Parameters, estimation, likelihood function

A summary of concepts needed for section 6.5: maximum likelihood estimators, from earlier sections 6.1-6.4

Suppose random variables $X_1, \ldots, X_n$ from a random sample from a discrete distribution or a continuous distribution for which the p.f. or the p.d.f. is $f(x|\theta)$, where the parameter θ belongs to some parameter space Ω.

Here the parameter θ describes everything which is unknown about the p.f. or p.d.f. of the data. The parameter θ is the object of statistical inference.

The assumption that $\theta \in \Omega$, where Ω is is some set of possible values, gives the precise description "to what degree" θ is unknown: there may be limitations on θ, expressed in $\theta \in \Omega$, but otherwise θ is unknown. The set Ω is called the parameter space. Every statistical procedure (tests, confidence intervals, estimates) enables statements and conclusions about θ, but always under the "a priori" assumption that $\theta \in \Omega$. For instance, we might test a null hypothesis $\theta \in \Omega_0$, where $\Omega_0 \subset \Omega$ etc.

Example 1. Let $X_1, \ldots, X_n$ be from a random sample from a Bernoulli distribution with unknown probability of success p. Here the p.f. for one observation is (where x is either 0 or 1)

$$\Pr(X_i = x) = p^x (1-p)^{1-x}.$$

Here the unknown parameter θ is $\theta = p$ and the parameter space may be $\Omega = (0, 1)$ (the open interval) or $\Omega = [0, 1]$ (the closed interval, where $p = 0$ and $p = 1$ are included). Thus the p.f. for one observation is

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}, \text{ if } x \in \{0, 1\}.$$

The joint p.f. for $X_1, \ldots, X_n$ is

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

Let us denote $\mathbf{x} = (x_1, \ldots, x_n)$ the vector of n observations. Then the joint p.f. for n observations is

$$f(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$$

and the parameter space Ω is as above.

Example 2. Let $X_1, \ldots, X_n$ be from a random sample from a normal distribution with unknown mean μ and variance σ^2. Assume that μ is not restricted: $-\infty < \mu < \infty$ but for the variance σ^2 it is natural to assume that it is positive: $\sigma^2 > 0$. Thus our parameter is two dimensional:

$$\theta = (\mu, \sigma^2)$$

and the parameter space is $\Omega = R \times (0, \infty)$ where R is the real line, $(0, \infty)$ is the set of all positive real numbers. The p.d.f. for one observation is

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Again, let $\mathbf{x} = (x_1, \ldots, x_n)$ be the vector of n observations; then the joint p.d.f. for n observations is

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

and the parameter space $\Omega = R \times (0, \infty)$ is as before.

The likelihood function. When we have observed values $\mathbf{x} = (x_1, \ldots, x_n)$ of a sample $X_1, \ldots, X_n$, we may consider the joint p.f. or p.d.f. as a function of θ,

$$L(\theta) = f(\mathbf{x}|\theta)$$

where $\mathbf{x}$ takes the observed values. Then $\mathbf{x}$ is "no longer" random, and $L(\theta)$ is a function of θ, defined for every value $\theta \in \Omega$. This is called the likelihood function.

Parameter estimation. This is a method of inference about θ, the simplest in some way, where one only tries to get a "best guess" of the unknown $\theta \in \Omega$, based on the data $\mathbf{x}$ at hand. The method of maximum likelihood (section 6.5) is one possible approach. Another one is Bayesian estimation.

Bayesian estimation. In Bayesian inference one assumes the parameter random, according to some probability distribution on the set of possible values Ω (the prior distribution). In conjunction with the data $\mathbf{x}$, one then calculates a posterior distribution. We have dealt with several examples of Bayesian inference, and will not further discuss estimation in this context (sections 6.2-6.4). We only mention that in Example 4.7.4, p. 225 we calculated a posterior distribution of the success probability p and found the best predictor of p in it (the conditional expectation $E(P|\mathbf{x})$. This is in fact an example of a Bayesian estimator.