XL CILAMCE
IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING
NOVEMBER 11-14, 2019
Praiamar Natal Hotel & Convention
Natal, RN-BRAZIL

# STATISTICAL INFERENCE TECHNIQUES APPLIED TO LARGE SAMPLES

**Hugo Vinícius Ferreira Azevedo**
**Eduardo Toledo de Lima Junior**
*hugovazevedo@gmail.com*
*limajunior@lccv.ufal.br*
*Scientific Computing and Visualization Laboratory, Federal University of Alagoas*
*Campus A. C. Simões, Tabuleiro dos Martins, 57072-900, Maceió, Alagoas, Brasil*

**Abstract.** The growing evolution of structural materials and analysis models demands a proper understanding of the safety levels adopted in the design practice. The uncertainties inherent to structural engineering problems can be evaluated from the statistical description of their design variables – dimensional, mechanical and loading ones – and incorporated into the analysis by using structural reliability models. The statistical characterization is a crucial part of the whole process, being carried out by inference techniques, as the goodness-of-fit (GoF) tests, which verify if sample data fits a theoretical distribution model, given a specified significance level. The GoF tests can be very sensitive to large samples - from the order of thousands of data, becoming unsuitable for this kind of analysis. Alternative techniques can be applied to handle large datasets. This is the case of the subsampling principle, which involves the random, non-biased withdrawal of subsamples from the original dataset, so that a part of the sample is used in the parameterization of the model and another part is used for the GoF test, in varying proportions, to be studied. In addition, the AIC and BIC (Akaike and Bayesian Information Criteria) values can be used as a preliminary indicative of the congruence between data sample and a theoretical distribution. It is proposed the analysis of two different samples, in order to apply some inference techniques, implemented in Python language. It is expected to contribute with the characterization of large samples for studies in data science and reliability analysis applied to engineering.

**Keywords:** Large Samples, Statistical Inference, Resampling, Kolmogorov-Smirnov.

## 1 Introduction

A fundamental premise for probabilistic and risk-based methods in engineering, is the proper understanding of the uncertainties associated with the design variables. For example, the characteristic compressive strength of concrete $(f_{ck})$ is evaluated from a batch, in which the dataset containing the resistance of each specimen is taken as a random variable (r.v.). The statistical characterization of the sample data has the purpose of modeling statistical distributions to represent the data, by using inference techniques, as the so-called Goodness-of-Fit (GoF) tests.

These tests provide some information about the fit between sample and theoretical model as, for instance, the p-value, whose higher values indicate a good agreement. There is a drawback on the usage of these tests, related to its sensitivity to the sample size. As shown in Lin et al. [1] and Baird and Harlow [2], as the sample size increases, the p-value are drastically reduced, even in cases where the good agreement between data and the candidate model is evident.

Therefore, this paper aims to explore alternative ways to handle large data series – from the order of thousands of data - including subsampling, bootstrap resampling and information criteria (IC).

## 2 Methodology

The idea behind of using subsampling techniques is to select a subset that is smaller in size than the original sample, but is representative enough for the GoF analysis. Two approaches are proposed, the first is to randomly remove a subset of *n* elements with repetition of the original sample, which is used for both parametrization and testing. A thousand realizations of this subsampling are performed using the Monte Carlo method, in order to statistically evaluate the p-value; and the second consists of randomly select part of the sample (ranging from 85% to 95%) to make the parametrization of the candidate model and the other part (ranging from 15% to 5%) to perform the GoF test.

The GoF test used in this work is the Kolmogorov-Smirnov (K-S) test. The test statistic $D_n$ is obtained by the comparison between the cumulative histogram of a sample $(S(x))$ and the cumulative frequency distribution of the candidate model $(F(x))$ on each sample point $x$. If the maximum difference between the theoretical and observed frequencies, defined in Eq. 1, is less than an expected critical value, defined for a certain sample size, the model fits the data, at an $\alpha$ significance level. This positive result is referred in statistics as the non-rejection of the null hypothesis $H_0$, which states that the data comes from the distribution model. A significance level of 5% is usually adopted as a reference value.

$$D_n = max|F(x) - S(x)|. \tag{1}$$

Another way to interpret the K-S test, is using the p-value associated with the distance $D_n$. The definition of the p-value is that it is "the probability, computed assuming that $H_0$ is true, that the test statistic would take a value as extreme or more extreme than that actually observed". In other words "the smaller the p-value, the stronger the evidence against $H_0$ provided by the data." (both quotes are from Moore [3]). If the p-value provided in the test is greater than the $\alpha$ significance level, the theoretical distribution is accepted for modeling the data.

As Burnham and Anderson [4] emphasize the importance of selecting models based on scientific principles, this work uses the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), both serve as a filter among the models to be tested in the K-S test. The analysis mode of both is the same, the model that obtains the lowest AIC or BIC value is the best candidate to represent the data sample. The Akaike Information Criteria is defined by Akaike [5] as:

$$AIC = -2logL(x_n|\hat{\boldsymbol{\theta}}) + 2p. \tag{2}$$

Where $p$ is the number of parameters to be estimated in the model, $L(x_n|\hat{\boldsymbol{\theta}})$ is the likelihood function, defined as $L(x_1, x_2, \ldots, x_n|\hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} f(x_n|\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}$ are the unknown parameters of each model. And the Bayesian Information Criteria, where $n$ is the sample size, is defined by Schwarz [6] as:

$$BIC = -2log f(x_n|\boldsymbol{\theta}) + plog(n). \tag{3}$$

Moreover, adopting the principle that p-values are random variables as showed in Murdoch et al. [7], another technique applied is the parametric analysis on how the p-values change as the subsample size increases, creating Confidence Intervals (CI) to ensure the tolerance of the p-value.

The dataset used in this work comprises 2 samples, each one with 3669 elements. The analyses are developed in *Python* language, with the use of the *Scipy* library (Jones et al. [8]) to parametrize and perform the GoF test on the samples. Considering the sample size, the K-S test rejects all the model distributions tested, due to its inherent limitation. Although the histograms, shown in Fig. 1, of the samples present a typical skewed behavior, predicted by classical distribution models, the GoF test provides fairly low p-values, for a lot of tests performed on the both samples. The goal is to achieve a proper inference about which distributions are most suitable for the data, still applying the K-S test.

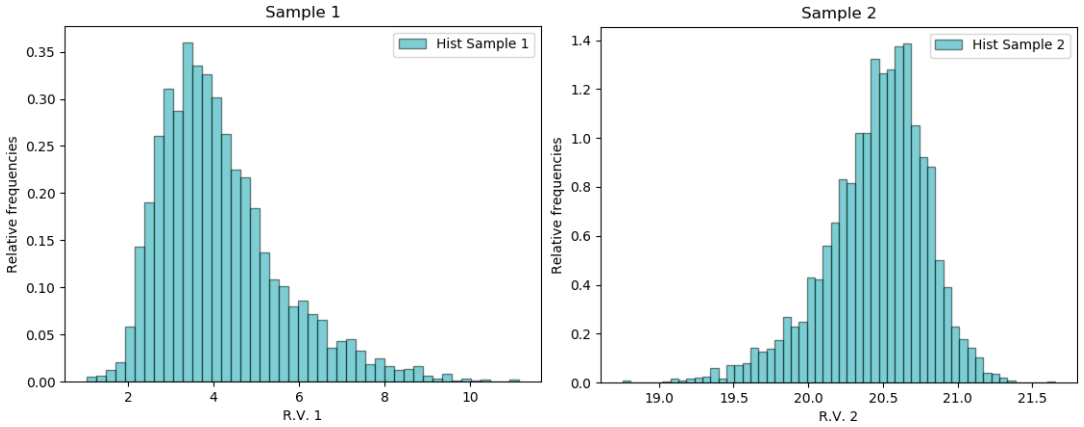


Figure 1. Histogram - Samples 1 and 2.

## 3 Results and Discussions

Akaike and Bayesian information criteria are used as a filter among the candidate distributions. The three distributions (among nine tested) with lower values of the IC are addressed to perform the GoF test. For Sample 1, the distributions Gumbel R (right-skewed Gumbel), Moyal and Maxwell are adopted as candidate models and for Sample 2, it is chosen Normal, Logistic and Gumbel L (left-skewed Gumbel).

In the first analysis, subsamples with different sizes - from 100 up to 1400 - are extracted from the original datasets, via bootstrap technique, and subject to the inference procedure considering the three best pre-ranked models. It is observed that AIC and BIC provide similar responses, hence, only the curves obtained with the BIC values are shown in Fig. 2.

For each subsample size, the parametrization and GoF test are evaluated a thousand times, enabling a statistical study of p-values. The average p-values, and corresponding CI are shown in Fig. 3 and Fig. 4.

It can be seen that the confidence bands narrow as the subsample size increases, indicating that the p-value estimate becomes more accurate, specially in the case of sample 2. Gumbel R and Moyal distributions stand out as good models for sample 1 data, with p-values over than 20%. However, the Maxwell model performs well in fitting the data for subsamples smaller than 500 elements, approximately, violating the reference value $\alpha = 5\%$ beyond this point. Regarding sample 2, Gumbel L and Logistic models are good surrogates of the data up to 1200 and 1000 elements in the subsample, respectively. The average

p-value for the Normal distribution remains above 5% up to 500 elements. As a characteristic inherent to this sample, the p-value decreases severely for the three candidate models, in the early branches of the curves, up to around 500 elements size.

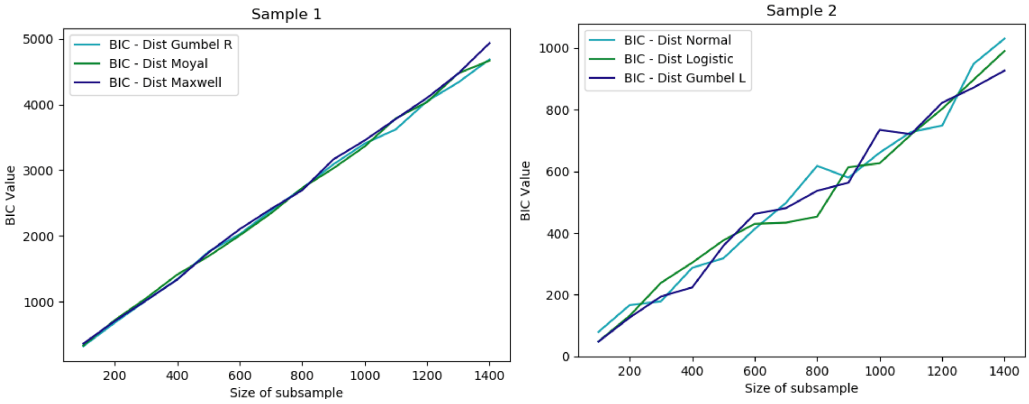


Figure 2. BIC - Sample 1 and Sample 2.

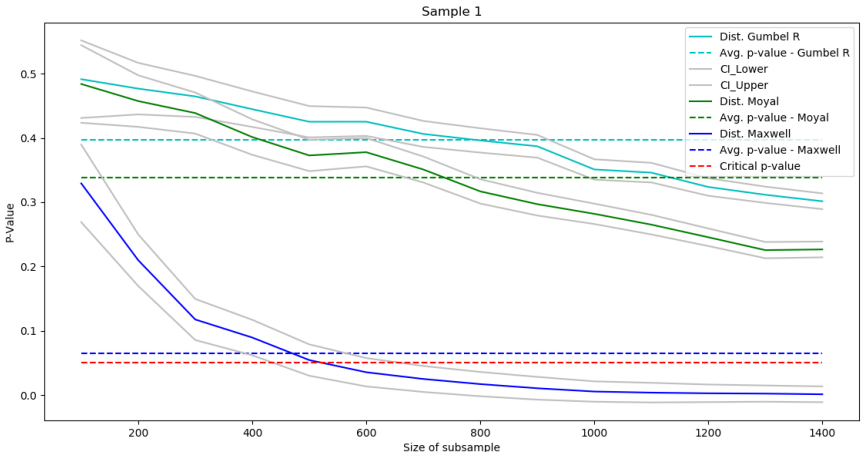


Figure 3. P-values for different subsample sizes - Sample 1.

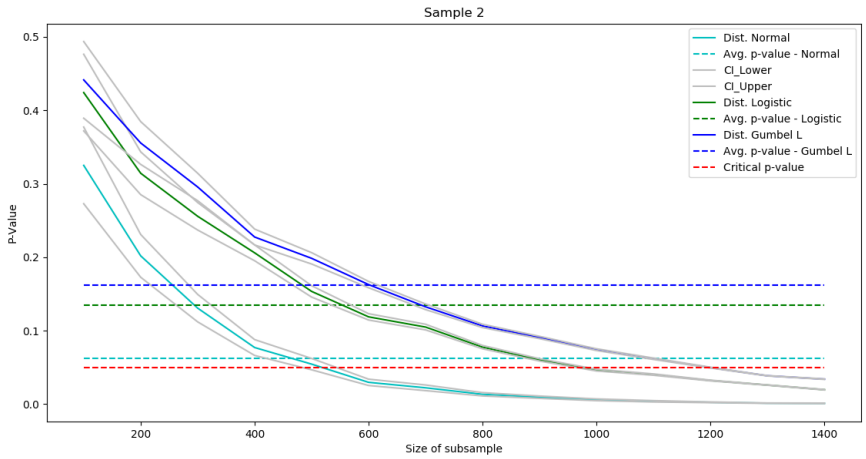


Figure 4. P-values for different subsample sizes - Sample 2.

The second approach, in which the original data is divided in two subsamples, for parametization and GoF test, confirms that using a small portion for the test could be a good alternative to infer which theoretical distribution fits the data, as demonstrated in Fig. 5. For both samples 1 and 2, there is a quasi-linear increase in the percentage of approved data with the reduction of test subsample size.

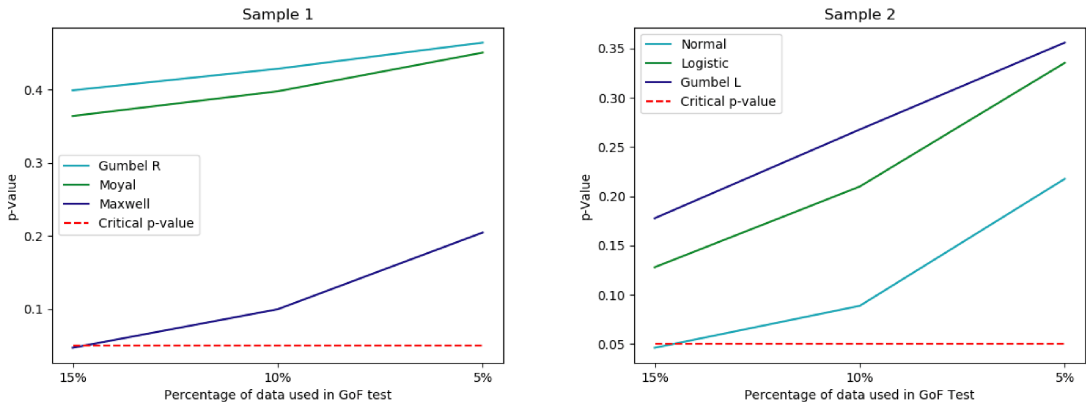


Figure 5. Resampling with divided data.

## 4 Conclusions

Some techniques based on resampling/subsampling were applied to the inference of large datasets. It is noted that the analysis based on Information Criteria is affected by the subsample size, not allowing to define the theoretical distribution more suitable to the data, among different candidate models.

However, by applying the bootstrap resampling with different sizes and the sample partition strategy, some coherent results were obtained for the two datasets tested, suggesting these techniques as viable protocols for the characterization of large samples.

An important question that arises is how to evaluate the size of the subsample so that it is adequate to represent the original data. The classical T-test and ANOVA procedures are limited to evaluate the equivalence between the mean and the variance, respectively, of the original and reduced samples, but do not address the distribution that generates the samples.

## Acknowledgements

To National Council for Scientific and Technological Development (CNPq) for the financial aid.

## References

[1] Lin, M., Lucas, H. C., & Shmueli, G., 2013. Too Big to Fail: Large Samples and the p -Value Problem. *Inf. Syst. Res.*, vol. 24, n. 4, pp. 906–917.

[2] Baird, G. L. & Harlow, L. L., 2016. Does One Size Fit All? A Case for Context-Driven Null Hypothesis Statistical Testing. *J. Mod. Appl. Stat. Methods*, vol. 15, n. 1, pp. 100–122.

[3] Moore, D. S., 1999. *The Basic Practice of Statistics*. W. H. Freeman \& Co., New York, NY, USA, 2nd edition.

[4] Burnham, K. P. & Anderson, D. R., 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.*, vol. 33, n. 2, pp. 261–304.

[5] Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, vol. 19, n. 6, pp. 716–723.

[6] Schwarz, G., 1978. Estimating the Dimension of a Model. *Ann. Stat.*, vol. 6, n. 2, pp. 461–464.

[7] Murdoch, D. J., Tsai, Y. L., & Adcock, J., 2008. P-values are random variables. *Am. Stat.*, vol. 62, n. 3, pp. 242–245.

[8] Jones, E., Oliphant, T., Peterson, P., et al., 2001–. SciPy: Open source scientific tools for Python.