. Based on a specific model we can construct the log-likelihood,

$$\log_e(\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_K \mid \text{data})) \equiv \log_e(\mathcal{L})$$

and K log-likelihood equations,

$$\frac{\partial \log(\mathcal{L})}{\partial \theta_1} = 0$$

$$\frac{\partial \log(\mathcal{L})}{\partial \theta_2} = 0$$

$$\vdots$$

$$\frac{\partial \log(\mathcal{L})}{\partial \theta_K} = 0$$

The solution of these equations gives the MLEs, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$.

The MLEs are almost always unique; in particular this is true of multinomial-based models.

In principle $\log(\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_K \mid \text{data})) \equiv \log(\mathcal{L})$ defines a “surface" in K-dimensional space, ideas of curvature still apply (as mathematical constructs). Plotting is hard for more than 2 parameters.

Sampling variances and covariances of the MLEs are computed from the log-likelihood,

$$\log(\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_K \mid \text{data})) \equiv \log(\mathcal{L})$$

based on curvature at the maximum. Actual formulae involve second mixed-partial derivatives of the log-likelihood, hence quantities like

$$\frac{\partial^2 \log(\mathcal{L})}{\partial \theta_1 \partial \theta_1} \quad \text{and} \quad \frac{\partial^2 \log(\mathcal{L})}{\partial \theta_1 \partial \theta_2}$$

evaluated at the MLEs.

Let $\hat{\Sigma}$ be the estimated variance-covariance matrix for the K MLEs; $\hat{\Sigma}$ is a K by K matrix. The inverse of $\hat{\Sigma}$ is the matrix of elements as below.

$$-\frac{\partial^2 \log(\mathcal{L})}{\partial\theta_i\,\partial\theta_i}$$

as the *i*th diagonal element, and

$$-\frac{\partial^2 \ell n(\mathcal{L})}{\partial\theta_i\,\partial\theta_j}$$

as the *i*, *j*th off-diagonal element.

(these mixed second partial derivatives are evaluated at the MLEs).

The use of log-likelihood functions (rather than likelihood functions) is deeply rooted in the nature of likelihood theory. Note also that LRT theory leads to tests which basically always involve taking $-2\times$ (log-likelihood at MLEs).

Therefore we give this quantity a symbol and a name: deviance,

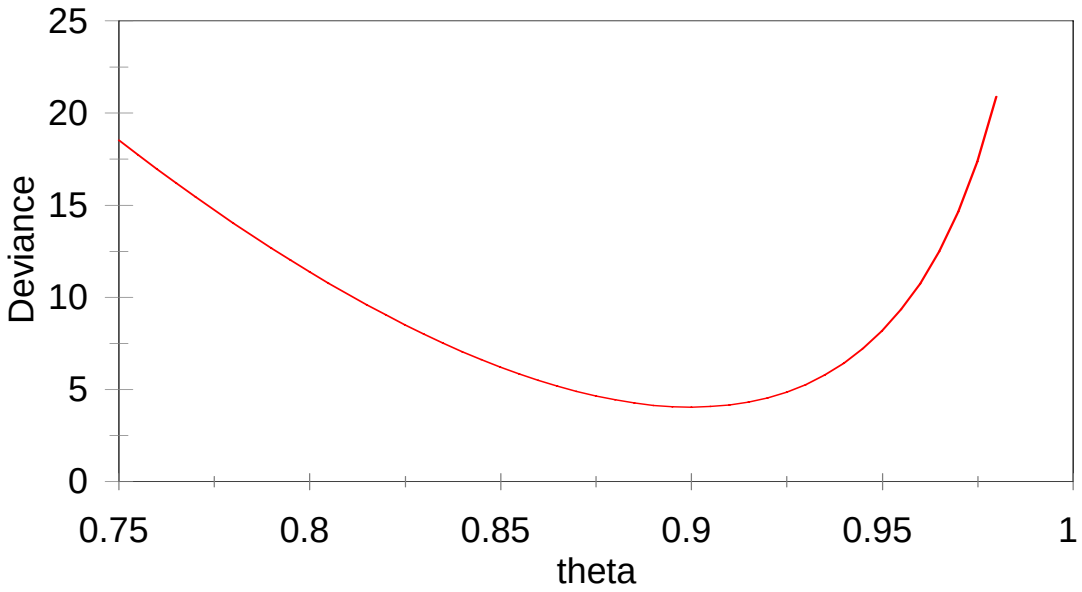
$$\textbf{deviance} = -2\log_e(\mathcal{L}(\hat{\theta})) + 2\log_e(\mathcal{L}_s(\hat{\theta})),$$

or

$$= -2\left(\log_e(\mathcal{L}(\hat{\theta})) - \log_e(\mathcal{L}_s(\hat{\theta}))\right),$$

evaluated at the MLEs for some model. Here, the first term is the log-likelihood, evaluated at its maximum point, for the model in question and the second term is the log-likelihood, evaluated at its maximum point, for the saturated model. The meaning of a saturated model will become clear in the following material; basically, in the multinomial models, it is a model with as many parameters as cells. This final term in the deviance can often be dropped, as it is often a constant across models.

The deviance for the saturated model $\equiv 0$. Deviance, like information, is additive. The deviance is approximately χ^2 with df = number of cells $-K$ and is thus useful is examining goodness-of-fit of a model. There are some ways where use of the deviance in this way will not provide correct results. *MARK* outputs the deviance as a measure of model fit and this is often very useful.


25
20
15
10
5
0
Deviance
0.75
0.8
0.85
0.9
0.95
1
theta