STATISTICAL PARAMETER ESTIMATION

Werner Gurker and **Reinhard Viertl**
Vienna University of Technology, Wien, Austria

Keywords: Confidence interval, consistency, coverage probability, confidence set, estimator, loss function, likelihood function, maximum likelihood, mean squared error, parameters, point estimator, risk function, sufficient statistic, unbiasedness, variance bound

Contents

Summary

The estimation of the parameters of a statistical model is one of the fundamental issues in statistics. Choosing an appropriate estimator, that is ‘best’ in one or another respect, is an important task, hence firstly several optimally criterions are considered. In practice, however, constructive methods of parameter estimation are needed. Some of the methods most frequently used are considered, the method of moments, linear

estimation methods, and the most important one, the method of maximum likelihood in some detail. At last, the closely related problem of interval estimation is considered.

1. Fundamental Concepts

1.1. Parameter and Estimator

All estimation procedures are based on a *random sample*, $X_1,\ldots,X_n$ from a *random variable* X. Let $f(x|\theta)$ denote the *probability mass function* (pmf), if X is a discrete, or the *probability density function* (pdf), if X is a continuous variable, where the form of the pmf or pdf is known, but the *parameter vector* (*parameter* for short) $\theta=(\theta_1,\ldots,\theta_k)$ is unknown. We call the set of possible values for the parameter θ the *parameter space* Θ, being a subset of $\mathbb{R}^k$.

Remark: As there are formally only slight (and quite obvious) differences between the discrete and the continuous cases, we focus on the latter for simplicity.

Let $T(x_1,\ldots,x_n)$ be a real-valued or vector-valued function whose domain includes the sample space χ of $\mathbf{X}=(X_1,\ldots,X_n)$. (Compare *Statistical Inference*.) Then the random variable $T(X_1,\ldots,X_n)$ is called a *statistic*, and its distribution is called the *sampling distribution* of *T*. Note that the value of a statistic can be computed from the sample alone and does not depend on any (unknown) parameters. The *sample mean* $\bar{X}$ and the *sample variance* S^2

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i, \qquad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2, \tag{1}$$

are well known examples of statistics. Our objective is to find statistics which will serve as estimators for the unknown parameter θ, or more generally for certain functions $\tau(\theta)$ of the parameters. Thus a very broad definition of an 'estimator' is the following.

Definition: Any statistic $T(X_1,\ldots,X_n)$ (from χ to $\tau(\Theta)$) is called an *estimator* of $\tau(\theta)$. When the actual sample values are implemented into *T* ,an *estimate t* of $\tau(\theta)$ results.

1.2. Mean Squared Error

There are various suggestions for a measure of 'closeness' of an estimator to the objective parameter function $\tau(\theta)$. For instance, one could consider the *probability* that the estimator *T* is close to $\tau(\theta)$,

$$P_\theta\left\{\left|T(X_1,\ldots,X_n)-\tau(\theta)\right|<\varepsilon\right\} \text{ for } \varepsilon>0 \tag{2}$$

or we could consider an *average* measure of closeness like the mean absolute deviation,

$$\mathrm{MAD}_T(\theta)=E_\theta\left[\left|T(X_1,\ldots,X_n)-\tau(\theta)\right|\right]. \tag{3}$$

What is mathematically more convenient is to consider an average squared deviation, the *mean squared error* (MSE),

$$\mathrm{MSE}_T(\theta)=E_\theta\left[\left(T(X_1,\ldots,X_n)-\tau(\theta)\right)^2\right]. \tag{4}$$

The MSE summarizes two properties of an estimator, its ‘precision’ and its ‘accuracy’, two important concepts in practical applications. By some simple transformations the MSE of an estimator *T* can be written as follows,

$$\mathrm{MSE}_T(\theta)=\mathrm{Var}\left(T(\mathbf{X})\right)+\left[\tau(\theta)-E_\theta\left(T(\mathbf{X})\right)\right]^2, \tag{5}$$

where $\mathrm{Var}(T(\mathbf{X}))$ denotes the variance of $T(\mathbf{X})$. The standard deviation, $\sqrt{\mathrm{Var}(T)}$, is a measure of the *precision* of an estimator (the smaller the variance, the greater the precision), that is a measure of its performance; the square root of the second term, $\left|\tau(\theta)-E_\theta(T(\mathbf{X}))\right|$ (not to be confused with the MAD), is a measure of how *accurate* the estimator is, that is how large on the average the error systematically introduced by using *T* is.

Though not stated explicitly, associated with all these measures is a certain concept of ‘loss’; the MSE, for instance, penalizes the deviations of an estimator from its objective function quadratically.

1.3. Loss and Risk

The estimation of a parameter θ can be regarded as some kind of a *decision problem*, that is to say, by using the estimator *T* we *decide*, given a specific sample $\mathbf{x}=(x_1,\ldots,x_n)$, on a specific value $\hat{\theta}$ for the unknown parameter, $\hat{\theta}=T(\mathbf{x})$. Clearly, $\hat{\theta}$ will generally be different from the true parameter value θ (and if not, we would not be aware of it).

Being involved in decision making in the presence of uncertainty, a certain kind of *loss*, $L\left(\theta,T(\mathbf{x})\right)$, will be incurred, meaning the ‘loss’ incurred, when the actual ‘state of nature’ is θ, but $T(\mathbf{x})$ is taken as the estimate of θ. Frequently it will be difficult to determine the actual loss function *L* over a whole region of interest (there are some rational procedures, however), so it is customary to analyse the decision problem using some 'standard' loss functions. Especially for estimation problems usually two loss functions are considered, the *squared error loss* and the *linear loss.*

The *squared error loss* is defined as (with $\mathbf{Q}$ a $k \times k$ known positive definite matrix)

$$L(\theta,\, \mathbf{a}) = (\theta - \mathbf{a})\ \mathbf{Q}\ (\theta - \mathbf{a})^T, \quad \mathbf{a} \in \mathbb{R}^k. \tag{6}$$

In case of a one-dimensional parameter, the loss function reduces to

$$L(\theta,\, a) = c(\theta - a)^2. \tag{7}$$

Frequently it will not be unreasonable to assume that the loss function is approximately linear (at least piecewise); for a one-dimensional parameter the *linear loss* can be written as (K_0 and K_1 are two known constants)

$$L(\theta, a) = \begin{cases} K_0(\theta - a) & \text{if} \quad a \le \theta \\ K_1(a - \theta) & \text{if} \quad a > \theta \end{cases} \tag{8}$$

If one regards over– and underestimation as being of equal (relative) importance, the loss function reduces to

$$L(\theta, a) = c\left|\theta - a\right| \tag{9}$$

Because the true 'state of nature' is not known (otherwise no decision would be required) the *actual* loss incurred will be unknown too. The usual way to handle this problem is to consider the 'average' or 'expected' loss incurred. Averaging over $\mathbf{X}$ alone leads to the classical (frequentist) notion of a 'risk' associated with a decision rule.

Definition: The *risk function* for the estimator *T* is defined as the expected value of the loss function,

$$R(\theta, T) = E\left[L(\theta, T(\mathbf{X}))\right] = \int_{\mathcal{X}} L(\theta, T(\mathbf{x}))\ f(\mathbf{x}|\theta)\ d\mathbf{x}. \tag{10}$$

The expectation is to be understood with respect to the distribution of $\mathbf{X} = (X_1, \ldots, X_n).$

Note that the mean squared error of an estimator, $\mathrm{MSE}_T(\theta),$ is the risk of the estimator with regard to a quadratic loss function.

Averaging over both, $\mathbf{X}$ and $\theta,$ leads to the *Bayes risk*. This approach requires the existence of a *prior distribution* for the parameter θ.

Definition: The *Bayes risk* for the estimator $T,$ with respect to the prior distribution π over the parameter space Θ, is defined as

$$r(\pi, T) = E\left[R(\theta, T)\right] = \int_{\Theta} R(\theta, T)\ \pi(\theta)\ d\theta \tag{11}$$

The expectation is to be understood with respect to the prior distribution π. Note that the Bayes risk is a number, not a function of θ. (Compare *Bayesian Statistics*.)

1.4. Sufficient Statistic

It is quite obvious that for an efficient estimation of a parameter or a function of a parameter rarely all the single sample values, $X_1,\ldots,X_n,$ have to be known, but that a few summarizing statistics (like the sample mean or the sample variance) will suffice, depending on the problem at hand. This intuitive concept can be formalized as follows.

Definition: A statistic $S(X_1,\ldots,X_n)$ is called a *sufficient statistic* for a parameter θ if the conditional distribution of $(X_1,\ldots,X_n)$ given $S = s$ does not depend on θ (for any value of s). S can also be a vector of statistics, $S = \left(S_1(X_1,\ldots,X_n),\ldots,S_k(X_1,\ldots,X_n)\right).$ In this case we say, that $S_i,\ i = 1,\ldots,k,$ are *jointly sufficient* for θ.

Though being quite intuitive the definition is not easy to work with. With the help of the *factorization theorem* however, the determination of sufficient statistics is much easier.

Theorem (Factorization Theorem): A statistic $S(X_1,\ldots,X_n)$ is a sufficient statistic for θ if and only if the joint density of $(X_1,\ldots,X_n)$ factors as

$$f(x_1,\ldots,x_n|\theta) = g(S(x_1,\ldots,x_n)|\theta)\ \ h(x_1,\ldots,x_n) \tag{12}$$

where the function h is nonnegative and does not depend on θ and the function g is nonnegative and depends on $x_1,\ldots,x_n$ only through $S(x_1,\ldots,x_n).$

1.5. Likelihood Function

The joint density function of a (random) sample $X_1,\ldots,X_n,$ is given by

$$f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{13}$$

Read in the usual way, $x_1,\ldots,x_n$ are mathematical variables, and θ is a fixed (but unknown) parameter value, which gave rise for the observations at hand. Turned the other way around, given that $\mathbf{x} = (x_1, \ldots, x_n)$, the function is called the *likelihood function*,

$$l(\theta | x_1, \ldots, x_n) = f(x_1, \ldots, x_n | \theta). \tag{14}$$

Frequently the (natural) logarithm of the likelihood function, $\ln l(\theta | x_1, \ldots, x_n)$, called the *log-likelihood function* is easier to work with.

The (log-) likelihood function is used to compare the plausibility of various parameter values, given the observations, $x_1, \ldots, x_n$, at hand. The most plausible value, the *maximum likelihood value*, plays a prominent role in parameter estimation (cf. Section 3.2).

What makes the likelihood function so important in parameter estimation is the fact, that it 'adjusts itself' even to rather complex *observational* situations. Consider for example the situation where fixed portions of the *sample space* χ are excluded from observation (called 'Type-I censoring'), a situation quite often encountered in reliability or survival analysis. Only failures in the interval $[a, b]$, for instance, are observed, failures smaller than a or larger than b are not observed, though we know their number, r and s, respectively. In this case the likelihood function is given by

$$l_1(\theta | \mathbf{x}) \propto \left[\int_{-\infty}^{a} f(x|\theta)\, dx \right]^r \cdot \prod_{i=r+1}^{n-s} f(x_{(i)} | \theta) \cdot \left[\int_{b}^{\infty} f(x|\theta)\, dx \right]^s \tag{15}$$

where $x_{(i)}$ denotes the *i*-th largest observation. A similar situation arises, when fixed portions of the *sample* are excluded from observation (called 'Type-II censoring'). If the smallest r and the largest s observations are excluded, the likelihood function is given by

$$l_2(\theta | \mathbf{x}) \propto \left[\int_{-\infty}^{x_{(r)}} f(x|\theta)\, dx \right]^r \cdot \prod_{i=r+1}^{n-s} f(x_{(i)} | \theta) \cdot \left[\int_{x_{(n-s)}}^{\infty} f(x|\theta)\, dx \right]^s. \tag{16}$$

Note that there is a fundamental difference. In the second case r and s are predetermined values, whereas in the first case they are to be observed as well (but they enter the likelihood function as if they were given in advance). In a certain sense the likelihood function adjusts itself to the different observational situations.

Apart from the difference mentioned above, the two likelihood functions are quite similar in appearance with respect to the parameter θ, being a parameter of the underlying stochastic model $f(x|\theta)$. So we could expect the conclusions drawn to be quite similar too. If, *by chance*, r and s coincide for the two cases, and

$a = x_{(r)}, b = x_{(n-s)}$, the conclusions drawn, though based on different observation schemes, should even be identical.

The foregoing example illustrates some aspects of a more general principle.

Likelihood Principle: If $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are two samples such that

$$l(\theta | \mathbf{x}) = C(\mathbf{x}, \mathbf{y})\, l(\theta | \mathbf{y}) \quad \text{for all } \theta \tag{17}$$

where $C(\mathbf{x}, \mathbf{y})$ is a constant not depending on θ, then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

1.6. Distributional Classes

Estimation procedures for distributions sharing some structural properties turn out to be quite similar. Moreover, the finding of 'optimal' estimators (and the demonstration of their optimality) becomes easier, if we can rely on certain properties of the underlying distribution. This is the purpose of the following definitions which cover a wide range of practically important distributions. The estimation problem for distributions not covered by any these classes usually will be more difficult.

1.6.1. (Log-) Location-Scale-Families

A cumulative distribution function (cdf) F is said to belong to the *location-scale-family* (LSF), if it can be written as

$$F(x; \mu, \sigma) = F_0\left(\frac{x-\mu}{\sigma}\right) \tag{18}$$

where $F_0$is a *base* (or *reduced*) cdf (not depending on parameters); μ is a *location parameter* (not necessarily the mean of X) and σ a *scale parameter* (not necessarily the standard deviation of X). The most important members of this class are the *normal* distributions, where the base is the cdf of the standard normal distribution, Φ.

The density function of a member of a LSF-class can be written as

$$f(x; \mu, \sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right) \tag{19}$$

where f_0 is the density function corresponding to F_0.

The cdf F is said to belong to the *log-location-scale-family* (LLSF), if it can be written as

$$F(x;\mu,\sigma) = F_0\left(\frac{\ln(x)-\mu}{\sigma}\right) \qquad (20)$$

F_0 again is the base (or reduced) cdf, and μ,σ are location and scale parameters, respectively. Now these terms are related to $\ln(X)$ instead of X. Important members of this class are the *lognormal* and the *Weibull* distributions.

The density function of a member of a LLSF-class now be written as

$$f(x;\mu,\sigma) = \frac{1}{x\sigma} f_0\left(\frac{\ln(x)-\mu}{\sigma}\right) \qquad (21)$$

where f_0 is the density function corresponding to F_0.

There are, however, practically important distributions not belonging to these classes; the *gamma* distributions, for instance, are neither LSF nor LLSF.

-
-
-

Bibliography

Bard, Y. (1974): *Nonlinear parameter Estimation*, New York: Academic Press.[Comprehensive and application oriented text for fitting models to data by different estimation methods]

Casella, G. and Berger, R.L. (1990): *Statistical Inference*, Pacific Grove: Wadsworth & Brooks/Cole. [Clear written introduction to the principles of data reduction, point estimation, hypotheses testing, and interval estimation and decision theory]

Hahn, G.J. and Meeker, W.O. (1991): *Statistical Intervals- A Guide for Practitioners*, New York: Wiley. [Application oriented presentation of confidence intervals, prediction intervals, and tolerance intervals]

Lehmann, E.L.(1993): *Theory of Point Estimation*, New York: Wiley. [High level text focusing on mathematical aspects of point estimation]

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974): *Introduction to the Theory of Statistics*, New York: McGraw-Hill. [Well written introduction to the methods of statistical estimation and other statistical techniques]

Pestman, W.R. (1998): *Mathematical Statistics-An Introduction*, Berlin: W. de Gruyter. [Well written basic text on mathematical aspects of statistics]

Biographical Sketches

Werner Gurker Born March 18, 1953, at Mauthen in Carinthia, Austria. Studies in engineering mathematics at the Technische Hochschule Wien. Receiving a Dipl.-Ing. degree in engineering mathematics in 1981. Dissertation in mathematics and Doctor of engineering science degree in 1988. Assistant professor at the Technische Hochschule Wien since 1995. Main interest and publications in statistical calibration and reliability theory.

Reinhard Viertl born March 25, 1946, at Hall in Tyrol, Austria. Studies in civil engineering and engineering mathematics at the Technische Hochschule Wien. Receiving a Dipl.-Ing. degree in engineering mathematics in 1972. Dissertation in mathematics and Doctor of engineering science degree in 1974. Appointed assistant at the Technische Hochschule Wien and promotion to University Docent in 1979. Research fellow and visiting lecturer at the University of California, Berkeley, from 1980 to 1981, and visiting Docent at the University of Klagenfurt, Austria in winter 1981 - 1982. Since 1982 full professor of applied statistics at the Department of Statistics, Vienna University of Technology. Visiting professor at the Department of Statistics, University of Innsbruck, Austria from 1991 to 1993. He is a fellow of the Royal Statistical Society, London, held the Max Kade fellowship in 1980, and is founder of the Austrian Bayes Society, member of the International Statistical Institute, president of the Austrian Statistical Society from 1987 to 1995. Invitation to membership in the New York Academy of Sciences in 1998. Author of the books *Statistical Methods in Accelerated Life Testing* (1988), *Introduction to Stochastics* in German language (1990), *Statistical Methods for Non-Precise Data* (1996). Editor of the books *Probability and Bayesian Statistics* (1987), *Contributions to Environmental Statistics* in German language (1992). Co-editor of a book titled *Mathematical and Statistical Methods in Artificial Intelligence* (1995), and co-editor of two special volumes of journals. Author of over 70 scientific papers in algebra, probability theory, accelerated life testing, regional statistics, and statistics with non-precise data. Editor of the publication series of the Vienna University of Technology, member of the editorial board of scientific journals, organiser of different scientific conferences.