Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

RECOGNITION OF CONTINUOUS PROBABILITY MODELS

Marcelo Tenório
Silvia Nassar
Paulo Freitas

Federal University of Santa Catarina
Computer Science and Statistics Department
Florianópolis, Brazil

Carlos Magno

Well Technology Engineering, CENPES
PETROBRAS SA
Rio de Janeiro, Brazil

ABSTRACT

It is well known that randomness is present in daily life and that often it is desirable to recognize inherent characteristics of this randomness. Probability theory describes a quantification of the uncertainty associated with this randomness. Based on probability theory, the present research describes an alternative methodology to the traditional statistical method of the recognition of the probabilistic models that best represent randomness. The main motivation of the methodology is to keep the largest possible amount of information present in the data. This methodology differs from the traditional statistical method, mainly in aspects related to the division of the data into classes when the data are continuous.

1 INTRODUCTION

The Petrobras Research Center (CENPES), in partnership with the Federal University of Santa Catarina (UFSC), developed a tool named E&PRisk to support decisions concerning drilling and completion of petroleum wells.

E&PRisk performs Monte Carlo simulation and related statistical analysis to assist in the evaluation of the total time necessary for, and the risks with, the construction of a well. It also assists in decision making with regard to the technological alternative to be used.

The simulated operation time of a petroleum well provides an estimate of the total time of its construction. The time of each operation is expressed by a probability distribution model that describes its randomness.

The following research was elaborated for the construction of a built-in tool for E&PRisk to recognize the probability distribution models that best represent the analyzed data.

This document presents some of the related statistical concepts, the proposed methodology, and its implementation. The document also describes the obtained results and their validation.

2 PROBABILITY

Probability theory analyzes processes that involve variability and randomness, applying mathematical models to facilitate analysis (Barbetta 2004).

Probability theory is empirical; the theory is based on observation. Probability theory describes what occurs in many proofs, and should utilize the results of many proofs in the effort to estimate probability (Moore and McCabe 1999, Montgomery and Runger 1999).

In probability, or probability models, there are two aspects to consider. The first is related to the intuition used to make decisions based on facts that have a high probability of occurring. For example, if the sky is cloudy, then there is a considerable chance of rain, and one should carry an umbrella. The second aspect is the inherent uncertainty of the decisions that can be taken with regard to a specific problem. For example, even if the sky is very cloudy, it is possible that it will not rain, at least while one is outside.

Some decisions become easier if it is possible to quantify the uncertainty associated with each fact. Probability theory allows a quantification associating uncertainty to one or more facts, therefore, is extremely useful in decision making (Barbetta 2004).

Thus, in modeling a process with uncertainty, it is necessary that the variability of the randomness be represented by a probability distribution of the associated variables.

2.1 Random Variable

A random variable is a variable, usually represented by *X*, that has an unique random numerical value for each result of an experiment. The word “random” indicates that the value is only known after the experiment (Triola 1998).

Quantitative variables are divided into two categories, discrete and continuous.

A discrete random variable admits a finite number of values or has a countable quantity of values (Triola 1998).

A continuous random variable can take an infinite number of values, and these values can be associated with measurements on a continuous scale, so there will be no gaps or interruptions (Montgomery and Runger 1999).

2.2 Frequency Distribution

One of the first steps of data mining is to calculate the frequency distribution of each variable, especially when there are a great number of observations (*n*).

The frequency distribution consists of data organization according to the occurrences of the different observed results (Barbetta 2004).

The frequencies can be presented in absolute, relative, or cumulative form. They are presented in a table or visualized in a graph format. For discrete variables, the column graph is the most used, and for continuous variables, the histogram is the main graphical presentation.

The traditional graphical presentation of discrete variables is to plot each value on the *x* axis and its frequency on the *y* axis. For continuous variables, generally, the total data width is divided into intervals, denominated as classes. Then the histogram is plotted with the minimum and maximum values of the classes on the *x* axis, and their respective frequencies on the *y* axis.

2.3 Probability Distribution

Besides identifying values of a random variable, frequently a probability can be attributed to each one of these values. When all values of a random variable and its respective probabilities are known, this creates a probability distribution.

A probability distribution represents the possible values and the probability of each value of a random variable (Triola 1998).

For a discrete random variable *X*, with possible values $x_1,...x_2,...,x_n$, the probability function is:

$$f(x_i) = P(X = x_i)\ . \qquad (1)$$

A graphical representation is illustrated in Figure 1.

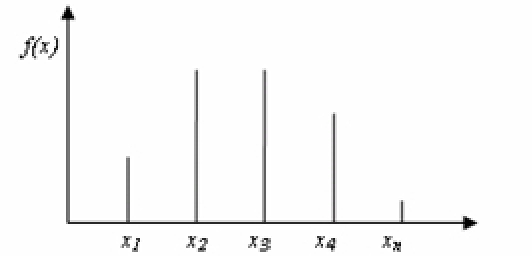


Figure 1: Graphical representation of the probability distribution of a discrete random variable *X*

The cumulative distribution function (CDF) is another form used to represent the probability distribution of a random variable. For a discrete random variable, the CDF is defined as:

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)\ . \qquad (2)$$

Its graphical representation is shown in Figure 2.

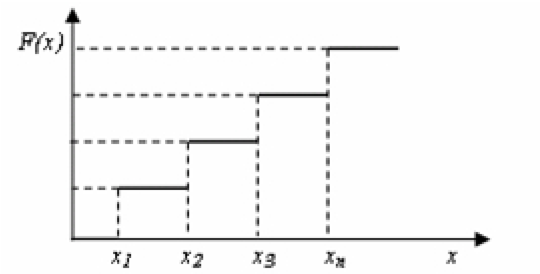


Figure 2: Graphical representation of the cumulative probability distribution of a discrete random variable *X*

For a continuous random variable *X*, the probability distribution is called the probability density function and is defined as:

$$f(x) \ge 0\ , \qquad (3)$$

$$\int_{-\infty}^{\infty} f(x)dx = 1\ . \qquad (4)$$

A histogram is an approximation of the probability density function, as illustrated in Figure 3. For each histogram interval, the bar area is equal to the relative frequency (ratio) of the variable values in the interval. The relative frequency is an estimate of the probability that the values contained in the interval will occur.

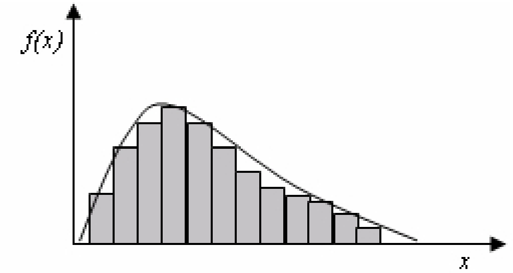


Figure 3: Histogram of the probability density function of a continuous random variable *X*

The CDF of a continuous random variable *X* on the interval $-\infty < x < \infty$ is:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx\ . \qquad (5)$$

Its graphical representation is illustrated in Figure 4.

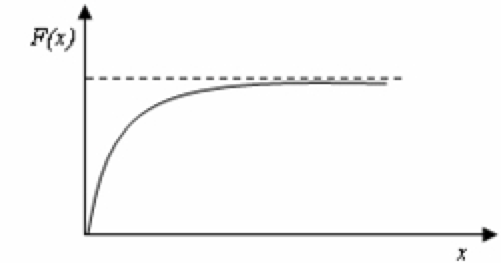


Figure 4: Graphical representation of the cumulative probability distribution of a continuous random variable *X*

Probability theory offers some theoretical models of probability distribution, for example, gamma, exponential, weibull, beta and normal (Law 1991, Jain 1991).

Thus, from a set of observed values of a variable, one seeks to discover which model of probability can best represent these data. This probability model recognition process is a statistical test called Goodness-of-fit Test.

3 GOODNESS-OF-FIT TEST

The statistical tests are classified into two categories, the parametric and nonparametric tests.

The parametric tests suppose that data follows a determinate probability distribution. However, the nonparametric tests are used when the assumptions needed to apply the parametric tests are not satisfied.

The goodness-of-fit tests are nonparametric tests. The objective of a goodness-of-fit test is to verify if the data from one sample behave according to a theoretical distribution (Barbetta 2004).

The chi-square goodness-of-fit test can be applied when one studies distributed data in classes and one has an interest in verifying whether the observed frequencies (sample data) in the *K* different classes $(O_i, i = 1,2,...,K)$ are significantly distinct from a set of *K* expected frequencies (probability distribution) $(E_i, i = 1,2,...,K)$. The hypotheses are:

$$H_0 : O_i = E_i \text{ for each } i = 1,2,...,K \text{ ,}$$
$$H_1 : O_i \neq E_i \text{ for some } i = 1,2,...,K \text{ .}$$

The result of this test, called χ^2, is a measure of the distance between the observed and the expected frequencies of each class (Barbetta 2004). Its expression is given by:

$$x^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}. \qquad (6)$$

If there is a fit (H_0 is true), the observed frequencies must be close to the expected ones, causing a small value for χ^2: the obtained variations would be only accidental. However, if there is no fit (H_1 is true), differences between the observed and the expected frequencies could be large, resulting in a large value for χ^2: it is less probable that the variations have been accidental.

3.1 The p-value

Given a hypothesis H_0 and a sample data set, the p-value reflects the tolerated probability of rejecting H_0 even when H_0 is, in fact, true. The p-value is called the significance and is obtained from the data sample. A very small p-value constitutes evidence against the H_0 hypothesis.

In this research, when it is desired to accept or reject some hypothesis, it is common to establish, in the research planning stage, the tolerable probability of incurring an error in rejecting H_0, even when H_0 is true. This value is known as the significance level of the test and is denoted by α. It is common to adopt the significance level $\alpha = 0.05$. But when more assurance is necessary in the affirmation of H_1, α can adopt low significance levels, such as $\alpha = 0.01$.

The following general decision rule of the statistical test is applied once the α significance level is established:

$p > \alpha$ ➔ accept H_0,
$p \leq \alpha$ ➔ reject H_0.

4 METHODOLOGY

We now present the concepts that support this research.

Traditional methods to select input for probability distributions include some difficult steps, especially those related to the choice of the number of intervals or classes (or, equivalently, their width) (Law 1991). If few classes are adopted, then the distribution is presented in a much reduced form, not evidencing some relevant characteristics of the variable. On the other hand, a distribution with many classes cannot enhance relevant aspects of the frequency distribution (Barbetta 2004).

To solve this difficulty, we opt to analyze the raw data, i.e., without grouping them. In this approach, the use of a cumulative distribution function (CDF) becomes obligatory. This function allows one to work with the individual values.

We made three changes to the traditional chi-square goodness-of-fit test.

First, when the observed data (*n*) increase, the distance, χ^2, also tends to increase, constituting evidence for the rejection of the tested model (tends to H_1). Therefore, from the observed data (*n*), a sample (*ng*) was extracted to calculate the distance, χ^2, seeking a balance between the data sample (*ng*) and the distance, χ^2, while keeping the data information intact. For example, see Table 1. The calculations of this procedure are demonstrated in the implementation section.

Table 1: Samples to calculate of the distance χ^2. Coefficient equals 95% and sample error margin equals 2.5%.

n	***ng***
10	10
30	29
50	48
60	58
80	76
100	94
150	137
200	177
500	377
900	568
1000	606
2000	869

Secondly, traditionally the chi-square goodness-of-fit test follows approximately a chi-square distribution with *K-1* degrees of freedom, where *K* is equal to the number of classes. The proposed methodology in this research does not use data grouped in classes, instead each class is represented by only one value, therefore the chi-square distribution has *n-1* degrees of freedom. But, as described above, a data sample was extracted (*ng*) from the set of observed values (*n*), therefore, the distribution contains *ng-1* degrees of freedom.

Thirdly and finally, to calculate the significance probability (p-value), the chi-square distribution was approximated by a normal distribution. It was verified that as the degrees of freedom (*ng-1*) increase, the chi-square distribution becomes symmetrical, tending toward a normal distribution.

5 IMPLEMENTATION

For each continuous variable *X* in the data sample, the following five steps are performed:

1. (Frequencies Distribution): Read the data sample in ascending order and calculate the frequency distribution, including the absolute observed frequency, absolute cumulative frequency, and relative cumulative frequency.
2. (Summary of the data): Calculate the size of the observed data (*n*), mean ($\bar{x}$), minimum, maximum, standard deviation (*s*) and variance (s^2).
3. (Probability distribution): In this research, the tested theoretical models are: uniform, exponential, triangular, normal and lognormal (Law 1991, Jain 1991). Using the CDF, the probability distribution of each model is calculated. For normal and lognormal models, which do not have a closed form CDF, a numerical integration method is used, known as Trapeze Rule to calculate the CDF (Leithold 1990).
4. (Goodness-of-fit Test): Before starting the test, it is necessary to do the following calculations for sample size:

$$cs = \left(\frac{z}{sem} \right)^2 * 0.25 \,, \qquad (7)$$

where:
cs = calculated size,
z = coefficient,
sem = sample error margin,

then:

$$ng = \frac{cs}{1 + \left(\frac{cs}{n} \right)} \,, \qquad (8)$$

where:
ng = sample size of goodness-of-fit test,
n = data sample size,
cs = calculated size,

and:

$$d = \frac{n}{ng} \,, \qquad (9)$$

where:
d = delta,
ng = sample size of goodness-of-fit test,
n = data sample size.

Starting the calculation of the χ^2 distance, as expressed in (6), the value of the absolute cumulative frequency is used as O_i. The value of the CDF is used as E_i, transforming it to the absolute form:

$$E_i = F(x) * n \,. \qquad (10)$$

where:
n = data sample size.

First, a sum (s_o_i) for the first O_i values, and another sum (s_e_i) for their respective E_i, are calculated. This occurs as long as:

$$s\_e_i < 5 \,. \qquad (11)$$

For later use, the stop position (*sp*) is saved from one of the sums (s_o_i or s_e_i). These sums are taken as the first term of the distance, χ^2.
The second term of distance, χ^2, is composed of the position values:

$$pv = sp + d \ . \tag{12}$$

where:
pv = position value,
sp = stop position in the vector,
d = delta.

The next terms are the position values given in (13), successively until the end of the data.

$$pv = pv + d \ , \tag{13}$$

where:
pv = position value,
d = delta.

5. (The p-value calculation): The p-value is calculated only for models with distance:

$$x^2 \leq u\_l \ , \tag{14}$$

for:

$$df = ng - 1 \ , \tag{15}$$

$$u\_l = \mu + \left(15 * \sqrt{\sigma^2}\right) , \tag{16}$$

$$\mu = df \ , \tag{17}$$

$$\sigma^2 = 2 * df \ , \tag{18}$$

where:
u_l = upper limit,
df = degrees of freedom,
ng = sample size of goodness-of-fit test,
μ = mean,
σ^2 = variance.

If the condition of (14) is not satisfied, the tested model is rejected. This means that, *p = 0.0001*. Otherwise, the distance, χ^2, is divided in *100* intervals of delta *0.01*u_l*, these intervals are assumed as the values of *x*-axis and applied to the normal density function with the parameters (mean and variance) of (17) and (18).

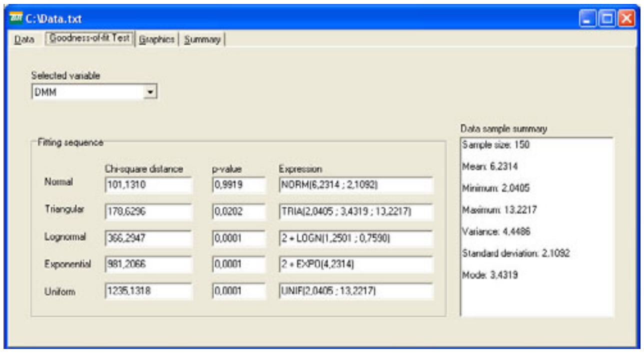
To obtain the area (*ar*), an integration sum of the results generated by the normal density function is obtained (numerical integration method, Trapeze Rule (Leithold 1990)). Finally, the p-value is obtained:

$$p = 1 - ar \ . \tag{19}$$

To conclude, the following hypotheses are elaborated:

H_0: there is no difference between the observed data and the theoretical model tested;
H_1: there is a difference.

Getting the p-value as a reference, the tested models are ordered from the best-fitting to the worst-fitting, as in Figure 5.



Selected variable: DMM

Fitting sequence

	Chi-square distance	p-value	Expression
Normal	101,1310	0,9919	NORM(6,2314 ; 2,1092)
Triangular	178,6296	0,0202	TRIA(2,0405 ; 3,4319 ; 13,2217)
Lognormal	366,2947	0,0001	2 + LOGN(1,2501 ; 0,7590)
Exponential	981,2066	0,0001	2 + EXPO(4,2314)
Uniform	1235,1318	0,0001	UNIF(2,0405 ; 13,2217)

Data sample summary
Sample size: 150
Mean: 6.2314
Minimum: 2.0405
Maximum: 13.2217
Variance: 4.4486
Standard deviation: 2.1092
Mode: 3.4319

Figure 5: Interface for the goodness-of-fit test from the software developed in this research

Remember that a very small p-value constitutes evidence against the hypothesis H_0 (Triola 1998). This means that the tested model does not represent the observed data.

To visualize the distribution graphs, a histogram is build with the relative cumulative frequencies and a line graph (over the histogram) with the results of the cumulative distribution function of the tested model, as can be seen in Figure 6. Thus it is possible to visualize and verify whether there is fitting or not.

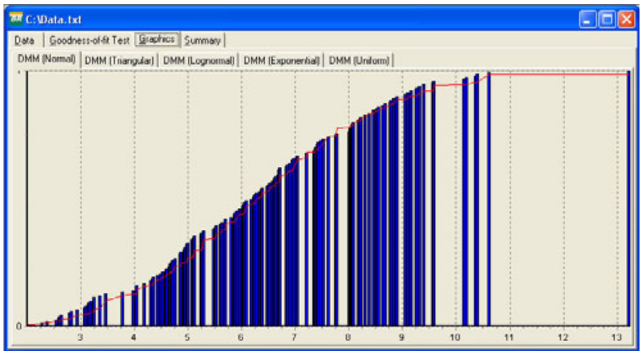


Figure 6: Graphics interface of the software developed in this research

As observed on Figures 5 and 6, software was developed (C++) based on the proposed methodology in this research. The objective of this software is to recognize probabilistic patterns of continuous data.

6 COMPARISON OF METHODOLOGIES

We performed a comparative test between our proposed methodology (PM) and the traditional chi-squared methodology. We utilized Input Analyzer, a statistical software application, to generate nine random samples (*n=30, 200, 2000*) for each model. Using the p-value as our measure of fit, we compared the results of PM, Input Analyzer, and Statistica (another statistical software application) on the data (Rockwell 2000, StatSoft 2001).

We present the results of this test in Table 2. Observe that the significance level $\alpha = 0.05$.

Table 2: Results of the test

		p-value		
Model (Parameter)	**Sample (*n*)**	**PM**	**Input**	**Statistica**
Uniform (2 ; 12)	30	0.0001	0.0545	0.0093
		0.2077	0.7500	0.0688
		0.9983	0.2560	0.1687
	200	1.0000	0.1500	0.8603
		0.0001	0.2570	0.0484
		1.0000	0.6000	0.5410
	2000	1.0000	0.7500	0.9064
		1.0000	0.7500	0.6208
		1.0000	0.6380	0.3872
Exponential (6)	30	0.9969	0.5700	0.5167
		0.9849	0.0050	0.9288
		0.9996	0.3780	0.8700
	200	1.0000	0.6790	0.2103
		1.0000	0.4860	0.6390
		1.0000	0.3180	0.1824
	2000	1.0000	0.7110	0.9109
		1.0000	0.7420	0.2306
		1.0000	0.4460	0.7406
Normal (8 ; 2)	30	0.9761	0.0050	0.0217
		0.9991	0.0050	0.2093
		0.9991	0.0050	0.7120
	200	1.0000	0.2670	0.5034
		1.0000	0.6850	0.3400
		1.0000	0.6820	0.1768
	2000	1.0000	0.1940	0.1914
		1.0000	0.3670	0.2672
		1.0000	0.6740	0.9208
Lognormal (0.6 ; 1)	30	0.9991	0.0050	0.0616
		0.9635	0.0050	0.7173
		0.0404	0.0050	0.0001
	200	1.0000	0.0050	0.5849
		0.9999	0.5790	0.8214
		0.0001	0.0194	0.1253
	2000	1.0000	0.6380	0.4164
		1.0000	0.1100	0.4456
		1.0000	0.0476	0.0506
Triangular (4 ; 10 ; 16)	30	0.0849	0.0651	Not applicable.
		0.6188	0.2810	
		0.0001	0.2870	
	200	0.0001	0.0230	
		0.6774	0.7500	
		0.0001	0.5040	
	2000	0.9998	0.2480	
		0.0001	0.3780	
		1.0000	0.2270	

Analyzing Table 2, the PM recognizes the data pattern of all cases of the exponential and normal models. The PM fails in some cases, for example, uniform, triangular, and lognormal models.

The gray cells indicate the cases where there are divergences between the applications tested. After a detailed analysis of these divergent cases, we observed that the random fluctuations in the process of the generation of the data influenced the estimates of the parameters. The PM has high sensitivity to the estimates of the models' parameters, because the PM works with the individual data (without grouping).

This high sensitivity is observed in the cases of the triangular model, which has three parameters. The PM failed to recognize the triangular model more often than any other model (i.e. p-value ≤ 0.05). In general, the models have one or two parameters. On the other hand, in the cases where the PM does not recognize the triangular pattern, it suggests the normal model, which it is acceptable.

It is possible to verify that in the majority of the tested models, the PM presents the best p-value. Remember that a very small p-value signifies rejection of the tested model.

7 FINAL REMARKS

This research details the study of a methodology, an alternative to the traditional method, for recognition of models of probability distribution, using the chi-square test.

The proposed methodology possesses the following characteristics:

1. The methodology works without grouping the data. In this sense, the methodology does not suffer some of the disadvantages that the traditional method presents, for example, the lack of representation of data (median values of each

class) and the difficulty of specifying a satisfactory quantity of classes to be utilized in the grouping of data.

2. The methodology utilizes the cumulative distributions. This characteristic is necessary because the cumulative form permits the methodology to work with individual values.
3. The methodology calculates the significance (p-value) by approximating the normal distribution instead of the chi-squared distribution. In this manner, the methodology has a mathematical advantage, because the normal function is easier to process than the chi-squared function. On the other hand, it has a disadvantage with respect to symmetry, which can be lacking for small sample sets.
4. The methodology permits a systematic selection of values to calculate the χ^2 distance. In addition to this advantage the methodology does not demonstrate difficulty in the grouping of data points. Furthermore, the systematic selection of data points to participate in the χ^2 distance calculation also differentiates this methodology from the traditional method.

Comparative tests were done with the objective of validating the precision of the implemented software with respect to others software (Rockwell 2000, StatSoft 2001) that also performs recognition of probability models using the traditional chi-squared method of data fitting.

We observed that the PM had excellent performance with respect to the normal, exponential, lognormal, and uniform distributions, compared to other software applications. However, the PM presented limitations with respect to the triangular distribution. We believe that this must be due to the estimates of the parameters of this distribution.

REFERENCES

Anderson, T. W. and Sclove, S. L., 1986. *An Introduction to the Statistical Analysis of Data*. Palo Alto: Scientific Press.

Barbetta, P. A. et al, 2004. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas S.A.

Bowker, A. H. and Lieberman, G. J., 1972. *Engineering Statistics*. New Jersey: Prentice-Hall.

Briec, W. et al, 2000. Returns to Scale on Nonparametric Deterministic Technologies: Simplifying Goodness-of-Fit Methods Using Operations on Technologies. *Journal of Productivity Analysis*, 14, 267–274.

Jankauskas, L. and McLafferty, S., 1995. Bestfit, Distribution Fitting Software by Palisade Corporation. *Winter Simulation Conference*. Newfield.

Jain, R., 1991. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley & Sons Inc, New York.

Johnson V., 2004. A Bayesian Chi-Squared Test for Goodness of Fit. *The University of Michigan Department of Biostatistics Working Paper Series*. Michigan.

Law, A. M., 1991. *Simulation Modeling and Analysis*. New York: McGraw-Hill.

Leithold, L., 1990. *The Calculus with Analytic Geometry*. New York: Harper & Row Publishers.

Moore, D. S. and McCabe, G. P., 1999. *Introduction to the Practice of Statistics*. New York: W.H. Freeman and Company.

Montgomery, D. C. and Runger, G. C., 1999. *Applied Statistics and Probability for Engineers*. New York: John Wiley & Sons, Inc.

Rockwell Software Inc, 2000. *Input Analyzer 6*. Milwaukee.

Romeu, J. L., 2003. The Chi-Square: a Large-Sample Goodness of Fit Test. *Start: Selected Topics in Assurance Related Technologies*, 10, 4, 1-6.

Knypstra S., 1998. *PQRS 3: Probabilities, quantiles and random samples*. Groningen.

StatSoft Inc, 2001. *Statistica 6: data analysis software system*. Tulsa.

Triola, M. F., 1998. *Elementary Statistics*. Boston: Addison Wesley Longman, Inc.

AUTHORS BIOGRAPHY

MARCELO TENÓRIO is an researcher at the Laboratory of Applied Statistics (LEA) and Performance Laboratory (PLab) at the Federal University of Santa Catarina (UFSC), Brazil. He received a master's degree in computer science from UFSC in 2005. He is a member of Brazilian Computer Society (SBC). His e-mail address is <marcelot@inf.ufsc.br> and his web address is <www.inf.ufsc.br/~marcelot>

SILVIA NASSAR is an associate professor at the Department of Computer Science at Federal University of Santa Catarina (UFSC), Brazil. She received a doctoral degree in production engineering from UFSC in 1996. Her research interests include simulation, artificial intelligence, Bayesian netwoks, and statistics. Her e-mail address is <silvia@inf.ufsc.br> and her web address is <www.inf.ufsc.br/~silvia>

PAULO FREITAS is an associate professor at the Department of Computer Science at Federal University of Santa Catarina (UFSC), Brazil. He received a doctoral degree in production engineering from UFSC in 1994. His research interests include simulation of computer systems for performance improvement, Monte Carlo methods, risk modeling and simulation, analysis for input modeling, and output analysis. He is a member of the Society for Computer Simulation (SCS) and the Brazilian Computer Soci-

ety (SBC). His e-mail address is <freitas@inf.ufsc.br> and his web address is <www.inf.ufsc.br/~freitas>

CARLOS MAGNO has been a Petrobras, SA, (Brazilian Energy Company) employee for 17 years. He is a doctoral candidate at the Federal University of Rio de Janeiro (UFRJ). His research interests include risk modeling and simulation, performance improvement, failure prediction, and artificial intelligence, all applied to Well Technology Engineering. His e-mail address is <cmcj@petrobras.com.br>