Part II

Statistics

1. Preliminaries

1. Introduction to Statistical Inference

General Setting Suppose that we are given n-numbers $x_1, \ldots, x_n$, which are believed to be generated from some random mechanism in the sense that they might as well have taken different values. The underlying generating mechanism may be intrinsically random like toss of a coin, but it may also be the case that we have a unknown deterministic mechanism, which we regard as random. We call such numbers *data*.

We begin statistical analysis of the data by postulating that they are realized values of n-random variables $X_1, \ldots, X_n$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. It is thus assumed that a particular outcome ω in the sample space Ω has incurred x_i's that we observe. If we define a random vector $X = (X_1, \ldots, X_n)'$, then

$$X(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_n(\omega) \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The *statistical inference* is the study of the distribution of the random vector X, or equivalently, the joint distribution of random variables $X_1, \ldots, X_n$.

The distribution of X is often assumed to belong to a certain family of distributions, which we denote by

$$\mathcal{P} = \{\mathrm{P}_\theta \,|\, \theta \in \Theta\}$$

Each member P_θ of the family $\mathcal{P}$ of distributions is indexed by $\theta \in \Theta$. The index θ is called the *parameter*, which is an element of the *parameter set* Θ. We often formulate the family in terms of the density p_θ, instead of the distribution P_θ, of X with respect to some base measure. The statistical inference with a parametric family of distributions is called the *parametric approach*. Of course, it is possible that we do not restrict ourselves to a particular family of distributions, and follow the *nonparametric approach*. Here we will consider mainly the former.

Statistical Inference The family of distributions is assumed to include as a member the true distribution, say P_{θ_0}. The study of the distribution of X therefore reduces to choose the member P_{θ_0} in the family $\mathcal{P}$. To see how such a study is usually done, we write

$$X : \Omega \ \rightarrow \ \mathcal{X}$$

where $\mathcal{X}$, which is $\mathbf{R}^n$ in most of our subsequent analyses, is sometimes referred to as the *state space*. The statistical inference is performed by constructing a function

$$\tau : \mathcal{X} \ \rightarrow \ \mathcal{T}$$

where $\Theta \subset \mathcal{T}$ in many cases. Such function τ or the corresponding random element $T = \tau(X)$ is called a *statistic*.

The statistical inference largely consists of two procedures – estimation of and hypothesis testing on θ. The problem of *estimation* is to find the true member P_{θ_0}, which is equivalent to finding the true value θ_0 of the parameter θ. It is done by a properly defined statistic τ. A statistic τ or the random element $T = \tau(X)$, if used to estimate θ, is called an *estimator*. The realized value $\tau(x)$ of T is an *estimate*. An estimator or estimate for the parameter θ is often denoted by $\hat{\theta}$. For the hypothesis testing, we partition Θ into two pieces and select one of them depending upon the value of the statistic τ. A statistic, used in this context, is called a *test statistic*.

2. Sufficiency

Let $T = \tau(X)$ be a statistic, and $\mathcal{P} = \{\mathrm{P}_\theta | \theta \in \Theta\}$ be a family of distributions of X. We define

Definition 1 *We say that T (or τ) is* sufficient *for $\mathcal{P}$ (or for θ) if the conditional distribution of X given $T = t$ does not depend on θ for all t.*

The distribution of X is unknown. It can be any member of the family $\mathcal{P}$. Therefore, the conditional distribution of X given $T = t$ would generally depend upon θ. However, if T is a sufficient statistic, it is uniquely determined irrespectively of the value of θ.

Remarks If T is sufficient, then we may write

$$p_\theta(x,t) = p(x|t)p_\theta(t)$$

Note that we omit the subscript θ in representing the conditional density $p(x|t)$, since it is independent of θ. We observe that

(a) The information of θ in the observation of X is concentrated in that of T. Usually, T is of lower dimension than X since the former is a function of the latter. Hence, the observation of T is less costly, though it includes the same amount of information on θ. Usefulness of a sufficient statistic lies in such data reduction.

(b) Once we know the value of T, then we may perform a random experiment and define a random variable $\tilde{X}$, say, such that the conditional distribution of $\tilde{X}$ given $T = t$ has density $p(x|t)$. We may design such a random experiment and a random variable, since $p(x|t)$ is independent of θ and known. Observed value $\tilde{x}$, though it is generally different from the original observation x, can obviously be regarded as an observation from the same distribution as X. We may therefore recover data in this sense from the observation of T.

Examples

(a) Suppose that $X \sim \mathbf{N}(0,\sigma^2)$ and $T = |X|$. For $T = t$, X can take only two values t and $-t$. Furthermore, since the distribution of X is symmetric about the origin, each point has conditional probability $1/2$. This is so, regardless of the value of σ^2. The statistic T is therefore sufficient. For the data recovery, we now consider a random experiment of flipping a fair coin, and a random variable $\tilde{X}$ that takes the value t if head is up and $-t$ if tail is up. The observed value of $\tilde{X}$ can be regarded as from the same distribution as X.

(b) Let X_1 and X_2 be independent Poisson(λ). We will show that $T = X_1 + X_2$ is sufficient. First, the joint density of X_1 and X_2 is

$$p_\lambda(x_1,x_2) = e^{-2\lambda}\frac{\lambda^{x_1+x_2}}{x_1!x_2!}\,\mathrm{I}\{x_1,x_2 = 0,1,\ldots\}$$

Since

$$\begin{aligned} p_\lambda(x_1, t) &= e^{-2\lambda} \frac{\lambda^t}{x_1!(t-x_1)!} \mathrm{I}\{x_1 = 0, \ldots, t, t = 0, 1, 2, \ldots\} \\ p_\lambda(t) &= e^{-2\lambda} \frac{(2\lambda)^t}{t!} \mathrm{I}\{t = 0, 1, \ldots\} \end{aligned}$$

the conditional density of X_1 given $T = t$ is given by

$$\begin{aligned} p(x_1|t) &= \frac{t!}{x_1!(t-x_1)!} \left(\frac{1}{2}\right)^t \mathrm{I}\{x_1 = 0, \ldots, t\} \\ &= \left(\begin{array}{c} t \\ x_1 \end{array} \right) \left(\frac{1}{2}\right)^{x_1} \left(\frac{1}{2}\right)^{t-x_1} \mathrm{I}\{x_1 = 0, \ldots, t\} \end{aligned}$$

so that the conditional distribution of X_1 given $T = t$ is Binomial$(t, \frac{1}{2})$. For the data recovery, consider a random experiment of t tosses of a fair coin and define random variables $\tilde{X}_1$ and $\tilde{X}_2$ as the numbers of head and tail, respectively.

(c) Let X_1 and X_2 be independent $\mathbf{N}(\mu, 1)$. Let us verify that $T = X_1 + X_2$ is sufficient. For this purpose, define $S = X_1 - X_2$. Then we can easily see that T and S are independent because the covariance of T and S are zero. Then the conditional distribution of S given T is equal to the distribution of S, and $S \sim \mathbf{N}(0, 2)$. From this we can deduce that the conditional distribution of X_1 and X_2 given T does not depend on μ.

The following Theorem provides a very convenient way of finding sufficient statistics. Let the distribution of X now be given by a family of densities

$$\mathcal{P} = \{p_\theta | \, \theta \in \Theta\}$$

Then

Theorem 1 (Factorization Theorem) *A statistic $T = \tau(X)$ is sufficient if and only if the density is factorized as*

$$p_\theta(x) = f(\tau(x), \theta) g(x)$$

Examples

(a) Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ), so that the joint density of $X_1, \ldots, X_n$ is given by

$$p_\lambda(x) = e^{-n\lambda} \frac{\lambda^{x_1 + \cdots + x_n}}{\prod_{i=1}^n x_i!}$$

We may simply write

$$\begin{aligned} \tau(x) &= \sum_{i=1}^n x_i \\ f(t, \lambda) &= e^{-n\lambda} \lambda^t \\ g(x) &= \frac{1}{\prod_{i=1}^n x_i!} \end{aligned}$$

to deduce that $\tau(x) = \sum_{i=1}^n x_i$ is sufficient for θ.

(b) Let $X_1, \ldots, X_n$ be i.i.d. $\mathbf{N}(\mu, \sigma^2)$ so that the joint density is given by

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{2\sigma^2}\mu^2\right) \end{aligned}$$

Therefore, it follows that

$$\tau(x) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)'$$

is sufficient for $\theta = (\mu, \sigma^2)'$.

(c) Let $X_1, \ldots, X_n$ be i.i.d. $\mathbf{N}(0, \sigma^2)$. It is easy to show that $\tau(x) = \sum_{i=1}^n x_i^2$ is sufficient for σ^2. However, it is not the only sufficient statistic. The statistics given by

$$\begin{aligned} \tau_1(x) &= (x_1, \ldots, x_n)' \\ \tau_2(x) &= (x_1^2, \ldots, x_n^2)' \\ \tau_3(x) &= (x_1^2 + \cdots + x_m^2, x_{m+1}^2 + \cdots + x_n^2)' \\ \tau_4(x) &= x_1^2 + \cdots + x_n^2 \end{aligned}$$

are all sufficient. For instance, if we let $T = \tau_1(X)$, then the conditional distribution of X given $T = t$ is t with probability 1. Therefore $T = \tau_1(X)$ is clearly sufficient.

Let T and S be two statistics. The last example demonstrates that

Lemma 2 *If $T = \varphi(S)$ for some function φ and T is sufficient, then S is also sufficient.*

Proof Obvious from the Factorization Theorem.

Remarks

(a) The result in the lemma also follows directly from the fact that the partition made by S is finer than that of T.

(b) If φ is many-to-one, T provides further reduction of data. Indeed, we say that a sufficient statistic T is *minimal* if it is a function of every sufficient statistic. A minimal sufficient statistic thus achieves the greatest reduction of data.

3. Exponential Families

We first introduce

Definition 2 *A family $\{\mathrm{P}_\theta | \, \theta \in \Theta\}$ of distributions is said to form an m-parameter* exponential family *if the distributions have densities of the form*

$$p_\theta(x) = \exp\left(\sum_{i=1}^{m} f_i(\theta)\tau_i(x) + g(\theta)\right) h(x)$$

The exponential families include many of the distributions that we know.

Remarks

(a) Note that for the exponential families

$$\tau(x) = (\tau_1(x), \ldots, \tau_m(x))'$$

is a sufficient statistic, which follows directly from the Factorization Theorem.

(b) If $X_1, \ldots, X_n$ are i.i.d. with density

$$p_\theta(x_i) = \exp\Big(f(\theta)\tau(x_i) + g(\theta)\Big)h(x_i)$$

then the density of $X = (X_1, \ldots, X_n)'$ is

$$p_\theta(x) = \exp\left(f(\theta)\sum_{i=1}^{n}\tau(x_i) + ng(\theta)\right) h(x_1)\cdots h(x_n)$$

from which we deduce that $\sum_{i=1}^{n} \tau(x_i)$ is sufficient.

Examples

(a) One parameter exponential family:

Poisson(λ)

$$\begin{aligned} p_\lambda(x) &= e^{-\lambda}\frac{\lambda^x}{x!} \\ &= \exp\Big(x\ln\lambda - \lambda\Big)\frac{1}{x!} \end{aligned}$$

Bernoulli(θ)

$$\begin{aligned} p_\theta(x) &= \theta^x(1-\theta)^{1-x} \\ &= \exp\left(x\ln\frac{\theta}{1-\theta} + \ln(1-\theta)\right) \end{aligned}$$

(b) Two-parameter exponential family:

$\mathbf{N}(\mu, \sigma^2)$

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \frac{\ln\sigma^2}{2}\right)\right) \end{aligned}$$

Gamma(α, β)

$$\begin{aligned} p_{\alpha,\beta}(x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{1}{\beta}x} \\ &= \exp\left((\alpha-1)\ln x - \frac{1}{\beta}x - \Big(\ln\Gamma(\alpha) + \alpha\ln\beta\Big)\right) \end{aligned}$$

4. Bayesian Approach

In Bayesian approach, unknown parameter θ is considered to be a realization of some random variable Θ. The distribution P_θ is now regarded as the conditional distribution of X given $\Theta = \theta$. Correspondingly, the density $p_\theta(x)$ of X is now written as $p(x|\theta)$. The density $p(\theta)$ of Θ, called *prior* density, reflects our subjective

belief on the true parameter value. The prior belief is modified according to the observation of x following Bayes' rule, i.e.,

$$p(\theta|x) = p(\theta)\frac{p(x|\theta)}{p(x)}$$

where $p(x) = \int p(x|\theta)p(\theta)\,d\theta$. The density $p(\theta|x)$ is called the *posterior* density.

Example Let X_i be i.i.d. Bernoulli(θ) and Θ is uniformly distributed on $[0,1]$. Then

$$\begin{aligned} p(x_i|\theta) &= \theta^{x_i}(1-\theta)^{1-x_i} \\ p(x|\theta) &= \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} \end{aligned}$$

and the posterior density is given by

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ &= \frac{1}{B\Big(\Sigma x_i + 1, n - \Sigma x_i + 1\Big)}\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}\,\mathrm{I}\{0 \le \theta \le 1\} \end{aligned}$$

which implies that the posterior distribution is $Beta\Big(\Sigma x_i + 1, n - \Sigma x_i + 1\Big)$.