

Ch8 Estimation of Parameters and Fitting of Probability Distributions

November 4, 2009

1 Introduction

Definition 1.1 *A point estimator is any function $W(X_1, \dots, X_n)$ of a sample; that is, any statistic is a point estimator.*

Note that an *estimator* is a function of the sample, while an *estimate* is the realized value of an estimator that is obtained when a sample is actually taken.

2 The Method of Moments

Let X_1, \dots, X_n be a sample from a population with pdf or pmf $f(x|\theta_1, \dots, \theta_k)$. Methods of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$\begin{aligned}m_1 &= \frac{1}{n} \sum_{i=1}^n X_i^1, & \mu'_1 &= EX^1 \\m_2 &= \frac{1}{n} \sum_{i=1}^n X_i^2, & \mu'_2 &= EX^2 \\& & \vdots & \\m_k &= \frac{1}{n} \sum_{i=1}^n X_i^k, & \mu'_k &= EX^k\end{aligned}$$

The population moment μ'_j will typically be a function of $\theta_1, \dots, \theta_k$, say $\mu'_j(\theta_1, \dots, \theta_k)$. The method of moments estimator $(\tilde{\theta}_1, \dots, \tilde{\theta}_k)$ of $(\theta_1, \dots, \theta_k)$ is obtained by solving the following system of

equations for $(\theta_1, \dots, \theta_k)$ in terms of (m_1, \dots, m_k) :

$$\begin{aligned} m_1 &= \mu'_1(\theta_1, \dots, \theta_k), \\ m_2 &= \mu'_2(\theta_1, \dots, \theta_k), \\ &\vdots \\ m_k &= \mu'_k(\theta_1, \dots, \theta_k). \end{aligned}$$

EX A (Normal method of moments) Suppose X_1, \dots, X_n are iid $N(\mu, \sigma^2)$. Let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. By the method of moments, we have

$$\begin{aligned} \bar{X} &= \mu, \\ \frac{1}{n} \sum X_i^2 &= \mu^2 + \sigma^2. \end{aligned}$$

Solving the systems yields the estimators

$$\tilde{\mu} = \bar{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum X_i^2 - \bar{X}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2.$$

EX B (Binomial method of moments) Let X_1, \dots, X_k be iid Binomial(k, p), that is,

$$P(X_i = x | k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \dots, k.$$

Here we assume that both k and p are unknown and we desire point estimators for both parameters. This model has been used to estimate crime rate for crimes that are known to have many unreported occurrences. By the method of moments, we have

$$\begin{aligned} \bar{X} &= kp \\ \frac{1}{n} \sum X_i^2 &= kp(1-p) + k^2 p^2 \end{aligned}$$

Solving for k and p yields the estimators

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n) \sum (X_i - \bar{X})^2} \quad \text{and} \quad \tilde{P} = \frac{\bar{X}}{\tilde{k}}.$$

Definition 2.1 Let $\tilde{\theta}_n$ be an estimate of a parameter θ based on a sample of size n . Then $\tilde{\theta}_n$ is said to be consistent in probability if $\tilde{\theta}$ converges in probability to θ as n approaches infinity; that is, for any $\epsilon > 0$,

$$P(|\tilde{\theta}_n - \theta| > \epsilon) \rightarrow 0,$$

as $n \rightarrow \infty$.

The weak law of large numbers implies that the sample moments converge in probability to the population moments. If the functions relating the estimates to the sample moments are continuous, the estimates will converge to the parameters as the sample moments converge to the population moments.

3 Maximum Likelihood Estimators

Recall that if X_1, \dots, X_n are an iid sample from a population with pdf or the frequency function $f(x|\theta_1, \dots, \theta_k)$, the likelihood function is defined by

$$L(\theta|\mathbf{x}) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\theta_1, \dots, \theta_k).$$

Definition 3.1 For each sample point \mathbf{x} , let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ , with \mathbf{x} held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample \mathbf{X} is $\hat{\theta}(\mathbf{X})$.

Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. However, there are two inherent drawbacks associated with the maximum likelihood estimation. The first problem is that of actually finding the global maximum and verifying that, indeed, a global maximum has been found. The second problem is that of numerical sensitivity. That is, how sensitive is the estimate to small changes in the data?

EX A (**Normal likelihood**) Let X_1, \dots, X_n be iid $N(\theta, 1)$, then

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

and

$$\log L(\theta|\mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2.$$

The equation $\frac{d}{d\theta} \log L(\theta|\mathbf{x}) = 0$ reduces to

$$\sum_{i=1}^n (x_i - \theta) = 0,$$

which has the solution $\hat{\theta} = \bar{x}$. To verify that \bar{x} is a global maximum of the likelihood function, we can use the following arguments. First, \bar{x} is the only zero of the first derivative. Second, verify that

$$\frac{d^2}{d\theta^2} \log L(\theta|\mathbf{x})|_{\theta=\bar{x}} < 0.$$

Thus, \bar{x} is the only extreme point in the interior and it is a maximum. To finally verify that \bar{x} is a global maximum, we must check the boundaries, $\pm\infty$. By taking limits it is easy to establish that the likelihood is 0 at $\pm\infty$. So $\hat{\theta} = \bar{x}$ is a global maximum and hence \bar{X} is the MLE.

EX B (Poisson distribution) If X follows a Poisson distribution with parameter λ , then

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

If X_1, X_2, \dots, X_n are iid, then the log likelihood is

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^n (X_i \log \lambda - \lambda - \log X_i!) \\ &= \log \lambda \sum_{i=1}^n X_i - n\lambda - \sum_{i=1}^n \log X_i! \end{aligned}$$

Setting the first derivative of the log-likelihood equal to zero, we find

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^n X_i - n = 0.$$

The MLE is then $\hat{\lambda} = \bar{X}$.

EX C (Multinomial cell probability) Suppose that X_1, \dots, X_m follow a multinomial distribution with the joint frequency function,

$$f(x_1, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}.$$

The log likelihood function is

$$l(p_1, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i.$$

To maximize this likelihood subject to the constraint, we introduce a Lagrange multiplier and maximize

$$l(p_1, \dots, p_m, \lambda) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i + \lambda \left(\sum_{i=1}^m p_i - 1 \right).$$

Setting the partial derivatives equal to zero, we have the following system of equations:

$$\hat{p}_j = -\frac{x_j}{\lambda}, \quad j = 1, \dots, m.$$

Summing both sides we have $1 = -n/\lambda$ or $\lambda = -n$. Therefore, we have

$$\hat{p}_j = \frac{x_j}{n},$$

which is an obvious set of estimates.

EX D Suppose that X_1, \dots, X_n form a random sample from a uniform distribution on the interval $[0, \theta]$ with $\theta > 0$. The joint pdf $f_n(\mathbf{x}|\theta)$ has the form

$$f_n(\mathbf{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

The MLE of θ must be a value of θ for which $\theta \geq \max\{x_1, \dots, x_n\}$ and that maximizes $1/\theta^n$ among all such values. Hence, the MLE of θ is

$$\hat{\theta} = \max\{X_1, \dots, X_n\}.$$

Limitations of Maximum Likelihood Estimation

Despite its intuitive appeal, the method of maximum likelihood is not necessarily appropriate in all problems.

- **Nonexistence of an MLE** Suppose that X_1, \dots, X_n form a random sample from a uniform distribution on the interval $(0, \theta)$. The pdf is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 < x_1, \dots, x_n < \theta \\ 0 & \text{otherwise} \end{cases}$$

It should be noted that the possible values of θ does not include the value $\max(x_1, \dots, x_n)$, because θ must be strictly greater than each observed value x_i ($i = 1, \dots, n$). Because θ can be chosen arbitrarily close to the value $\max(x_1, \dots, x_n)$ but cannot be chosen equal to this value, it follows that the MLE of θ does not exist.

- **Non-uniqueness of an MLE** Suppose that X_1, \dots, X_n form a random sample from a uniform distribution on the interval $[\theta, \theta + 1]$, where the value of the parameter θ is unknown ($-\infty < \theta < \infty$). The joint pdf $f_n(\mathbf{x}|\theta)$ is

$$f_n(\mathbf{x}|\theta) = \begin{cases} 1 & \text{for } \theta \leq x_i \leq \theta + 1, i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

This is equivalent to

$$f_n(\mathbf{x}|\theta) = \begin{cases} 1 & \text{for } \max(x_1, \dots, x_n) - 1 \leq \theta \leq \min(x_1, \dots, x_n) \\ 0 & \text{otherwise} \end{cases}$$

Thus, it is possible to select as an MLE any value of θ in the interval

$$\max(x_1, \dots, x_n) - 1 \leq \theta \leq \min(x_1, \dots, x_n).$$

In this example, the MLE is not uniquely specified.

Properties of Maximum Likelihood Estimators

Invariance Suppose that the variables X_1, \dots, X_n form a random sample from a distribution for which either the pf or the pdf is $f(x|\theta)$. The parameter θ is unknown, and it may be one-dimensional or a vector of parameters. Let Ω be the space of θ . Let $g(\theta)$ be an arbitrary function of the parameter, and let G be the image of Ω under the function g . For each $t \in G$, define

$$G_t = \{\theta : g(\theta) = t\}.$$

Define the MLE of $g(\theta)$ to be \hat{t} where

$$L^*(\hat{t}) = \max_{t \in G} L^*(t) = \max_{t \in G} \max_{\theta \in G_t} \log f_n(\mathbf{x}|\theta).$$

Theorem 3.1 *Let $\hat{\theta}$ be an MLE of θ and let $g(\theta)$ be a function of θ . Then an MLE of $g(\theta)$ is $g(\hat{\theta})$.*

PROOF: Since $L^*(t)$ is the maximum of $\log f_n(\mathbf{x}|\theta)$ over θ in a subset of Ω , we know

$$L^*(t) \leq \log f_n(\mathbf{x}|\hat{\theta})$$

for all $t \in G$. Let $\hat{t} = g(\hat{\theta})$. Note that $\hat{\theta} \in G_{\hat{t}}$. Since $\hat{\theta}$ maximizes $f_n(\mathbf{x}|\theta)$ over all θ , it also maximizes $\log f_n(\mathbf{x}|\theta)$ over $\theta \in G_{\hat{t}}$. Hence,

$$L^*(\hat{t}) = \log f_n(\mathbf{x}|\hat{\theta})$$

and $\hat{t} = g(\hat{\theta})$ is an MLE of $g(\theta)$. \square

EX E Suppose that the variables X_1, \dots, X_n form a random sample from a normal distribution for which both μ and σ^2 are unknown. We have found that the MLEs of μ and σ^2 are

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

From the invariance property, we can conclude that the MLE of σ is

$$\hat{\sigma} = \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

The MLE of $E(X^2)$ is

$$\bar{X}_n^2 + \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Consistency Under mild conditions, which are typically satisfied in practical problems, the sequence of MLE's is a consistent sequence of estimators of θ . In other words, in most problems the sequence of MLE's converges in probability to the unknown value of θ as $n \rightarrow \infty$.

Numerical Computation In many problems there exists a unique MLE, but this MLE can not be expressed as an explicit algebraic function of the observations in the sample. In this case, $\hat{\theta}$ can be determined by numerical computation.

EX F (Gamma distribution) Since the density function of a gamma distribution is

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad 0 \leq x < \infty,$$

the log likelihood function of an iid sample, X_1, \dots, X_n , is

$$\begin{aligned} l(\alpha, \lambda) &= \sum_{i=1}^n [\alpha \log \lambda + (\alpha - 1) \log X_i - \lambda X_i - \log \Gamma(\alpha)] \\ &= n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^n \log X_i - \lambda \sum_{i=1}^n X_i - n \log \Gamma(\alpha) \end{aligned}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= n \log \lambda + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \frac{\partial l}{\partial \lambda} &= \frac{n\alpha}{\lambda} - \sum_{i=1}^n X_i \end{aligned}$$

Setting the second partial equation to zero, we find

$$\hat{\lambda} = \frac{n\hat{\alpha}}{\sum_{i=1}^n X_i} = \frac{\hat{\alpha}}{\bar{X}}.$$

Substituting it into the first partial equation, we obtain

$$n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0.$$

This equation can be solved by a numerical method.

4 The Bayesian Approach to Parameter Estimation

In the classical approach the parameter, θ , is thought to be an unknown, but fixed, quantity. A random sample, X_1, \dots, X_n , is drawn from a population indexed by θ and, based on the observed

values in the sample, knowledge about the value of θ is obtained. In the Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution (called the prior distribution). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (hence the name prior distribution). A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the posterior distribution. This updating is done with the use of Bayes' Rule, hence the name Bayesian statistics.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\mathbf{x}|\theta)$, then the posterior distribution is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})},$$

where $m(\mathbf{x})$ is the marginal distribution of \mathbf{X} , that is,

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta.$$

EX A (Binomial Bayes estimation) Let X_1, \dots, X_n be iid Bernoulli(p). Then $Y = \sum X_i$ is binomial (n, p). We assume the prior distribution on p is beta(α, β). The joint distribution of Y and p is

$$\begin{aligned} f(y, p) &= \left[\binom{n}{y} p^y (1-p)^{n-y} \right] \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \right] \\ &= \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1}. \end{aligned}$$

The marginal of Y is

$$f(y) = \int_0^1 f(y, p) dp = \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y + \alpha)\Gamma(n - y + \beta)}{\Gamma(n + \alpha + \beta)}.$$

The posterior is

$$f(p|y) = \frac{f(y, p)}{f(y)} = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(y + \alpha)\Gamma(n - y + \beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1},$$

which is $beta(y+\alpha, n-y+\beta)$. A natural estimate for p is the mean of the posterior distribution, which would give us the Bayes estimator of p ,

$$\hat{p}_B = \frac{y + \alpha}{\alpha + \beta + n}.$$

EX B (Normal Bayes estimators) Let $X \sim N(\theta, \sigma^2)$, and suppose that the prior on θ is $N(\mu, \tau^2)$.

The posterior distribution of θ is also normal, with mean variance by

$$E(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu,$$

and

$$\text{Var}(\theta|x) = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

5 Conjugate Prior Distributions

For each of the most popular statistical models, there exists a family of distributions for the parameter with a very special property. If the prior distribution is chosen to be a member of that family, then the posterior distribution will also be a member of that family. Such a family of distributions is called a conjugate family. Choosing a prior distribution from a conjugate family will typically make calculation of the posterior distribution particularly simple.

Definition 5.1 *let \mathcal{F} denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a conjugate family for \mathcal{F} if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π , and all $x \in \mathcal{X}$.*

Sampling from a Bernoulli Distribution

Theorem 5.1 *Suppose that X_1, \dots, X_n form a random sample from a Bernoulli distribution for which the value of the parameter θ is unknown ($0 < \theta < 1$). Suppose also that the prior distribution of θ is a beta distribution with given parameters α and β ($\alpha > 0$ and $\beta > 0$). Then the posterior distribution of θ given that $X_i = x_i$ ($i = 1, \dots, n$) is a beta distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$.*

PROOF: Let $y = \sum_{i=1}^n x_i$. The likelihood function of \mathbf{x} is

$$f_n(\mathbf{x}|\theta) = \theta^y(1 - \theta)^{n-y}.$$

The prior pdf satisfies the following relation

$$\xi(\theta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1},$$

for $0 < \theta < 1$. It follows that the posterior of θ is

$$\xi(\theta|\mathbf{x}) \propto \theta^{\alpha+y-1}(1 - \theta)^{\beta+n-y-1},$$

for $0 < \theta < 1$; that is, a beta distribution with parameters $\alpha + y$ and $\beta + n - y$. \square

The family of beta distributions is called a conjugate family of prior distributions for samples from a Bernoulli distribution. Any parameters of the family of prior distributions (such as α and β in the above theorem) are called *prior hyperparameters* in order to distinguish them from parameters like θ . The corresponding parameters of the posterior distributions ($\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$ in the above theorem) are called posterior hyperparameters.

EX A Suppose that the proportion θ of defective items in a large shipment is unknown, the prior distribution of θ is a uniform distribution on the interval $[0, 1]$, and items are to be selected at random from the shipment and inspected until the variance of the posterior distribution of θ has been reduced to the value 0.01 or less.

The uniform distribution in the interval $[0, 1]$ is a beta distribution for which $\alpha = 1$ and $\beta = 1$. Therefore, after y defective items and z nondefective items have been obtained, the posterior distribution of θ will be a beta distribution with $\alpha = y + 1$ and $\beta = z + 1$. Therefore, the variance of the posterior distribution of θ will be

$$V = \frac{(y + 1)(z + 1)}{(y + z + 2)^2(y + z + 3)}.$$

Sampling is to stop as soon as $V \leq 0.01$.

Sampling from a Poisson Distribution

Theorem 5.2 *Suppose that X_1, \dots, X_n form a random sample from a Poisson distribution for which the value of the mean θ is unknown ($\theta > 0$). Suppose also that the prior distribution of θ is a gamma distribution with given parameters α and β ($\alpha > 0$ and $\beta > 0$). Then the posterior distribution of θ , given that $X_i = x_i$ ($i = 1, \dots, n$), is a gamma distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n$.*

PROOF: Let $y = \sum_{i=1}^n x_i$. The likelihood function $f_n(\mathbf{x}|\theta)$ satisfies the relation

$$f_n(\mathbf{x}|\theta) \propto e^{-n\theta} \theta^y.$$

The prior pdf of θ has the form

$$\xi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}, \quad \text{for } \theta > 0.$$

Since the posterior pdf $\xi(\theta|\mathbf{x})$ has the form

$$\xi(\theta|\mathbf{x}) \propto \theta^{\alpha+y-1} e^{-(\beta+n)\theta}, \quad \text{for } \theta > 0,$$

that is, the posterior of θ is a gamma distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n$. \square

EX B Consider a Poisson distribution for which the mean θ is unknown, and suppose that the prior pdf of θ is as follows:

$$\xi(\theta) = \begin{cases} 2e^{-2\theta} & \text{for } \theta > 0 \\ 0 & \text{for } \theta \leq 0 \end{cases}$$

This prior pdf is the pdf of a gamma distribution with parameters $\alpha = 1$ and $\beta = 2$. Therefore, the posterior of θ is a gamma distribution with parameters $y + 1$ and $n + 2$. The variance of the posterior distribution is

$$V = \frac{y + 1}{(n + 2)^2}.$$

Sampling from a Normal Distribution

Theorem 5.3 *Suppose that X_1, \dots, X_n form a random sample from a normal distribution for which the value of the mean θ is unknown ($-\infty < \theta < \infty$) and the value of the variance σ^2 is known ($\sigma^2 > 0$). Suppose also that the prior distribution of θ is a normal distribution with given values of the mean μ and the variance ν^2 . Then the posterior distribution of θ given that $X_i = x_i$ ($i = 1, \dots, n$) is a normal distribution for which the mean μ_1 and variance ν_1^2 are as follows:*

$$\mu_1 = \frac{\sigma^2 \mu + n\nu^2 \bar{x}_n}{\sigma^2 + n\nu^2}$$

and

$$\nu_1^2 = \frac{\sigma^2 \nu^2}{\sigma^2 + n\nu^2}.$$

PROOF: The likelihood function $f_n(\mathbf{x}|\theta)$ has the form

$$f_n(\mathbf{x}|\theta) \propto \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right].$$

The prior pdf $\xi(\theta)$ has the form

$$\xi(\theta) \propto \exp \left[-\frac{1}{2\nu^2} (\theta - \mu)^2 \right].$$

It follows that the posterior pdf $\xi(\theta|\mathbf{x})$ has the form

$$\begin{aligned} \xi(\theta|\mathbf{x}) &\propto \exp \left\{ -\frac{1}{2} \left[\frac{n}{\sigma^2} (\theta - \bar{x}_n)^2 + \frac{1}{\nu^2} (\theta - \mu)^2 \right] \right\} \\ &\propto \exp \left[-\frac{1}{2\nu_1^2} (\theta - \mu_1)^2 \right]. \end{aligned}$$

The proof is completed. \square

The mean μ_1 of the posterior distribution of θ can be rewritten as

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + n\nu^2}\mu + \frac{n\nu^2}{\sigma^2 + n\nu^2}\bar{x}_n,$$

that is, μ_1 is a weighted average of the mean μ of the prior distribution and the sample mean \bar{x}_n .

EX C Suppose that observations are to be taken at random from a normal distribution for which the value of the mean θ is unknown and the variance is 1, and that the prior distribution of θ is a normal distribution for which the variance is 4.

It follows from Theorem 5.3 that after n observations have been taken, the variance ν_1^2 of the posterior distribution of θ will be

$$\nu_1^2 = \frac{4}{4n + 1}.$$

Sampling from an Exponential Distribution

Theorem 5.4 *Suppose that X_1, \dots, X_n from a random sample from an exponential distribution for which the value of the parameter θ is unknown ($\theta > 0$). Suppose also that the prior distribution of θ is a gamma distribution with given parameters α and β ($\alpha > 0$ and $\beta > 0$). Then the posterior distribution of θ given that $X_i = x_i$ ($i = 1, \dots, n$) is a gamma distribution with parameters $\alpha + n$ and $\beta + \sum_{i=1}^n x_i$.*

PROOF: Let $y = \sum_{i=1}^n x_i$. Then the likelihood function $f_n(\mathbf{x}|\theta)$ is

$$f_n(\mathbf{x}|\theta) = \theta^n e^{-\theta y}.$$

The prior pdf $\xi(\theta)$ has the form

$$\xi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}, \quad \text{for } \theta > 0$$

It follows, therefore, that the posterior pdf $\xi(\theta|\mathbf{x})$ has the form

$$\xi(\theta|\mathbf{x}) \propto \theta^{\alpha+n-1} e^{-(\beta+y)\theta}, \quad \text{for } \theta > 0,$$

that is, a gamma distribution with parameters $\alpha + n$ and $\beta + y$. \square

Improper Prior Distributions

EX A Bortkiewicz (1898) counted the numbers of Prussian soldiers killed by horsekick in 14 army units for each of 20 years, a total of 280 counts. Of the 280 counts, 144 of them were 0, 91

of them were 1, 32 of them were 2, 11 counts were 3, and 2 counts were 4. Suppose that we were going to model the 280 counts as a random sample of Poisson random variables X_1, \dots, X_{280} with mean θ . A conjugate prior would be a member of the gamma family with prior hyperparameters α and β . Theorem 5.2 says that the posterior distribution of θ would be a gamma distribution with the posterior hyperparameters $\alpha + 196$ and $\beta + 280$, since the sum of the 280 counts equals 196. If we let $\alpha = 0$ and $\beta = 0$, we get the improper prior pdf $\xi(\theta) = \theta^{-1}$ for $\theta > 0$. Pretending as if this really were a prior pdf, the resulting posterior is a gamma distribution with parameters 196 and 280.

EX B Suppose X_1, \dots, X_{23} are iid samples drawn from a normal distribution with parameter θ and variance 0.25. A conjugate prior for θ would be a normal distribution with mean μ and variance ν^2 . Suppose the average of the 23 samples is 4.15, so the posterior distribution of θ would be a normal distribution with mean $\mu_1 = (0.25\mu + 23 \times 4.15\nu^2)/(0.25 + 23\nu^2)$ and variance $\nu_1^2 = (0.25\nu^2)/(0.25 + 23\nu^2)$. If we let $\nu^2 \rightarrow \infty$, we get $\mu_1 \rightarrow 4.15$ and $\nu_1^2 \rightarrow 0.25/23$. Having infinite variance for the prior distribution of θ is like saying that θ is equally likely to be anywhere on the real line. If we use an improper “normal distribution” prior with variance ∞ , the calculation in Theorem 5.3 would yield a posterior distribution that is a normal distribution with mean \bar{x}_n and variance σ^2/n . The improper prior pdf in this case is $\xi(\theta) = 1$.

The Bayes Estimate for Large Samples

Suppose that the proportion θ of defective items in a large shipment is unknown, and that the prior distribution of θ is a uniform distribution on the interval $[0, 1]$. Suppose also that the value of θ must be estimated, and that the squared error loss function is used. Suppose, finally, that in a random sample of 100 items from the shipment, exactly 10 items are found to be defective. Since the uniform distribution is a beta distribution with parameters $\alpha = 1$ and $\beta = 1$, and since $n = 100$ and $y = 10$ for the given sample, the Bayes estimate is $\delta^*(\mathbf{x}) = 11/102 = 0.108$.

Next, suppose the prior of θ is Beta(1,2) with the pdf

$$\xi(\theta) \propto 2(1 - \theta), \quad 0 < \theta < 1.$$

The resulting Bayes estimate is $\delta^*(\mathbf{x}) = 11/103 = 0.107$.

The two prior distributions are quite different. Nevertheless, because the number of observations in the sample is so large ($n = 100$), the Bayes estimates with respect to the two different priors

are almost the same. Furthermore, the values of both estimates are very close to the observed proportion of defective items in the sample, which is $\bar{x}_n = 0.1$.

Consistency of the Bayes Estimator

A sequence of estimators that converges to the unknown value of the parameter being estimated, as $n \rightarrow \infty$, is called a consistent sequence of estimators.

The practical interpretation of this result is as follows: When large number of observations are taken, there is high probability that the Bayes estimator will be very close to the unknown value of θ .

For example, let X_1, \dots, X_n be a random sample (given θ) from a Bernoulli distribution with parameter θ . Suppose that we use a conjugate prior for θ . It follows from the law of large numbers (Section 4.8) that \bar{X}_n converges in probability to θ as $n \rightarrow \infty$. Since

$$\delta^*(\mathbf{X}) - \bar{X}_n \rightarrow 0, \quad \text{in probability}$$

as $n \rightarrow \infty$, it can also be concluded that $\delta^*(\mathbf{X})$ converges in probability to the unknown value of θ as $n \rightarrow \infty$.

6 Sufficient Statistics

Definition of a Statistic

Let X_1, \dots, X_n be a random sample drawn from a distribution for which the pf or pdf is $f(x|\theta)$. Any real-valued function $T(X_1, \dots, X_n)$ of the observations in the random sample is called a statistic. Three examples of statistic are the sample mean \bar{X}_n ; the maximum Y_n of the values X_1, \dots, X_n ; and the function $T(X_1, \dots, X_n)$, which has the constant value 3 for all values of X_1, \dots, X_n .

In an estimation problem, we can say that an estimator of θ is a statistic whose value can be regarded as an estimate of the value of θ .

Definition of a Sufficient Statistic

Definition 6.1 *A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of the sample \mathbf{X} given the value of $T(\mathbf{X})$ does not depend on θ .*

To use this definition to verify that a statistic $T(\mathbf{X})$ is a sufficient statistic for θ , we must verify

that $f(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ does not depend on θ . Since $\{\mathbf{X} = \mathbf{x}\}$ is a subset of $\{T(\mathbf{X}) = T(\mathbf{x})\}$,

$$\begin{aligned} f(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) &= \frac{f(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{f(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{f(\mathbf{X} = \mathbf{x})}{f(T(\mathbf{X}) = T(\mathbf{x}))} = \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}, \end{aligned}$$

where $f(\mathbf{x}|\theta)$ is the joint pf/pdf of the sample \mathbf{X} and $q(t|\theta)$ is the pmf/pdf of $T(\mathbf{X})$. Thus, $T(\mathbf{X})$ is a sufficient statistic for θ if and only if, for every \mathbf{x} , the above ratio is constant as a function of θ .

Theorem 6.1 *If $f(\mathbf{x}|\theta)$ is the joint pdf or pmf of \mathbf{X} and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every \mathbf{x} in the sample space, the ratio $f(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant as a function of θ .*

EX A (Binomial sufficient statistic) Let X_1, \dots, X_n be iid Bernoulli random variables with parameter θ , $0 < \theta < 1$. Then $T(\mathbf{X}) = X_1 + \dots + X_n$ is a sufficient statistic for θ . Note that $T(\mathbf{X})$ counts the number of X_i 's that equal 1, so $T(\mathbf{X}) \sim \text{Bi}(n, \theta)$. The ratio of pmfs is thus

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{\prod \theta^{x_i} (1-\theta)^{1-x_i}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \quad (\text{define } t = \sum_{i=1}^n x_i) \\ &= \frac{\theta^{\sum x_i} (1-\theta)^{\sum(1-x_i)}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\sum x_i}} \end{aligned}$$

Since this ratio does not depend on θ , by Theorem 6.1, $T(\mathbf{x})$ is a sufficient statistic for θ .

EX B (Normal sufficient statistic) Let X_1, \dots, X_n be iid $N(\mu, \sigma^2)$, where σ^2 is known. We wish to show that the sample mean, $T(\mathbf{X}) = \bar{X}$, is a sufficient statistic for μ .

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

Recall that the sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$. Thus, the ratio of pdf is

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

which does not depend on μ . By Theorem 6.1, the sample mean is a sufficient statistic for μ .

Theorem 6.2 (Factorization Theorem) Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of a sample \mathbf{X} . A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points \mathbf{x} and all parameter points θ ,

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}). \quad (1)$$

PROOF: We give the proof only for discrete distributions. Suppose $T(\mathbf{X})$ is a sufficient statistic. Choose $g(t|\theta) = P_\theta(T(\mathbf{X}) = t)$ and $h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$. Because $T(\mathbf{X})$ is sufficient, the conditional probability $h(\mathbf{x})$ does not depend on θ . Thus,

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}) \end{aligned}$$

So factorization (1) has been exhibited. Also, the last two lines above imply that $g(T(\mathbf{x})|\theta)$ is the pmf of $T(\mathbf{X})$.

Now assume the factorization (1) exists. Let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. Define $A_{T(\mathbf{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$. Then

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} \quad (\text{density transformation}) \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta) \sum_{A_{T(\mathbf{x})}} h(\mathbf{y})} \quad (\text{since } T \text{ is a constant on } A_{T(\mathbf{x})}) \\ &= \frac{h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} h(\mathbf{y})} \end{aligned}$$

Since the ratio does not depend on θ , by Theorem 6.1, $T(\mathbf{X})$ is a sufficient statistic for θ . \square

To use the Factorization Theorem to find a sufficient statistic, we factor the joint pdf of the sample into two parts. One part does not depend on θ and it constitutes the $h(\mathbf{x})$ function. The other part depends on θ , and usually it depends on the sample \mathbf{x} only through some function $T(\mathbf{x})$ and this function is a sufficient statistic for θ .

EX A (Normal sufficient statistic) Let X_1, \dots, X_n be iid $N(\mu, \sigma^2)$, where σ^2 is known. The pdf can be factored as

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

We can define

$$g(t|\theta) = \exp\{-n(t - \mu)^2/(2\sigma^2)\}$$

by defining $T(\mathbf{x}) = \bar{x}$, and

$$h(\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\sum_{i=1}^n (x_i - \bar{x})^2/(2\sigma^2)\right\}.$$

Thus, by the Factorization Theorem, $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ .

EX B Sampling from a Poisson Distribution Suppose that X_1, \dots, X_n form a random sample from a Poisson distribution for which the value of the mean θ is unknown. We shall show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ .

The pf can be factorized as follows:

$$f_n(\mathbf{x}|\theta) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{1}{\prod_{i=1}^n x_i!} e^{-n\theta} \theta^{\sum_{i=1}^n x_i}.$$

It follows that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ .

EX C Sampling from a Uniform Distribution Suppose that X_1, \dots, X_n form a random sample from a uniform distribution on the interval $[0, \theta]$, where θ is unknown. We shall show that $T(\mathbf{X}) = \max(X_1, \dots, X_n)$ is a sufficient statistic for θ .

The joint pdf of X_1, \dots, X_n is

$$f_n(\mathbf{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \leq x_i \leq \theta, i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

We note that $f_n(\mathbf{x}|\theta)$ can be written as

$$f_n(\mathbf{x}|\theta) = \frac{1}{\theta^n} I[\max(x_1, \dots, x_n) \leq \theta],$$

where $I[\cdot]$ be an indicator function. It follows that $T(\mathbf{X}) = \max(X_1, \dots, X_n)$ is a sufficient statistic for θ .

Jointly Sufficient Statistics

In almost every problem in which θ is a vector, as well as in many problems in which θ is one-dimensional, there does not exist a single statistic T that is sufficient. In such a problem it is necessary to find two or more statistics T_1, \dots, T_k that together are jointly sufficient statistics.

Suppose that for each possible value (t_1, \dots, t_k) of (T_1, \dots, T_k) , the conditional joint distribution of (X_1, \dots, X_n) given $(T_1, \dots, T_k) = (t_1, \dots, t_k)$ does not depend on θ . Then T_1, \dots, T_k are called *jointly sufficient statistics* for θ .

Theorem 6.3 Factorization criterion for jointly sufficient statistics. *The statistics T_1, \dots, T_k are jointly sufficient statistics for θ if and only if the joint pdf or joint pf $f_n(\mathbf{x}|\theta)$ can be factored as follows for all values of $\mathbf{x} \in \mathbb{R}^n$ and all values of $\theta \in \Omega$:*

$$f_n(\mathbf{x}|\theta) = u(\mathbf{x})v[T_1(\mathbf{x}), \dots, T_k(\mathbf{x}), \theta].$$

Here the functions u and v are nonnegative, the function u may depend on \mathbf{x} but does not depend on θ , and the function v will depend on θ but depends on \mathbf{x} only through the k statistics $T_1(\mathbf{x}), \dots, T_k(\mathbf{x})$.

EX A (Normal sufficient statistic, both parameters unknown) Assume that X_1, \dots, X_n are iid $N(\mu, \sigma^2)$, and that both μ and σ^2 are unknown so the parameter vector is $\theta = (\mu, \sigma^2)$. Let $T_1(\mathbf{x}) = \bar{x}_n$ and $T_2(\mathbf{x}) = s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$.

$$\begin{aligned} f_n(\mathbf{x}|\theta) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]/(2\sigma^2)\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-(n(t_1 - \mu)^2 + (n-1)t_2)/(2\sigma^2)\right\}. \end{aligned}$$

Let $u(\mathbf{x}) = 1$. By the factorization theorem, $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\bar{X}, S^2)$ is a sufficient statistic for (μ, σ^2) .

Suppose now that in a given problem the statistics T_1, \dots, T_k are jointly sufficient statistics for some parameter vector θ . If k other statistics T'_1, \dots, T'_k are obtained from T_1, \dots, T_k by a one-to-one transformation, then T'_1, \dots, T'_k are also jointly sufficient statistics for θ .

EX B Another pair of jointly sufficient statistic for the parameters of a normal distribution Assume that X_1, \dots, X_n are iid $N(\mu, \sigma^2)$, and that both μ and σ^2 are unknown so the parameter vector is $\theta = (\mu, \sigma^2)$. Let $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$ and $T_2(\mathbf{x}) = \sum_{i=1}^n X_i^2$. By the factorization theorem, the statistics T_1 and T_2 are jointly sufficient statistics for μ and σ^2 .

EX C Suppose that X_1, \dots, X_n form a random sample from a uniform distribution on the interval $[a, b]$, where a and b are unknown. The joint pdf is

$$f_n(\mathbf{x}|a, b) = \begin{cases} \frac{1}{(b-a)^n} & \text{for } \min(x_1, \dots, x_n) \geq a \text{ and } \max(x_1, \dots, x_n) \leq b \\ 0 & \text{otherwise} \end{cases}$$

That is,

$$f_n(\mathbf{x}|a, b) = \frac{1}{(b-a)^n} I[\min(x_1, \dots, x_n) \geq a] I[\max(x_1, \dots, x_n) \leq b].$$

Since this expression depends on \mathbf{x} only through the values of $\min(x_1, \dots, x_n)$ and $\max(x_1, \dots, x_n)$, it follows that the statistics $T_1 = \min(x_1, \dots, x_n)$ and $T_2 = \max(x_1, \dots, x_n)$ are jointly sufficient statistics for a and b .

The Rao-Blackwell Theorem

Theorem 6.4 *If T is sufficient for θ , the maximum likelihood estimate is a function of T .*

PROOF: From the factorization theorem, the likelihood is $g(T, \theta)h(\mathbf{x})$, which depends on θ only through T . To maximize this quantity, we need only maximize $g(T, \theta)$. \square

Theorem 6.5 (*Rao-Blackwell Theorem*) *Let $\hat{\theta}$ be an estimator of θ with $E(\hat{\theta}^2) < \infty$ for all θ . Suppose that T is sufficient for θ , and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then, for all θ ,*

$$E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2.$$

The inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

PROOF: We first note that, from the property of iterated conditional expectation

$$E(\tilde{\theta}) = E[E(\hat{\theta}|T)] = E(\hat{\theta}).$$

Therefore, to compare the mean squared error of the two estimators, we need only compare their variances. Since,

$$\text{Var}(\hat{\theta}) = \text{Var}[E(\hat{\theta}|T)] + E[\text{Var}(\hat{\theta}|T)],$$

or

$$\text{Var}(\hat{\theta}) = \text{Var}(\tilde{\theta}) + E[\text{Var}(\hat{\theta}|T)],$$

$\text{Var}(\hat{\theta}) \geq \text{Var}(\tilde{\theta})$ unless $E[\text{Var}(\hat{\theta}|T)] = 0$, which is the case only if $\hat{\theta}$ is a function of T , which would imply $\hat{\theta} = \tilde{\theta}$. \square