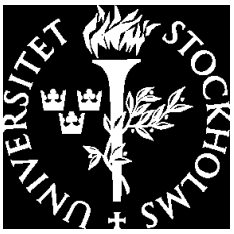


Derivation of Probability Distributions for Risk Assessment

Adam Duracz

Master´s Thesis in Computer Science at Stockholm University, Sweden 2006

Derivation of Probability Distributions for Risk Assessment

Adam Duracz

Master´s Thesis in Computer Science (20 credits)
Single Subject Courses
Stockholm University year 2006
Supervisor at Nada was Henrik Eriksson
Examiner was Stefan Arnborg

TRITA-CSC-E 2006:042
ISRN-KTH/CSC/E–06/042–SE
ISSN-1653-5715

Department of Numerical Analysis and Computer Science
Royal Institute of Technology
SE-100 44 Stockholm, Sweden

Abstract

Derivation of Probability Distributions for Risk Assessment

Risk assessment is more and more widely applied in different areas. The essence of a risk assessment is to estimate the consequences of adverse events and their probability. For a quantitative judgment of the risks, it is necessary to estimate the uncertainty of the variables that govern the events. The uncertainty is commonly expressed as probability distributions.

One of the main problems for the practical application of risk assessments is that the needed probability distributions, usually, are not readily available. These have to be derived from other existing information and knowledge. Several methods have been proposed in the literature for the derivation of probability distributions. The choice of an appropriate method depends on what information and knowledge is available. Some of these methods are used widely while others, being unconventional or still under development, have been less commonly applied.

There is a lack of a software package that integrates those of the existing distribution fitting methods that are most relevant to environmental risk assessment, facilitating the choice of one or more methods depending on the situation. This is particularly true with regard to the methods most appropriate in the field of environmental risk assessment, where data is scarce. In this work we will attempt to fill this gap by developing a prototype of such an application.

Sammanfattning

Härledning av sannolikhetsfördelningar för riskanalys

Säkerhetsanalys används i ett ökande antal områden, vilket har gjort ämnet till en viktig del av många förr orelaterade sysslor. Det centrala i en säkerhetsanalys är att uppskatta följderna av oönskade händelser och deras sannolikheter. För att göra en kvantititiv uppskattning av säkerheten behövs uppskattningar av osäkerheten i de variabler som styr händelserna. Osäkerheten uttrycks vanligen i form av sannolikhetsfördelningar.

Ett av de huvudsakliga problemen som uppstår under tillämpning av säkerhetsanalys är att de nödvändiga sannolikhetsfördelningarna vanligtvis inte är tillgängliga. Dessa måste härledas från annan information och kunskap som finns till hands. Flera metoder för härledning av sannoliketsfördelningar har föreslagits i litteraturen. Valet av metod beror på den information och kunskap som är tillgänglig. Vissa metoder används vitt medan andra okonventionella eller fortfarande mognande metoder är mindre populära.

Ingen mjukvaruapplikation på marknaden inbegriper de existernade metoder som är lämpliga för miljörelaterad säkerhetsanalys, där tillgången till data ofta är liten. En del av detta examensarbete är att skapa en prototyp av sådan applikation för att underlätta valet av en eller flera metoder beroende på önskemål.

Preface

Traditionally, risk assessment has dealt with qualitative measures of probability. Ordinal scales of risk such as {*low*, *medium*, *high*} have been used to express the probability of an event, an approach that carries two major weaknesses. First, an ordinal cannot be reliably evaluated outside of its context. For example, using the above scale, the ordinal *low* can both be used to describe the risk of suffering a heart attack for a healthy 50 year old and to describe the same risk for a newly born child. Clearly the underlying probabilities differ greatly, but because the chosen scale lacks sufficient resolution, it can produce an unreliable assessment of risk. Secondly, it is hard to perform any kind of calculations based on qualitative measures of probability. One can try to convert an ordinal scale into an interval scale by replacing the ordinals with numbers or by creating rules for combining the ordinals, but, as mentioned above, this will always include a subjective judgment and can thus not serve as a reliable method of estimating risk.

The solution to this is *quantitative* or *probabilistic* risk assessment. It replaces the ordinal scales of the qualitative approach with probabilities, the manipulation of which has a solid base in the field of probability theory. Worth mentioning, however, is that because a decision maker often proceeds based on a qualitative measure of risk, the problem of mapping ordinals to quantities remains and still plays a role in the assessment, though its use has been reduced to a minimum.

The goal of this work is to study and implement in a software application two methods for the derivation of probability distributions, as well as to identify the situations in which the methods could be used. Focus will be put on methods most suitable for the distribution fitting problems encountered during environmental impact assessment modelling, particularly of radionucleids pertaining to nuclear waste storage facilities.

This work is divided into two parts. Part 1 provides an account of the methods that were studied, and a discussion of their applicability in various circumstances. Part 2 includes examples of how these methods can be applied to commonly encountered distributionfitting problems and a description of the software tool that was developed to assist an assessor in the choice and use of the aforementioned methods.

Acknowledgements

I would like to thank my supervisors Rodolfo Avila at Facilia and Henrik Eriksson at KTH[1], my examiner Stefan Arnborg at KTH, Erik Johansson, Per-Anders Ekström and Robert Broed at Facilia[2], Joanna Tyrcha and Daniel Thorburn at SU[3], Scott Ferson at Ramas[4], Oskar Sandberg at Chalmers[5], Christoffer Gottlieb, my father Andrzej Duracz and my brother Jan Duracz for their help and patience.

[1] Royal Insitute of Technology in Stockholm `http://www.kth.se/`
[2] `http://www.facilia.se`
[3] Stockholm University `http://www.su.se`
[4] `http://www.ramas.com/`
[5] Chalmers University of Technology `http://www.chalmers.se/`

Contents

Part I

Theory

Chapter 1

Quantifying uncertainty

A key activity when performing a quantitative risk assessment is the estimation of the uncertainty present in a scrutinized variable. The height of a human, for instance, has an average value, but there are also minimum and maximum constraints on this quantity. Therefore, to fully describe its behavior one needs a function, mapping classes of heights to their respective probabilities. If we were describing the uncertainty present in the number of sunny days in a year, this function would map each integer between 0 and 365 to a probability. The major difference in these two cases is that the first quantity is continuous and the second is discrete. Thus, in the first case, it is not meaningful to talk about a probability for a given specific height, since there is an infinite number of such heights between the minimum and the maximum values, and therefore (given the classical definition of probability[1]) the probability of each height is equal to zero. Instead, when describing the uncertainty of a continuous variable, one uses the concept of a probability density function (PDF) that maps intervals of possible events to probabilities. In the discrete case, this function is called the probability mass function (PMF). We will concentrate on the continuous case in this work, since it has more applications in the field of interest.

1.1 Distribution Functions

So, the quantification of an uncertainty boils down to the choice of a distribution function. In chapter 2 we will cover how to choose this function, given various sets of available information and knowledge. First, however, we classify a number of functions that are useful when describing uncertainties.

1.1.1 Parametric

In many important cases, the underlying process generating the values of a variable can be described using a simple mathematical function. Because many variables in nature are due to the same type of underlying process, a handful of functions

has proven to be sufficient to describe an incredible amount of different quantities. These functions exhibit different shapes depending on the values you assign to their parameters. These families of functions, indexed by their parameters, are commonly referred to as parametric distribution functions (DFs). The most notable parametric DF is the Gaussian distribution. Also known as the Normal distribution or bell curve, it has a wide range of interesting properties which is why it is the most commonly used DFs. One of these properties is that, under certain conditions, an infinite sum of DFs converges to a Gaussian[2]. This is often given as a motivation for choosing a Gaussian function to describe a property which is thought to be a result of a summation of a large number of factors (each with its own DF). However the simple mathematical form 1.1 of the Gaussian, making it easy to manipulate analytically, surely has also contributed to its popularity.

$$p(x) = \frac{e^{\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{1.1}$$

It may seem counterintuitive that a quantity such as the height H of a human could be described by such a simple function. And there is a basis to this doubt, as H is the composite of many of its subclasses - the height of women, men, the height of basketball players and the height of newly born children, among many others. Each of these classes has its own mode and if we were to approximate H as the composite of only a few of them, we should be able to detect these modes in the DF of H. However, due to the near infinite number of such classes, the actual shape of the DF of H is well approximated by a Gaussian curve.

Perhaps the greatest advantage of using a parametric DF to describe the uncertainty of a variable is the way it allows you to represent your observed data in a very compact and portable form. Instead of having to deal with large collections of data points, which can exist in many different formats and are difficult to analyze and manipulate directly, one only needs to keep track of a the name of a DF and the values of its parameters. A detailed review of the most significant parametric distributions is available in Evans, Hastings and Peacock[3].

1.1.2 Nonparametric

Despite the positive qualities of parametric DFs, there are situations when the available information or the process behind the studied quantity simply do not permit the selection of a sufficiently similar (this similarity is formalized in chapter 3) parametric DF. One example is when the number of classes and the variance of the elements of each class is small, and the modes of the individual classes are far apart. Then the DF of the combined quantity will exhibit more than one mode. Because of the unimodal character of the commonly used DF, this situation requires the use of a nonparametric or empirical DF. The name nonparametric is actually somewhat misleading as these functions, in fact, do have parameters. However, as

opposed to parametric DFs, whose parameters represent important statistics of the observed data, the nonparametric DFs use the entire data set as parameters.

Perhaps the most familiar and intuitive of the nonparametric DFs is the frequency histogram, a bar chart which approximates the PDF. It can be derived from the histogram by scaling its bars to make its integral equal to unity. The frequency histogram is an invaluable visual tool when assessing the accuracy of a parametric approximation, but suffers from a number of drawbacks. Its construction requires the arbitrary assignment of bar widths and bar positions, which means that unless one has access to a very large amount of data, the shape of the DF varies heavily as the bar widths and positions are altered. Methods for inferring these parameters from data exist[4], but will not be covered in this work. Unless these methods are used, this means that the construction of a frequency diagram requires interaction from the assessor, changing the parameters of the diagram until the most likely shape can be established.

A more robust nonparametric distribution function is the empirical cumulative distribution function (ECDF). Known as the Kaplan-Meier estimate[5], the ECDF is computed by ordering the observations and plotting them against their observed quantiles. This method has the disadvantage that many relatively diverse DF shapes tend to have similar looking CDFs.

There is a simple way of obtaining the empirical PDF (EPDF) of a variable from its ECDF. It uses the fact that the CDF is the primitive function of the PDF, and thus one can obtain the PDF from the CDF by differentiating. However, since differentiation is a noise-amplifying operation, the resulting EPDF is very jagged and needs considerable smoothing for many areas of application, particularly if a plot is to be constructed to compare the EPDF with a parametric distribution fitted to the data set.

An alternate approach, which introduces the smoothing from the start and thus, under certain conditions, produces a PDF very similar to the true one is Parzen window or kernel density estimation (KDE). Here, a symmetric PDF or kernel is chosen, and a copy of this function is centered about each observed data point. The EPDF is then equal to the sum of the kernels divided by the number data points. In the case when the modeled variable is smooth, the KDE gives relatively good estimate without the need of interaction from the assessor. This method is discussed further in section 2.2.

Chapter 2

Choosing a distribution function

Choosing a function to estimate the density of interest is a multi-tiered process which should never be done automatically without the conscious scrutiny of the assessor. This is particularly true when data sets are small. No amount of positive indications, whether in the form of passed hypothesis tests or apparent graphical fit should convince the assessor to choose a distribution without first asserting its theoretical feasibility. This includes checking that known constraints such as minimum, maximum, skewness and tail size are not broken. All theoretical goodness-of-fit (GoF) tests break down as the sample size approaches zero, and even with large samples, the tests all have their drawbacks which should be carefully studied before using the tests to choose a DF.

A highly interesting method for choosing distribution functions that did not fit into the scope of this paper is Bayesian inference. This method utilizes information that may be available regarding the studied density by updating a prior distribution (based on the belief and information available before the data set was observed) with the observed data. It is particularly relevant to situations where data is in short supply due to the fact that even a single data point can be taken into account. The traditional approach for performing Bayesian inference is explored extensively in DeGroot[6]. A new, numerical method based on a modified version of Markov-Chain Monte Carlo simulation is covered in Green[7].

2.1 Parametric methods

The task of choosing a specific parametric function for a given sample is twofold. First, optimal parameters are found for each family of distributions, and then the fit of these optimal distributions is assessed to find the most appropriate distribution. We will begin by reviewing two of the most commonly used methods of finding optimal parameters for a given family of distributions.

The maximum likelihood estimate (MLE) is calculated by maximizing the likelihood function of the sample with respect to the parameters of the chosen distribution function. The likelihood of an event is an extension of the concept of a

probability that describes how likely it is to obtain the observed sample given the parameters of the scrutinized distribution function.

The parameters that maximize the likelihood of a sample are generally very good estimates of the sought parameters, and the MLE carries many desirable properties, especially for large samples. Among these is the fact that if p is a one-to-one function and $\hat{\theta}$ is the MLE of θ then $p(\hat{\theta})$ is the MLE of $p(\theta)$[2], and that as the sample size increases, the MLE converges in probability to the minimum variance unbiased estimators (MVUE[1])[2]. However, this method also bears some significant drawbacks:

1. It requires the identification of a model-specific likelihood function.
2. A solution may not exist in the parameter space S, or this maximum could be attained on the boundary of S [2].
3. If a solution exists, it only represents a stationary point. This means that the solution could correspond to a saddle point rather than a global (or even local) extremum. This entails calculating the second derivative from both sides of the solution to determine if the solution is indeed a maximum [2].
4. To ensure that the found maximum is a global one, all local maxima have to be compared [2].
5. With small samples (less than 10), the estimates can be heavily biased[2].

Another method, often used to calculate the initial values for the optimization required to find other estimates, is the least squares method (LSM). Here, we minimize the sum of the absolute square residuals of the sample:

$$R^2 = \sum [y_i - F(x_i)]^2 \tag{2.1}$$

where the x_i denote the data points and y_i the value of the ECDF at x_i.

2.2 Nonparametric methods

The construction of a histogram requires the grouping of observed values to form each bar in the diagram. If B is the number of bars and N is the number of observations, and if we call the number of possible heights a bar can assume in a diagram its resolution, then a data set of 20 observations plotted as a histogram with five bars gives a resolution of four. Just like the ability to discern detail in a digital photograph depends on its resolution, the ability to distinguish between different shapes is dependent on the resolution of the histogram. Consequently, with a data size of less than about 10–15 points a histogram loses most of its utility, remaining somewhat useful for detecting features such as skewness in the data.

[1] Defined as the estimator that has minimal variance for all possible values of the parameters.

[2] `http://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm`

The KDE approach has the disadvantage of artificially smoothing the resulting PDF, so it gives very bad approximations of functions which feature discontinuities (such as the uniform distribution, if viewed as a function on $\mathbb{R}$). However, compared to the frequency histogram, it requires only one parameter to be arbitrarily chosen – the spread parameter of the kernel. This value can be estimated from data by minimizing the approximate mean integrated square error (AMISE) of the estimator (K is the kernel):

$$AMISE(\lambda) = \frac{1}{4}\lambda^4 (\int_t t^2 K(t)dt)^2 \int_x (f''(x))^2 dx + \frac{1}{n\lambda}\int_t K(t)^2 dt \qquad (2.2)$$

2.3 The maximum entropy method

The concept of entropy was first formulated in the form commonly used today by Ludwig Boltzmann as part of the theory of thermodynamics. Claude Shannon generalized it (into what is now called Boltzmann-Gibbs-Shannon differential entropy) as a measure of the disorder[3] in a system for use in information theory in his 1948 paper *A Mathematical Theory of Communication*[8]. Assuming that $\{E_i\}$ is a set of events, that $|\{E_i\}| = n$, that $P = \{p_i\}$ is the set of probabilities corresponding to these events, and that H is defined on P, the entropy function is defined as the unique[4] function H satisfying the following three characteristics.

1. H is continuous in the p_i.
2. If for all i, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

The meaning of the third characteristic is illustrated in Figure 2.1.

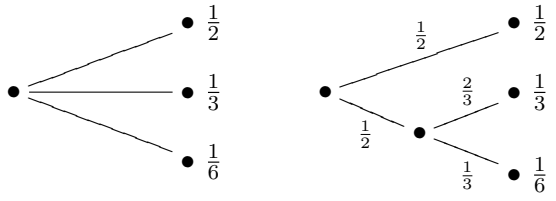


Figure 2.1. Decomposition of a choice from three possibilities.

The left diagram represents an event space containing three events E_1, E_2 and E_3, corresponding to probabilities $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}$ and $p_3 = \frac{1}{6}$, respectively. The

[3] Given probabilities for a set of events, it measures the uncertainty we have about the outcome.
[4] Up to a scalar constant, which is customarily set to one.

diagram to the right represents the situation when the occurrence of E_2 and E_3 is dependent on another event, call it E_4, with probability $p_4 = \frac{1}{2}$. An example of such a situation is if our event space consists of a simplified weather forecast, where we let E_1 represent clear skies, E_2 rainfall and E_3 snowfall. Then we could decompose E_2 and E_3 by introducing another event, E_4, representing downfall, having probability $p_4 = p_2 + p_3 = \frac{1}{2}$. Now, the third characteristic of H means that

$$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3}) \tag{2.3}$$

where the coefficient $\frac{1}{2}$ is due to the fact that the second event only occurs half of the time (it only snows or rains half of the time).

By a theorem presented in the same paper, the only function satisfying these three constraints in the case of a continuous PDF $p_c(x)$ is

$$H(X) = -\int_{x \in X} p_c(x) log(p_c(x)) dx \tag{2.4}$$

and in the case of a discrete PMF $p_d(x)$

$$H(X) = -\sum_{x \in X} p_d(x) log(p_d(x)) dx \tag{2.5}$$

The principle of using the entropy concept to derive probability distributions was introduced by E.T. Jaynes in 1957[9], and differs strongly from the other methods described in this work. The principle infers a distribution function that preserves the maximal level of uncertainty (entropy) given presupposed constraints on the modeled variable. This means that the choice of any other distribution will require making additional assumptions unsupported by the given constraints[5]. The maximum entropy (ME) distribution thus constitutes a mathematically well founded choice of model when there is a lack of knowledge and data.

A direct derivation of the ME distribution from equations 2.4 and 2.5 involves solving a system of nonlinear equations, the solution of which involves variational calculus using the Lagrange multiplier method. A simpler way, applicable under certain conditions when the constraints are on the expected value, is to use the following theorem, first proved by Boltzmann[11].

Suppose S is a closed subset of $\mathbb{R}$, $f_1, \ldots, f_n$ are measurable functions and $a_1, \ldots, a_n$ are real numbers. We consider the class K of all continuous random variables which are supported on S and which satisfy the n expected value conditions:

$$E(f_j(X)) = a_j, \text{ for } j = 1, ..., n \tag{2.6}$$

If there is a member in K whose density function is positive everywhere in S, and if there exists a maximal entropy distribution for K, then its probability density $p(x)$ is of the form:

[5] A strict, decision theoretic explanation is available in[10]

$$p(x) = ce^{\sum_{j=1}^{n} \lambda_j f_j(x)}, \text{ for all } x \in S \tag{2.7}$$

To derive the maximum entropy distribution, we seek to identify c and $\{\lambda_i\}$. Suppose we want to do this given the following two constraints (E denotes the expected value function):

$$S = [0, \infty] \tag{2.8}$$

$$E(x) = \mu \tag{2.9}$$

Equations 2.8 and 2.9 correspond to integral equations 2.10 and 2.11, respectively

$$\int_S p(x)dx = 1 \tag{2.10}$$

$$\int_S xp(x)dx = \mu \tag{2.11}$$

where

$$p(x) = ce^{\lambda x} \tag{2.12}$$

Solving for c and lambda yields that $c = -\mu = \lambda$, which implies $p(x) = \mu e^{-\mu x}$, an exponential distribution function with mean μ.

Applying this theorem to the most commonly encountered sets of expected value constraints yields the maximum entropy distributions[6] presented in Table 2.1 [12][13].

Table 2.1. Constraints and corresponding maximum entropy distributions.

Case	Available constraints	Maximum entropy distribution
1	Mean	Exponential
2	Lower bound = 0 and a quantile	Exponential
3	Lower bound > 0 and a quantile	Shifted exponential or gamma
4	Range	Uniform
5	Range and mean	Beta
6	Mean and standard deviation	Normal
7	Range, mean and standard deviation	Beta
8	Mean rate of occurrence	Poisson

The process of deriving the ME distribution given the constraints stated in the left column of Table 2.1 differs from case to case. For cases 1, 4 and 5, it is a matter of mapping the constraints to the parameters of the distribution functions. Let x_q

[6]Here, *shifted exponential* denotes an exponential disribution plus a constant shift value.

be the estimate of the known quantile. In cases 2 and 3 the mean parameter of the exponential distribution is calculated as $\mu = \frac{-x_q}{ln(1-q)}$ [11]. In case 7 there are formulas for deriving the parameters of the beta distribution from the constraints[13]. If, however, we are only given a range and a mean value (case 8), then the standard deviation that maximizes the entropy of the resulting beta distribution is chosen[11].

When using the ME method, one should be aware of the fact that not all classes of distributions contain one that maximizes the entropy for a given set of constraints. For instance, the class of continuous distributions on $\mathbb{R}$ with mean equal to zero contains members with arbitrarily high entropy, and is thus an example of a class where no member has maximal entropy. This is illustrated in Figure 2.2, where *sigma* denotes the σ parameter of the normal distribution.

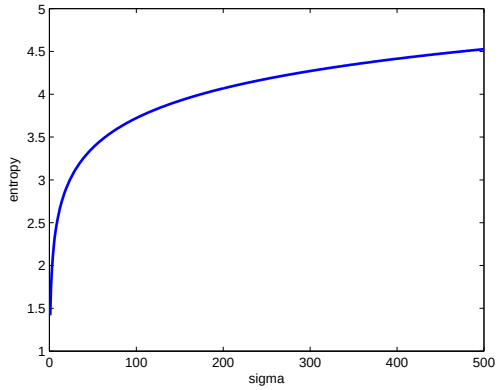


Figure 2.2. Entropy of Normal(1,σ) where $1 \leq \sigma \leq 500$.

As with all methods for dealing with the issue of lack of information the MEM does not give you more information than you supplied. As Conrad [8] points out, we cannot assume that a coin is fair just because our knowledge about it is limited. Thus the ME distribution should be carefully studied to see if additional constraints become apparent. If this is the case, they should be added to the list of constraints and the process repeated to verify them. Used in this way, the method can be valuable for eliciting information from the assessor which may not be evident initially.

Chapter 3

Assessing the approximation

When we have arrived at a distribution, using one or several of the methods mentioned in chapter 2, we need to verify that the function adheres to the observations and to any additional knowledge that the assessor might possess but did not already include during the approximation process. While different types of verification schemes are available for different approximation methods, some approaches are common to all methods.

3.1 Using hypothesis testing

If we want to verify the precision (also known as goodness) of a parametric approximation (also called a fit), we can employ methods based on so called goodness-of-fit (GoF) statistics. These are functions that, in some measure, calculate the distance between the empirical distribution and the hypothesized parametric DF. The result, in the form of a yes or no answer given at a specific significance level. This is achieved by formulating a statistical hypothesis test, where the null hypothesis states that the observed data can be assumed to be drawn from the given distribution. The three test statistics that are most popular today are χ^2, Kolmogorov-Smirnov and Anderson-Darling.

The χ^2 test statistic measures the same value that the χ^2 parameter estimation technique seeks to minimize - the sum of the square residuals of the sample. This is an intuitive test which is easy to understand graphically. Among the disadvantages of this statistic is that it relies on a histogram approximation of the empirical DF, with the sensitivity to the binning parameters and sample size that this entails.

The Kolmogorov-Smirnov (K-S) test statistic is even simpler to grasp, it is simply the maximum distance between the ECDF (calculated as $\{n(i)/N\}$ where $n(i)$ is the n:th order statistic and N is the sample size) and the hypothesized distribution. Perhaps the main advantage of the K-S test statistic is that it remains relatively reliable even when the sample size is small, since it does not require the construction of a histogram from the observed data. Because K-S measures the maximum distance between the hypothesized CDF and the ECDF, it is not very

sensitive to discrepancies in the tails. The Anderson-Darling (A-D) test statistic is a modified version of the K-S statistic, tweaked to give a better assessment of the fit in the tails of the distribution. If we let N denote the sample size, the A-D statistic is calculated as

$$A^2 = -N - S \tag{3.1}$$

where

$$S = \sum_{i=1}^{N} \frac{2i-1}{N} [log_e(F(Y_i)) + log_e(1 - F(Y_{N+1-i}))] \tag{3.2}$$

A disadvantage of A-D is that, to give a good result, it needs to be modified[1] for each family of distributions and also for different sample sizes.

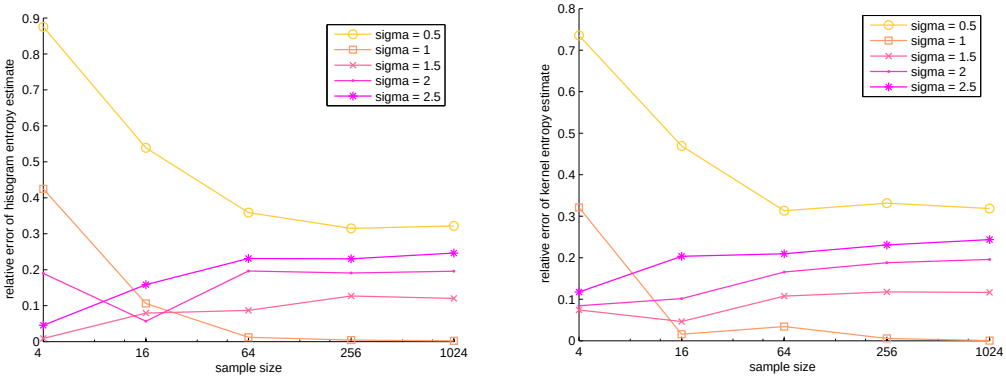
3.2 Using entropy

A common way of utilizing GoF tests based on test statistics is to rank distributions found by parameter estimation techniques by the lowest significance level at which the distribution passes the test. This assumes that the results of these tests are directly comparable and ignores the fact that the tests often have non-overlapping conditions under which they break down. At a certain sample size, K-S might give an accurate result, while χ^2 might give a very rough approximation of the actual fit. This can produce a very misleading ranking and thus lead to the choice of the wrong distribution. Another approach, which has recently gained popularity, is to estimate the fit by comparing the entropy of the hypothesized distribution with the empirical entropy, calculated from the EPDF. Since the entropy of a DF is only dependent on its shape and not on its location, it is very important to combine this approach with a visual inspection of the compared DFs.

To obtain an estimate of the empirical entropy we need to use one of the available methods for calculating the EPDF. Each of these methods has its specific drawbacks, but they all have a tendency to misestimate the tails of a distribution.

Table 3.1 on page 16 compares the true entropy of a few common distributions, calculated analytically, to their approximations (based on a KDE of the density function) and lists the relative error (calculated as analytic entropy divided by absolute error) for each of these. The entropy estimates which correspond to unbounded distributions are about 10 times more precise than those for unbounded distributions. The estimates corresponding to bounded distributions converge monotonously with an increasing sample size. The theoretical motivation for this is that unbounded distributions or, more accurately, distributions with long tails are hard to approximate accurately using sample-based density estimation methods. The tails are regions where, by definition, the frequency is small. Thus the amount of information about the PDF in these regions is small. This means that even with a very large number

[1] `http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm`



(a) Histogram method (b) KDE method

Figure 3.1. Convergence properties of two empirical entropy estimation methods.

of samples the contribution to the shape of the tails is insufficient to render a good entropy estimate.

For samples drawn from distributions with a high spread (e.g. normal distributions with $\sigma \ggg 1$), empirical density estimation methods underestimate the tails, and thus the empirical entropy is biased downward. When the sample comes from a distribution with a very small spread (e.g. normal distributions with $\sigma \lll 1$) the tails are overestimated and thus the resulting empirical entropy is biased upward. This can be seen in Figure 4.1, which also indicates that the two compared entropy estimations methods have similar convergence properties.

A possible solution to the problem of estimating the empirical entropy of unbounded distributions is to work with a restriction of function to a bounded subset of its support. This way, at least the entropy of the restricted function could be calculated precisely.

In Table 4.1, MC refers to Monte Carlo simple random sampling and LH to Latin Hypercube sampling.

Table 3.1. Accuracy of a KDE based numerical entropy approximation.

Sampling	Distribution	Sample	Entropy (Analytical)	(Numerical)	(Error)
MC	Normal(1,1)	10	1.4189	1.3015	0.0827
MC	Normal(1,1)	100	1.4189	1.3706	0.0340
MC	Normal(1,1)	1000	1.4189	1.3812	0.0266
MC	Normal(1,2)	10	1.7655	1.8182	0.0298
MC	Normal(1,2)	100	1.7655	2.0546	0.1637
MC	Normal(1,2)	1000	1.7655	2.1092	0.1947
MC	Normal(1,5)	10	2.2237	2.7934	0.2562
MC	Normal(1,5)	100	2.2237	2.9058	0.3067
MC	Normal(1,5)	1000	2.2237	3.0286	0.3620
LH	Normal(1,1)	10	1.4189	1.1094	0.2181
LH	Normal(1,1)	100	1.4189	1.3753	0.0307
LH	Normal(1,1)	1000	1.4189	1.4311	0.0086
LH	Normal(1,2)	10	1.7655	1.8175	0.0295
LH	Normal(1,2)	100	1.7655	1.9777	0.1202
LH	Normal(1,2)	1000	1.7655	2.0676	0.1712
LH	Normal(1,5)	10	2.2237	2.9931	0.3460
LH	Normal(1,5)	100	2.2237	2.9590	0.3307
LH	Normal(1,5)	1000	2.2237	3.0569	0.3747
MC	Triang(0,2,4)	10	1.1931	0.9489	0.2047
MC	Triang(0,2,4)	100	1.1931	1.1059	0.0731
MC	Triang(0,2,4)	1000	1.1931	1.2057	0.0110
MC	Triang(0,6,12)	10	2.2918	2.6420	0.1528
MC	Triang(0,6,12)	100	2.2918	2.2045	0.0381
MC	Triang(0,6,12)	1000	2.2918	2.3043	0.0055
MC	Uniform(0,10)	10	2.3979	2.5883	0.0794
MC	Uniform(0,10)	100	2.3979	2.4297	0.0133
MC	Uniform(0,10)	1000	2.3979	2.3879	0.0042

Part II

Application

Chapter 4

Case study

To evaluate the methods presented in this work, a number of parameter sample collections (landscape model parameters used in simulations at SKB[1])used in investigations at Facilia™were studied to deduce which methods were most relevant. As is common in the area of environmental risk assessment, where sample collection is expensive, the number of available data points per parameter was low. If a few extreme cases are ignored (sample sizes of over 100 and less than three) the average number of samples is about nine. As explained in section 2.2, this is not enough for a reliable parametric estimate of the underlying density (at least not unless there is prior knowledge about the distribution of the variable), so for most of the studied parameters the ME method had to be used.

Below follow three examples of variables that are typical input parameters in environmental risk assessment models, and a recommended process for the derivation of probability distributions to represent them.

4.1 Example 1

The suspended particle concentration in water is an example of a variable for which it is easy to obtain many samples. In this particular case 40 data points are available, which is enough to obtain a fair MLE, assuming that the samples are representative of the variable.

Knowledge of the variable also indicates that the assumption that the variable has a log-normal distribution, and that its theoretical minimum is zero. This information is useful to rule out distribution families that fail to match the theoretical characteristics (e.g. non-negativity) of the variable, but cannot be rejected using GoF tests.

Table 4.1 contains the 40 available sample values. Table 4.2 contains estimates, derived using the maximum likelihood method, for the parameters of a set of distribution functions commonly encountered in risk assessment. Also included in the latter table are the K-S statistics corresponding to each applicable distribution.

[1] `http://www.skb.se/`

Table 4.1. Sample for Example 1, suspended particle concentration in water.

0.314	0.261	0.318	0.350	0.143	0.140	0.162	0.159
0.331	0.504	0.286	0.218	0.215	0.174	0.368	0.321
0.395	0.308	0.545	0.187	0.198	0.193	0.177	0.257
0.442	0.345	0.219	0.405	0.255	0.286	0.203	0.282
0.288	0.229	0.267	0.218	0.200	0.521	0.130	0.216

The results have been ranked by these statistics (distributions for which the null hypothesis was rejected at significance level 0.05 are marked N/A), and their order indicates that a reasonable choice of a distribution function for this example is a Log-Normal(0.271,0.107)[2]. Because our prior assumption regarding the disribution of the particle concentration was that it was log-normal, and given that its K-S statistic is low and visual fit closely matches the relative histogram we can safely choose this distribution to model our variable.

Table 4.2. Distributions and statistics for Example 1.

Distribution	Rank	K-S Statistic
Beta$(5.369, 14.029)$	5	0.1213
Exponential	N/A	N/A
Extreme Value$(0.222, 0.082)$	2	0.0724
Gamma$(7.7159, 0.0357)$	4	0.1143
Log-Normal$(0.2711, 0.1070)$	1	0.0694
Normal$(0.259, 0.094)$	6	0.1310
Uniform$(0.13, 0.545)$	N/A	N/A
Weibull$(0.310, 2.815)$	3	0.0840

In figure 4.1, the distributions presented in table 4.2 are plotted against a frequency histogram approximation of the data set in table 4.1.

[2]Log-Normal with $\mu = 0.271$, $\sigma = 0.107$

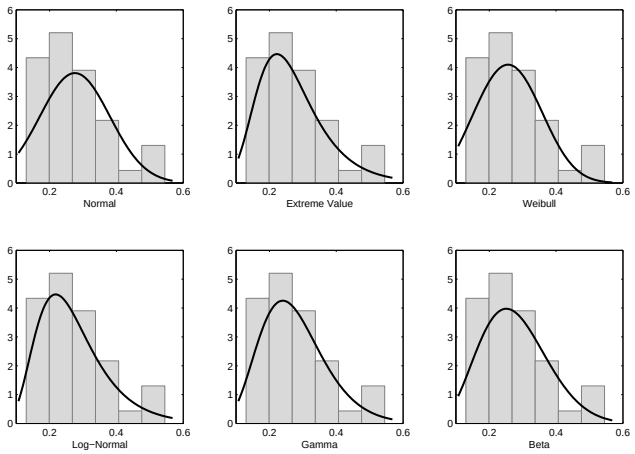


Figure 4.1. Distributions fitted to sample in Table 4.1.

4.2 Example 2

The porosity of a layer of sediment is harder to sample than the aforementioned concentration and thus the number of available data samples is usually smaller for this type of variable. Also, since it is a property of a solid material, its probability density has a greater tendency of being a composite of a number of different densities, each corresponding to the material making up the different slices of the sediment layer. This has dire consequences for the quality of a parametric fit made based on the data, since the total density will be multimodal (one mode per slice), which is clearly visible in Figure 4.2.

In this example, the number of data samples is 14 – a combination of two data sets each containing seven samples (see Table 4.3). To properly interpret this, the assessor needs to decide whether the modelled variable has other component densities which were omitted during sampling.

If this is the case, then by looking at the relative frequency histogram plot in Figure 4.2 we can see that a reasonable density assignment for this variable would have two modes and a nonzero density in between these. This function can be represented in several ways, one of them is a KDE truncated at the minimum and maximum observed or theoretical values. If we have reason to consider that the density in between the modes has similar characteristics to the density around the two observed modes a better choice is to assign a uniform distribution with minimum and maximum parameters derived from the observations.

If, however, the assessor has firm grounds to believe that the true density lack any modes other than the observed ones, we could assign an empirical density estimate based on the histogram, to ensure that simulations based on the distribution stay true to the observed data.

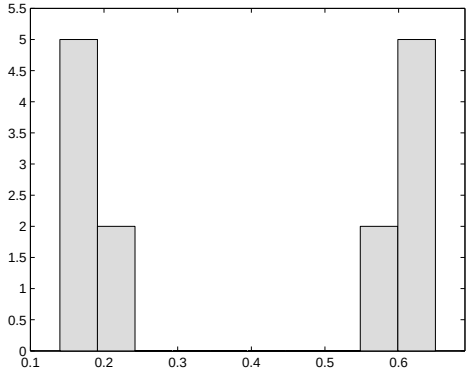


Figure 4.2. Histogram of porosity.

Table 4.3. Sample for Example 2, porosity of a layer of sediment.

Data set	Observed values						
1	0.222	0.183	0.163	0.170	0.207	0.166	0.140
2	0.550	0.650	0.600	0.550	0.600	0.600	0.600

4.3 Example 3

When we are trying to model a variable based on even fewer data samples empirical density estimates become unreliable and a better method for choosing a function to represent the data is the ME method. Since the input for this method is not a data sample but rather a set of statistics we first need to discern which statistics we think we can derive from the available data set. Using the definitions of each statistic we could, of course, simply plug in our data to obtain any statistic, but this would imply that we believe that the data set is representative of the true underlying statistic, which is seldom the case for such small data samples.

Table 4.4. Sample for Example 3, biomass of an animal species for a given area.

0.55	0.00019	0.00235	0.01793

In this example, the values in Table 4.4 constitute our data set (the annual biomass productivity of an animal, such as the number of kilograms of elk per square kilometer per year). Our prior assumptions include that our data seems to fairly represent the mean value but not necessarily the range and standard deviation of the variable, and that its theoretical minimum is zero. Since there is no assumed upper limit we set this to infinity and applying the theorem presented in section 2.3 we get that the maximum entropy distribution given our constraints is an exponential distribution with mean equal to the observed mean, approximately 0.143.

Applying the K-S GoF test to this example gives us a p-value (observed significance value) of 0.046, which means that our hypothesis that the sample came from an Exponential(0.143) should be rejected at the 95 % significance level (a p-value greater than 0.05 is required for us not to reject it). However, considering the small sample size, we could choose to ignore this result.

Chapter 5

The software tool

A part of this work was to develop a software tool to facilitate the use of the above described methods. The choice of computation engine fell on MATLAB®[1] due to its wide-spread use as an engine and language for mathematical calculations making it common among potential users of the tool. For the graphical user interface, the Swing® API in Java®[2] was chosen over the built-in capabilities of MATLAB, as it was presumed that it would facilitate data exchange between the tool and Excel® (commonly used for storing sample data) and other relevant applications[3].

The tool consists of a collection of MATLAB scripts and a Java application that calls the scripts in order to perform its calculations. The Statistical Toolbox for MATLAB was used to provide functionality such as MLE parameter estimation and the K-S GoF hypothesis test. The KDE Toolbox for MATLAB was used to provide KDE related functionality.

5.1 The user interface

Illustrated in Figure 5.1, the user interface is divided into four main parts. The top of the screen hosts a series menu items as well as buttons to facilitate access to the most commonly used functionality, such as the creation, saving and opening of files. The menus follow standard graphical user interface conventions to reduce the learning curve.

The center of the screen is empty when a new document is created. When the user performs a fit, this area is filled with a tabbed panel containing the following plots:

- A probability density plot, containing plots of the estimated PDFs. If the user is inspecting a data fit, the plot also contains a relative frequency histogram and a KDE plot.

[1] http://www.mathworks.com
[2] http://java.sun.com
[3] Using copy/paste and libraries such as POI (http://jakarta.apache.org/poi).

- A cumulative density plot, where the estimates CDFs are plotted against the ECDF.
- A plot of the survival function $s(x) = 1 - p(x)$ for each estimated PDF $p(x)$.
- A plot of the hazard function $h(x) = p(x)/s(x)$ for each estimated PDF.
- A plot of the cumulative hazard function $H(y) = \int_{-inf}^{y} p(x)dx$.

To the left of the plot area, two tables inside a tabbed panel provide the possibility of entering and retrieving data. The input and output tables can be used either manually using a keyboard to enter data or using the clipboard to copy/paste or drag and drop data to and from any compatible application, including Excel, OpenOffice and MATLAB. To the right of the plotting area is a column of collapsible panels. The topmost of these is where the user selects the desired fitting method.

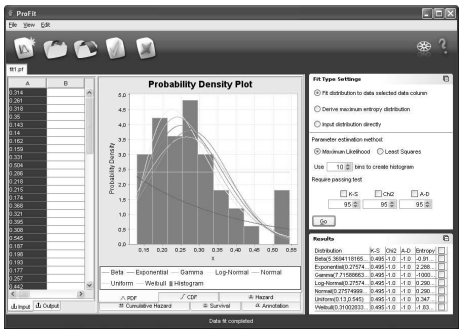


Figure 5.1. User interface of the ProFit software tool.

- If the *data fit method* is chosen, the panel expands to reveal a panel containing controls for setting the various specifics of the data fit, including the parameter estimation method and the statistic by which the resulting distributions should be ranked. When the user presses the *Go* button, a new panel is added below, containing a table of those distributions that were compatible with the selected data set, their estimated parameters, entropy and K-S, χ^2 and A-D test statistics. At this point the center area is also populated with the relevant plots.
- If the *maximum entropy method* is chosen, the top panel in the column expands to make space for a table corresponding to Table 2.1. When the user selects a row in this table, a new panel is added below containing a list of input boxes corresponding to the parameters specified by the constraints specified on this row. The user can either enter the values of these constraints manually or choose (using a check box situated to the right of the input box for the value) to derive the values from the sample currently selected in the data input table.

- Finally, if the *direct input method* is chosen, the user is presented with a list of all the supported distributions. Upon selecting one of these, a panel containing text input boxes for the parameters of the distribution is added below.

The application was designed with the intention of working in tandem with other applications, as an interface for choosing distributions for simulation or sensitivity analysis. Thus, if the application is started in the correct mode, the menu at the top of the screen also includes a button to accept the currently selected distribution and export this to the host application.

Chapter 6

Conclusions

The process of assigning probability distributions to data is complicated and includes many subtle pitfalls which is why many experts in the field advise that someone with a deep knowledge of statistics should be involved if this is possible. If it is not possible then being cautious and double checking results using graphical plots and against theoretical constraints can help make the process more reliable.

Given a large data sample, well representative of its underlying variable, the task is straight-forward. A parametric fit is first attempted followed by goodness-of-fit tests to discern which DF should be chosen. If no parametric distribution from the candidate set has a sufficiently good fit, a nonparametric distribution is deduced, allowing if not a compact representation of the data, at least the ability to use the modeled variable in probabilistic simulations.

When the sample is smaller the process changes, and becomes more demanding of the assessor. Empirical plots of the data reveal whether there is enough information in the sample to make a reliable parametric fit. If not, a decision needs to be made of what statistics can be deduced from the sample. Using these, the MEM is used to deduce a distribution.

Finally, if the sample is too small for a parametric fit and there is a significant knowledge about a prior distribution for the modeled variable, a Bayesian update can be performed to incorporate the observations into the posterior distribution. Implementing this in the software application would be highly useful in the context of environmetal risk assessment.

Figure 6.1 on page 30 presents a flow chart of the process described in this work. It does not include all the elements of the procedure, but summarizes the most important steps and choices and the flow of data throughout the process.

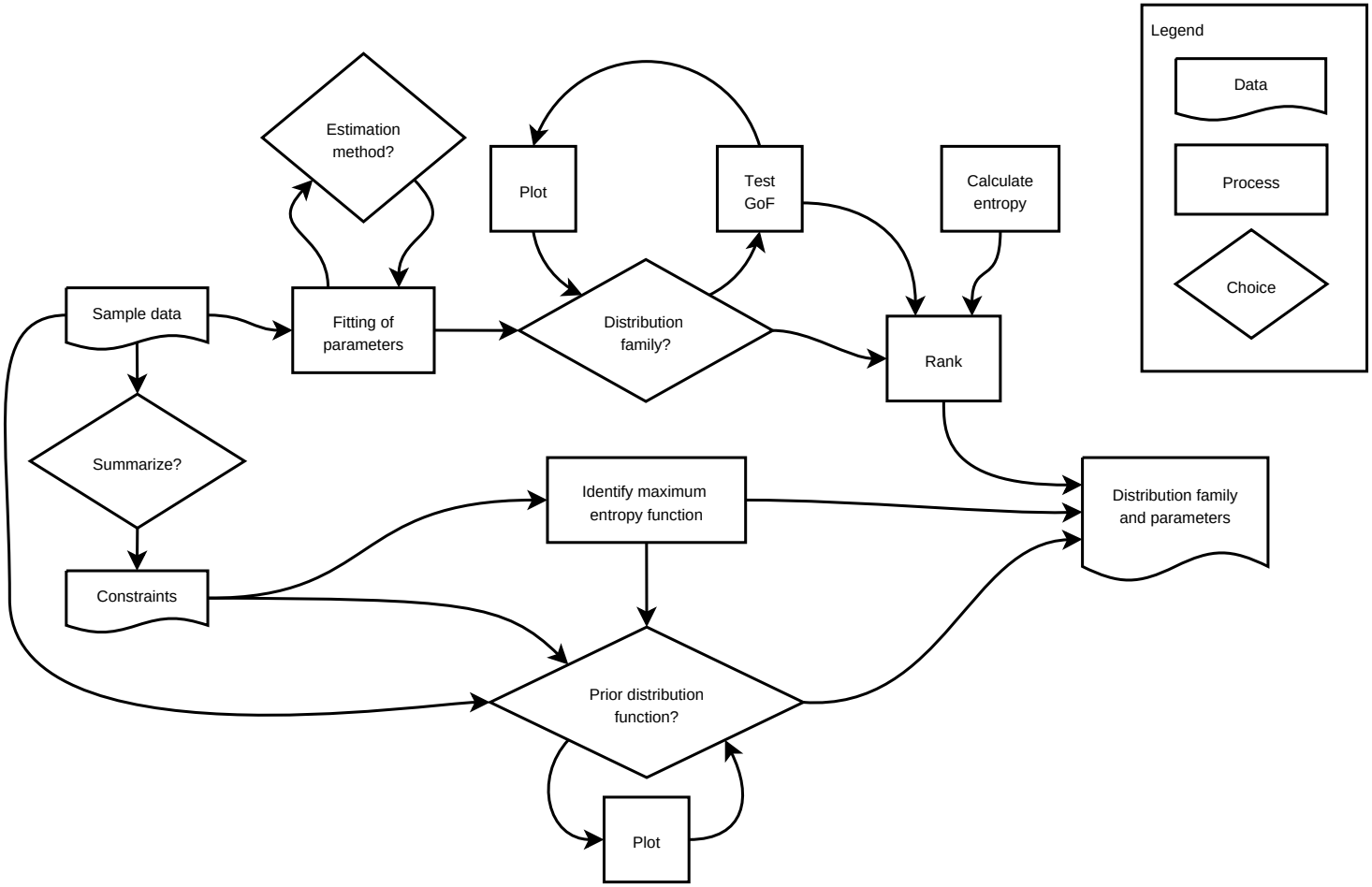


Figure 6.1. Flow chart of distribution fitting

Dictionary

CDF Cumulative density function $f(x) = \{P(x) \leq x\}$.
Continuous A mapping which preserves the topological structure of the mapped space.
Density function The statistical function that shows how the density of possible observations in a population is distributed. It is the derivative $f(x)$ of the CDF $F(x)$ of a random variable, if $F(x)$ is differentiable. Geometrically, $f(x)$ is the ordinate of a curve such that $f(x)dx$ yields the probability that the random variable will assume some value within the range dx of x.
Discrete piecewise constant.
ECDF Empirical cumulative distribution function.
Empirical Based on observed data.
Histogram A bar chart diagram, first used by A.M. Guerry in 1833 where the bars' heights are equal to the frequency of the value at their centres in the plotted data sample.
Kernel In the context of density estimation, a function centered at each data point, corresponding to the bars in a histogram.
Mode A value around which a large number of can be observations can be expected.
Model A mathematical representation of a quantity or set of quantities defined by a function over a parameter space.
Order statistic An element of the ordered set of observations.
PDF See density function
Robust Not overly sensitive to changing conditions.
Significance level The probability of making a Type I error, or rejecting the null hypothesis when it is actually true.
Support The subset of the domain of a function outside of which the function is equal to zero.
Skewness Asymmetry.
Unimodal Having one mode.
Unbiased An estimator that neither over- nor under-estimates the approximated quantity.

Bibliography

[1] Pierre-Simon Laplace 1825 (translated by A.I. Dale). *Philosophical Essay on Probabilities*. Springer-Verlag, 1995.

[2] Ajit C. Tamhane and Dorothy D. Dunlop. *Statistics and Data Analysis: From Elementary to Intermediate*. Prentice-Hall Inc., 1999.

[3] Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. John Wiley and Sons, 2000.

[4] M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51(1):59, 1997.

[5] Paul Meier E. L. Kaplan. Data-based choice of histogram bin width. *Journal of the American Statistical Association*, 53(282):457, 1958.

[6] MH DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[7] Peter Green. Reversible jump markov chain monte carlo computation and bayesian model determination. 1995.

[8] Keith Conrad. Probability distributions and maximum entropy. 2005.

[9] Edwin Thompson Jaynes. Information theory and statistical mechanics, I and II. *Physical Reviews*, 106 and 108:620–630 and 171–190, 1957.

[10] Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *The Annals of Statistics*, 2004.

[11] R.C. Lee and W.E. Wright. *Development of human exposure-factor distributions using maximum-entropy inference*, volume 4. 1994.

[12] Milton Edward Harr. *Reliability-Based Design in Civil Engineering*. McGraw-Hill, 1987.

[13] Michael A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730, 1974.

TRITA-CSC-E 2006:042
ISRN-KTH/CSC/E–06/042–SE
ISSN-1653-5715

www.csc.kth.se