1 The Likelihood Function and Sufficiency

1.1 The likelihood function

Consider taking a sample from some infinite population, where an observed value of some variable is associated with each sample member. (Note the theoretical concept of an *infinite population*, an idealization which may refer to a situation where observations may be taken repeatedly without limit, as in a sequence of tosses of a coin, or which may be regarded as an approximation to a very large population.) Let the vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ denote the resulting observations from a sample of size n. We envisage that the *statistical experiment* of taking such a sample of size n could be carried out repeatedly, each time obtaining a new set of observations, which would vary in a random manner from repetition to repetition. Hence the use of the terminology of a *random variable* for each of the individual observed values X_i and *random vector* for the vector of observations $\mathbf{X}$.

We set up a *statistical model* for the observations in the form of a family $\mathcal{F}$ of probability distributions. Let $\mathcal{X}$ denote the *range*, the set of all possible values that the random vector $\mathbf{X}$ can take. The random vector $\mathbf{X}$ will be assumed to have a (joint) probability density function (p.d.f.) $\{f(\mathbf{x};\theta),\ \mathbf{x} \in \mathcal{X})\}$ that depends upon a vector of q parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$. Some or all of the parameters will be unknown, and the purpose of carrying out the sampling experiment will be to make inferences about the unknown parameters. Let Θ denote the *parameter space*, the set of all possible values that the vector of parameters θ can take. The family $\mathcal{F}$ of probability distributions with which we are concerned may thus be written as

$$\mathcal{F} = \{f(\mathbf{x};\theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}.$$

The vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ will represent the observed sample values obtained on a particular occasion when the experiment has been carried out. We shall distinguish between random variables (r.v.s), denoted by upper case letters, and realizations of the r.v.s, the values observed on a particular occasion, denoted by lower case letters, with a similar distinction between the notation $\mathbf{X}$ for a random vector and the notation $\mathbf{x}$ for a particular realization of $\mathbf{X}$. However, the distinction will often become blurred.

- We shall use the same notation $f(\mathbf{x};\theta)$ for both continuous and discrete random variables. In the discrete case, $f(\mathbf{x};\theta)$ is the probability mass function (p.m.f.).
- Usually the sample will be a *random sample*. Given that we are sampling from an infinite population, this means that, for any given θ, $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables (i.i.d. r.v.s), so that their joint p.d.f. (or p.m.f.), now written as $f_{\mathbf{X}}(\cdot)$, may be factorized as

$$f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta),$$

where on the right hand side in the above expression $f(x_i;\theta)$ now represents the marginal p.d.f. of X_i. The joint p.d.f. $f_{\mathbf{X}}(\cdot;\theta)$ is distinct from the marginal p.d.f. $f(\cdot;\theta)$, but we shall for simplicity of notation usually drop the subscript $\mathbf{X}$.

Given the vector of parameters θ, the joint p.d.f. $f(\mathbf{x};\theta)$ as a function of $\mathbf{x}$ describes the probability law according to which the values of the observations $\mathbf{X}$ vary from repetition to repetition of the sampling experiment. The study of the properties of such p.d.f.s falls within the realm of distribution theory. In the theory of statistical inference we are concerned with a different type of problem. Given a particular vector of sample values $\mathbf{x}$ observed on some particular occasion, we wish to make inferences about the unknown parameters θ.

- Inference is an example of the standard scientific procedure of *inductive reasoning*, where we attempt to formulate general conclusions from a specific set of data. Probability and distribution theory is concerned with *deductive reasoning*: starting from a set of axioms that are assumed to hold in general, we predict specific consequences.

A fundamental role in the theory of statistical inference is played by what is known as the likelihood function. Given a particular vector of observed values $\mathbf{x}$, where $\mathbf{x} \in \mathcal{X}$, the *likelihood function* $L(\theta;\mathbf{x})$ is the function $f(\mathbf{x};\theta)$ considered as a function of θ. Thus, given $\mathbf{X} = \mathbf{x}$,

$$L(\theta;\mathbf{x}) = f(\mathbf{x};\theta), \qquad \theta \in \Theta.$$

With this inversion of the roles of $\mathbf{x}$ and θ, we may think of the likelihood function as an expression of the relative likelihood of the various possible values of θ as having given rise to the observed $\mathbf{x}$. Given our statistical model, we are comparing how good an explanation the different values of θ provide for the observed value of $\mathbf{x}$. We avoid use of the term "probability" with regard to θ, because in most approaches to problems of inference we think of θ not as a random variable but as fixed, albeit unknown. Note that, although $f(\mathbf{x};\theta)$ as a function of $\mathbf{x}$ is a p.d.f., $L(\theta;\mathbf{x})$ as a function of θ will in general not be a p.d.f.

There will be some arbitrariness in when we use the notation $L(\theta;\mathbf{x})$ and when $f(\mathbf{x};\theta)$. We shall use the former when we particularly wish to emphasize that we are considering the likelihood function, given an observed value of $\mathbf{x}$, rather than the p.d.f. It will often be convenient to use the *log likelihood function*, $\ln L(\theta;\mathbf{x})$.

Example 1

Consider a random sample $X_1, X_2, \ldots, X_n$ of size n from a $N(\mu,\sigma^2)$ distribution. In this case, $\mathcal{X} = \mathbb{R}^n$, $\theta = (\mu,\sigma^2)$ and $\Theta = \mathbb{R} \times (0,\infty)$. The joint p.d.f. is given by

$$\begin{aligned} f(\mathbf{x};\mu,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(s_x^2 + n(\bar{x}-\mu)^2)\right] \end{aligned}$$

where $\bar{x}$ is the sample mean and s_x^2 the sample corrected sum of squares,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus, given $\mathbf{x}$, we may write the likelihood function as

$$L(\mu, \sigma^2; \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(s_x^2 + n(\bar{x} - \mu)^2)\right) \tag{1}$$

and the log likelihood function as

$$\ln L(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(s_x^2 + n(\bar{x} - \mu)^2). \tag{2}$$

1.2 Sufficient statistics

Definition: A *statistic* $T \equiv T(\mathbf{X})$ is a real-valued or vector-valued function of the sample observations.

Mathematically speaking, a statistic T is a random variable or a random vector. Statistically speaking, a statistic may be regarded as a summary or reduction of the sample data. A statistic is a characteristic of the sample (note that it must not depend upon the parameter values), whereas a parameter is a characteristic of the population from which the sample is drawn. Given a particular vector of observed sample values $\mathbf{x}$, the corresponding value of the statistic will be denoted by $t \equiv T(\mathbf{x})$. For example, the sample mean and corrected sum of squares, considered either separately or jointly, are statistics.

Definition: A statistic T is said to be a *sufficient statistic* for θ if the conditional distribution of $\mathbf{X}$ given T does not depend upon θ.

The idea of a sufficient statistic is that it captures all the information about θ contained in the sample. Given the value of T, knowledge of $\mathbf{X}$ provides no further information about θ. We shall see later that this interpretation really does correspond to the formal definition above.

We give two results which are useful in checking whether a statistic is sufficient. They hold for both continuous and discrete distributions, but we shall give proofs for the discrete case only. The first result, expressed in Theorem 1, may be used in its own right to establish whether a statistic is sufficient, but it is more important as a stepping stone to Theorem 2 (The Factorization Theorem), which provides the standard method of checking for sufficiency.

Note first that in the discrete case the p.m.f. $q(t;\theta)$ of $T \equiv T(\mathbf{X})$ is given by

$$\begin{aligned} q(t;\theta) &= \mathbb{P}(T = t;\theta) \\ &= \sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} f(\mathbf{x};\theta). \end{aligned} \tag{3}$$

Theorem 1 *Let $f(\mathbf{x};\theta)$ denote the p.d.f. of a random vector $\mathbf{X}$ and $q(\cdot;\theta)$ the p.d.f. of the statistic $T(\mathbf{X})$. Then T is a sufficient statistic for θ if and only if for all $\mathbf{x} \in \mathcal{X}$ the ratio*

$$\frac{f(\mathbf{x};\theta)}{q(T(\mathbf{x});\theta)}$$

does not depend upon θ.

Proof. From the definition of conditional probability,

$$\begin{aligned}
\mathbb{P}(\mathbf{X}=\mathbf{x}|T=T(\mathbf{x});\theta) &= \frac{\mathbb{P}(\mathbf{X}=\mathbf{x}, T=T(\mathbf{x});\theta)}{\mathbb{P}(T=T(\mathbf{x});\theta)} \\
&= \frac{\mathbb{P}(\mathbf{X}=\mathbf{x};\theta)}{\mathbb{P}(T=T(\mathbf{x});\theta)} \\
&= \frac{f(\mathbf{x};\theta)}{q(T(\mathbf{x});\theta)}.
\end{aligned}$$

Hence

$$\mathbb{P}(\mathbf{X}=\mathbf{x}|T=t;\theta) = \begin{cases} f(\mathbf{x};\theta)/q(T(\mathbf{x});\theta) & \text{if } t = T(\mathbf{x}) \\ \\ 0 & \text{if } t \neq T(\mathbf{x}) \end{cases}. \tag{4}$$

From the definition of sufficiency, T is sufficient for θ if and only if $\mathbb{P}(\mathbf{X}=\mathbf{x}|T=t;\theta)$ does not depend on θ. Thus the result of the theorem follows immediately from Equation (4).

Example 2

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli r.v.s with parameter θ such that $\theta \in (0,1)$, i.e., for $i = 1, \ldots, n$,

$$\begin{aligned}
\mathbb{P}(X_i = 1) &= \theta, \\
\\
\mathbb{P}(X_i = 0) &= 1-\theta.
\end{aligned}$$

We may write

$$f(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \qquad x_i = 0, 1.$$

These Bernoulli r.v.s may be regarded as corresponding to n independent trials, where the probability of "success" at each trial is θ and the probability of "failure" is $1-\theta$; $X_i = 1$ if and only if there is a success at the ith trial.

Define the statistic T by

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

Thus T represents the total number of successes in the n trials and has the binomial $\text{Bin}(n,\theta)$ distribution. Writing $t = \sum_i x_i$,

$$f(\mathbf{x};\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t}.$$

Hence

$$\begin{aligned}\frac{f(\mathbf{x};\theta)}{q(T(\mathbf{x});\theta)} &= \frac{\theta^t(1-\theta)^{n-t}}{\left(\begin{array}{c} n \\ t \end{array}\right)\theta^t(1-\theta)^{n-t}} \\ &= \left(\begin{array}{c} n \\ t \end{array}\right)^{-1}.\end{aligned}$$

This expression does not depend upon θ. Hence, applying Theorem 1, T is sufficient for θ. Knowledge of the total number of successes in the n trials is sufficient for making inferences about θ. Any further information regarding the outcomes of the individual trials is irrelevant.

Theorem 2 (The Factorization Theorem) *Let $f(\mathbf{x};\theta)$ denote the p.d.f. of a random vector $\mathbf{X}$. A statistic T is sufficient for θ if and only if there exist functions $g(t;\theta)$ and $h(\mathbf{x})$ such that*

$$f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x}) \qquad \mathbf{x}\in\mathcal{X}, \theta\in\Theta. \tag{5}$$

Proof. Firstly, suppose that T is a sufficient statistic. Since by Theorem 1 the ratio $f(\mathbf{x};\theta)/q(T(\mathbf{x});\theta)$ does not depend upon θ, we may write it as $h(\mathbf{x})$, so that

$$\frac{f(\mathbf{x};\theta)}{q(T(\mathbf{x});\theta)} = h(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{X}, \theta\in\Theta. \tag{6}$$

Writing $q(t;\theta) = g(t;\theta)$ and rearranging Equation (6), we obtain the factorization (5).

Conversely, suppose that the factorization of Equation (5) holds. Writing $T(\mathbf{x}) = t$ and using Equation (3),

$$\begin{aligned}\frac{f(\mathbf{x};\theta)}{q(T(\mathbf{x});\theta)} &= \frac{g(t;\theta)h(\mathbf{x})}{\sum_{\{\mathbf{y}:T(\mathbf{y})=t\}} g(t;\theta)h(\mathbf{y})} \\ &= \frac{h(\mathbf{x})}{\sum_{\{\mathbf{y}:T(\mathbf{y})=t\}} h(\mathbf{y})}.\end{aligned}$$

Since this ratio does not depend upon θ, by Theorem 1, T is a sufficient statistic for θ.

Example 1 (continued)

For a random sample from a $N(\mu,\sigma^2)$ distribution, recalling Equation (1) and taking $T(\mathbf{x}) = (\bar{x}, s_x^2)$, we can write

$$f(\mathbf{x};\mu,\sigma^2) = g(\bar{x}, s_x^2;\mu,\sigma^2)h(\mathbf{x}),$$

where

$$g(\bar{x}, s_x^2;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(s_x^2 + n(\bar{x}-\mu)^2)\right)$$

and

$$h(\mathbf{x}) = 1.$$

Hence, by the Factorization Theorem, $(\bar{x}, s_x^2)$ is a sufficient statistic for (μ,σ^2).

- Here and in many other applications of the factorization criterion we may simply take $h(\mathbf{x}) = 1$.
- It should be noted that which statistics are sufficient depends upon the family of distributions being used to model the data.

We consider another example with a feature that requires some special care and which results in a very different sufficient statistic. We use the notation Uniform(a, b) for the (continuous) uniform distribution on the interval (a, b).

Example 3

Consider a random sample $X_1, X_2, \ldots, X_n$ of size n from a Uniform$(0, \theta)$ distribution, where $\theta > 0$. Each individual observation X_i has p.d.f. given by

$$f(x_i; \theta) = \begin{cases} \theta^{-1} & \text{for } 0 < x_i < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Here, the parameter space $\Theta = (0, \infty)$. It is the specification of the sample space $\mathcal{X}$ that requires more care. We could say $\mathcal{X} = (0, \theta)^n$, with the sample space depending on the value of the parameter θ. However, it will clarify matters with regard to the question of sufficiency if we take $\mathcal{X} = (0, \infty)^n$ and rewrite

$$f(x_i; \theta) = \theta^{-1} I_{(0,\theta)}(x_i)$$

and

$$f(\mathbf{x}; \theta) = \theta^{-n} I_{(0,\infty)} \left(\min_{1 \leq i \leq n} x_i \right) I_{(-\infty,\theta)} \left(\max_{1 \leq i \leq n} x_i \right), \tag{7}$$

where I_A denotes the indicator function of the set A,

$$I_A(\mathbf{x}) = \begin{cases} 1 & \text{for } \mathbf{x} \in A \\ \\ 0 & \text{otherwise.} \end{cases}$$

It follows from Equation (7) and the factorization theorem that a sufficient statistic T is given by

$$T(\mathbf{X}) = \max_{1 \leq i \leq n} X_i.$$

The corresponding likelihood function is given by

$$L(\theta; \mathbf{x}) = \theta^{-n} I_{[t,\infty)}(\theta) I_{(0,\infty)} \left(\min_{1 \leq i \leq n} x_i \right)$$

where $t = \max_{1 \leq i \leq n} x_i$.

1.3 The sufficiency and likelihood principles

The sufficiency principle

Given the family $\mathcal{F}$ of distributions, if T is a sufficient statistic for θ then any inference about θ should depend on the vector of observations $\mathbf{X}$ only through the value of T. In other words, if $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ are such that $T(\mathbf{x}) = T(\mathbf{y})$ then any inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

There are several different approaches to the philosophy of statistical inference, many of which (if not most) accept the sufficiency principle and the following form of the likelihood principle. These two principles are in fact more or less equivalent in a good many of scenarios.

The likelihood principle

Given the family $\mathcal{F}$ of distributions, if $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ are such that the ratio of the likelihood functions

$$\frac{L(\theta; \mathbf{x})}{L(\theta; \mathbf{y})}$$

is constant as a function of θ, for $\theta \in \Theta$ then any inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

We may note that the ratio of the likelihood functions being constant is equivalent to the difference of the log likelihoods being constant.

According to the likelihood principle, for making inferences about θ it is sufficient to know the likelihood function up to an arbitrary multiplicative constant or the log likelihood up to an arbitrary additive constant.

Theorem 3 *The likelihood principle implies the sufficiency principle.*

Proof. Assume that the likelihood principle is valid. Given the family $\mathcal{F}$ of distributions, let T be a sufficient statistic for θ and let $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ be such that $T(\mathbf{x}) = T(\mathbf{y})$. Let t denote the common value of $T(\mathbf{x})$ and $T(\mathbf{y})$. From the Factorization Theorem, there exist functions $g(\cdot)$ and $h(\cdot)$ such that

$$f(\mathbf{x}; \theta) = g(t; \theta)h(\mathbf{x}), \qquad \theta \in \Theta$$

and

$$f(\mathbf{y}; \theta) = g(t; \theta)h(\mathbf{y}), \qquad \theta \in \Theta.$$

Hence the ratio

$$\frac{L(\theta; \mathbf{x})}{L(\theta; \mathbf{y})} = \frac{f(\mathbf{x}; \theta)}{f(\mathbf{y}; \theta)} = \frac{h(\mathbf{x})}{h(\mathbf{y})}$$

is constant as a function of $\theta \in \Theta$. It follows from the likelihood principle that any inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

Thus if the likelihood principle is valid then so is the sufficiency principle.

In general, for a given family $\mathcal{F}$ of distributions, there will be an abundance of sufficient statistics available to us. We shall prefer to use those that represent the maximal possible reduction of the sample data consistent with keeping all information relevant to making inferences about θ. Such statistics are known as minimal sufficient statistics.

Definition

A sufficient statistic T is said to be a *minimal sufficient statistic* if, for any other sufficient statistic S, T is a function of S.

Minimal sufficient statistics are not unique – any one to one function of a minimal sufficient statistic is also minimal sufficient. As we have seen, in sampling from a normal distribution, $(\bar{x}, s_x^2)$ is a sufficient statistic for (μ, σ^2). In fact, it is minimal sufficient, but then, for example, so is $(\sum x_i, \sum x_i^2)$.

1.4 Exponential families of distributions

A family of univariate p.d.f.s, $\{f(x;\theta) : \theta \in \Theta\}$, is said to be a *(k-parameter) exponential family* if it can be expressed in the form

$$f(x;\theta) = h(x) \exp\left(\sum_{j=1}^{k} \eta_j(\theta) t_j(x) - A(\theta)\right). \tag{8}$$

In Equation (8) we require that $h(x) \geq 0, t_1(x), \ldots, t_k(x)$ are all real-valued functions of x and that $A(\theta), \eta_1(\theta), \ldots, \eta_k(\theta)$ are all real-valued functions of the possibly vector-valued parameter θ. Usually, although not necessarily, $k = q$, the dimension of the vector θ of parameters.

There are many results concerning exponential families that can be established and utilized. As a consequence, the theory of statistical inference from a random sample of observations is simplified appreciably when the family of distributions from which it is drawn is of the exponential type. Many of the standard families of distributions commonly used in statistics are exponential families, e.g., the normal, gamma, binomial and Poisson families.

Example 4

Consider the p.d.f. $f(x;\theta)$ of the binomial distribution $\text{Bin}(n, \theta)$ with the value of n given and the parameter $\theta \in (0, 1)$ unknown,

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

We may rewrite this as

$$\begin{aligned} f(x;\theta) &= \binom{n}{x} (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^x \\ &= \binom{n}{x} \exp\left[x \ln\left(\frac{\theta}{1-\theta}\right) + n \ln(1-\theta)\right], \end{aligned}$$

which is in the form of Equation (8) with $k = 1$,

$$h(x) = \binom{n}{x}, \qquad A(\theta) = -n \ln(1-\theta),$$

$$\eta_1(\theta) = \ln\left(\frac{\theta}{1-\theta}\right), \qquad t_1(x) = x.$$

It is sometimes convenient to reparametrize the exponential family specified by Equation (8), using the $\eta_j(\theta) \equiv \eta_j$ as parameters and writing $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$, to obtain what is known as the *canonical form*,

$$f(x;\eta) = h(x) \exp\left(\sum_{j=1}^{k} \eta_j t_j(x) - a(\eta)\right).$$

The $\eta_j,\ j = 1, \ldots, k$ are known as the *natural parameters* or *canonical parameters* for the family.

The natural parameter for the binomial $\text{Bin}(n, \theta)$ distribution of Example 4 is the *log-odds*,

$$\eta = \ln\left(\frac{\theta}{1-\theta}\right) \qquad \eta \in \mathbb{R}.$$

If a random sample $X_1, X_2, \ldots, X_n$ of size n is taken from a k-parameter exponential family with p.d.f. as given in Equation (8) then the joint p.d.f. of the sample is given by

$$\prod_{i=1}^{n} f(x_i;\theta) = \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left(\sum_{j=1}^{k} \eta_j(\theta) \sum_{i=1}^{n} t_j(x_i) - nA(\theta)\right).$$

It follows from the Factorization Theorem that the *canonical statistic*

$$\left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is sufficient for θ (or for η).