0.1 Introduction

R. A. Fisher, a pioneer in the development of mathematical statistics, introduced a measure of the amount of information contained in an observation from \( f(x|\theta) \). Fisher information can be obtained by differentiating the identity

\[
1 = \int_{-\infty}^{\infty} f(x|\theta) \, dx
\]

with respect to \( \theta \). Assuming certain smoothness conditions concern differentiating under the integral the integral sign, we first obtain

\[
0 = \frac{\partial}{\partial \theta} 1 = \frac{\partial}{\partial \theta} \int f(x|\theta) \, dx = \int \frac{\partial}{\partial \theta} f(x|\theta) \, dx = \int \frac{\partial}{\partial \theta} \ln(f(x|\theta)) f(x|\theta) \, dx
\]

where the last equality follows after multiplying and dividing by \( f(x|\theta) \). This last condition states that

\[
E \left[ \frac{\partial}{\partial \theta} \ln(f(X_1|\theta)) \right] = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \ln(f(x|\theta)) f(x|\theta) \, dx = 0
\]

Fisher information is defined as the variance of \( \frac{\partial}{\partial \theta} \ln(f(X_1|\theta)) \) or

\[
I_1(\theta) = E \left[ \left( \frac{\partial}{\partial \theta} \ln(f(X_1|\theta)) \right)^2 \right]
\]

**Remark:** The information about \( \theta \) in a random a sample of size \( n \) from \( f(x|\theta) \) is

\[
I_n(\theta) = n E[\left( \frac{\partial}{\partial \theta} \ln(f(X_1|\theta)) \right)^2]
\]

It is sometimes convenient to use an alternative form for Fisher information that involves second derivatives.

\[
I_1(\theta) = E[- \frac{\partial^2}{\partial \theta^2} \ln(f(X_1|\theta))]
\]

This is obtained by differentiating a second time to obtain

\[
0 = \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} \ln(f(x|\theta)) f(x|\theta) \, dx
\]

\[
= \int \frac{\partial^2}{\partial \theta^2} \ln(f(x|\theta)) f(x|\theta) \, dx + \int \frac{\partial}{\partial \theta} \ln(f(x|\theta)) \frac{\partial}{\partial \theta} \ln(f(x|\theta)) f(x|\theta) \, dx
\]
Example Let $X_1, ..., X_n$ be a random sample of size $n$ from a normal distribution with known variance.

(a) Evaluate the Fisher information $I_1(\mu)$.
(b) Evaluate the alternative form of Fisher information

$$I_1(\mu) = E[-\frac{\partial^2}{\partial \mu^2} \ln( X_1 | \mu )]$$

Solution (a) Since the probability density function of a single observation is

$$\frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$$

we have

$$\frac{\partial}{\partial \mu} \ln f( x_1 | \mu) = \frac{\partial}{\partial \mu} \left( -\frac{1}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} (x_1 - \mu)^2 \right)$$

$$= \frac{1}{\sigma^2} (x_1 - \mu)$$

and

$$\left( \frac{\partial}{\partial \mu} \ln f( x_1 | \mu) \right)^2 = \frac{1}{\sigma^2} \frac{(x_1 - \mu)^2}{\sigma^2}$$

But, $E[(X_1 - \mu)^2]/\sigma^2 = var(X_1)/\sigma^2 = 1$ so

$$I_1(\mu) = \frac{1}{\sigma^2}$$

(b) Taking the second partial derivative gives

$$\frac{\partial^2}{\partial \mu^2} \ln f( x_1 | \mu) = \frac{\partial^2}{\partial \mu^2} \frac{(x_1 - \mu)}{\sigma^2} = -\frac{1}{\sigma^2}$$

so $I_1(\mu) = 1/\sigma^2$ which agrees with the original calculation involving only the first derivative.

Maximum Likelihood Estimation

0.2 Maximum Likelihood Estimation

A very general approach to estimation, proposed by R. A. Fisher, is called the method of maximum likelihood. To set the ideas, we begin with a special
0.2. MAXIMUM LIKELIHOOD ESTIMATION

Figure 1: The two possible distributions for $X$

case. Suppose that one of two distributions must prevail. For example, let $X$ take the possible values 0, 1, 2, or 3 with probabilities specified by distribution 1 or with probabilities specified by distribution 2 (see Table 1 and Figure 1).

Table 1. Two Possible Distributions for $X$

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p (x)$</td>
<td>.0625</td>
<td>.2500</td>
<td>.3750</td>
<td>.2500</td>
<td>.0625</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p (x)$</td>
<td>.2401</td>
<td>.4116</td>
<td>.2646</td>
<td>.0756</td>
<td>.0081</td>
</tr>
</tbody>
</table>

The first is the binomial distribution with $p = .5$ and the second the binomial with $p = .4$ but this fact is not important to the argument.

If we observe $X = 3$ should our estimate of the underlying distribution be distribution 1 or distribution 2? Suppose we take the attitude that we will
select the distribution for which the observed value \( x = 3 \) has the highest probability of occurring. Because this calculation is done after the data are obtained, we use the terminology of maximizing likelihood rather than probability. For the first distribution \( P[X = 3] = .2500 \) and for the second distribution \( P[X = 3] = .0756 \) so we estimate that the first distribution is the distribution that actually produced the observation 3.

If, instead, we observed \( X = 1 \) the estimate would be distribution 2 since .4116 is larger than .2500.

Let us take this example a step further and assume that \( X \) follows a binomial distribution with \( n = 3 \) but that \( 0 \leq p \leq 1 \) is unknown. The count \( X \) then has the distribution

\[
\binom{n}{x} p^x (1 - p)^{4-x} \quad \text{for} \quad x = 0, 1, 2, 3, 4
\]

If we again observe \( X = 3 \), we evaluate the binomial distribution at \( x = 3 \) and obtain

\[
4p^3(1 - p)^{4-3} \quad \text{for} \quad 0 \leq p \leq 1
\]

which is a function of \( p \). We now vary \( p \) to best explain the observed result. This curve, \( L(p) \), is shown in Figure 2.

We take the value at which the maximum occurs as our estimate. Using calculus, the maximum occurs at the value of \( p \) for which the derivative is zero.

\[
\frac{d}{dp} 4p^3(1 - p) = 4(3p^2 - 4p^3) = 0
\]

so our estimate is \( \hat{p} = .75 \). To review, this value maximizes the after the fact probability of observing the value 3.

More generally, a random sample of size \( n \) is taken from a distribution that depends on a parameter \( \theta \). The random sample produces \( n \) values \( x_1, x_2, ..., x_n \) which we substitute into the joint probability distribution, or probability density function, and then study the resulting function of \( \theta \).

\textit{Definition.} The function of \( \theta \) that is obtained by substituting the observed values of the random sample \( X_1 = x_1, ..., X_n = x_n \) into the joint probability distribution or the density function for \( X_1, X_2, ..., X_n \)

\[
L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta)
\]

is called the \textit{likelihood function for} \( \theta \).
We often simplify the notation and write \( L(\theta) \) with the understanding that the likelihood function does depend on the values \( x_1, x_2, \ldots, x_n \) from the random sample.

Given the values \( x_1, x_2, \ldots, x_n \) from a random sample, one distinctive feature of the likelihood function is the value or values of \( \theta \) at which it attains its maximum.

**Definition.** A statistic \( \hat{\theta}(x_1, \ldots, x_n) \) is a **maximum likelihood estimator** of \( \theta \) if \( \hat{\theta} \) is a value for the parameter that maximizes the likelihood function \( L(\theta | x_1, \ldots, x_n) \).

The maximum likelihood estimators satisfy a general **invariance**. Specifically, if \( g(\theta) \) is a continuous one-to-one function of \( \theta \) and \( \hat{\theta} \) is the maximum likelihood estimator of \( \theta \), the maximum likelihood estimator of \( g(\theta) \) is obtained by simple substitution.

\[
\text{maximum likelihood estimator of } g(\theta) = g(\hat{\theta})
\]
**Example** As in Example, let \( X_1, \ldots, X_n \) be a random sample of size \( n \) from the Poisson distribution

\[
f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

Obtain the maximum likelihood estimator of \( P[X_1 = 0] = e^{-\lambda} \).

From Example the maximum likelihood estimator of \( \lambda \) is \( \hat{\lambda} = \bar{x} \). Consequently, by the invariance property, the maximum likelihood estimator of \( e^{-\lambda} \) is \( e^{-\bar{x}} \).

### 0.3 Large Sample Distributions of MLE’s

When the likelihood has continuous partial derivatives and other regularity conditions hold, the distribution of the MLE (maximum likelihood estimator) \( \hat{\theta} \) tends to a normal distribution as the sample sizes increases. To simplify the statement, we assume that the maximum likelihood estimator is uniquely defined.

**Theorem** [Asymptotic Normality] Let \( \theta \) be the prevailing value of the parameter and let \( \hat{\theta} \) be the maximum likelihood estimator. Under suitable regularity conditions,

\[
P_{\theta}[\sqrt{n}(\hat{\theta} - \theta) \leq y] - \int_{-\infty}^{y} \frac{I_1^{1/2}(\theta)}{\sqrt{2\pi}} \exp(-\frac{1}{2}I_1(\theta)u^2)du
\]

That is, \( \sqrt{n}(\hat{\theta} - \theta) \) converges in distribution to a normal distribution having mean 0 and variance \( 1/I_1(\theta) \), where \( I_1(\theta) \) is the Fisher information in a sample of size one evaluated at the prevailing parameter. The result holds with \( I_1(\hat{\theta}) \) or even the empirical information

\[
\text{empirical information} = \frac{1}{n}(\frac{\partial}{\partial \theta} \ln f( X|\theta))^2
\]

\[
= (\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f( X_i|\theta))^2 = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f( X_i|\theta)
\]

which can be evaluated at the prevailing \( \theta \) or the MLE \( \hat{\theta} \).

* A large sample approximate confidence interval for \( \theta \) is

\[
( \hat{\theta} - z_{\alpha/2} \frac{1}{\sqrt{n I(\theta)}}, \hat{\theta} + z_{\alpha/2} \frac{1}{\sqrt{n I(\theta)}} )
\]
Some students may be interested in the idea of Proof. This can be ignored.

Idea of Proof of Normal Limit Recall the expansion \( h(u) = h(u_0) + h'(u_0)(u - u_0) + h''(u_0)(u - u_0)^2/2 \).

\[
\frac{\partial}{\partial \theta} \ln f(X|\theta) \Big|_{\hat{\theta}} = \frac{\partial}{\partial \theta} \ln f(X|\theta) + \frac{\partial^2}{\partial \theta^2} \ln f(X|\theta) \big|_{\hat{\theta}}(\theta - \hat{\theta}) + \frac{\partial^3}{\partial \theta^3} \ln f(X|\theta) \big|_{\hat{\theta}}(\theta - \hat{\theta})^2/2
\]

The term on the left side vanishes under appropriate conditions. The first term on the right is the sum of \( n \) independent and identically distributed random variables having mean 0 and variance \( I_1(\theta) \). Dividing both sides by \( \sqrt{n} \), we conclude from the central limit theorem that this first term converges in distribution to a normal distribution with mean 0 and variance \( I_1(\theta) \).

The second term on the right hand side can be written as

\[
\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln f(X|\theta) \sqrt{n}(\theta - \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(X_i|\theta) \sqrt{n}(\theta - \hat{\theta})
\]

The sum converges to \(-I_1(\theta)\) by the law of large numbers so the result follows. We neglect the last term on the right.

**Remark 2** In the vector parameter case, the limiting distribution is the multivariate normal distribution with mean vector 0 and covariance matrix \( I_1(\theta_1, \theta_2) \).

The regularity conditions do not hold for the uniform distributions \( f(x; \theta) = 1/\theta \) for \( 0 < x < \theta \).