Chapter 7
Maximum Likelihood: Basic Results

7.1 Basic Properties

As we have seen once we have an estimator and its sampling distribution we can easily obtain confidence intervals and tests regarding the parameter. We now develop the theory of estimation focusing on the method of maximum likelihood, which for parametric models is the most widely used method. This will also supply us with a collection of statistical methods for important problems.

For comparing two values of a parameter, θ_2 vs θ_2, a natural role is played by the likelihood ratio

$$\mathscr{L}\mathscr{R}(\theta_2, \theta_1; x) = \frac{f(x;\theta_2)}{f(x;\theta_1)}$$

According to the Law of Likelihood the likelihood ratio represents the statistical evidence in the data for comparing θ_2 to θ_1.

The **score function** is defined by

$$s(\theta; x) = \frac{\partial \ln[f(x;\theta)]}{\partial\theta}$$

The score function plays a major role in the theory of maximum likelihood estimation.

Example. Consider n iid normal random variables with parameters θ, σ^2 where σ^2 is known. Then

$$f(x;\theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}$$

C.A. Rohde, *Introductory Statistical Inference with the Likelihood Function*, DOI 10.1007/978-3-319-10461-4_7

and

$$f^{'}(x;\theta) = f(x;\theta)\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta)$$

It follows that

$$\frac{f^{'}(x;\theta)}{f(x;\theta)} = \frac{n(\overline{x} - \theta)}{\sigma^2}$$

As a random variable we have that the score function has expected value 0 and variance n/σ^2 when evaluated at the true θ.

Because of the Law of Likelihood a natural estimate of θ is that value of θ which maximizes the likelihood or the log of the likelihood.

Assuming that $\ln[f(x;\theta)]$ is differentiable with respect to θ the maximum likelihood estimate is then the solution to

$$\frac{\partial \ln[f(x;\theta)]}{\partial\theta} = 0$$

which is called the **likelihood** or **score** equation. If there are r parameters we differentiate with respect to each and equate to 0, obtaining r equations. Note that one needs to check the second derivative to ensure a maximum.

Example 1 (Binomial).. If X is binomial with parameter θ then

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad x = 0, 1, \ldots, n$$

First note that if $x = 0$ then $f(0;\theta) = (1-\theta)^n$ and in this case $\widehat{\theta} = 0$. If $x = n$ then $f(n;\theta) = \theta^n$ and in this case $\widehat{\theta} = 1$.

For $x = 1, 2, \ldots, n-1$ we have that

$$\ln[f(x;\theta)] = \ln[\binom{n}{x}] = x\ln(\theta) + (n-x)\ln(1-\theta)$$

and

$$\frac{\partial \ln[f(x;\theta)]}{\partial\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x - n\theta}{\theta(1-\theta)}$$

It follows that

$$\widehat{\theta} = \frac{x}{n} \text{ for } x = 0, 1, \ldots, n$$

Note that it is unbiased with variance $\theta(1-\theta)/n$ so that it is also consistent.

Example 2.. Let $Y_1, Y_2, \ldots, Y_n$ be iid each normal with mean μ and variance σ^2. Then we have

$$\begin{aligned} f(y;\theta) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

It follows that the log likelihood is given by

$$\ln[f(y;\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Thus we have that

$$\frac{\partial \ln[f(x;\theta)]}{\partial \mu} = \frac{1}{\sigma^2} n(\overline{y}-\mu)$$

and

$$\frac{\partial \ln[f(x;\theta)]}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^4}$$

and it follows that

$$\widehat{\mu} = \overline{y} \text{ and } \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

7.2 Consistency of Maximum Likelihood

1. Consider the case where there are only two possible values of the parameter θ_2 and θ_1.
2. Also suppose that we have n observations which are realized values of independent and identically distributed random variables having density $f(x;\theta_2)$ or $f(x;\theta_1)$.

   The maximum likelihood estimate is defined by

$$\widehat{\theta} = \begin{cases} \theta_2 \text{ if } f(x_1, x_2, \ldots, x_n; \theta_2) \geq f(x_1, x_2, \ldots, x_n; \theta_1) \\ \theta_1 \text{ otherwise} \end{cases}$$

1. Assume with no loss of generality that θ_2 is the true value of the parameter.
2. The maximum likelihood estimator is consistent if

$$\mathbb{P}_{\theta_2}(\widehat{\theta} = \theta_2) \quad \longrightarrow \quad 1$$

We note that $\widehat{\theta} = \theta_2$ if and only if

$$\frac{f(x_1, x_2, \ldots, x_n; \theta_2)}{f(x_1, x_2, \ldots, x_n; \theta_1)} = \prod_{i=1}^{n} \frac{f(x_i; \theta_2)}{f(x_i; \theta_1)} > 1$$

Equivalently

$$\sum_{i=1}^{n} \ln \left[\frac{f(x_i; \theta_2)}{f(x_i; \theta_1)} \right] > 0$$

Now note that the random variables

$$Y_i = \ln \left[\frac{f(X_i; \theta_2)}{f(X_i; \theta_1)} \right] \quad i = 1, 2, \ldots, n$$

are independent and identically distributed.

Moreover

$$\begin{aligned}
\mathbb{E}_{\theta_2}(Y_i) &= \int \ln \left[\frac{f(x; \theta_2)}{f(x; \theta_1)} \right] f(x; \theta_2) \lambda(dx) \\
&= -\int \ln \left[\frac{f(x; \theta_1)}{f(x; \theta_2)} \right] f(x; \theta_2) \lambda(dx) \\
&> -\int \left[\frac{f(x; \theta_1)}{f(x; \theta_2)} - 1 \right] f(x; \theta_2) \lambda(dx) \\
&= 0
\end{aligned}$$

By the law of large numbers we have that

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \quad \xrightarrow{p} \quad \mathbb{E}_{\theta_2}(Y) > 0$$

and hence

$$\mathbb{P}_{\theta_2}(\widehat{\theta} = \theta_2) \quad \longrightarrow \quad 1$$

i.e., $\widehat{\theta}$ is consistent:

1. The same proof holds provided the parameter space Θ is finite.
2. The more general case where Θ is an interval requires more delicate arguments and is of technical, not statistical interest.

7.3 General Results on the Score Function

We know that

$$\int f(x;\theta) d\lambda(x) = 1$$

for any density function $f(x;\theta)$. Recall that for a function g we write

$$\int g(x;\theta) d\lambda(x) = \begin{cases} \int g(x;\theta) dx \ g \text{ continuous} \\ \sum g(x;\theta) \ g \text{ discrete} \end{cases}$$

Assuming that we can differentiate under the integral or summation sign, we have that

$$\int \frac{\partial f(x;\theta)}{\partial \theta} d\lambda(x) = 0$$

Now note that

$$\frac{\partial f(x;\theta)}{\partial \theta} = \frac{\partial \ln[f(x;\theta)]}{\partial \theta} f(x;\theta)$$

It follows that

$$\mathbb{E}_\theta \left\{ \frac{\partial \ln[f(x;\theta)]}{\partial \theta} \right\} = 0$$

Thus the expected value of the score function is 0.

If we differentiate again we have that

$$\int \frac{\partial^2 f(x;\theta)}{\partial \theta^2} \lambda(x) = 0$$

Noting that

$$\begin{aligned} \frac{\partial^2 f(x;\theta)}{\partial \theta^2} &= \frac{\partial}{\partial} \left[\frac{\partial f(x;\theta)}{\partial \theta} \right] \\ &= \frac{\partial}{\partial} \left[\frac{\partial \ln[f(x;\theta)]}{\partial \theta} f(x;\theta) \right] \end{aligned}$$

we see that

$$\frac{\partial^2 f(x;\theta)}{\partial \theta^2} = \left[\frac{\partial^2 \ln[f(x;\theta)]}{\partial \theta^2} \right] f(x;\theta)$$

$$+\left[\frac{\partial \ln[f(x;\theta)]}{\partial\theta}\right]\frac{\partial f(x;\theta)}{\partial\theta}$$

The right-hand side may be written as

$$\left[\frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta^2}\right] f(x;\theta) + \left[\frac{\partial \ln[f(x;\theta)]}{\partial\theta}\right]^2 f(x;\theta)$$

It follows that

$$\mathbb{E}_\theta\left\{\left[\frac{\partial \ln[f(x;\theta)]}{\partial\theta}\right]^2\right\} = -\mathbb{E}_\theta\left\{\left[\frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta^2}\right]\right\}$$

and hence

$$\mathbb{V}_\theta\left\{\frac{\partial \ln[f(x;\theta)]}{\partial\theta}\right\} = -\mathbb{E}_\theta\left\{\left[\frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta^2}\right]\right\}$$

The quantity

$$-\mathbb{E}_\theta\left\{\left[\frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta^2}\right]\right\}$$

is called the (expected) **Fisher information** and

$$-\left[\frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta^2}\right]$$

is called the (observed) **Fisher information**.

7.4 General Maximum Likelihood

1. Let X be a random variable with density $f(x;\theta)$.
2. Assume that the parameter space Θ is an interval and that $f(x;\theta)$ is sufficiently smooth so that derivatives with respect to θ are defined and that differentiation under a summation or integral is allowed.
3. Finally assume that the range of X does not depend on θ.

Under weak regularity conditions it follows from the previous section that

$$\mathbb{E}_\theta\left\{\left[\frac{\partial \ln[f(X;\theta)]}{\partial\theta}\right]\right\} = 0$$

$$\mathbb{E}_\theta\left\{\left[\frac{\partial \ln[f(X;\theta)]}{\partial\theta}\right]^2\right\} = -\mathbb{E}_\theta\left\{\left[\frac{\partial^2 \ln[f(X;\theta)]}{\partial\theta^2}\right]\right\}$$

Thus the random variable

$$U(\theta) = \left[\frac{\partial \ln[f(X;\theta)]}{\partial \theta}\right]$$

i.e., the **score function** has expected value and variance given by

$$\mathbb{E}_\theta[U(\theta)] = 0 \ \ , \mathbb{V}_\theta[U(\theta)] = i(\theta)$$

where

$$i(\theta) = -\mathbb{E}_\theta \left\{ \left[\frac{\partial^2 \ln[f(X;\theta)]}{\partial \theta^2}\right] \right\}$$

is the expected Fisher information for a sample size of one.

Example. If X is normal with mean θ and variance σ^2 with σ^2 known then

$$\ln[f(x;\theta)] = -\frac{1}{2}\ln[2\pi\sigma^2] - \frac{1}{2\sigma^2}(x-\theta)^2$$

and hence

$$\frac{\partial \ln[f(x;\theta)]}{\partial \theta} = \frac{x-\theta}{\sigma^2}$$

and

$$\frac{\partial^2 \ln[f(x;\theta)]}{\partial \theta^2} = -\frac{1}{\sigma^2}$$

so Fisher's information is

$$i(\theta) = \frac{1}{\sigma^2}$$

Example. If X is Bernoulli θ then

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}$$

and hence

$$\ln[f(x;\theta)] = x\ln(\theta) + (1-x)\ln(1-\theta)$$

It follows that

$$\frac{\partial \ln[f(x;\theta)]}{\partial\theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

and

$$\frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta^2} = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

so Fisher's information is

$$i(\theta) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

If we have a random sample $X_1, X_2, \ldots, X_n$ from $f(x;\theta)$ and if

$$u_i(\theta) = \frac{\partial \ln[f(x_i;\theta)]}{\partial\theta}$$

then

$$\overline{U}(\theta) = \frac{1}{n}\sum_{i=1}^{n} U_i(\theta)$$

is the sample mean of n iid random variables with expected value 0 and variance $i(\theta)$. It follows that

$$\sqrt{n}\,\overline{U} \quad \xrightarrow{d} \quad \mathrm{N}[0, i(\theta)]$$

by the central limit theorem.

Define the maximum likelihood estimate of θ as that value of θ which maximizes $f(\mathbf{x};\theta)$ or equivalently $\ln[f(\mathbf{x};\theta)]$.

Thus we solve

$$\frac{\partial \ln[f(\mathbf{x};\theta)]}{\partial\theta} = 0$$

or when $f(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ we solve

$$u(\theta) = \sum_{i=1}^{n} u_i(\theta) = 0$$

Since we can write, using Taylor's theorem,

$$u(\widehat{\theta}) = u(\theta) + \frac{du(\theta)}{d\theta}(\widehat{\theta} - \theta) + v(\theta^*)\frac{(\widehat{\theta} - \theta)^2}{2}$$

where

$$v(\theta^*) = \left.\frac{d^2u(\theta)}{d\theta^2}\right|_{\theta=\theta^*}$$

and θ^* is between θ and $\widehat{\theta}$.

Since $u(\widehat{\theta}) = 0$ we have

$$(\widehat{\theta} - \theta)\left[\frac{du(\theta)}{d\theta} + v(\theta^*)\frac{(\widehat{\theta} - \theta)}{2}\right] = -u(\theta)$$

It follows that

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{1}{\sqrt{n}}u(\theta)}{\left[-\frac{1}{n}\frac{du(\theta)}{d\theta} - \frac{1}{n}v(\theta^*)\frac{(\widehat{\theta}-\theta)}{2}\right]}$$

Application of the results of the preceding section shows that

$$\sqrt{n}(\widehat{\theta} - \theta) \quad \xrightarrow{d} \quad \mathrm{N}(0, [i(\theta)]^{-1})$$

where $i(\theta)$ is Fisher's information for a sample of size 1.

7.5 Cramer-Rao Inequality

If $t(x)$ is any unbiased estimator of θ i.e.

$$\mathbb{E}[t(X)] = \theta$$

then

$$\int t(x)f(x;\theta)d\lambda(x) = \theta$$

Assuming that we can differentiate under the integral or summation sign, we have that

$$\int t(x)\frac{\partial \ln[f(x;\theta)]}{\partial\theta}f(x;\theta)d\lambda(x) = 1$$

and hence

$$\mathbb{C}\left\{t(X), \left[\frac{\partial \ln[f(X;\theta)]}{\partial\theta}\right]\right\} = 1$$

It follows that

$$\mathbb{V}[t(X)]\mathbb{V}\left\{\frac{\partial \ln[f(X;\theta)]}{\partial\theta}\right\} \geq 1$$

or

$$\mathbb{V}[t(X)] \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the expected Fisher information. Thus the smallest variance for an unbiased estimator is the inverse of Fisher's information. This result is called the Cramer–Rao inequality.

Since $1/I(\theta)$ is the large sample variance of the maximum likelihood estimator we have the result that the method of maximum likelihood produces estimators which are asymptotically efficient, i.e., have smallest variance.

7.6 Summary Properties of Maximum Likelihood

1. Maximum likelihood have the **equivariance** property: i.e., the maximum likelihood estimate of $g(\theta)$, $\widehat{g(\theta)}$, is $g(\widehat{\theta})$.
2. Under weak regularity conditions maximum likelihood estimators are **consistent**, i.e.,

$$\widehat{\theta} \xrightarrow{p} \theta$$

3. Maximum likelihood estimators are **asymptotically normal**:

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, v(\theta_0))$$

   where $v(\theta_0)$ is the inverse of **Fisher's information**.
4. Maximum likelihood estimators are **asymptotically efficient**, i.e., in large samples

$$\mathbb{V}(\widehat{\theta}) \leq \mathbb{V}(\widetilde{\theta})$$

   where $\widetilde{\theta}$ is any other consistent estimator which is asymptotically normal.

The regularity conditions under which the results on maximum likelihood estimators are true consist of conditions of the form:

(i) The range of the distributions cannot depend on the parameter.
(ii) The first three derivatives of the log likelihood function with respect to θ exist are continuous and have finite expected values as functions of X.

7.7 Multiparameter Case

All of the results for maximum likelihood generalize to the case where there are p parameters $\theta_1, \theta_2, \ldots, \theta_p$. Let

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}$$

If the pdf is given by

$$f(\mathrm{x}; \boldsymbol{\theta})$$

the maximum likelihood or score equation is

$$\frac{\partial \ln[f(\mathrm{x}; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ln[f(\mathrm{x};\boldsymbol{\theta})]}{\partial \theta_1} \\ \frac{\partial \ln[f(\mathrm{x};\boldsymbol{\theta})]}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ln[f(\mathrm{x};\boldsymbol{\theta})]}{\partial \theta_p} \end{bmatrix} = \mathbf{0}$$

Fisher's information matrix

$$\mathcal{I}(\boldsymbol{\theta})$$

has $i - j$ element given by

$$-\frac{\partial^2 \ln[f(\mathrm{x}; \boldsymbol{\theta})]}{\partial \theta_i \partial \theta_j}$$

Note that it is a $p \times p$ matrix.

Under regularity conditions, similar to those for the single parameter case we have

1. The maximum likelihood estimate of $g(\boldsymbol{\theta})$, $\widehat{g(\boldsymbol{\theta})}$, is $g(\widehat{\boldsymbol{\theta}})$.
2. Maximum likelihood estimators are **consistent**, i.e.,

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$$

3. Maximum likelihood estimators are **asymptotically normal**:

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \;\approx\; \mathrm{N}(0, \mathbf{V}_n(\boldsymbol{\theta}_0))$$

where $\mathbf{V}_n(\boldsymbol{\theta}_0)$ is the inverse of Fisher's information matrix. We can replace $\boldsymbol{\theta}_0$ by $\widehat{\boldsymbol{\theta}}$ to use this result to determine confidence intervals.

7.8 Maximum Likelihood in the Multivariate Normal

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ be independent each having a multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e.,

$$f_{\mathbf{Y}_i}(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}}[\det(\boldsymbol{\Sigma})]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right\}$$

The joint density is thus

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}[\det(\boldsymbol{\Sigma})]^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right\}$$

We will show that the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$$

and

$$\boldsymbol{\Sigma} = \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^\top$$

i.e., the $j - k$ element of $\mathbf{S}$ is

$$\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)(y_{ik} - \overline{y}_k)$$

essentially the sample covariance between the jth and kth variable.

The first step is to note that

$$\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$$

can be written as

$$\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}}) + n(\overline{\mathbf{y}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu})$$

or

$$n\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] + n(\overline{\mathbf{y}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu})$$

where the trace of a square matrix, tr(A), is the sum of the diagonal elements, i.e.,

$$\text{tr}(A) = \sum_{i=1}^{p} a_{ii}$$

Thus the joint density $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ = can be written as

$$(2\pi)^{-\frac{np}{2}}[\det(\boldsymbol{\Sigma})]^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] - \frac{n}{2}(\overline{\mathbf{y}} - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu})\right\}$$

It follows immediately that the maximum likelihood estimate of $\boldsymbol{\mu}$ is $\overline{\mathbf{y}}$ and the joint density at $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$ is thus

$$f_{\mathbf{Y}}(\mathbf{y}; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = (2\pi)^{-\frac{np}{2}}[\det(\mathbf{S})]^{-\frac{n}{2}} \exp\left\{-\frac{np}{2}\right\}$$

The ratio

$$\frac{f_{\mathbf{Y}}(\mathbf{y}; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})}{f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

is thus equal to

$$\frac{[\det(\mathbf{S})]^{-\frac{n}{2}} \exp\left\{-\frac{np}{2}\right\}}{[\det(\boldsymbol{\Sigma})]^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] - \frac{n}{2}(\overline{\mathbf{y}} - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu})\right\}}$$

which is greater than or equal to

$$\det(\boldsymbol{\Sigma}^{-1}\mathbf{S})^{-\frac{n}{2}} \exp\left\{-\frac{np}{2} + \frac{n}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\}$$

This ratio is greater than or equal to 1 if and only its logarithm is greater than or equal to 0. The logarithm is

$$\frac{n}{2}\left\{-\ln\left[\det(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right] - p + \text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\}$$

If $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the characteristic roots of $\boldsymbol{\Sigma}^{-1}\mathbf{S}$ then it can be shown that

1. $\lambda_i \geq 0$ for each i
2. $\det(\boldsymbol{\Sigma}^{-1}\mathbf{S}) = \prod_{i=1}^{p}\lambda_i$
3. tr$(\boldsymbol{\Sigma}^{-1}\mathbf{S}) = \sum_{i=1}^{p}\lambda_i$

It follows that the log of the ratio is greater than or equal to

$$\frac{n}{2}\left\{-\sum_{i=1}^{p}\ln(\lambda_i) - \frac{p}{2} + \sum_{i=1}^{p}\lambda_i\right\}$$

or

$$\frac{n}{2}\left\{\sum_{i=1}^{p}[\lambda_i - 1 - \ln(\lambda_i)]\right\}$$

which is greater than or equal to zero since

$$a - 1 - \ln(a) \geq 0 \text{ for any positive real number}$$

Thus the maximum likelihood estimators for the multivariate normal are

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{y}} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \mathbf{S}$$

We usually use

$$\frac{n}{n-1}\mathbf{S}$$

as the estimator so that the estimated components of $\boldsymbol{\Sigma}$ are exactly the sample covariances and variances.

7.9 Multinomial

Suppose that $X_1, X_2, \ldots, X_k$ have a multinomial distribution, i.e.,

$$f(x_1, x_2, \ldots, x_k; \theta_1, \theta_2, \ldots, \theta_k) = n! \prod_{i=1}^{k} \frac{\theta_i^{x_i}}{x_i!}$$

where

$$0 \leq x_i \leq n \text{ each } i = 1, 2, \ldots, k \text{ and } \sum_{i=1}^{k} x_i = n$$

and

$$0 \leq \theta_i \leq 1 \text{ each } i = 1, 2, \ldots, k \text{ and } \sum_{i=1}^{k} \theta_i = 1$$

Note that

$$\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i \text{ and } x_k = n - \sum_{i=1}^{k-1} x_i$$

The maximum likelihood estimates of the θ_i are found by taking the partial derivatives of the log likelihood with respect to θ_i for $i = 1, 2, \ldots, k-1$ where the log likelihood is

$$\ln[f(\mathbf{x}, \boldsymbol{\theta}] = \ln(n!) - \sum_{i=1}^{k} \ln(x_i!) + \sum_{i=1}^{k} x_i \ln(\theta_i)$$

Since $\theta_k = 1 - \theta_1 - \theta_2 - \cdots - \theta_{k-1}$ we have

$$\frac{\partial \ln[f(\mathbf{x}, \boldsymbol{\theta})]}{\partial \theta_i} = \frac{x_i}{\theta_i} - \frac{x_k}{\theta_k}$$

for $i = 1, 2, \ldots, k-1$. It follows that the maximum likelihood estimates satisfy

$$x_i \widehat{\theta}_k = \widehat{\theta}_i x_k \text{ for } i = 1, 2, \ldots, k-1$$

Summing from $i = 1$ to $k - 1$ yields

$$(n - x_k)\widehat{\theta}_k = (1 - \widehat{\theta}_k)x_k$$

and hence

$$n\widehat{\theta}_k = x_k$$

so that

$$\frac{x_i x_k}{n} = \widehat{\theta}_i x_k \text{ or } \widehat{\theta}_i = \frac{x_i}{n}$$

The second derivatives of the log likelihood are given by

$$\frac{\partial^2 \ln[f(\mathbf{x}, \boldsymbol{\theta})]}{\partial \theta_i^2} = -\frac{x_i}{\theta_i^2} - \frac{x_k}{\theta_k}$$

which has expected value

$$-\frac{n\theta_i}{\theta_i^2} - \frac{n\theta_k}{\theta_k^2} = -\frac{n}{\theta_i} - \frac{n}{\theta_k}$$

and

$$\frac{\partial^2 \ln[f(\mathbf{x}, \boldsymbol{\theta})]}{\partial \theta_i \partial \theta_j} = -\frac{x_k}{\theta_k^2}$$

which has expected value

$$-\frac{n\theta_k}{\theta_k^2} = -\frac{n}{\theta_k}$$

Thus Fisher's information matrix, $\mathcal{I}(\boldsymbol{\theta})$, is given by

$$\mathcal{I}(\boldsymbol{\theta}) = n \begin{bmatrix} \frac{1}{\theta_1} + \frac{1}{\theta_k} & \frac{1}{\theta_k} & \cdots & \frac{1}{\theta_k} \\ \frac{1}{\theta_k} & \frac{1}{\theta_2} + \frac{1}{\theta_k} & \cdots & \frac{1}{\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\theta_k} & \frac{1}{\theta_k} & \cdots & \frac{1}{\theta_{k-1}} + \frac{1}{\theta_k} \end{bmatrix}$$

Fisher's information can be written in matrix form as

$$n\left[\mathbf{D}(\boldsymbol{\theta})^{-1} + \frac{1}{\theta_k}\mathbf{1}\mathbf{1}^\top\right]$$

where $\mathbf{D}(\boldsymbol{\theta})$ is a $k-1 \times k-1$ matrix with diagonal elements $\theta_1, \theta_2, \ldots, \theta_{k-1}$ and $\mathbf{1}$ is a $k-1$ column vector with each element equal to 1.

The general theory of maximum likelihood then implies that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \quad \xrightarrow{d} \quad \mathrm{N}\left(\mathbf{0}, [i(\boldsymbol{\theta})]^{-1}\right)$$

where $i(\boldsymbol{\theta})$ is Fisher's information matrix with $n = 1$.

It is easy to check that

$$[i(\boldsymbol{\theta})]^{-1} = \mathbf{D}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^\top$$

or

$$[i(\boldsymbol{\theta})]^{-1} = \begin{bmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_{k-1} \\ -\theta_2\theta_1 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{k-1}\theta_1 & -\theta_{k-1}\theta_2 & \cdots & \theta_{k-1}(1-\theta_{k-1}) \end{bmatrix}$$

which we recognize as the variance covariance matrix of $X_1, X_2, \ldots, X_{k-1}$

Standard maximum likelihood theory implies that

$$n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top [i(\boldsymbol{\theta}]\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \quad \xrightarrow{d} \quad \chi^2(k-1)$$

Now note that

$$n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top [i(\boldsymbol{\theta}](\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

is equal to

$$n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \left[\mathbf{D}(\boldsymbol{\theta})^{-1} + \frac{1}{p_k}\mathbf{1}\mathbf{1}^\top\right](\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

and hence to

$$n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathbf{D}(\boldsymbol{\theta})^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{n}{\theta_k}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathbf{1}\mathbf{1}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

This last expression simplifies to

$$n\sum_{i=1}^{k-1}\frac{(\widehat{\theta}_i - \theta_i)^2}{\theta_i} + \frac{n}{\theta_k}\left[\sum_{i=1}^{k-1}(\widehat{\theta}_i - \theta_i)\right]^2$$

which in turn simplifies to

$$\sum_{i=1}^{k-1}\frac{(x_i - n\theta_i)^2}{n\theta_i} + \frac{n}{\theta_k}(\theta_k - \widehat{\theta}_k)^2$$

and to

$$\sum_{i=1}^{k-1}\frac{(x_i - n\theta_i)^2}{n\theta_i} + \frac{(x_k - n\theta_k)^2}{n\theta_k}$$

This finally reduces to

$$\sum_{i=1}^{k}\frac{(x_i - n\theta_i)^2}{n\theta_i}$$

Noting that $\mathbb{E}(X_i) = n\theta_i = E_i$ this last formula may be written as

$$\sum_{i=1}^{k}\frac{(\mathrm{X}_i - \mathrm{E}_i)^2}{\mathrm{E}_i}$$

which is called Pearson's chi-square statistic. For large n, it has a chi-square distribution with $k - 1$ degrees of freedom.