9. Point estimation

Remember that the random variables X_1, X_2, …, X_n satisfying two conditions: they are independent and have a common PMF/PDF $f(x;\theta)$ are called a **random sample (RS)** or simply **sample** of size n. A specific set of observed values $x_1, x_2, \ldots, x_n$ is a set of **sample values** assumed by the sample.

A **statistic** is any function $h(\cdot)$ of a given sample X_1, X_2, …, X_n for which the value can be determined once the sample values $x_1, x_2, \ldots, x_n$ have been observed.

The notation $f(x;\theta)$ aims to stress that the PMF/PDF under consideration depends on a parameter θ varying within a given range Θ. The precise value of parameter θ is unknown; our aim is to "estimate" it from the sample X_1, X_2, …, X_n. This means that we want to determine a function $\theta^*(\cdot)$ depending on the sample but not on θ (a statistic) which we could take as a projected value of θ. If the experiment that yielded the data set were to be repeated, we would obtain different values $x_1, x_2, \ldots, x_n$. The function $\theta^*(\cdot)$ when applied to the new data set would yield a different value for θ. Such a function will be called an **estimator** of θ; its particular value is often called an **estimate**.

We thus see that an estimate is itself a random variable possessing a probability distribution, which depends both on the functional form defined by $\theta^*(\cdot)$ and on the distribution of the underlying random variable X with PMF/PDF $f(x;\theta)$.

The domain of statistics that emerges is called **parametric estimation**. The problem of parameter estimation is one class in the broader topic of **statistical inference** in which our object is to make inferences about various aspects of the underlying distribution $f(x;\theta)$ (called **population distribution**) on the basis of observed sample values.

It is important to note that a statistic, being a function of random variables, is a random variable. When used to estimate a distribution parameter, its statistical properties, such as mean, variance, and distribution, give information concerning the quality of this particular estimation procedure.

9.1. Sample mean and sample variance

The information of a RS is usually summarized by a handful of statistics. Certain statistics play an important role in statistical estimation theory; the most important of these are the sample mean and sample variance. Some properties of these statistics are discussed below.

The sample mean $\bar{X}$ is frequently used to estimate the value of μ (the mean of the population). Sometimes we also need to estimate the value of the distribution's variance σ^2; this is done by **sample variance** defined by:

$$S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n\ -\ 1).$$

(taking its square root, one gets S, the **sample standard deviation**).

Note that the numerator is the sum of squares of individual deviations from the sample mean; the definition intentionally avoids using the distribution mean μ, as its value is usually unknown.

To find the expected value of S^2, we transform its numerator first:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left(X_i - \mu - (\overline{X} - \mu)\right)^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2$$

(note that $\overline{X} - \mu$, being free of i, is considered as a constant by the summation).

It follows

$$E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = n\ \mathrm{var}(X) - 2\sum_{i=1}^{n}\mathrm{Cov}(\bar{X}, X_i) + n\ \mathrm{var}(\bar{X}) =$$

$$= n\ \mathrm{var}(X) - 2n\ \mathrm{Cov}(\bar{X}, X_1) + \mathrm{var}(X),$$

$\mathrm{Cov}(\bar{X}, X_1)$, $\mathrm{Cov}(\bar{X}, X_2)$, $\mathrm{Cov}(\bar{X}, X_3)$, ... must have the same value, and $\mathrm{Cov}(\bar{X}, X_1) = n^{-1}\sum_{i=1}^{n}\mathrm{Cov}(X_i, X_1) = n^{-1}\ \mathrm{var}(X)$.

This implies that $E(S^2) = (n\sigma^2 - 2\sigma^2 + \sigma^2)/(n - 1) = \sigma^2$.

Thus, S^2 is a so called unbiased estimator of the distribution's variance σ^2 (meaning it has the correct expected value). Precise definition will be given later.

Does this imply that $S \equiv (n\ -\ 1)^{-1/2}\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}$ (the sample standard deviation) has the expected value of σ? The answer is "no", S is a (slightly) biased estimator of the population's standard deviation (the exact value of the bias depends on the shape of the corresponding distribution).

This S is useful when estimating the value of the population mean μ. We know that $\bar{X}$ is the unbiased estimator of μ, having the standard deviation of $\sigma/\sqrt{n}$. We would like to express this as $\mu \approx \bar{X} \pm \sigma/\sqrt{n}$ (the so called confidence interval for estimating μ) but since we ordinarily don't know the exact value of σ either, we have to substitute its estimator S, thus: $\mu \approx \bar{X} \pm S/\sqrt{n}$. Later we investigate these issues in more detail.

To be able to say anything more about $\bar{X}$ and S^2, we need to know the distribution form which we are sampling. We will assume that the distribution is normal, with mean μ and variance σ^2.

The distribution of $\bar{X}$ must also be normal (with mean μ and standard deviation of $\sigma/\sqrt{n}$, as we already know) for any sample size n (not just "large").

Regarding S^2, one can show that it is independent of $\bar{X}$, and that the distribution of $(n-1)S^2/\sigma^2$ is χ^2_{n-1}. The proof of the independence is rather complex and so is omitted here. It is follows that $(\bar{X} - \mu)/(S/\sqrt{n})$ has the t_{n-1}

distribution because $(\bar{X} - \mu)/(S/\sqrt{n}) = [(\bar{X} - \mu)/(\sigma/\sqrt{n})]/[(S/\sqrt{n})/(\sigma/\sqrt{n})]$, where $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has standard normal distribution and $(S/\sqrt{n})/(\sigma/\sqrt{n}) = \sqrt{(n-1)\,S^2/\sigma^2}\,/\sqrt{n-1}$.

The sample mean and sample variance are examples of point estimators.

9.2. Properties of the estimators

You could see that some important families of PMFs/PDFs depend on a parameter (or several parameters forming a vector). For instance, Poisson PMFs are parameterized by $\lambda > 0$, and so are exponential PDFs, normal PDFs are parameterized by pair μ and σ^2, where μ is the mean and σ^2 is the variance. The "true" value of a parameter (or several parameters) is considered unknown and we will have to develop the means to make a judgment about what it is.

For example, it is well known that the number of hops by a bird before it takes off is described by a geometric distribution. Similarly, emission of alpha-particles by radioactive material is described by a Poisson distribution (this follows immediately if one assumes that the emission mechanism works independently as time progresses). However, the parameter of the distribution may vary with the type of bird or the material used in the emission experiment (and also other factors).

We observe a sample of values of a given number n of iid RVs (X_1, X_2, …, X_n) with a common PMF/PDF $f(x;\theta)$. The joint PDF/PMF of the random vector $\mathbf{X}$ is denoted by $f_{\mathbf{X}}(\mathbf{x};\theta)$ and is given by the product

$$f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta),\ \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Here, and below vector $\mathbf{x}$ is a sample value of $\mathbf{X}$. The subscript $\mathbf{X}$ in notation $f_{\mathbf{X}}(\mathbf{x};\theta)$ are often omitted.

In principle, any function of $\mathbf{x}$ can be considered as an estimator, but in practice we want it to be "reasonable". We therefore need to develop criteria for which estimator is good and which bad.

Example 9.1**.** Let X_1, X_2, …, X_n be iid and X_i has Poisson distribution with parameter λ.

Consider the sample mean $\bar{X} = \bar{X}_n$ as an estimator of parameter λ.

We already knew that the sample mean has the following useful properties:

(i) The random value $\bar{X}$ is grouped around the true value of the parameter: $E(\bar{X}) = \lambda$.

This property is called **unbiasedness**.

(ii) $\bar{X}$ approaches the true value as $n \to \infty$: $\bar{X} \xrightarrow{P} \lambda$ $\left(\bar{X} \xrightarrow{a.s.} \lambda\right)$ – the weak (strong) LLN.

Property (ii) is called **consistency** (**strong consistency**).

(iii) For large n, $\sqrt{n}\dfrac{\bar{X}-\lambda}{\sqrt{\lambda}} \xrightarrow{F} \mathrm{N}(0,1)$ (the CLT).

This property is often called **asymptotic normality**.

We are also able to see that $\bar{X}$ has another important property:

(iv) $\bar{X}$ has the minimal mean square error in a wide class of estimators λ^*:

$E\left(\bar{X}-\lambda\right)^2 \le E\left(\lambda^*-\lambda\right)^2$. The proof is below.

Give some general definitions.

Definition 9.1. An estimator $\theta^* = \theta^*(\mathbf{x})$ of a parameter θ is **consistent (strongly consistent)** if as $n \to \infty$ for all $\theta \in \Theta$

$$\theta^* \xrightarrow{P} \theta \quad \left(\theta^* \xrightarrow{a.s.} \theta\right).$$

It is thus a large-sample concept and is a good quality for an estimator to have. Usually we are not interesting in the estimators that are not consistent.

How to check the consistency of the estimator? We can use the Chebyshov's inequality, from which follows that if $E\left(\theta^*\right) \to \theta$ and $\mathrm{var}\left(\theta^*\right) \to 0$ as $n \to \infty$ then θ^* is a consistent estimator. It is important to note that it gives a sufficient but not necessary condition for consistency.

Every "decent" estimator must be consistent; but that by itself does not make it particularly "good". For example $\hat{\mu} = \left(X_2 + X_4 + ... + X_n\right)/(n/2)$ (for n even) is a consistent estimator of μ. Yet, we are wasting one half of our sample, which is unacceptable.

Definition 9.2. An estimator $\theta^* = \theta^*(x)$ of a parameter θ is **unbiased** if for all $\theta \in \Theta$

$$E\left(\theta^*\right) = \theta.$$

If only $E\left(\theta^*\right) \to \theta$ as $n \to \infty$ then θ^* is called **asymptotically unbiased.**

The value $b\left(\theta^*\right) = E\left(\theta^*\right) - \theta$ is called the **bias** of the estimator θ^*, usually proportional to $1/n$ for asymptotically unbiased estimators.

Unbiasedness is a desirable property of the estimator but making an estimator unbiased (or at least asymptotically so) is not enough to make it even acceptable. Consider estimating μ of a distribution by taking $\hat{\mu} = X_1$ (the first observation only), throwing away X_2, …, X_n. We get a fully unbiased estimator which is evidently unacceptable, since we are wasting nearly all the information contained in our sample. Also sometimes biased estimators have less mean square errors, so unbiasedness can be outweighed by other considerations (see Soong, Example 9.5, p. 272).

Definition 9.3. The **mean square error** of an estimator $\theta^* = \theta^*(\mathbf{x})$ is

$$\mathrm{MSE}(\theta^*) = E(\theta^* - \theta)^2.$$

Show that $\mathrm{MSE}(\theta^*) = b^2(\theta^*) + \mathrm{var}(\theta^*)$.

The minimum value of the mean square error of an estimator is often a criteria the "best" estimator.

A "good" estimator should not only be unbiased, but it should also have a variance which is as small as possible.

The result concerning the problem is given by the Cramér–Rao (CR) inequality, or CR bound. When an estimator achieves this bound, it is automatically the "best". The relevant details are summarized in the following

Theorem (Cramér-Rao inequality). *Assume that a PDF/PMF* $f(x;\theta)$ *depends smoothly on parameter* θ. *Take an unbiased estimator* $\theta^*(\mathbf{X})$ *of* θ. *Let some regularity conditions on* $f(x;\theta)$ *and* $\theta^*(\mathbf{X})$ *hold. Then for any such estimator, the following bound holds:*

$$\mathrm{var}(\theta^*) \geq \frac{1}{n\mathrm{I}(\theta)},$$

where $\mathrm{I}(\theta) = E\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 = -E\left(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right)$.

The quantity $\mathrm{I}(\theta)$ is often called the **Fisher information** and features in many areas of probability theory and statistics.

The analog theorem can be formulated for the estimates with some fixed bias.

C.R. Rao (1920–) is an Indian mathematician who studied in India and Britain (he took his Ph.D. at Cambridge University and was Fisher's only formal Ph.D. student in statistics), worked for a long time in India and currently lives and works in the USA. Rao's contributions in statistics are now widely recognized.

The CR inequality is named also after C.H. Cramér (1893–1985), a prominent Swedish analyst, number theorist, probabilist and statistician, and C.R. Rao. One story is that the final form of the inequality was proved by Rao, then a young (and inexperienced) lecturer at the Indian Statistical Institute, overnight in 1943 in response to a student enquiry about some unclear places in his presentation.

Remark. The regularity conditions imply interchanging the derivation $\frac{\partial}{\partial\theta}$ and the integration (or summation) in the equalities

1) $\frac{\partial}{\partial\theta}\int_R f(x;\theta)dx = 0 \left(\text{or } \frac{\partial}{\partial\theta}\sum_i f(x_i;\theta) = 0 \right)$ the equality holds as $\int_R f(x;\theta)dx = 1 \left(\text{or } \sum_i f(x_i;\theta) = 1 \right)$;

2) $1 = \frac{\partial}{\partial\theta}\theta = \frac{\partial}{\partial\theta}E(\theta^*)$, as $E\left(\theta^*(\mathbf{X})\right) = \int_{R^n} \theta^*(\mathbf{x}) f(\mathbf{x};\theta)d\mathbf{x}$ $\left(\text{or } E\left(\theta^*(\mathbf{X})\right) = \sum_{\mathbf{x}} \theta^*(\mathbf{x}) f(\mathbf{x};\theta) \right)$.

Emphasize the fact: to hold the conditions estimated parameter θ should not appear in the limits of the distribution (for example for continuous uniform distribution the conditions do not hold).

Based on this CR bound we define the so called efficiency of an unbiased estimator $\hat{\theta}$ as the ratio of the theoretical variance bound $\text{CRV}_\theta = \frac{1}{n\text{I}(\theta)}$ to the actual variance of θ^*, thus:

$$eff\left(\theta^*\right) = \frac{\text{CRV}_\theta}{\text{var}\left(\theta^*\right)}$$

usually expressed in percent (we know that its value cannot be bigger that 1, i.e. 100 %). An estimator $\hat{\theta}$ whose variance is as small as CRV_θ is called **efficient.** An estimator which reaches 100 % efficiency only in the $n \to \infty$ limit is called **asymptotically efficient.**

Here we run into two difficulties:

1. The variance of an estimator is, in general, a function of the unknown parameter (to see that, go back to the S^2 example), so we are comparing functions, not values. It may easily happen that two unbiased estimators have variances such that one estimator is better in some range of θ values and worse in another.

2. Even when the "best" estimator exists, how do we know that it does and, more importantly, how do we find it? CR inequality does not give us the method of obtaining the estimators.

Examples:

1. How good is $\bar{X}$ as an estimator of μ of the Normal distribution N(μ, σ)?

Solution: We know that its variance is σ^2/n. To compute the CR bound we do

$$-\frac{\partial^2}{\partial\mu^2}\ln f(x) = \frac{\partial^2}{\partial\mu^2}\left(\frac{\ln 2\pi}{2} + \ln\sigma + \left(x-\mu\right)^2 / \left(2\sigma^2\right) \right) = \frac{1}{\sigma^2}.$$

Thus CRV equals σ^2/n implying that $\bar{X}$ is the best (unbiased) estimator of μ.

2. Let us find the efficiency of $\bar{X}$ to estimate the mean β of the exponential distribution, with $f(x) = 1/\beta\, e^{-x/\beta}$ for $x > 0$.

Solution: $-\frac{\partial^2}{\partial\beta^2}\ln f(x) = -\frac{\partial^2}{\partial\beta^2}\left(\ln\beta + x/\beta\right) = -\frac{1}{\beta^2} + \frac{2x}{\beta^3}$,

$E\left(-\frac{1}{\beta^2} + \frac{2X}{\beta^3}\right) = \frac{1}{\beta^2}$. It follows CRV = β^2/n.

We know that $E(\bar{X}) = \beta$ and $\text{var}(\bar{X}) = \beta^2/n$. Conclusion: $\bar{X}$ is the best estimator of β.

We must point out that efficient estimators exist only under certain conditions.

It is not always possible to calculate MSE (or asymptotic MSE). In these cases the following definition can help us.

Definition 9.4. An estimator $\theta^* = \theta^*(\mathbf{x})$ of a parameter θ is **asymptotically normal with a coefficient** $\text{var}(\theta)$ if

$$\sqrt{n}(\theta^* - \theta)/\sqrt{\text{var}(\theta)} \xrightarrow{F} N(0,1).$$

The asymptotic normality is an important property of the sequence of the estimators. It can be used not only for comprising the estimators (choosing the "best" one, i.e. with the minimum coefficient), but for construction of confidence intervals and hypotheses testing.

Example 9.2. A frequent case is where $X_1, X_2, \ldots, X_n$ are iid and X_i are $N(\mu, \sigma^2)$. When speaking of normal samples, one usually distinguishes three situations:

(i) the mean μ is unknown and variance σ^2 known (say, $\sigma^2 = 1$);

(ii) μ is known (say, equal to 0) and $\sigma^2 > 0$ unknown;

(iii) neither μ nor σ^2 is known.

In cases (i) and (iii), an estimator for μ is the sample mean $\bar{X}$ with $E(\bar{X}) = \mu$ (unbiasedness) and normal distribution $N(\mu,\sigma^2/n)$.

In case (ii), an unbiased estimator for σ^2 is S^2, the distribution of $(n-1)S^2/\sigma^2$ is χ^2_{n-1}.

In case (iii), the pair $(\bar{X}, S^2)$ can be taken as an estimator for vector (μ, σ^2) and we obtain joint unbiasedness, joint consistency and joint asymptotic normality.

What about the methods of obtaining the estimators? The following concept is very useful.

9.3. Sufficient statistics

Remind that a statistic is an arbitrary function of sample vector **x** or its random counterpart **X**.

We call a function T of **x** (possibly, with vector values) a **sufficient statistic** for parameter $\theta \in \Theta$ if the conditional distribution of random sample **X** given $T(\mathbf{X})$ does not depend on θ.

The significance of this concept is that the sufficient statistic encapsulates all knowledge about sample **x** needed to produce a "good" estimator for θ.

The most efficient way to check the sufficiency is to use the **factorization criterion**.

The factorization criterion is a general statement about sufficient statistics. It says: *T is sufficient for* θ *iff the PMF/PDF* $f_{\mathrm{X}}(\mathbf{x},\theta)$ *can be written as a product* $g(T(\mathbf{x}),\theta)\,h(\mathbf{x})$ *for some functions g and h.*

The proof in the discrete case is straightforward (you can do it as an exercise). In the continuous case we need some elements of measure theory.

The idea behind the factorization criterion goes back to a 1925 paper by R.A. Fisher (1890–1962), the outstanding UK applied mathematician, statistician and genetist.

We can take the criterion as the definition of sufficient statistic.

Clearly, sufficient statistics are not unique. So the next useful step is to consider a **minimal sufficient statistic.** Any sufficient statistic is a function of the minimal one. In other words, the minimal sufficient statistic represents the least amount of detail we should know about sample **x**. Any further suppression of information about the sample would result in the loss of sufficiency.

In all examples below, sufficient statistics are minimal.

Examples:

1. *Bernoulli distribution*: $f_{\mathbf{X}}(\mathbf{x},\theta) = p^{x_1+x_2+\ldots x_n}(1-p)^{n-(x_1+x_2+\ldots x_n)}$ is a function of p and of a single combination of the sample values, namely $\sum_{i=1}^{n} x_i$. A sufficient statistic for estimating p is thus $\sum_{i=1}^{n} X_i$.

2. *Normal distribution* (the mean μ is unknown and variance σ^2 known):

$$f_{\mathrm{X}}(\mathbf{x},\theta) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_{i=1}^{n} x_i^2/\left(2\sigma^2\right)\right)\times\exp\left(\left(-n\mu^2+2\mu\sum_{i=1}^{n} x_i\right)/\left(2\sigma^2\right)\right),$$

where the first factor (to the left of $\times$) contains no μ and the second factor is a function of only a single combination of the sample values, namely their sum. This leads to the same conclusion as in the previous example.

When neither μ nor σ^2 is known $T(\mathbf{x}) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$ is sufficient statistic for (μ, σ^2).

3. *Exponential distribution*: $f_{\mathbf{X}}(\mathbf{x},\theta)=\exp\left(-\lambda\sum_{i=1}^{n}x_i\right)/\lambda^n$. A sufficient statistic for estimating λ is $\sum_{i=1}^{n}X_i$.

4. *Poisson distribution*: in example 9.1 the sample mean $\bar{X}$ is a sufficient statistic for λ:

$$f_{\mathbf{X}}(\mathbf{x},\theta)=e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}=e^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}/\prod_{i=1}^{n}x_i!.$$

And we know how to make the statistics in previous examples into unbiased estimators.

4. The function $X_{(n)}=\max(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for estimating θ of the uniform on $[0, \theta]$ distribution:

$$f_{\mathbf{X}}(\mathbf{x},\theta)=\theta^{-n}I\left(\mathbf{x}: X_{(n)}\le\theta\right),$$

where $I(\cdot)$ is an indicator.

Is $X_{(n)}$ an unbiased estimator of θ? Consider the distribution function

$$F_{X_{(n)}}(x)=P\left(X_{(n)}<x\right)=\prod_{i=1}^{n}P\left(X_i<x\right)=x^n\theta^{-n}\ \text{for}\ x\in[0,\theta],$$

PDF $f_{X_{(n)}}(x)=F'_{X_{(n)}}(x)=nx^{n-1}\theta^{-n}$ for $x\in[0,\theta]$,

and $E\left(X_{(n)}\right)=\int_0^{\theta}xf_{X_{(n)}}(x)dx=n\int_0^{\theta}\frac{x^n}{\theta^n}dx=\frac{n}{n+1}\theta$.

So $X_{(n)}$ is asymptotically unbiased and $\theta^*=\frac{n+1}{n}X_{(n)}$ is unbiased estimator of θ.

In practice if we find the sufficient statistic all we have to do to convert it into the best possible estimator of θ is to make it unbiased (by some transformation, which is usually easy to design).

The only difficulty with the approach arises when a sufficient statistic does not exist (try finding it for the Cauchy distribution).

One can resort to using one of the following two techniques for finding an estimator.

9.4. Method of moments

The oldest systematic method of point estimation − **method of moments** − was proposed by K. Pearson (1894) and was extensively used by him and his co-workers. It was neglected for a number of years because of its general lack of optimum properties. The method of moments is simple in concept.

Consider a selected probability PMF/PDF $f(x;\boldsymbol{\theta})$ for which parameters $\boldsymbol{\theta}=(\theta_1,...,\theta_m)^T$ are to be estimated based on sample X_1, X_2, …, X_n. The theoretical or population moments of distribution $f(x,\boldsymbol{\theta})$ are

$$\alpha_j = \int_{-\infty}^{\infty} x^j f(x;\boldsymbol{\theta})dx,\ j \geq 1.$$

They are, in general, functions of the unknown parameters $\alpha_j = \alpha_j(\theta_1,...,\theta_m)$. However, sample moments of various orders can be found from the sample by

$$M_j = \sum_{i=1}^{n} X_i^j / n,\ j \geq 1.$$

The method of moments suggests that, in order to determine estimators $\hat{\boldsymbol{\theta}}=(\hat{\theta}_1,...,\hat{\theta}_m)^T$ from the sample, we equate a sufficient number of sample moments to the corresponding population moments. By establishing and solving as many resulting moment equations as there are parameters to be estimated, estimators for the parameters are obtained. Hence, the procedure for determining $\hat{\theta}_1,...,\hat{\theta}_m$ consists of the following steps:

1. Let

$$\alpha_j(\hat{\theta}_1,...,\hat{\theta}_m) = M_j,\ j = \overline{1,m}. \qquad (9.1)$$

These yield *m* moment equations with *m* unknowns $\hat{\theta}_1,...,\hat{\theta}_m$.

2. Solve for $\hat{\theta}_1,...,\hat{\theta}_m$ from this system of equations. These are called the **moment estimators** for $\theta_1,...,\theta_m$.

Remark. It is not necessary to consider *m* consecutive moment equations as indicated by (9.1), any convenient set of *m* equations that lead to the solution for $\hat{\theta}_1,...,\hat{\theta}_m$ is sufficient. Lower-order moment equations are preferred, however, since they require less manipulation of observed data.

An attractive feature of the method of moments is that the moment equations are straightforward to establish, and there is seldom any difficulty in solving them. However, a shortcoming is that such desirable properties as unbiasedness or efficiency are not generally guaranteed for estimators so obtained.

Consistency of moment estimators can be established under general conditions (see Soong, p. 279).

The advantage of the method is that it requires only the moments of population, the knowledge of its distribution are not necessary.

Examples:

1. Normal distribution: $f(x;\mu,\sigma^2)=\left(2\pi\sigma^2\right)^{-1/2}\exp\left(-\left(x-\mu\right)^2/\left(2\sigma^2\right)\right)$, estimate parameters $\theta_1=\mu$ and $\theta_2=\sigma^2$.

Following the method of moments, we need two moment equations, and the most convenient ones are obviously $\alpha_1=\mu$ and $\alpha_2=\sigma^2+\mu^2$, corresponding sample moments $M_1=\sum_{i=1}^{n}x_i/n$ and $M_2=\sum_{i=1}^{n}x_i^2/n$.

Now, we have the system $\hat{\alpha}_1=\sum_{i=1}^{n}x_i/n$ and $\hat{\alpha}_2=\sum_{i=1}^{n}x_i^2/n$.

Hence, the first of these moment equations gives $\hat{\mu}=\sum_{i=1}^{n}x_i/n=\overline{X}$.

The properties of this estimator have already been discussed above. It is unbiased and has minimum variance among all unbiased estimators for μ. We see that the method of moments produces desirable results in this case.

The second moment equation gives $\hat{\alpha}_2=\hat{\sigma}^2+\hat{\mu}^2=\sum_{i=1}^{n}x_i^2/n$ or

$$\hat{\sigma}^2=\hat{\alpha}_2-\hat{\mu}^2=\sum_{i=1}^{n}x_i^2/n-\overline{X}^2=\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2/n=\frac{n-1}{n}S^2.$$

This, as we have shown, is a biased estimator for σ^2.

2. Uniform on $(0,\theta)$ distribution. We wish to estimate parameter θ from a sample of size *n*.

The density function is $f(x,\theta)=1/\theta,\ x\in[0,\theta]$ and the first moment is $\theta/2$.

It follows from the method of moments that, on letting $\overline{X}=\hat{\theta}/2$ we obtain

$$\hat{\theta}=2\overline{X}. \qquad (9.2)$$

Upon little reflection, the validity of this estimator is somewhat questionable because, by definition, all values assumed by population value are supposed to lie within interval $(0,\theta)$. However, we see from (9.2) that it is possible that some of the samples are greater than $\hat{\theta}$. Intuitively, a better estimator might be the *n*th-order statistic $X_{(n)}$. This is the outcome following the method of maximum likelihood, to be discussed below.

9.5. Method of maximum likelihood

First introduced by R. Fischer in 1922, the method of maximum likelihood has become the most important general method of estimation from a theoretical point of view.

Consider the joint distribution $f_{\mathbf{x}}(\mathbf{x};\theta)=\prod_{i=1}^{n} f(x_i;\theta)$ where, for simplicity, is the only parameter to be estimated from a set of sample values $x_1, x_2, \ldots, x_n$. We call it the **likelihood function** and denote $L(\mathbf{x};\theta)$. When the sample values are given, likelihood function L becomes a function of a single variable θ. The estimation procedure for θ based on the method of maximum likelihood consists of choosing as an estimate of θ the particular value of that maximizes $L(\cdot)$. The maximum of $L(\cdot)$ occurs in most cases at the value of θ where $\mathrm{d}L(\mathbf{x};\theta)/\mathrm{d}\theta$ is zero. Hence, in a large number of cases, the **maximum likelihood estimate (MLE)** of θ based on sample values $x_1, x_2, \ldots, x_n$ can be determined from

$$\mathrm{d}L(\mathbf{x};\hat{\theta})/\mathrm{d}\hat{\theta}=0.$$

Since function L is always nonnegative and attains its maximum for the same value of θ as $\ln L$, it is generally easier to obtain MLE by solving

$$\mathrm{d}\ln L(\mathbf{x};\hat{\theta})/\mathrm{d}\hat{\theta}=0 \tag{9.3}$$

because $\ln L$ is in the form of a sum rather than a product.

Equation (9.3) is referred to as the **likelihood equation**. The desired solution is one where root is a function of $x_1, x_2, \ldots, x_n$ if such a root exists. When several roots of (9.3) exist, the MLE is the root corresponding to the global maximum of L or $\ln L$.

By choosing a value of that maximizes L, or $\ln L$, we in fact say that we prefer the value of $\hat{\theta}$ that makes as probable as possible the event that the sample values indeed come from the population.

The extension to the case of several parameters is straightforward. In the case of m parameters the MLEs of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$, are obtained by solving simultaneously the system of likelihood equations

$$\partial \ln L(\mathbf{x};\hat{\boldsymbol{\theta}})/\partial\,\hat{\theta}_j=0,\quad j=\overline{1,m}. \tag{9.4}$$

The universal appeal enjoyed by maximum likelihood estimators stems from the optimal properties they possess when the sample size becomes large. Under mild conditions imposed on the PMF/PDF of population, the MLEs are asymptotically unbiased, efficient and normal. However, these important properties are large-sample properties.

Unfortunately, very little can be said in the case of a small sample size; it may be biased and nonefficient.

Let us also make an observation on the solution procedure for solving likelihood equations. Although it is fairly simple to establish (9.3) or (9.4), they are frequently highly nonlinear in the unknown estimates, and close-form solutions for the MLE are sometimes difficult, if not impossible, to achieve. In many cases, iterations or numerical schemes are necessary.

Examples:

1. Let us consider again the normal distribution $\ln L = \ln f(\mathbf{x}, \mu, \sigma^2) = -\frac{n \ln 2\pi}{2} - n \ln \sigma - \sum_{i=1}^{n} (x_i - \mu)^2 / (2\sigma^2)$. Let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ as before; the likelihood equations are

$$\frac{\partial \ln f(x, \hat{\theta}_1, \hat{\theta}_2)}{\partial \hat{\theta}_1} = \frac{\partial \ln f(x, \hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\mu}} = 2 \sum_{i=1}^{n} (x_i - \hat{\mu}) / (2\hat{\sigma}^2) = 0,$$

$$\frac{\partial \ln f(x, \hat{\theta}_1, \hat{\theta}_2)}{\partial \hat{\theta}_2} = \frac{\partial \ln f(x, \hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \sum_{i=1}^{n} (x_i - \hat{\mu})^2 / (2\hat{\sigma}^4) = 0.$$

Solving the above equations simultaneously, the MLEs of μ and σ^2 are found to be $\hat{\mu} = \sum_{i=1}^{n} x_i / n = \overline{X}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \overline{X})^2 / n$ which coincide with their moment estimators in this case.

2. Consider uniform on $(0, \theta)$ distribution. Likelihood function $L = f(\mathbf{x}, \theta) = \theta^{-n}$ if $X_{(n)} \le \theta$ is monotonic decreasing and therefore takes the maximum value at point $\hat{\theta} = X_{(n)}$.

This estimator is seen to be different from that obtained by using the moment method and, as we already commented, it is a more logical choice.

Let us also note that we did not obtain the result by solving the likelihood equation. The likelihood equation does not apply in this case as the maximum of *L* occurs at the boundary and the derivative is not zero there.

9.6. Exercises

1. Consider a sample of size 3 from N(μ, σ). What is the relative efficiency of $(X_1 + 2X_2 + X_3)/4$ (obviously unbiased) with respect to $\overline{X}$ when estimating μ?

The **relative efficiency** of $\hat{\theta}_1$ compared to $\hat{\theta}_2$ is the ratio $\frac{\mathrm{var}(\hat{\theta}_2)}{\mathrm{var}(\hat{\theta}_1)}$.

2. For uniform distribution U(*a*,*b*) estimate both *a* and *b* by the method of moments.

3. For U(*a*,*b*) estimate both *a* and *b* by the likelihood method. What can you say about the relative efficiency of the estimators compared with the ones of the previous exercise?

4. For binomial distribution estimate both *n* and *p* by the method of moments.

5. Consider a sample from a distribution with a finite variance σ^2. Proof that the sample mean is asymptotically normal with the coefficient σ^2.

6. Consider uniform on $(0,\theta)$ distribution. Are the estimators $\hat{\theta}_1 = 2\overline{X}$ and $\hat{\theta}_2 = X_{(n)}$ of the parameter θ asymptotically normal?

7. Find the MLE for the parameter of Poisson distribution. How good is $\overline{X}$ in estimating λ of the Poisson distribution? Proof the asymptotic normality of the estimator and find the coefficient.

8. Let $X_1, X_2, \ldots, X_n$ be a sample from Poisson distribution with parameter λ. Consider X_1 as the estimator of the parameter. Is the estimator unbiased? Is it consistent?

9. Distribution given by $f(x) = 3x^2\theta e^{-\theta x^3}$ for $x > 0$; estimate θ by the likelihood method.

10.* Distribution given by $f(x) = \frac{2x}{a}e^{-x^2/a}$ for $x > 0$; estimate a by the method of moments. Find the expectation. Is the estimator unbiased? Construct the unbiased estimator.

Hint: use Gamma function.

11. Consider consistency and unbiasedness of two estimators of the mean $\hat{\theta}_1 = \overline{X} + \frac{1}{n}$ and $\hat{\theta}_2 = \overline{X} \,/\left(\sigma/\sqrt{n}\right)$.