9.07 Introduction to Statistics for Brain and Cognitive Sciences
Emery N. Brown

Lecture 9: Likelihood Methods

I. Objectives

1. **Review the concept of the joint distribution of random variables.**
2. **Understand the concept of the likelihood function.**
3. **Understand the concept of maximum likelihood estimation.**
4. **Understand the key properties of maximum likelihood estimation.**
5. **Understand the concept of Fisher information.**
6. **Understand how to compute the uncertainty in maximum likelihood estimates.**

II. Joint Probability Density or Distribution of a Set of Random Variables

We recall that two events E_1 and E_2 are independent if

$$\Pr(E_1 \bigcap E_2) = \Pr(E_1)\Pr(E_2) \tag{9.1}$$

Intuitively, this statement says that knowledge about E_1 gives no information about E_2. In general a set of events $E_1, \ldots, E_n$ is independent if

$$\Pr(E_1 \bigcap E_2 \ldots \bigcap E_n) = \prod_{i=1}^{n} \Pr(E_i) \tag{9.2}$$

It follows from **Lecture 1** and **Lecture 5** that if $x = x_1, \ldots, x_n$ is a sample of independent, identically distributed observations from a pmf or a pdf $f(x_i \mid \theta)$ then the joint probability density of $x = x_1, \ldots, x_n$ is

$$f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{9.3}$$

Equation 9.3 follows directly from the definition of a *pmf* for a discrete random variable. Using a *cdf* of a continuous random variable which is differentiable, it is also easy to show. Equation 9.3 is essential for our definition of the likelihood function of a sample of independent observations.

III. The Likelihood Function

A. Definition

Definition 9.1 (Likelihood Function). Given $x = x_1, \ldots, x_n$ is an independent, identically distributed sample from a *pdf* or *pmf* $f(x_i \mid \theta)$. Let $f(x \mid \theta)$ denote the joint *pdf* or *pmf* of the sample $x = (x_1, \ldots, x_n)$ defined in Eq. 9.3. Given $X = x$ is observed, the function of θ defined as

$$L(\theta) = f(x \mid \theta) = \prod_{k=1}^{n} f(x_k \mid \theta) \tag{9.4}$$

is called the **likelihood function** for the parameter θ. The likelihood function or likelihood provides an objective means of assessing the "information" in a sample of data about the model parameter θ. We view the data as fixed and now study $L(\theta) = f(x \mid \theta)$ as a function of θ. For a given model it summarizes all the information in the sample about θ. Use of the likelihood will allow us to overcome some of the ambiguities we faced with the method-of-moments.

B. Examples of Likelihoods

Example 2.1 (continued). Binomial Likelihood Function. In our learning experiment, we observed 22 correct responses from the animal in 40 trials. We derived the pmf of the 22 observations in 40 trials as the binomial distribution

$$f(22 \mid p) = \binom{40}{22} p^{22}(1-p)^{18}. \tag{9.5}$$

(Notice that this distribution does not have exactly the same form as Eq. 9.3, because we allowed for all the possible ways the sample could be realized. It turns out that in our likelihood analysis, this distinction will not matter because the binomial coefficient does not alter the likelihood). Since the unknown parameter is p we have the likelihood of p is

$$L(p) = f(22 \mid p) = \binom{40}{22} p^{22}(1-p)^{18}. \tag{9.6}$$

In general, for the binomial *pmf* with n trials and probability of a correct response p, the likelihood

$$L(p) = f(x \mid p) = \binom{n}{x} p^{x}(1-p)^{n-x}. \tag{9.7}$$

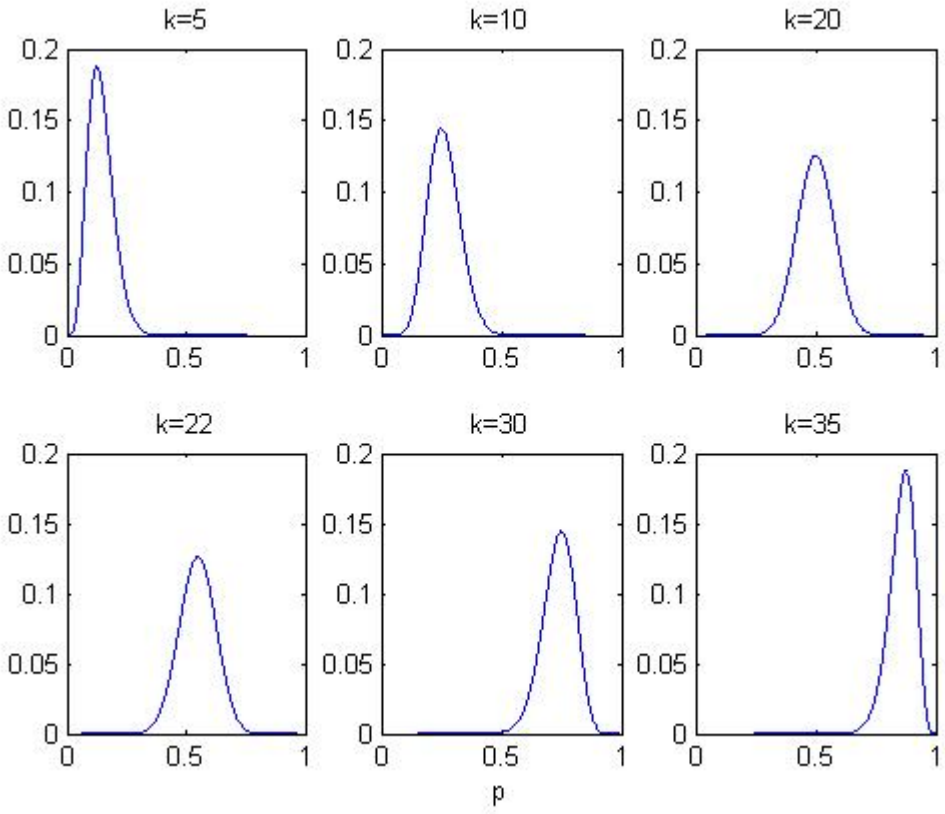

Figure 9A. Binomial Likelihood Function L(p) for n = 40 and k = 5, 10, 22, 30, 35.

Plotting these elementary likelihood functions is useful to gain insight into what type of information they provide. We can predict the shape of the binomial likelihood function based on discussions of the beta distribution in **Lecture 3** (Eq. 3.30, Figure 3J).

Example 3.2 Quantile Release Hypothesis (Poisson Likelihood). We used a Poisson pmf to model the quantile release of acetylcholine at the synapse. If we record the number of quanta released in n non-overlapping time intervals of a specified length, then the likelihood function for the Poisson parameter λ, which is the expected number of quanta released, is

$$L(\lambda) = f(x \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \tag{9.8}$$

We see in this example, a very important feature of the likelihood. Namely, that it has reduced or summarized the data from all n observations to simply $S_n = \sum_{i=1}^{n} x_i$, the sum of the n observations. For a large class of probability models, this is a typical feature.

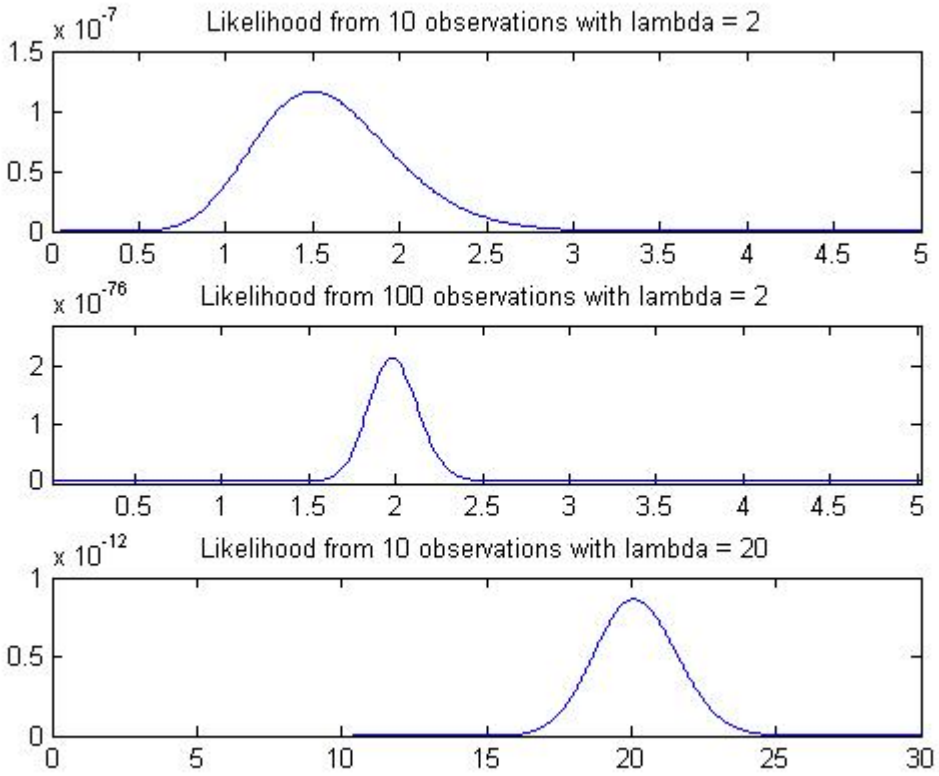

Figure 9B Poisson Likelihood Functions $L(\lambda)$.

The shape of the Poisson likelihood could be predicted based on results in **Lecture 3**. Note that the essential features of this likelihood are described by

$$L(\lambda) \propto \lambda^{S_n} e^{-n\lambda} \tag{9.9}$$

which is proportional to a gamma probability density (Eq. 3.26) with $\alpha = (S_n - 1)$ and $\beta = n$ (Figure 3l).

Example 3.2 (continued) Magnetoencephalogram Noise Data (Gaussian Likelihood Function). If we view the MEG data as a random sample from a Gaussian probability model then in this problem there are two unknown parameters: μ and σ^2.

$$
\begin{aligned}
L(\mu,\sigma^2) &= f(x \mid \mu,\sigma^2) \\
&= \prod_{i=1}^{n} f(x_i \mid \mu,\sigma^2) \\
&= \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}
\end{aligned}
\tag{9.10}
$$

If we expand the sum in the exponent, we see that the data have been compressed into two distinct terms: $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$. That is, we can rewrite Eq. 9.10 as

$$
L(\mu,\sigma^2) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n}{2}\left[\log(2\pi\sigma^2) + \frac{\mu^2}{\sigma^2}\right]\right\}
$$

A crucial feature is that the coefficients on these two statistics involve the unknown parameters. This data reduction is an important consequence of the likelihood analysis. These statistics are called **sufficient statistics** in that it is possible to show that they contain all the information in the data about the parameters μ and σ^2. We state this concept formally.

Definition 9.2 A statistic $T(X)$ is **sufficient** for a parameter θ if the probability density of the data conditional on the statistic is independent of θ. That is, if $f(X_1,\ldots,X_n \mid T(X),\theta) = h(X)$, then $T(X)$ is sufficient.

If there are two parameters, then the smallest number of sufficient statistics is two. Similarly, in the case of the Poisson likelihood above, the number of sufficient statistics was one consistent with the fact that there is only one parameter to estimate. There are very precise theorems that tell us how to identify the sufficient statistics for parameters in a probability model. We will not discuss them here. Details can be found in Pawitan (2001), DeGroot and Schervish (2002) and in Rice (2007).

Likelihood analyses in general are based on the sufficient statistics. When method-of-moments and likelihood analyses differ, it is usually because the sample moments are not the sufficient statistics for the parameters to be estimated. We will illustrate this point below when we define maximum likelihood estimation.

C. Maximum Likelihood Estimation

Ideally, we would like to analyze the likelihood function for every problem. This gets to be especially challenging when the number of parameters is large. Therefore, we would like to derive a summary that gives us a sense about the shape of the likelihood. This is what the **maximum likelihood estimate** and the **Fisher information** provide. We describe them in turn.

Definition 9.3. Maximum Likelihood Estimation. For each sample point $x = (x_1, ..., x_n)$ let $\hat{\theta}(x)$ be a parameter value of which $L(\theta) = L(\theta \mid x)$ attains a maximum as a function of θ for fixed x. $\hat{\theta}(x)$ is a **maximum likelihood (ML) estimator (MLE)** of the parameter θ.

In particular, if $L(\theta)$ is differentiable, we can consider $\frac{\partial L(\theta)}{\partial \theta} = 0$ and check the conditions on $\frac{\partial^2 L(\theta)}{\partial \theta^2}$ to be sure that the estimate defines a maximum. Remember that this would mean that for a one-dimensional problem, verifying that the second derivative was negative and for the $d-$dimensional problem verifying the condition that the determinant of the Hessian is negative definite. In likelihood analyses, it is usually easier to work with $\log L(\theta)$ instead of $L(\theta)$. The function $\frac{\partial \log L(\theta)}{\partial \theta}$ is called the **score function**. Hence, computing the MLE can often be formulated as finding $\hat{\theta}$ which solves the score equation

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0. \tag{9.11}$$

Two challenging problems for finding maximum likelihood estimates of parameters in complex high dimensional models is finding the solution to Eq. 9.11 numerically and establishing that this solution achieves a global maximum.

Example 9.1. Suppose that $x_1, \ldots, x_n$ are independent, identically distributed observations from a gamma distribution with parameters α and β. If α is known then we have the likelihood function is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{k=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_k^{\alpha-1} e^{-\beta x_k} \\ &= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \exp\{(\alpha-1)\sum_{k=1}^{n} \log x_k - \beta \sum_{k=1}^{n} x_k\} \end{aligned} \tag{9.12}$$

and the log likelihood function is

$$\log L(\alpha, \beta) = -n \log \Gamma(\alpha) + n\alpha \log \beta + (\alpha - 1)\sum_{k=1}^{n} \log(x_k) - \beta \sum_{k=1}^{n} x_k \tag{9.13}$$

Differentiating with respect to β

$$\frac{\partial \log L(\alpha, \beta)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{k=1}^{n} x_k \tag{9.14}$$

and solving yields

$$0 = \frac{\alpha}{\beta} - \bar{x} \tag{9.15}$$

$$\hat{\beta} = \frac{\alpha}{\bar{x}}. \tag{9.16}$$

Notice also that if α is known, then the second derivative of the log likelihood evaluated at $\hat{\beta}$ is

$$\left.\frac{\partial^2 \log L(\alpha,\beta)}{\partial \beta^2}\right|_{\hat{\beta}} = -\frac{n\alpha}{\hat{\beta}^2} = -\frac{n\bar{x}}{\alpha}$$

This is clearly negative since α is positive and $\bar{x}$ is positive. Hence, the log likelihood (and the likelihood) is maximized at $\hat{\beta}$. If $\alpha = 1$ we have $\hat{\beta} = \bar{x}^{-1}$ is the maximum likelihood estimate for an exponential model.

If α is unknown, then there is no closed form solution for either α or β. The estimates must then be found numerically. To see this we differentiate Eq. 9.13 with respect to α to obtain the log likelihood for α

$$\frac{\partial \log L(\alpha,\beta)}{\partial \alpha} = -n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + n\log\beta + \sum_{k=1}^{n}\log(x_k) \tag{9.17}$$

Substituting for β from Eq. 9.16 we obtain

$$\frac{\partial \log L(\alpha)}{\partial \alpha} = -n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + n\log\frac{\alpha}{\bar{x}} + \sum_{k=1}^{n}\log(x_k) \tag{9.18}$$

Setting the left hand side of Eq. 9.18 equal to zero and solving for the MLE of α requires a numerical procedure such as Newton's method. Plausible starting values for computing the ML estimates can be obtained from the method-of-moments estimates. We can see in this case that the method-of-moments estimates and the maximum likelihood estimates are different. The distinction between the maximum likelihood and the method-of-moments estimates comes about because for the gamma distribution, the sufficient statistics for α and β are $\sum_{k=1}^{n}\log(x_k)$ and $\sum_{k=1}^{n}x_k$. This is suggested by how the likelihood summarizes the data sample as indicated by the terms in the log likelihood function in Eq. 9.13. Given the definition of the sufficient statistics, this shows that the simple method-of-moments estimates are not "efficient" (i.e., they must have some loss of information). We will make this concept more precise below.

Example 3.2 (continued) Let $x_1, \ldots, x_n$ be a random sample of MEG background noise believed to obey a Gaussian probability density with unknown μ and σ^2. It is easy to show that $\bar{x} = n^{-1}\sum_{k=1}^{n}x_k$ is the maximum likelihood (ML) estimate of μ and $\hat{\sigma}^2 = n^{-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$ is the maximum likelihood estimate of σ^2. This is straightforward to show by differentiating the

Gaussian log likelihood, equating the gradient to zero and solving for μ and σ^2. Here the method-of-moments and the maximum likelihood estimates are the same. In this case, the sufficient statistics are the first two sample moments.

Example 9.2. Inverse Gaussian Distribution. If $x_1, \ldots, x_n$ is a random sample from an inverse Gaussian distribution with parameters μ and λ, then the likelihood is

$$\begin{aligned} L(\mu,\lambda) &= \prod_{k=1}^{n}\left(\frac{\lambda}{2\pi x_k{}^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(x_k-\mu)^2}{2x_k\mu^2}\right\} \\ &= \prod_{k=1}^{n}\left(\frac{\lambda}{2\pi x_k{}^3}\right)^{\frac{1}{2}} \exp\left\{-\sum_{k=1}^{n}\frac{\lambda(x_k-\mu)^2}{2x_k\mu^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{\lambda}{\mu^2}\sum_{k=1}^{n}x_k + \lambda\sum_{k=1}^{n}x_k^{-1} - n\left[\log\left(\frac{\lambda}{2\pi}\right)+2\frac{\lambda}{\mu}\right]+\sum_{k=1}^{n}\log x_k^3\right]\right\}. \end{aligned} \tag{9.19}$$

Recall that for this probability model the mean is μ and the variance is μ^3/λ. What are the maximum likelihood estimates of μ and λ? Are they the same as the method-of-moments estimate? The maximum likelihood estimates are

$$\hat{\mu}_{ML} = n^{-1}\sum_{i=1}^{n} x_i \tag{9.20}$$

$$\hat{\lambda}_{ML}{}^{-1} = n^{-1}\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\hat{\mu}_{ML}}\right) \tag{9.21}$$

whereas the method-of-moments estimates are

$$\hat{\mu}_{MM} = n^{-1}\sum_{i=1}^{n} x_i \tag{9.22}$$

$$\hat{\lambda}_{MM} = \bar{x}^3/\hat{\sigma}^2. \tag{9.23}$$

Can you venture a guess as to what are the sufficient statistics for μ and λ? (Remember the maximum likelihood estimates are always functions of the sufficient statistics. If you can derive the last line in Eq. 9.19, then the extra credit problem on Homework Assignment 7 is straightforward).

Invariance Property of the Maximum Likelihood Estimator. Suppose that a distribution is indexed by a parameter θ. Let $T(\theta)$ be a function of θ, then if $\hat{\theta}_{ML}$ is the maximum likelihood estimator of θ then $T(\hat{\theta}_{ML})$ is the maximum likelihood estimate of $T(\theta)$. For example, if $\theta = \lambda$, the scale parameter of the inverse Gaussian distribution the MLE $\hat{\lambda}_{ML}$ is given in Eq. 9.21. If we are interested in $T(\lambda) = \cos(\lambda)$ then the MLE of $T(\lambda)$ is $T(\hat{\lambda}_{ML})$. This is one of the most important properties of maximum likelihood estimates.

D. Fisher Information (Pawitan, 2001).
The MLE provides a point (or single number summary) estimate. "In general a single number is not sufficient to represent a likelihood function. If the log likelihood is well approximated by a quadratic function, then we need at least two quantities to represent it: the location of its maximum the (MLE) and the curvature at the maximum." The curvature will be the **Fisher information**. When this quadratic approximation is valid we call the likelihood function **regular**. Fortunately, when sample sizes are sufficiently large, the likelihood function generally does become regular. To restate this critical requirement, regular problems are those in which we can approximate the log likelihood function around the MLE by a quadratic function.

To define the Fisher information we first recall how we compute the Taylor series expansion of a function. The Taylor series expansion of a function $f(x)$ about a point x_0 is defined as

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)(x - x_0)^2}{2!} + ... \tag{9.24}$$

Let us assume that our log likelihood function is differentiable and let us expand the log likelihood function in a Taylor series about the ML estimate $\hat{\theta}_{ML}$. This gives

$$\log L(\theta) = \log L(\hat{\theta}_{ML}) + \log L(\hat{\theta}_{ML})'(\theta - \hat{\theta}_{ML}) + \frac{\log L(\hat{\theta}_{ML})''(\theta - \hat{\theta}_{ML})^2}{2!} + ... \tag{9.25}$$

Because $\log L(\hat{\theta}_{ML})' = 0$ since $\hat{\theta}_{ML}$ is the ML estimate of θ, we obtain

$$\log L(\theta) \approx \log L(\hat{\theta}_{ML}) + \frac{\log L(\hat{\theta}_{ML})''(\theta - \hat{\theta}_{ML})^2}{2!}. \tag{9.26}$$

Therefore, we can approximate $L(\theta)$ for θ close to $\hat{\theta}$ as

$$L(\theta) \approx L(\hat{\theta}_{ML}) \exp\{-\tfrac{1}{2} I(\hat{\theta}_{ML})(\theta - \hat{\theta}_{ML})^2\} \tag{9.27}$$

where

$$I(\hat{\theta}_{ML}) = -\log L(\hat{\theta}_{ML})''. \tag{9.28}$$

is the **observed Fisher information**. It is a number because it is evaluated at the MLE. $I(\hat{\theta}_{ML})$ measures the curvature of the likelihood around the MLE. Eq. 9.27 implies that near the MLE the likelihood can be approximated as a Gaussian distribution. We can gain intuition about the meaning of the Fisher information by considering the following Gaussian example.

Example 3.2 (continued). If we assume σ^2 is known, then ignoring irrelevant constants we get

$$\log L(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{9.29}$$

and

$$\frac{\partial \log L(\mu)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \tag{9.30}$$

$$\frac{\partial^2 \log L(\mu)}{\partial \mu^2} = -\frac{n}{\sigma^2} \tag{9.31}$$

Thus $I(\hat{\mu}) = \frac{n}{\sigma^2}$. Hence, $Var(\hat{\mu}) = n^{-1}\sigma^2 = I(\hat{\mu})^{-1}$. Thus a key point is: A large amount of information implies smaller variance.

Example 2.1 (continued). For the binomial model of the learning experiment, the log likelihood and score functions (ignoring irrelevant constants) are

$$\log L(p) = k \log p + (n-k)\log(1-p) \tag{9.32}$$

$$\frac{\partial \log L(p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} \tag{9.33}$$

The MLE is

$$\hat{p} = \frac{k}{n} \tag{9.34}$$

and

$$I(p) = -\frac{\partial^2 \log L(p)}{\partial p^2} = \frac{k}{p^2} + \frac{n-k}{(1-p)^2} \tag{9.35}$$

so that at the MLE the observed Fisher information is

$$I(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}. \tag{9.36}$$

See the **Addendum 1 to Lecture 9** to understand the details of the last step.

Remark 9.1. In principle, we can judge the quadratic approximation to the log likelihood by plotting the true log likelihood with the quadratic approximation superimposed.

Remark 9.2. For the Gaussian distribution from **Example 3.2**, the quadratic approximation is exact. A practical rule for nearly all likelihood applications: a reasonably regular likelihood means $\hat{\theta}_{ML}$ is approximately Gaussian.

Remark 9.3. The result in Eq. 9.27 suggests that an approximate 95% confidence interval for θ may be constructed as

$$\hat{\theta}_{ML} \pm 2se(\hat{\theta}_{ML}) \tag{9.37}$$

where $se(\hat{\theta}_{ML}) = [I(\hat{\theta}_{ML})]^{-\frac{1}{2}}$.

Remark 9.4. Reporting $\hat{\theta}_{ML}$ and $I(\hat{\theta}_{ML})$ is a practical alternative to plotting the likelihood function in every problem.

The quantity we derived above was the observed Fisher information. It is an estimate of the Fisher information which we define below.

Definition 9.4. If $x = x_1, ..., x_n$ is a sample with joint probability density $f(x|\theta)$, then the **Fisher information** in x for the parameter θ is

$$\begin{aligned} I(\theta) &= E[\frac{\partial \log f(x|\theta)}{\partial \theta}]^2 \\ &= \int [\frac{\partial \log f(x|\theta)}{\partial \theta}]^2 f(x|\theta) dx. \end{aligned} \tag{9.38}$$

Equivalently,

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}] \\ &= -\int \frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} f(x|\theta) dx. \end{aligned} \tag{9.39}$$

Now that we have defined the Fisher information we can use it to make an important general statement about the performance of an estimator.

Theorem 9.1 (Cramer-Rao Lower Bound). Suppose $x = x_1, \ldots, x_n$ is a sample from $f(x|\theta)$, and $T(x)$ is an estimator of θ and $E_\theta[(T(x))]$ is a differentiable function of θ. Suppose also that

$$\frac{d}{d\theta} \int h(x) f(x|\theta) dx = \int h(x) \frac{df(x|\theta)}{d\theta} dx, \tag{9.40}$$

for all $h(x)$ with $E_\theta |h(x)| < \infty$. Then

$$Var(T(x)) \geq \frac{(\frac{dE_\theta(T(x))}{d\theta})^2}{E_\theta(\frac{\partial \log f(x|\theta)}{\partial \theta})^2}. \tag{9.41}$$

Equation 9.41 is the **Cramer-Rao Lower Bound (CRLB)** for the estimator $T(x)$. For a proof see Casella and Berger (1990).

CRLB gives the lowest bound on the variance of an estimator. If the estimate is unbiased, then the numerator is 1 and the denominator is the Fisher information. If θ is a $d \times 1$ vector then the Fisher information is a $d \times d$ matrix given by

$$I(\theta) = E_\theta[(\frac{\partial \log f(x|\theta)}{\partial \theta})^T \frac{\partial \log f(x|\theta)}{\partial \theta}] = -E_\theta[\frac{\partial^2 \log f(x|\theta)}{\partial \theta \partial \theta^T}]. \quad (9.42)$$

We will make extensive use of the Fisher information to derive confidence intervals for our estimates and to make inferences in our problems.

E. Criteria for Evaluating an Estimator

If $T(x)$ is an estimator of θ, then there are several criteria that can be used to evaluate its performance. Four commonly used criteria to evaluate the performance of estimators are:

1. Mean-Squared Error (MSE)

$$E_\theta[T(x) - \theta]^2$$

The smaller the MSE the better the estimator.

2. Unbiasedness

$$E_\theta(T(x)) = \theta$$

The bias of an estimator is $b_T(\theta) = E(T(x)) - \theta.$

An estimator is unbiased if its expected value is the parameter it estimates.

3. Consistency

$$T(x) \xrightarrow{p} \theta \text{ as } n \to \infty$$

Consistency means that as the sample size increases the estimator converges in mean-squared error, probability or almost surely to the true value of the parameter.

4. Efficiency Achieves a minimum variance (Cramer-Rao Lower Bound).

The efficiency of an estimator is usually characterized in terms of its variance relative to another estimator. The Cramer-Rao Lower bound gives a characterization of how small the variance of an estimator can be.

F. Summary of Key Properties of Maximum Likelihood Estimation

1. Maximum likelihood estimates are generally biased.

2. Maximum likelihood estimates are consistent, hence, they are asymptotically unbiased.

3. Maximum likelihood estimates are asymptotically efficient.

4. The variance of the maximum likelihood estimate may be approximated by the reciprocal (inverse) of the Fisher information.

5. The maximum likelihood estimate $\hat{\theta}_{ML}$ is asymptotically Gaussian with mean θ and variance $I(\theta)^{-1}$.

6. If $\hat{\theta}_{ML}$ is the maximum likelihood estimate of θ then $h(\hat{\theta}_{ML})$ is the maximum likelihood estimate of $h(\theta)$.

7. Given that $\hat{\theta}_{ML} \sim N(\theta, I(\theta)^{-1})$, by the Invariance Principle for the MLE and a Taylor Series expansion we have

$$h(\hat{\theta}_{ML}) \approx N(h(\theta), h'(\theta)^2[I(\theta)]^{-1}) \tag{9.43}$$

and an approximate 95% confidence interval for $h(\theta)$ is

$$h(\hat{\theta}_{ML}) \pm 1.96 h'(\hat{\theta}_{ML})[I(\hat{\theta}_{ML})]^{-\frac{1}{2}}. \tag{9.44}$$

See **Addendum 2** to **Lecture 9** for an explanation and illustration of this result.

III. Summary

The likelihood function and maximum likelihood estimation are two of the most important theoretical and practical concepts in statistics.

Acknowledgments

I am grateful to Uri Eden for making the figures, to Jim Mutch for careful proofreading and comments and to Julie Scott for technical assistance.

Textbook References

Casella, G, Berger RL. *Statistical Inferenc*e. Belmont, CA: Duxbury, 1990.

DeGroot MH, Schervish MJ. *Probability and Statistics*, 3rd edition. Boston, MA: Addison Wesley, 2002.

Pawitan Y. *In All Likelihood: Statistical Modeling and Inference Using Likelihood.* London: Oxford, 2001.

Rice JA. *Mathematical Statistics and Data Analysis*, 3rd edition. Boston, MA, 2007.

Literature Reference

Kass RE, Ventura V, Brown EN. Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology* 2005, 94: 8-25.