International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:9, No:11, 2015

A Brief Study about Nonparametric Adherence Tests

Vinicius R. Domingues, Luan C. S. M. Ozelim

***Abstract*—**The statistical study has become indispensable for various fields of knowledge. Not any different, in Geotechnics the study of probabilistic and statistical methods has gained power considering its use in characterizing the uncertainties inherent in soil properties. One of the situations where engineers are constantly faced is the definition of a probability distribution that represents significantly the sampled data. To be able to discard bad distributions, goodness-of-fit tests are necessary. In this paper, three non-parametric goodness-of-fit tests are applied to a data set computationally generated to test the goodness-of-fit of them to a series of known distributions. It is shown that the use of normal distribution does not always provide satisfactory results regarding physical and behavioral representation of the modeled parameters.

***Keywords*—**Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von-Mises, Nonparametric adherence tests.

I. Introduction

AS in most branches of science, the use of probabilistic and statistical methods has become extremely important in the development of modern Geotechnics. In particular, the concept of reliability of a venture has attracted considerable attention in recent years, boosting therefore the study of statistics for engineers.

It is known that one of the key points of the correct modeling of geotechnical data is the choice of a probability distribution representing the behavior of the data analysis. In general, the normal distribution has been the default choice of the vast majority of engineers. This can be explained by the wide applicability of this distribution; however, there are cases where the use of such random variable does not preserve the meaning and the physical behavior of the variables under consideration. Then arises the question on which distribution best fits a given sample of the target population. To provide answers to this question, adherence tests are used.

II. Adherence Tests

Adherence tests are statistical tests used to measure how well a given probability distribution is able to model the data set being analyzed. Adherence tests are also commonly called goodness of fit tests [1]. In general this type of testing is based on a hypothesis test, in which the null hypothesis, H0, is that data considered follow a given distribution test, while the alternative hypothesis, H1, considers that the data do not follow that distribution. There are two large groups of adherence tests with respect to prior knowledge of the parameters and distribution of data, namely parametric and non-parametric adherence tests.

A. Parametric Adherence Tests

Parametric adherence tests are those in which the distribution of the studied population is known or selected in some way and not into question, so that the hypotheses to be tested only involve population parameters. For example, in the case of an ANOVA, the rigid assumptions for the variables under comparison (normal variables, for example) incur in this type of testing not providing answers on how well other distributions fit to the data.

B. Nonparametric Adherence Tests

Nonparametric adherence tests, on the other hand, are valid tests for a broad range of distributions, so that their application provides an evaluation of the hypothesis of a given random variable being distributed in a distribution different from Normal. Still, this type of adherence test provides the means to as-certain the distribution that best fits the data analysis.

In this article, it will be explored the use of three known adherence tests in the evaluation of the probability distribution that best fits a set of data generated computationally. The three tests to be explored are: Kolmogorov-Smirnov test; Anderson-Darling test and test Cramer-Von-Mises.

III. Considered Adherence tests

The three adherence tests considered belong to the class of tests that uses the empirical distribution function (EDF) in the calculation of their statistics. Thus, it is necessary to first define this concept.

A. Empirical Distribution Function

Consider the ordered data set {x1, x2, x3, ..., xN}, whose empirical cumulative distribution function, Fn(X), one wants to calculate. Mathematically, the EDF may be given by:

$$F_n(x) = \frac{1}{N}\sum_{i=1}^{N}\left\lceil \frac{1+sign\left(x-x_i\right)}{2} \right\rceil \quad (1)$$

where ⌈x⌉ indicates the ceiling mathematical function that provides the greatest integer less than x and sign (x) is the sign function, in which the result is +1 if x is positive and -1 otherwise. In other words, the EDF a variable x is ranges from 0 to 1 and is increased by 1 / N when X passes each value in the ordered set of data. The EDF being de-fined, the adherence tests of interest can be studied.

B. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov belongs to the highest class of EDF based statistics, given the fact that it works with the

Vinicius R. Domingues and Luan C. S. M. Ozelim are with the Department of Civil and Environmental Engineering, University of Brasilia, Brasilia 70910-900, Brazil (e-mail: vinicius.rdomingues@gmail.com, luanoz@gmail.com).

biggest difference between the empirical distribution Fn(x) and the hypothetical F(x).

The hypothetical cumulative distribution comes from the test distribution, over which the null hypothesis of the hypothesis test resides. Thus, in the case of the Kolmogorov-Smirnov test, the null hypothesis is that the data are distributed according F (x) and the alternative hypothesis is that the data do not follow such distribution

Mathematically, the KS statistics of the Kolmogorov-Smirnov test may be defined as:

$$KS = \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right| \quad (2)$$

In order to be able to accept or reject the null hypothesis, the value of KS and its distribution should be assessed. Note that, intuitively, if the distribution pattern approximates the empirical distribution, the value of KS should be small. On the other hand, when KS is big, it is an indicative that the test does not characterize the distribution and the variable of interest. Put in another way, H0 can be rejected if the statistic value KS is greater than or equal to a given limit value, KSmax, the last of which depend on the confidence level adopted, α. Another important concept that arises from this evaluation process is called the p-value, which gives the probability of obtaining a statistic at least as extreme as the one calculated, assuming that the null hypothesis is true. Thus, one can accept H0 if the p-value associated is greater than the significance level.

In the present paper, the software Mathematica will be used to evaluate the statistics and p-values of the three tests considered.

C. Cramer-Von-Mises Test

Unlike the Kolmogorov-Smirnov test, Cramer-Von Mises test is a quadratic test of the empirical distribution function. This designation stems from the fact that this test works with the squared differences between the empirical and the hypothetical distributions [2]. This test has been applied to study a wide variety of problems in science. In [3], a Cramer–von Mises type test based on local time of switching diffusion process has been studied. On the other hand, in [4] a comparison between the Cramer-von Mises test and adaptive tests has been performed. In the present paper, on the other hand, we restrict our attention to the traditional test.

Mathematically, the CM statistics of the Cramèr-Von-Mises test may be defined as:

$$CM = N\int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 dF(x) \quad (3)$$

By knowing the data and the target distribution, the use of (3) becomes ready.

As in the case of the Kolmogorov-Smirnov test, the Cramer-Von-Mises test provides results which confirm whether or not H0, according to the statistical distribution of the CM or according to the p-value.

D. Anderson-Darling Test

As the Cramer-Von-Mises test, Anderson-Darling test is a quadratic test of the empirical distribution function. Furthermore, unlike what happens in (3), a weight is given to each observation inside the integral [5]. This test has been modified and applied to several distributions, such as power-law types [6] and extreme-value distributions [7]. On the other hand, as the case of the CM test, we restrict our analysis to the classical Anderson-Darling test.

Mathematically, the AD statistics of the Anderson-Darling test may be defined as:

$$AD = N\int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 F(x)\left(1 - F(x)\right) dF(x) \quad (4)$$

Interestingly, the main difference between Ander-son-Darling and Cramèr-Von-Mises tests is that the first gives greater weight to the data coming from the tails of the distributions. Similarly, to the other tests already mentioned, the hypothesis H0 should be rejected by comparison with extreme values.

Being defined the adherence tests to be used; the next step is to characterize a process of generating random data through the software Mathematica and subsequent application of the adherence tests.

IV. Randomizes Samples Generation and Adherence Tests Application

In the present paper, three randomized samples of 10^4 elements will be considered. In order to evaluate the applicability of the adherence tests, each one of the randomized samples will display a tendency that can be found in the practice of geotechnical engineering. The considered tendencies are: variables distributed according to a normal distribution; variables with asymmetric distribution and symmetric variable with long tail.

A. Normal Sample (D1)

According to a normal distribution, the Random Variate function of the software Mathematica was used for the generation of the distributed sample. The histogram of the respective generated data can be found in Fig. 1

The histogram is shown in Fig. 1 was generated from a normal distribution with a mean of 2 and standard deviation of 3.

Mathematically, one can define the probability density function of a normal variable with mean μ and standard deviation σ by means of the following equation:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \quad (5)$$

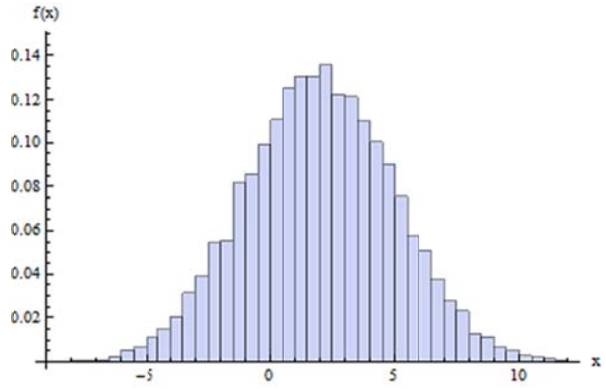

Fig. 1 Random sample from a normal distribution with 10^4 elements

B. Asymmetric Sample (D2)

In order to generate the sample that follows a skewed distribution the RandomVariate function of the software Mathematica was also used. It was considered that the data follow a Levy distribution for the generation process. The histogram of the respective generated data can be found in Fig. 2. In this case the Levy distribution with location parameter 4 and dispersion parameter 2 was used as source.

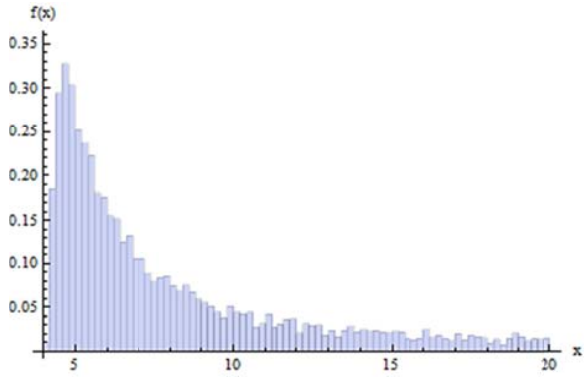

Fig. 2 Random sample from a Levy distribution with 104 elements

Mathematically, one can define the probability density function of a Levy variable with location parameter μ and dispersion parameter σ from (6):

$$f(x)=\frac{e^{-\frac{\sigma}{2(x-\mu)}}\left(\frac{\sigma}{x-\mu}\right)^{3/2}}{\sqrt{2\pi\sigma}},x>\mu \tag{6}$$

C. Sample with Long Tail (D3)

Finally, for the generation of the sample distributed according to a distribution with long tail the RandomVariate function of the software Mathematica was used again. The histogram of the data generated can be found in Fig. 3.

To generate the data presented in Fig. 3, a student's t distribution with location parameter 0, scale parameter 3 and 4 degrees of freedom was used.

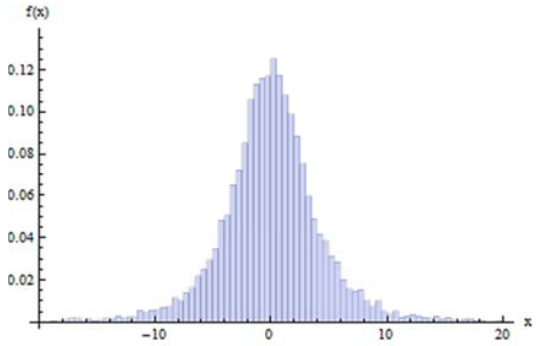

Fig. 3 Random sample of a Student's t distribution with 104 elements

Mathematically, it is possible to define the probability density function of a t-student variable with location parameter μ, scale parameter σ and ν degrees of freedom, through (7):

$$f(x)=\frac{\left(\frac{\nu}{\nu+\frac{(x-\mu)^2}{\sigma^2}}\right)^{\frac{1+\nu}{2}}}{\sqrt{\nu}\sigma Beta\left[\frac{\nu}{2},\frac{1}{2}\right]} \tag{7}$$

where Beta [x, y] is the special function defined by the following integral:

$$Beta[x,y]=\int_0^1 t^{x-1}(1-t)^{y-1}dt \tag{8}$$

As the samples and its source distributions are set, it is possible to proceed to the data analysis and application of the adherence tests.

D. Basic Data Analysis

Based on the generated data it is possible to calculate the average, standard deviation and skewness for each one of the considered sets. Table I contains those information for each data group.

TABLE I
BASIC DATA ANALYSIS

Data	*Average*	*Standard deviation*	*Skewness*
D1	1.99	3.01	0.02
D2	6037	228169	74.92
D3	1.01	4.19	0.13

As expected, the D1 and D3 data are practically symmetric, while the D2 data is extremely asymmetrical. Still, according to the distributions used to generate the data, averages and standard deviations are compatible.

After the basic data analysis, the next step is the determination of which distributions are going to be tested according to the goodness-of-fit.

E. Distributions to Be Tested

As previously argued, to use the adherence tests the test distributions must be determined. For simplicity, the classes of test distributions will be the same used for the generation of data that is the Normal, Levy and Student's t distributions.

Each of the data will be tested against three distributions whose parameters were previously determined through the maximum likelihood estimation provided in Mathematica software. The parameters for each distribution to be tested are shown in Table II.

TABLE II
PARAMETERS AND DISTRIBUTIONS TO BE TESTED

Data	*Normal* (μ, σ)	*Levy* (μ, σ)	*T-student* (μ, σ, ν)
D1	(2, 3)	(-9.9, 10.8)	(2, 3, 412)
D2	(6037, 2.3 10^5)	(4, 2)	(10^{11}, 10^{11}, 27368)
D3	(1, 4.2)	(-29, 29)	(1, 3, 4)

F. Adherence Tests

By applying the adherence tests discussed in this article on the generated data and using as test distributions those presented in Table II, Tables III-V can be generated.

TABLE III
GOODNESS-OF-FIT TESTS RESULTS (P-VALUES) FOR THE D1 DATA CONSIDERING THE TEST DISTRIBUTIONS SHOWN IN TABLE II

Distribution	*Kolmogorov-Smirnov*	*Cramèr-Von-Mises*	*Anderson-Darling*
Normal	0.92	0.91	0.93
Levy	0	0	0
T-student	0.94	0.92	0.95

It is worth noting that in cases where the D1 data set is considered, adherence tests clearly show that only the Normal and Student's t distributions can represent the sample. It is further noted that the fact that both distributions mentioned fit well to the data follows from the consideration that the Normal distribution is the limiting case of t distribution when it has a large number of degrees of freedom. Thus, it can be noted that adherence tests used are accurate on the characterization of the distribution process from which the D1 data set belongs.

TABLE IV
GOODNESS-OF-FIT TESTS RESULTS (P-VALUES) FOR THE D2 DATA CONSIDERING THE TEST DISTRIBUTIONS SHOWN IN TABLE II

Distribution	*Kolmogorov-Smirnov*	*Cramèr-Von-Mises*	*Anderson-Darling*
Normal	0	0	0
Levy	0.22	0.23	0.30
T-student	0	0	0

By looking at Table IV, it is easy to perceive that all tests suggest the Levy distribution as the best distribution of the data. In fact, the p-value is 0 for the other distributions, implying the rejection of H0 for cases of Normal and Student's t distributions. Therefore, all the tests are effective in characterizing the D2 set of data.

TABLE V
GOODNESS-OF-FIT TESTS RESULTS (P-VALUES) FOR THE D2 DATA CONSIDERING THE TEST DISTRIBUTIONS SHOWN IN TABLE II

Distribution	*Kolmogorov-Smirnov*	*Cramèr-Von-Mises*	*Anderson-Darling*
Normal	0	0	0
Levy	0	0	0
T-student	0.52	0.54	0.64

It is easy to notice from the Table V that the adherence tests indicate that the D3 data are distributed according to a Student's t distribution. This fact shows, once again, that the considered goodness-of-fit tests are effective as regards the determination of the probability distribution of a data set.

V. CONCLUSIONS

The role of Statistics in exact sciences has been more considered with passing of time. Especially in the geotechnical engineering field, the inherent variability of soil parameters finds a great ally in this area of the knowledge.

One of the most common situations when an engineer is analyzing a set of data is the proper choice of probability distribution that represents it. This choice taken must be as reliable as it can possibly be, so that the physical behavior of the modeled variable is not lost.

The adherence tests come as powerful allies to help determine which distribution is better adjusted to the database. Therefore, it is imperative that there is complete knowledge of them for a successful statistic modeling process.

One of the aims of this paper was a discussion about some the features of three of the most known non-parametric adherence testes, such as: Kolmogorov-Smirnov test, Cramèr-Von-Mises test and Anderson-Darling test.

From computational experiments, it was shown how these tests can be used to determine the probability distribution that adapts better to the data in analysis. This way, it is believed that a contribution to the diffusion of traditional statistics tools in the field of geotechnical engineering has been done.

ACKNOWLEDGMENTS

The authors would like to thank FAP-DF for the financial support, provided as grants (number 193.000.564/2015) to present the paper at the conference.

REFERENCES

[1] Torman, V. B. L.; Coster, R.; Riboldi, J. 2012. Normalidade de variáveis: métodos de verifcação e comparação de alguns testes não paramétricos por simulação, Rev. HCPA, Porto Alegre, v.32, n.2, p.227-234.
[2] Anderson, T. W. and Darling, D. A. 1954. A Test of Goodness-of-Fit. Journal of the American Statistical Association, 49: 765–769.
[3] Gassem, A. 2011. On Cramér–von Mises type test based on local time of switching diffusion process. Journal of Statistical Planning and Inference, Vol 141(4), P. 1355–1361.
[4] Inglot, T. and Ledwina, T. 2004. On consistent minimax distinguishability and intermediate efficiency of Cramér–von Mises test. Journal of Statistical Planning and Inference, Vol. 124 (2), P. 453–474.
[5] D'Agostino, R. B. and Stephens, M. A.. 1986. Goodness-of-Fit Techniques. New York: Marcel Dekker. ISBN 0-8247-7487-6, 1986.
[6] Coronel-Brizio, H. F. and Hernández-Montoya, A. R. 2010. The Anderson–Darling test of fit for the power-law distribution from left-

censored samples. Physica A: Statistical Mechanics and its Applications, Vol. 389 (17), P. 3508–3515.

[7] Heo, J-H., Shin, H., Nam, W., Om, J., Jeong, C. 2013. Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter. Journal of Hydrology, Vol. 499 (30), P. 41–49.