Point Estimation Methods with Applications to Item Response Theory Models

F Bartolucci and L Scrucca, Università degli Studi di Perugia, Perugia, Italy

Glossary

Bayesian inference – A method of statistical inference which exploits both the information arising from the data, summarized by the likelihood function, and the experimenter's belief before observing the data, represented by the prior distribution. The name Bayesian comes from the use of the Bayes' theorem, from the work of Reverend Thomas Bayes, in the inferential process.

Estimator/estimate – An estimator is any sample statistic which is used to estimate an unknown population parameter; the sample mean is an example of estimator of the population mean. An estimate is the value of the estimator that is obtained for an observed sample.

Item response theory (IRT) – A body of theory concerning models for the analysis of data collected by the administration of test items aimed at measuring a certain latent trait of interest, such as the ability of a student in a certain field. IRT models are based on the assumption that the probability that a subject responds correctly to an item is a function of the latent trait.

Likelihood function – The likelihood function is the probability or the density of the observed sample expressed as a function of the parameters of interest. It plays a central role in statistical inference and, in particular, for the maximum likelihood estimation method.

Parameter – A parameter is an index used to represent a certain population characteristic. The mean of a variable, such as the income of every resident in a certain region, is an example of parameter. Typically, the parameter is unknown and, then, it has to be estimated. In the classical approach to inference, the parameter is a fixed unknown constant. On the contrary, in Bayesian inference, the parameter is a random variable with a probability distribution representing the uncertainty on its knowledge.

Point estimation – A statistical inference method consisting of assigning a single value, called point estimate, to each unknown parameter.

Random sample – A subset of the population units which is selected by a random mechanism. It is the basis of any method of statistical inference.

Statistic – Any mathematical function used to summarize the information in the sample data.

Statistical inference – The process of using the information in a sample to draw conclusions about the population from which the sample is drawn.

Introduction

The general goal of statistical inference is to provide some conclusions about the distribution of a variable of interest (X) in a certain population on the basis of a random sample, that is, a subset of the population units randomly chosen. Often, the interest is on a particular statistical index describing the distribution of X, which is called parameter (θ); examples are the average income of the population living in a certain region or the variance of the number of employees in the population (in the sense of set) of firms operating in a certain field. In a parametric setting, we assume that the distribution of X is described by a certain statistical model, that is, a family of distributions indexed by the parameter θ, $\{f(x;\theta), \theta \in \Theta\}$, where $f(x;\ \theta)$ denotes a probability mass function (pmf) or a probability density function (pdf) and Θ is the parameter space. Note that θ may also be a vector of parameters of suitable dimension; however, we will not use a special notation for this case. Provided that the statistical model holds, knowledge of the true value of θ is equivalent to the full knowledge of the distribution of interest.

In this setting, a random sample of size n may be defined as the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, whose components are mutually independent and identically distributed with marginal distribution $f(x;\ \theta_0)$, where θ_0 denotes the true value of the parameter of interest. Note that, in the presence of a finite population, sample units have to be drawn with replacement in order to have independence between the random variables X_i, $i = 1, \ldots, n$. Another important concept for what follows is that of sample space, which is the set of all the possible samples of size n. This set may be discrete or continuous according to the nature of the support of X. An observed sample is denoted by $x = (x_1, \ldots, x_n)$. Because of the independence of the sample units, its pmf or pdf is simply given by $f(x) = \prod_{i=1}^{n} f(x_i)$. The information on θ contained in the sample is typically summarized by a sample statistic, which is defined as a mathematical function of the random sample, that is, $T = t(\boldsymbol{X})$. Examples of sample statistics are the sample mean, $\bar{X} = \sum_{i=1}^{n} X_i/n$, and the sample variance, $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$. Since $\boldsymbol{X}$ is a random vector, any sample statistic is a random variable with a distribution which is referred to as the sampling distribution.

Different inferential methods are available in the statistical literature. For a unitary description of these methods and related approaches, see Barnett (1999) and Casella and Berger (2002). In the following, we deal, in particular, with point estimation, which consists of assigning to θ a single value out from the parameter space.

Point Estimation Methods

An estimator of the unknown parameter θ may be defined as any sample statistic $T = t(\boldsymbol{X})$ used to estimate θ. An estimate $t = t(x)$ is the value of the estimator for an observed sample. The statistical theory on point estimation mainly focuses on methods of evaluating estimators and methods of finding estimators.

Evaluation of an Estimator

Methods of evaluating an estimator are necessary because, in principle, different estimators can be used for the same parameter.

The most well-known criterion to compare different estimators is based on the mean squared error (MSE) which, for an estimator T of θ, is defined as the expected value of its (squared) error, that is, $\mathrm{MSE}_\theta(T) = \mathrm{E}_\theta[(T-\theta)^2]$. According to this criterion, the estimator T is better than another estimator T^* if the MSE of the first is uniformly smaller than that of the second, that is, $\mathrm{MSE}_\theta(T) \leq \mathrm{MSE}_\theta(T^*)$ for all $\theta \in \boldsymbol{\Theta}$, with at least one value of θ for which the inequality strictly holds. In this case, we say that T is more efficient than T^*. Note that other measures of error, such as $\mathrm{E}_\theta(|T - \theta|)$, are possible to evaluate the performance of an estimator. However, MSE is usually preferred because it is easier to treat analytically. Moreover, it may be decomposed as follows:

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + \mathrm{B}_\theta(T)^2$$

where $\mathrm{Var}_\theta(T) = \mathrm{E}_\theta\{[T - \mathrm{E}_\theta(T)]^2\}$ is the variance of T and $\mathrm{B}_\theta(T) = \mathrm{E}_\theta(T) - \theta$ is its bias. The first is a measure of dispersion of the distribution of the estimator around its mean and the second is a measure of its systematic error. Both quantities need to be small in magnitude in order to have a small MSE.

It has to be clear that comparing two estimators on the basis of the MSE is not always possible since it may happen that, between T and T^*, the first is better for certain values of θ, whereas the second is better for other values of θ. Then, in real problems, we cannot expect to find the most efficient estimator, that is, the estimator which has the minimum MSE, for all θ in $\boldsymbol{\Theta}$, among all the possible estimators of the same parameter.

The indeterminacy problem described above is usually dealt with by restricting the class of possible estimators to that of the unbiased estimators. An estimator T of θ is said to be unbiased if its expected value is always equal to the parameter value or, equivalently, it always has null bias; in symbols,

$$\mathrm{E}_\theta(T) = \theta \text{ or } \mathrm{B}_\theta(T) = 0, \forall\theta \in \Theta$$

Examples of unbiased estimators are $\bar{X}$, when the parameter of interest is the population mean (μ), and S^2, when the parameter of interest is the population variance (σ^2).

For any unbiased estimator, the MSE is equal to its variance and then, within the class of the unbiased estimators of the same parameter, that with minimum MSE corresponds to the uniformly minimum variance unbiased estimator (UMVUE), that is, the unbiased estimator T, such that

$$\mathrm{Var}_\theta(T) \leq \mathrm{Var}_\theta(T^*), \ \forall\theta \in \Theta$$

for any other unbiased estimator T^* of θ. Different methods are available in the statistical literature to check when an estimator of a certain parameter is the UMVUE. Fundamental results on this topic are represented by the Cramer–Rao inequality, the Rao–Blackwell theorem, and the Lehman–Scheffé theorem. For a detailed illustration of these results and the related theory, see Lehmann and Casella (1998).

Statistical theory on point estimation is also focused on the asymptotic properties of an estimator, which concern its performance when the sample size grows to infinity. The most important of these properties is known as consistency. In particular, an estimator T of θ is consistent if

$$\lim_{n\to\infty} P_\theta(|T_n - \theta| < \epsilon) = 1, \quad \forall\epsilon > 0$$

where the subscript n has been added to T in order to recall that the distribution of the estimator depends on the sample size. Intuitively, this property says that as the sample size grows and, then, the available amount of information increases, we expect the estimator to attain values closer and closer to the true parameter value. Related properties are those of asymptotic unbiasedness and consistency in MSE. We refer to Cox and Hinkley (1974) and Lehmann (1999) for a detailed description.

Methods of Finding Estimators

Often, intuition can lead us to find an estimator for the parameter of interest. However, more methodical ways of finding estimators are needed. In the following, we discuss three general methods: method of moments, maximum likelihood method, and, under a different inferential approach, Bayesian estimation. The first two methods are formulated following a frequentist approach, which only exploits the information contained in the sample. This is in contrast with the Bayesian approach, which also exploits *a priori* information, that is, the knowledge or subjective opinions about the parameter of interest before any data have been observed.

Method of Moments

One of the oldest and simplest methods of finding estimators is the method of moments, which dates back to K. Pearson in the late 1800s. Suppose that the random sample is drawn from a statistical model depending on a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$ of dimension k. This method consists of equating the first k population moments to the corresponding sample moments and solving the resulting system of equations with respect to θ. The population and the sample moments of order r are defined, respectively, as

$$\mu_r(\theta) = \int_{-\infty}^{+\infty} x^r f(x;\theta)\mathrm{d}x \text{ and } M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

Thus, the system of equations to be solved with respect to θ is

$$\begin{gathered}\mu_1(\theta) = M_1\\ \vdots\\ \mu_k(\theta) = M_k\end{gathered}$$

In general, the estimators obtained by the method of moments are consistent, asymptotically unbiased, and have asymptotic normal distribution. However, their efficiency can usually be improved upon. For a detailed description of this method see Bowman and Shenton (1985).

Maximum Likelihood Method

The maximum likelihood method is the most popular technique for deriving estimators. It is based on the likelihood function which, for an observed sample $\boldsymbol{x}$, is defined as the probability (or density) of $\boldsymbol{x}$ expressed as a function of θ; in symbols

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

This function provides a measure of plausibility of each possible value of θ on the basis of the observed data. Then, the method at issue consists of estimating θ through the value of θ which maximizes $L(\theta)$ since this corresponds to the parameter value for which the observed sample is most likely. The estimate found in this way, that is,

$$\hat{\theta} = \hat{\theta}(\boldsymbol{x}) \text{ such that } L(\hat{\theta}) = \sup_{\theta\in\Theta} L(\theta)$$

is the maximum likelihood estimate (mle) of θ. When expressed as a function of the random sample $\boldsymbol{X}$, we have the maximum likelihood estimator (MLE) $\hat{\theta}(\boldsymbol{X})$. Obviously, this method of finding estimators is in agreement with the likelihood principle, which says that the inferential conclusions on θ based on two different sampling schemes must be the same when these schemes give rise to proportional likelihood functions; see Casella and Berger (2002) for a formal definition of this principle.

In order to find the mle of θ, we have to solve an optimization problem. It is usually simpler to maximize the log-likelihood

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

instead of $L(\theta)$. Since the logarithmic function is monotonic increasing, the two maximization problems are equivalent. In the single-parameter case ($k = 1$), the usual procedure to maximize $\ell(\theta)$ is based on solving the likelihood equation

$$\frac{\partial\ell(\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{f'(x_i;\theta)}{f(x_i;\theta)} = 0$$

In order to be sure that the found root is the global maximum of $\ell(\theta)$, we also have to verify that

$$\frac{\partial^2\ell(\theta)}{\partial\theta^2} = \sum_{i=1}^{n} \frac{f''(x_i;\theta)f(x_i;\theta) - f'(x_i;\theta)f'(x_i;\theta)}{f(x_i;\theta)^2} < 0$$

and evaluate $\ell(\theta)$ at the boundary of the parameter space.

In the multiparametric case, in which $\theta = (\theta_1, \ldots, \theta_k)$ with $k > 1$, the problem is usually more complex as it is based on the solving the system of linear equations

$$\frac{\partial\ell(\theta)}{\partial\theta_i} = 0, \quad i = 1, \ldots, k$$

An explicit solution for this system is seldom available and, therefore, we need to use iterative algorithms to maximize the log-likelihood. Starting from an initial guess, these algorithms update θ in an appropriate way until convergence, that is, until a stationary point of $\ell(\theta)$ is found. The most popular of these algorithms is known as the Newton–Rapshon algorithm. At the $(t + 1)$-th iteration, it updates θ as

$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{J}(\theta^{(t)})^{-1} s(\theta^{(t)})$$

where $\theta^{(t)}$ is the current value of the parameter, $s(\theta)$ is the score vector, that is, the first derivative vector of $\ell(\theta)$, whereas $\mathcal{J}(\theta)$ is the observed information matrix, that is, the second derivative matrix of $\ell(\theta)$ with negative sign. A variant of the Newton–Raphson algorithm is the Fisher-scoring algorithm which, at each iteration, updates θ as

$$\theta^{(t+1)} = \theta^{(t)} + I(\theta^{(t)})^{-1} s(\theta^{(t)})$$

where $I(\theta)$ is the Fisher information, that is, the expected value of the observed information matrix. Using the observed or the Fisher information matrix is equivalent if $f(x;\ \theta)$ belongs to the regular exponential family. Through these algorithms it is usually possible to find a local maximum of the log-likelihood; however, in general, it is not guaranteed that this also corresponds to the global maximum. When there are more than one local maxima, a crucial point is that of the choice of the starting values for the algorithm; often, the method of moments is used to obtain reasonable starting values.

One of the main reasons of the great popularity of the maximum likelihood method is that the resulting estimator has many interesting properties. These properties hold under certain regularity conditions which are not very restrictive, albeit quite technical. In the following, we provide a brief summary of these properties; for a detailed description we refer to Cox and Hinkley (1974), Azzalini (1996), Lehmann and Casella (1998), Lehmann (1999), and Severini (2000).

One of the most important properties is the invariance property, according to which if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. Note that this property also holds when $\tau(\theta)$ is not a one-to-one function of θ. Moreover, the maximum likelihood method is in agreement with the so-called sufficiency principle. This is because the estimator depends on the sample only through a sufficient statistic for θ, which represents the only relevant information on the parameter; thus, if θ is the only parameter of interest, further information contained in the sample can be discarded. Another property is that if an UMVUE exists and suitable regularity conditions hold, then the MLE coincides with this estimator.

The previous properties are finite-sample properties. Estimators formulated with the maximum likelihood method also have interesting asymptotic properties. First, $\hat{\theta}_n$ is a consistent estimator of θ. The consistency of the MLE also holds for the estimator $\tau(\hat{\theta})$ of $\tau(\theta)$. Moreover, the MLE is an asymptotically efficient estimator, in the sense that its variance tends to the lower bound of the Cramér–Rao inequality, and, for large samples, the standard deviation of the MLE may be approximated by the square root of the diagonal elements of the inverse of the information matrix. Finally, a very important property is that the estimator has asymptotic normal distribution. This allows us to easily construct confidence intervals and to test hypotheses on the parameter of interest.

Bayesian Estimators

In the frequentist approach, the parameter θ is considered an unknown, but fixed, quantity and only the information coming from the sampling data is relevant for inference. Thus, we do not take into account the prior belief on the parameter. On the contrary, in the Bayesian approach, the parameter θ is considered as a random variable with a certain probability distribution, referred to as the prior distribution, whose role is that of representing the experimenter's belief before observing the data. Within this approach, the prior distribution is updated on the basis of the likelihood function through the Bayes' theorem. The resulting distribution is referred to as the posterior distribution and summarizes the information in both the prior distribution and in the data.

Let $\pi(\theta)$ denote the prior distribution and let $f(\boldsymbol{x}\,|\,\theta)$ denote the conditional distribution of the sample given θ. Note that $f(\boldsymbol{x}\,|\,\theta)$ is equivalent to the likelihood function discussed earlier. According to the Bayes' theorem, we can derive the posterior distribution of θ on the basis of the observed sample $\boldsymbol{x}$ as

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

where $m(\boldsymbol{x})$ is the marginal pdf (or pmf) of the data, which is given by

$$m(\boldsymbol{x}) = \int_{-\infty}^{+\infty} f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta$$

The mean, which is often used to summarize the posterior distribution of the parameter, is a Bayesian point estimator of θ. Sometimes, the mode or the median of the posterior distribution is used instead.

A great variety of textbooks on Bayesian inference now exist. For an accessible introduction we suggest Ghosh *et al.* (2006), whereas for more advanced treatment, we suggest Gelman *et al.* (2003) and Robert and Casella (2005) for related computational methods.

Item Response Theory Models

Item response theory (IRT) models are tailored to the analysis of data arising from the administration of a questionnaire made of a series of items which measure a common latent trait, that is, a characteristic of the human being which is not directly observable; for a review see the work by Hambleton and Swaminathan (1996). The main application of these models is in educational assessment, where the latent trait corresponds to a certain type of ability of an examinee.

In the following, after a brief summary of the main assumptions of IRT models for dichotomously scored items, we focus on maximum likelihood methods for their estimation. For an updated and detailed description of these methods, see Baker and Kim (2004).

Suppose that a questionnaire made of r items is administered to a sample of n subjects, and let y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, r$, be a categorical response variable for the jth item administered to the ith subject. Depending on the context, y_{ij} will denote a random variable or one of its realizations. In educational assessment, the response variable is often binary; in particular, y_{ij} is equal to 1 if subject i responds correctly to item j and to 0 otherwise.

With reference to the binary case, the main assumptions of IRT models may be summarized as follows:

- *Unidimensionality*. For each subject i, the responses to the r items depend on the same latent parameter ϕ_i, which is unidimensional.
- *Local independence*. For each subject i, the responses to the r items are independent given ϕ_i.
- *Monotonicity*. The probability of a correct response of the ith subject to the jth item, in symbols $p_j(\phi_i) = p(y_{ij} = 1 \mid \phi_i)$, is a monotonic increasing function of ϕ_i.

The probabilities $p_j(\phi_i)$ are modeled via a function which is usually referred to as the item characteristic curve (ICC) or item response function. Different models arise according to the assumed parametrization of the ICC. The most well known is the one parameter logistic (1PL) parametrization, according to which

$$p_j(\phi_i; \beta_j) = \frac{e^{\phi_i - \beta_j}}{1 + e^{\phi_i - \beta_j}}$$

where β_j is the difficulty level of item j. The resulting model is usually referred to as the Rasch model (Rasch, 1960). More general parametrizations were proposed by Birnbaum (1968) and rely on further parameters to describe the item characteristics.

Maximum Likelihood Estimation

In the following, we describe three different methods based on the maximum likelihood paradigm for the estimation of the above IRT models for dichotomously scored items.

Joint maximum likelihood method

The joint maximum likelihood (JML) method consists of maximizing the likelihood of the model, corresponding to the probability of the observed data matrix, with respect to the ability and item parameters jointly.

Under the assumption of local independence, the joint distribution of the response vector $y_i = (y_{i1}, \ldots, y_{ir})$ may be expressed as

$$p(y_i|\phi_i, \psi) = \prod_j p_j(\phi_i; \psi_j)^{y_{ij}} [1 - p_j(\phi_i; \psi_j)]^{1 - y_{ij}}$$

where ψ is the vector of the item parameters, which depend on the IRT model of interest. Under the 1PL model, it only contains the parameters β_j.

Assuming that the response vectors for the subjects in the sample are independent each other, the conditional probability of observing the response matrix Y with elements y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, r$, is equal to

$$p(Y|\phi, \psi) = \prod_i \prod_j p_j(\phi_i; \psi_j)^{y_{ij}} [1 - p_j(\phi_i; \psi_j)]^{1 - y_{ij}}$$

where ϕ is the vector of the ability parameters. Then, for an observed matrix of responses Y, the JML method consists of estimating the model parameters by maximizing the joint likelihood

$$L_{\mathcal{J}}(\phi, \psi) = p(Y|\phi, \psi)$$

with respect to ϕ and ψ.

For the Rasch model, the joint likelihood becomes

$$L_{\mathcal{J}}(\phi, \psi) = \frac{e^{\sum_i t_i \phi_i - \sum_j s_j \beta_j}}{\prod_i \prod_j (1 + e^{\phi_i - \beta_j})}$$

where $t_i = \sum_j y_{ij}$ is the number of correct responses provided by subject i and $s_j = \sum_i y_{ij}$ is the number of subjects who responded correctly to item j. The corresponding log-likelihood is equal to

$$\ell_{\mathcal{J}}(\phi, \psi) = \sum_i t_i \phi_i - \sum_j s_j \beta_j - \sum_i \sum_j \log(1 + e^{\phi_i - \beta_j})$$

For each IRT model, the joint likelihood is invariant with respect to certain transformations of the item and ability parameters. Then, suitable constraints have to be put on these parameters in order to make the model identifiable. The following constraints may alternatively be used for the Rasch model:

- $\beta_1 = 0$: in this way, the first item is taken as a reference item and, then, β_j, $j = 2, \ldots, r$, has to be interpreted as the difficulty of item j with respect to the first one.
- $\sum_i \phi_i = 0$: in this way, the average ability of the subjects is fixed at 0 and ϕ_i, $i = 1, \ldots, n$, is interpreted as the ability of subject i with respect to the average of the group.

These constraints are equivalent in the sense that the maximum value of the likelihood that may be reached under each of them is the same, which is also equal to the unconstrained maximum of the likelihood. Further constraints need to be used for more sophisticated models.

The algorithm to maximize $\ell_{\mathcal{J}}(\phi, \psi)$ that is commonly used is based on a series of Newton–Raphson steps to be performed until convergence. Explicit expressions for the derivatives used in these steps are available; see Hambleton and Swaminathan (1996) and Baker and Kim (2004). We note that the point at convergence does not depend on the starting values chosen for the parameters since, provided that the JML estimate exists, the log-likelihood is a strictly concave function over the full parameter space.

It has to be recalled that the JML estimate is not ensured to exist. A set of conditions on the matrix Y which ensures this existence was derived by Fischer (1981). A necessary condition is that there exist neither subjects who respond correctly or incorrectly to all items nor items to which all subjects respond correctly or incorrectly. These subjects and items must be removed from the dataset when they exist. Moreover, the JML estimator is not consistent as n grows to infinity with r fixed. This does not mean that the estimates obtained by using this method are unreliable. Indeed, if the number of subjects and that of items are large enough, the bias of the JML estimator is expected to be low.

Conditional maximum likelihood

The conditional maximum likelihood (CML) method may be only applied to the Rasch model and it is typically used to estimate its difficulty parameters. The method is based on the maximization of the conditional likelihood of these parameters given a set of minimal sufficient statistics for the ability parameters.

Let $\beta = (\beta_1, \ldots, \beta_r)$ denote the vector of item parameters under the Rasch model. First of all, consider that the conditional probability that subject i attains score t_i, given the ability parameter ϕ_i, is equal to

$$p(t_i|\phi_i,\beta) = \sum\nolimits_{y\in\mathcal{Y}(t_i)} p(y_i = y|\phi_i) = \frac{e^{\phi_i t_i}}{\prod_j (1+e^{\phi_i-\beta_j})} q_i$$

where $q_i = \sum_{y\in\mathcal{Y}(t_i)} \exp\{-\sum_j y_j \beta_j\}$, with $\mathcal{Y}(t)$ denoting the set of all the binary vectors $y = (y_1, \ldots, y_r)$ with elements having sum t. These vectors have t elements equal to 1 and $r - t$ elements equal to 0. Consequently, for each subject i, the conditional probability of the response configuration y_i, given the sufficient statistics t_i, is equal to

$$p(y_i|t_i,\phi_i,\beta) = \frac{p(y_i|\phi_i,\beta)}{p(t_i|\phi_i,\beta)} = \frac{e^{-\sum_j y_{ij}\beta_j}}{q_i} \qquad [1]$$

which does not depend on ϕ_i and, then, may be denoted by $p(y_i|t_i,\beta)$.

The conditional likelihood exploited within the CML method is

$$L_C(\beta) = \prod_i p(y_i|t_i,\beta)$$

Note that the subjects who respond correctly or incorrectly to all the items do not contribute to this likelihood. Then, the corresponding log-likelihood may be expressed as

$$\ell_C(\beta) = -\sum_j s_j^* \beta_j - \sum_i d_i \log q_i \qquad [2]$$

where $d_i = 1\{0 < t_i < r\}$ is a dummy variable equal to 1 if subject i contributes to the likelihood and to 0 otherwise and $s_j^* = \sum_i d_i y_{ij}$. An identifiability problem also arises in this case; for this reason, we must use the constraint $\beta_1 = 0$ or, alternatively, $\sum_j \beta_j = 0$.

In order to maximize $\ell_C(\beta)$, we may use a Newton–Raphson algorithm. Efficient methods to compute the conditional log-likelihood and its derivatives, in a way which is viable even when r is large, are available in the statistical and psychometric literatures (Formann, 1986; Gustafson, 1980; Liou, 1994).

Certain conditions need to be fulfilled for the existence of the CML estimate (Fischer, 1981). In most situations, a condition which ensures this existence is that there do not exist items to which no subjects or all subjects respond correctly. Finally, we have to stress that the main advantage of the CML method is that the resulting estimator is consistent for r fixed as n grows to infinity. For a detailed description of the asymptotic properties of this estimator, see Andersen (1970); see also Andersen (1972).

Marginal maximum likelihood

The marginal maximum likelihood (MML) method is tailored to IRT models formulated under the assumption that the ability is a random parameter. The method consists of maximizing the likelihood corresponding to the manifest probability of the observed responses, that is, the marginal probabilities of these responses once the ability parameters have been integrated out.

Let $f(\phi_i)$ denote the distribution for the ability parameters. The manifest distribution of y_i is given by

$$p(y_i|\psi) = \int_{\Re} p(y_i|\phi_i,\psi) f(\phi_i) \mathrm{d}\phi_i$$

Moreover, the manifest probability of Y is $p(Y|\psi) = \prod_i p\ (y_i|\psi)$ and, for an observed matrix of responses, it corresponds to the marginal likelihood on which the MML method is based. This is denoted by $L_M(\psi, \eta)$, where η is the vector of parameters on which $f(\phi_i)$ possibly depends. The corresponding log-likelihood is denoted by $\ell_M(\psi,\eta)$. The latter is maximized by a version of the EM algorithm of Dempster *et al.* (1977) which is based on the complete likelihood, that is, the likelihood that we could compute if we knew the ability level of each subject. The MML estimator of the item parameters is denoted by $\hat{\psi}_M$ and that of the parameters for the distribution of the ability is denoted by $\hat{\eta}_M$.

A crucial assumption for the implementation of the MML method concerns the distribution of the ability. The most common assumption is that the ability has a normal distribution with unknown mean and variance, which are estimated together with the item parameters. More advanced solutions consist of assuming that the ability has a discrete distribution with a suitable number of support points and weights. The first paper discussing this kind of approach is that of Bock and Lieberman (1970). Further developments are due to Bock and Aitkin (1981), Thissen (1982), Lindsay *et al.* (1991), Pfanzagl (1993), and Schilling and Bock (2005).

Example

In order to provide an illustration of the maximum likelihood estimation methods outlined above, we describe an application based on a dataset which was provided by the Educational Testing Service. The dataset concerns the responses of 1510 students to 12 publicly released items on mathematics collected in 1996 within a project called National Assessment of Educational Progress (NAEP); for details, see Bartolucci and Forcina (2005).

By using the JML method, for the Rasch model we obtained the estimates of the item and ability parameters which are displayed in **Table 1**. Note that we adopted the identifiability constraint $\beta_1 = 0$ and, consequently, we took the first item as a reference item.

The estimates of the item parameters allow us to evaluate the difficulty level of each item. In particular, we can conclude that the last item is the most difficult, whereas the fourth item is the easiest. Moreover, the standard errors and the confidence intervals may be used to assess whether the difficulty level of each item is significantly different from that of the first item. We observe that the second item is not significantly different from the first in terms of difficulty since the corresponding confidence interval includes zero. The same cannot be said about, for example, the third item. Finally, the estimates of the ability parameters can be used to assess the students to whom the questionnaire was administered. We conclude, for instance, that the second and third students have the same ability level as they provided the same number of correct responses, and this ability level is higher than that of the first student.

As an illustration of the CML method, we show in **Table 2** the estimates of the item parameters obtained with this method.

Table 1 JML estimates of the item and ability parameters of the Rasch model for the NAEP dataset

	Estimate	*SE*	*95% Conf.int.*	
β_1	0.000	–	–	–
β_2	−0.051	0.097	−0.241	0.138
β_3	0.755	0.093	0.574	0.936
β_4	−1.140	0.111	−1.357	−0.923
β_5	1.672	0.092	1.491	1.853
β_6	0.014	0.096	−0.175	0.202
β_7	0.724	0.093	0.542	0.905
β_8	1.305	0.092	1.125	1.485
β_9	0.365	0.094	0.181	0.549
β_{10}	0.574	0.093	0.391	0.756
β_{11}	2.697	0.098	2.505	2.888
β_{12}	2.751	0.098	2.558	2.944
ϕ_1	−0.080	0.674	−1.400	1.241
ϕ_2	1.193	0.662	−0.104	2.491
ϕ_3	1.193	0.662	−0.104	2.491
ϕ_4	0.770	0.649	−0.501	2.041
ϕ_5	−0.080	0.674	−1.400	1.241
⋮	⋮	⋮	⋮	⋮
ϕ_{1510}	2.158	0.750	0.689	3.626

The results are similar to those previously obtained, except that only the estimates of the item parameters are now available. This is because the ability parameters are removed by conditioning on the corresponding sufficient statistics. By comparing the results in **Table 2** with those in **Table 1**, we note that the CML estimates are very similar to the JML estimates and, therefore, the items are equivalently ranked. This is because, with 12 items, we can also obtain reliable estimates with the JML method. To this regard we recall that, in contrast to the JML estimator, the CML estimator is in general consistent.

Finally, we applied the MML method based on a discrete latent distribution for the ability. In this case, we have to choose the number of support points of the latent distribution, which is equivalent to the number of latent classes in the population. Following a standard practice in this field, we make this choice by using the Bayesian Information Criterion of Schwarz (1978) which is based on the minimization of the index:

$$\mathrm{BIC} = -2\ell_M(\hat{\psi}, \hat{\eta}) + \log(n)\#\text{parameters}$$

In practice, we fit the model with an increasing number of classes until this index does not start do decrease. Then, we choose the number of classes of the model with the smallest BIC. For the NAEP data, this procedure leads to the results shown in **Table 3**.

We clearly choose three latent classes. A probability and an ability level are associated to each of these classes. The estimates of these parameters are shown in **Table 4**.

Table 2 CML estimates of the item parameters of the Rasch model for the NAEP dataset

	Estimate	*SE*	*95% Conf.int.*	
β_1	0.000	–	–	–
β_2	−0.047	0.092	−0.229	0.134
β_3	0.691	0.088	0.517	0.864
β_4	−1.040	0.106	−1.247	−0.833
β_5	1.521	0.088	1.349	1.693
β_6	0.013	0.092	−0.168	0.193
β_7	0.662	0.089	0.489	0.836
β_8	1.191	0.088	1.019	1.363
β_9	0.334	0.090	0.158	0.511
β_{10}	0.525	0.089	0.351	0.700
β_{11}	2.427	0.092	2.246	2.607
β_{12}	2.474	0.093	2.292	2.655

Table 3 Selection of the number of classes for the latent class Rasch model applied to the NAEP dataset

#Classes	$\ell_M(\hat{\psi}, \hat{\eta})$	*#Parameters*	*BIC*
1	−11009	12	22106
2	−10242	14	20586
3	−10166	16	20450
4	−10163	18	20458

Table 4 MML estimates of the ability parameters of the Rasch model for the NAEP dataset (three latent classes)

Class	*Ability*	*Probability*
1	−0.645	0.165
2	0.970	0.457
3	2.432	0.378

We can observe that the classes are well separated in terms of ability level. The first class, which includes subjects with the lowest ability level, is also the smallest in the population (it has the smallest probability), whereas the other two classes have a comparable size. The corresponding estimates of the item parameters are, in practice, equal to those obtained with the CML method which are displayed in Table 2. This is in agreement with the theory of Lindsay *et al.* (1991) about the equivalence between CML and MML estimation methods for the Rasch model which holds under certain regularity conditions.

See also: Ability Testing; Generalized Linear Mixed Models; Item Response Theory; Latent Class Models; Model Selection; Rasch Models.

Bibliography

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of Royal Statistical Society, B* **32**, 283–301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, B* **34**, 42–54.

Azzalini, A. (1996). *Statistical Inference Based on the Likelihood*. London: Chapman and Hall.

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques*, 2nd edn. New York: Marcel Dekker.

Barnett, V. (1999). *Comparative Statistical Inference*. New York: Wiley.

Bartolucci, F. and Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika* **70**, 31–43.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R. (eds.) *Statistical Theories of Mental Test Scores*, pp 397–479. Reading, MA: Addison-Wesley.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **35**, 179–197.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika* **35**, 179–197.

Bowman, K. O. and Shenton, L. R. (1985). Method of Moments. In Kotz, S. and Johnson, M. L. (eds.) *Encyclopedia of Statistical Sciences,* vol. 5, pp 467–473. New York: Wiley.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edn. Pacific Groove, CA New York.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Duxbury: Chapman and Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B* **39**, 1–38.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika* **46**, 59–77.

Formann, A. K. (1986). A note on the computation of the second-order derivatives of the elementary symmetric functions in the Rasch model. *Psychometrika* **51**, 335–339.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edn. New York: Chapman and Hall/CRC.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer.

Gustafson, J. E. (1980). A solution of the conditional estimation problem for long tests in the rasch model for dichotomous items. *Educational and Psychological Measurement* **40**, 377–385.

Hambleton, R. K. and Swaminathan, H. (1996). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Nijhoff.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York: Springer.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd edn. New York: Springer.

Lindsay, B., Clogg, C., and Grego, J. (1991). Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86**, 96–107.

Liou, M. (1994). More on the computation of higher-order derivatives of the elementary symmetric functions in the rasch model. *Applied Psychological Measurement* **18**, 53–62.

Pfanzagl, J. (1993). A case of asymptotic equivalence between conditional and marginal maximum likelihood estimators. *Journal of Statistical Planning and Inference* **35**, 301–307.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Intitute for Educational Reserch.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. New York: Springer.

Schilling, S. and Bock, R. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika* **70**, 533–555.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. New York: Oxford University Press.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* **47**, 175–186.