CHAPTER 2. ESTIMATION

1 Introduction

1. Remark. Estimation is the name given to the statistical procedure by which numerical values obtained from a sample are used as an approximation for the corresponding unknown parameters for the parent population. Example are: the mean and variance of the population values (MT130); the regression parameters in a regression problem (MT230); and, the estimates of the factor effects in analysis of variance (MT230).

2. Formulation. The random variables $X_1, X_2, ..., X_n$ are called *a random sample of size n* from the population $f(x)$ if $X_1, X_2, ..., X_n$ are independent random variables and the pdf or pf of each X_i is the same function $f(x)$. Alternatively, $X_1, X_2, ..., X_n$ are called *independent and identically distributed (iid) random variables* with the common pdf or pf $f(x)$.

From Ch.1, the joint pdf or pf of $X_1, X_2, ..., X_n$ is given by

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

In particular, if the population pdf or pf is a member of a parametric family with pdf or pf $f(x, \theta)$, then the joint pdf or pf is

$$f(\mathbf{x}, \theta) = f(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

The parameter θ may be a single real number or a vector whose coordinates are all the parameters of interest in the problem under discussion. We write Θ for the set of all possible (sensible) values of θ.

Along with the sample random variables we have their observed numerical values $x_1, x_2, ..., x_n$.

3. Examples. (i) Let $X_1, X_2, ..., X_n$ be a random sample from a Binomial distribution B(1,θ). Then $\Theta = \{\theta : 0 \leq \theta \leq 1\}$.

(ii) If $X_1, X_2, ..., X_n$ are a random sample from a normal distribution N(μ,σ^2), then

$\Theta = \{\theta = (\mu, \sigma) : -\infty < \mu < +\infty, \sigma > 0\}$.

4. Notation. We simplify the writing by putting:

$\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{X} = (X_1, X_2, ..., X_n)$, so that we write

$F(x_1, x_2, ..., x_n, \theta) = F(\mathbf{x}, \theta)$ and $f(x_1, x_2, ..., x_n, \theta) = f(\mathbf{x}, \theta)$.

5. Aim. Our goal is to assign a numerical value for θ, or the coordinates of θ in the multi-parameter case, and thereby obtain information about the nature of the distribution of the values of the parent population.

As in any area of mathematical application, the first question to be asked about an estimate is "what is the accuracy?". Bearing in mind that, in our context, the estimate is based on a random sample, the question should be re-expressed in terms of what reliance

can be placed on the numerical value obtained. A measured response to this question is provided by probability calculations involving the distribution of the sample random variables. This involves the assumption that the distribution has a particular functional form which contains the unknown parameter θ.

6. Definition. Let $g(\theta)$ be a function of the parameter θ.

(i) A (point) **estimator** of $g(\theta)$ is a function $\widehat{g}(\mathbf{X})$ of the sample variables $\mathbf{X} = (X_1, X_2, ..., X_n)$.

(ii) A (point) **estimate** of $g(\theta)$ is a function $\widehat{g}(\mathbf{x})$ of the observed values $\mathbf{x} = (x_1, x_2, ..., x_n)$.

[Thus, $\widehat{g}(\mathbf{x})$ is the observed value of the statistic $\widehat{g}(\mathbf{X})$.]

7. Examples. (i) Let $X_1, X_2, ..., X_n$ be a random sample from the binomial distribution B$(1, \theta)$ - so that $f(\mathbf{x}, \theta) = \theta^t(1-\theta)^{n-t}$, where $t = x_1 + x_2 + ... + x_n$. Then:

(a) With $g(\theta) = \theta$ (the population mean), take as an estimate $\widehat{g}(\mathbf{x}) = \widehat{\theta}(\mathbf{x}) = \overline{x} = t/n$ with corresponding estimator $\widehat{\theta}(\mathbf{X}) = \overline{X}$; and,

(b) with $g(\theta) = \theta(1-\theta)$ (the population variance), take as an estimate $\widehat{g}(\mathbf{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2$ with corresponding estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

(ii) Let $X_1, X_2, ..., X_n$ be a random sample from the normal N(μ,σ^2) distribution. Here, we put $\theta = (\mu, \sigma)$. Then:

(a) with $g(\theta) = \mu$ we may take as an estimate $\widehat{g}(\mathbf{x}) = \overline{x}$ with estimator $\overline{X}$; and,

(b) with $g(\theta) = \sigma^2$ we may take as an estimate $\widehat{g}(\mathbf{x}) = \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ with estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

8. Remark. The definition above of an estimator neither specifies any method of finding estimators nor offers any guidance on 'desirable' properties we should seek for them. Also, the definition makes no mention of any correspondence between an estimator and the parameter it is to estimate. In order to select a useful estimator from the range of possibilities allowed by the definition, the first requirement is a set of criteria by which different estimators may be compared.

9. Definition. An estimator $\widehat{g}(\mathbf{X})$ of $g(\theta)$ is **unbiased** if $\mathrm{E}(\widehat{g}(\mathbf{X})) = g(\theta)$ for all $\theta \in \Theta$. An estimator that is not unbiased is **biased**.

10. Examples. (i) All the estimators in 2.1.7 are unbiased.

(ii) Let $X_1, X_2, ..., X_n$ be a random sample from the uniform distribution U(0,θ). Then $\widehat{g}_1(\mathbf{X}) = 2\overline{X}$ and $\widehat{g}_2(\mathbf{X}) = \frac{n+1}{n}Y_n$, where $Y_n = \max\{X_1, X_2, ..., X_n\}$ are two unbiased estimators of the parameter θ (here, $g(\theta) = \theta$) - see Exercises 1. Further, also from Exercises 1,

$$\mathrm{var}(\widehat{g}_1(\mathbf{X})) = 4\mathrm{var}(\overline{X}) = \frac{\theta^2}{3n} \qquad \text{and} \qquad \mathrm{var}(\widehat{g}_2(\mathbf{X})) = \frac{(n+1)^2}{n^2}\mathrm{var}(Y_n) = \frac{\theta^2}{n(n+2)}.$$

Hence, since $n \geq 1$, $\mathrm{var}(\widehat{g}_2(\mathbf{X})) \leq \mathrm{var}(\widehat{g}_1(\mathbf{X}))$.

11. Principle. Generally, we will look for unbiased estimators and, given a choice of estimators, we prefer the estimator with smaller variance. [Recall that the variance measures the spread of the values of the random variable, so that a small variance suggests that the values are concentrated around the mean.]

12. Remark. In direct contradiction to the principle, some statisticians would assert a preference for a biased estimator, **of known bias**, over an unbiased estimator with larger variance.

2 Methods of Estimation

Two general methods by which estimators may be determined are now described. As the examples and exercises demonstrate, neither method will automatically produce an unbiased estimator. The third method of estimation - least squares estimation - is covered in MT230.

1. Method 1: The method of moments

For this method we assume that we have a simple random sample. Hence, the sample random variables $X_1, X_2, ..., X_n$ have a common distribution and, for a positive integer r, the (common) finite moment $m_r = \mathrm{E}(X_1^r)$, is assumed to exist. The method of moments is to **estimate** m_r by the corresponding sample moment $\widehat{\psi}_r = \frac{1}{n}\sum_{i=1}^n x_i^r$ with **estimator** $\widehat{\Psi}_r = \frac{1}{n}\sum_{i=1}^n X_i^r$. By definition, the moment estimators are unbiased estimators for the population moments. The **method of moments estimators** are found by equating the first r sample moments to the corresponding population moments and solving the resulting system of equations w.r.t. unknown parameters.

2. Examples. The following examples show how the method of moments may be used to obtain estimators for other population parameters, by the device of equating the moments to their expected values.

(i) Suppose that $X_1, X_2, ..., X_n$ are a random sample from $U(0, \theta)$.

(ii) Suppose that $X_1, X_2, ..., X_n$ are iid having unknown mean μ and variance σ^2.

(iii) Suppose that $X_1, X_2, ..., X_n$ are a random sample from the uniform distribution $U(a, b)$ where a and b are unknown.

3. Method 2: The Maximum Likelihood Principle

A frequently used method, available when the joint distribution $F(\mathbf{x},\theta)$ and joint p.d.f./p.f. $f(\mathbf{x},\theta)$ take a specified form, is to proceed as follows.

Define the **likelihood function** $\mathrm{L}(\theta,\mathbf{x}) = f(\mathbf{x}, \theta)$ - regarded as a function of θ for fixed values of $\mathbf{x} = (x_1, x_2, ..., x_n)$.

The principle of maximum likelihood is to choose as the maximum likelihood estimate (MLE) the value $\widehat{\theta}(x_1, x_2, ..., x_n) = \widehat{\theta}(\mathbf{x})$ such that

$$\mathrm{L}(\widehat{\theta}(\mathbf{x}), \mathbf{x}) = \sup\{\mathrm{L}(\theta, x_1, x_2, ..., x_n): \theta \in \Theta\} = \sup\{\mathrm{L}(\theta,\mathbf{x}): \theta \in \Theta\}$$

with the corresponding statistic $\widehat{\theta}(\mathbf{X})$ being the ML estimator.

4. Remarks. (i) An intuitive understanding of the principle is most readily seen in the discrete case, with an approximation argument extending the idea to the continuous

variable case. For discrete sample variables the likelihood function is simply Pr $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$, and, with $x_1, x_2, ..., x_n$ fixed, the principle asserts that $\widehat{\theta}(\mathbf{x})$ should be taken to maximize the probability of obtaining the sample values $x_1, x_2, ..., x_n$ which actually are obtained.

(ii) In most cases, especially when differentiation is to be used, it is convenient to work with the natural logarithm of L(θ,**x**), that is with the function log L(θ,**x**) known as the **log likelihood**. This is possible because the log function is strictly increasing on $(0, \infty)$, which implies that the extrema of L(θ,**x**) and log L(θ,**x**) coincide.

(iii) If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} \log L(\theta, \mathbf{x}) = 0, \quad i = 1, \ldots, k.$$

Note the the solutions of the above equations are only *possible candidates* for the MLE since the first derivative being zero is only a necessary condition for a maximum, not a sufficient condition. Furthermore, the zeroes of the first derivatives only locate extreme points of a function in the interior of the domain of a function. If the extrema occur on the boundary, the first derivative may not be zero.

5. Example. Suppose that the sample random variables $X_1, X_2, ..., X_n$ are a random sample from the binomial B(1,θ) distribution.

6. Example. Suppose that the sample variables are a random sample from a population distributed N(μ, σ^2) where μ and σ are unknown.

7. Example. Suppose the sample variables $X_1, X_2..., X_n$ are a random sample from a population having the uniform distribution $U(0, \theta)$ with p.d.f. $f(x, \theta) = 1/\theta$ for $0 < x \leq \theta$, and is zero otherwise.