A Statistical Toolkit for Data Analysis

S. Donadio, S. Guatelli, B. Mascialino, M.G. Pia, INFN, University of Genova, Italy
A. Pfeiffer, A. Ribon, CERN, Geneve, Switzerland
P. Viarengo, IST, National Cancer Research Institute, Genova, Italy

Abstract

The present project aims to develop an open-source and object-oriented software Toolkit for statistical data analysis. Its statistical testing component contains a variety of Goodness-of-Fit tests, from Chi-squared to Kolmogorov-Smirnov, to less known, but generally much more powerful tests such as Anderson-Darling, Goodman, Fisz-Cramer-von Mises, Kuiper, Tiku. Thanks to the component-based design and the usage of the standard abstract interfaces for data analysis, this tool can be used by other data analysis systems or integrated in experimental software frameworks.
This Toolkit has been released and is downloadable from the web. In this paper we describe the statistical details of the algorithms, the computational features of the Toolkit and describe the code validation.

INTRODUCTION

Statistical methods play a significant role throughout the life-cycle of physics experiments, being an essential component of physics analysis. In spite of this, only a few basic tools for statistical analysis were available in the public domain FORTRAN libraries for physics. Nowadays the situation is unchanged even among the libraries of the new generation. For this reason, we decided to launch a new project, with the aim of creating an open-source, up-to-date and sophisticated object-oriented statistical Toolkit for physics data analysis.
In this paper we will focus our attention on a specific component of the statistical Toolkit, that is made-up by a collection of Goodness-of-Fit (**GoF**) [1] tests. Its aim is to provide a wide set of algorithms to test whether the distributions of two variables are compatible.

GOODNESS-OF-FIT TESTING

The applications of statistical comparisons of distributions in physics are manyfold:

- regression testing, in various phases of the software life-cycle,
- validation of simulation through comparison to experimental data,
- comparison among different experimental distributions,
- comparison between experimental data and theoretical functions,
- monitoring detector behavior with respect to a reference in online DAQ.

Classical statistical-inference techniques are based on fairly specific assumptions regarding the nature of the underlying distribution. Usually both its form and some parameter values must be explicitly stated in the hypothesis and this requires a certain level of knowledge about what is going to be compared. When this is not the case, an alternative set of statistical techniques is available: distribution-free or non-parametric procedures. Non-parametric testing, in fact, allows the formulation of an hypothesis which is not a statement about parameter values.
Non-parametric statistics include **Goodness-of-Fit** testing. These tests measure the compatibility of a random sample with a theoretical probability distribution function or between the empirical distributions of two different populations coming from the same theoretical distribution. From a general point of view, the aim may consist also in testing whether the distributions of two random variables are identical against the alternative that they differ in some way.
More in particular, in Goodness-of-Fit testing, the null hypothesis can be concerned only with the form of the population:

$$\mathbf{H_0} : \mathbf{F} = \mathbf{G} \qquad \text{for all } x$$

against an alternative broad one

$$\mathbf{H_1} : \mathbf{F} \neq \mathbf{G} \qquad \text{for some } x.$$

In this kind of tests the acceptance of the null hypothesis $\mathbf{H_0}$ means that the researcher will be able to specify the distribution analysed. Since the alternative includes differences in location, scale, other parameters, form or any combination of these, the rejection of the null hypothesis will not provide much specific information.

Chi-squared Test

With the purpose of quantifying the measure of the deviation between two distributions, many software toolkits for physics data analysis solve the problem by means of the well known and wide-spread Chi-squared test. This test was introduced to describe discrete distributions, but it can be useful also in case of unbinned distributions. In this case the researcher must group data into classes, sacrificing in this way a good deal of the information conveyed by the distribution itself. In spite of the fact that this

test has a general applicability, it must be noticed that the Chi-squared asymptotic distribution is *not* valid if the theoretical frequencies involved in the computation are lower that 5. For these reasons, a powerful and up-dated statistical Toolkit for physics data analysis should supplement the Chi-squared test with other statistical tests, involving individual sample values.

The Kolmogorov-Smirnov Family

A common alternative to the Chi-squared tests includes tests based on Kolmogorovs Empirical Distribution Function (*edf*) definition [2]. These tests are: **Kolmogorov-Smirnov** [2, 3], **Goodman** [4] and **Kuiper** [5]. In any case, the test statistics is a linear function of the maximum vertical distance between the *edf*s of the two distributions. Tests belonging to the Kolmogorov-Smirnov family can be applied only to *continuous* distributions. Some other limitations are related with the fact that these tests tend to be more sensitive near the center of the distribution with respect to the tails.

The Anderson-Darling Family

The Anderson-Darling family of tests measures the integrated quadratic deviation of the two *edfs* suitably weighted by a weighting function $\psi(F(x))$. Different mathematical formulations of the weighting function ψ define the **Anderson-Darling** [6], **Fisz-Cramer-von Mises** [7, 8, 9] and **Tiku** [10] test statistics. These tests can be performed on both binned or unbinned data and they are satisfactory for symmetric and right-skewed distributions. It must be pointed out the fact that these tests give more weight to the tails than the test belonging to the Kolmogorov-Smirnov type.

Power of the statistical tests

Dealing with a non-parametrical set of tests a *proper* evaluation about the power of these tests cannot be made. In general, the Chi-squared test, for its simplicity, is the least powerful one because of information loss due to data grouping (binning). On the other hand, all the tests based on the supremum statistics are more powerful than the Chi-squared one, focusing only on the maximum deviation between the two EDFs. The most powerful tests are the ones containing a weighting function, as the comparison is made all along the range of x, rather than looking for a marked difference at one point [11].

GOF TOOLKIT ARCHITECTURE

The Toolkit has been developed following a rigorous software process (*United Software Development Process*), mapped onto the **ISO 15504** guidelines. With the aim of guaranteeing the quality of the product, the software development follows a spiral approach and the software life cycle is iterative-incremental, based on a User Requirements Document and providing Requirements Traceability.
The project adopts a solid architectural approach to offer the functionality and the quality needed by the user, to be maintainable over a large time scale and to be extensible, accommodating in this way future evolutions of the user requirements.
The component-based design of the Toolkit adopting both object-oriented techniques and generic programming, facilitates the re-use of the Toolkit as well as its integration in other data analysis frameworks.

Core component

The main features of the core component of the **GoF** Toolkit are:

- the Toolkit distinguishes input distributions on the basis of their type, as binned and unbinned data must be treated in different ways from a statistical point of view,
- the whole comparison process is managed by one object (*ComparatorEngine*), which is parametrised on the distribution type and on the algorithm selected by the user.

The comparison returns to the user a statistics comparison result object, giving access to the computed value of the test statistics, the number of degrees of freedom and the quality of the comparison (p-value).
Every algorithm contained in the **GoF** Toolkit is specialised for only one kind of distribution (binned or unbinned). In this way the user can access only those algorithms whose applicability conditions fit the kind of distribution he/she deals with.
The object-oriented design allows for an easy extension of the **GoF** Toolkit to new algorithms without interfering with the existing code.

User layer

From the user's point of view, the object-oriented techniques adopted together with the standard **AIDA**(*Abstract Interfaces for Data Analysis*) [12] interfaces shield the user from the complexity of both the architecture of the core components and the computational aspects of the mathematical algorithms implemented. The user layer manages the interaction between the user and the core statistical component. All the user has to do is to choose the most appropriate algorithm (in practice writing one line of code) and to run the comparison. This implies that the user does not need to know statistical details of any algorithm, he/she also does not have to know the exact mathematical formulation of the distance nor of the asymptotic probability distribution he/she is computing. Therefore the user can concentrate on the choice of the algorithm relevant for his/her data. As an example, if the user tries to apply the Kolmogorov-Smirnov comparison to binned data, the **GoF** will not run

the comparison, as the class *KolmogorovSmirnovComparisonAlgorithm* is defined to work only on unbinned distributions.

SOFTWARE TESTING

On the basis of the rigorous software process that the project adopted, the **GoF** Toolkit code has undergone a test process, consisting in unit, integration and system tests. Testing focuses primarily on the evaluation or assessment of quality of the software product, guaranteeing the correctness and robustness of the software. It involves: (1) finding and documenting defects in software quality, (2) validating the software product functions as designed and (3) validating that the requirements have been implemented appropriately.
Unit testing involved every class of the **GoF** Toolkit; integration testing was also performed on every complete statistics algorithm included in the **GoF** Toolkit, with the aim of validating the whole **GoF** statistics process of comparison.
All the tests performed are distributed as part of any public Toolkit release. Moreover, test result summaries, demonstrating the correct functionality of the Toolkit, are included as part of the documentation of a Toolkit release and are available on the web [1].
At the user layer level of the architecture, another set of tests verifies the integration of the elements in the Comparison package and the correct functioning of the GoF component as a whole. The testing strategy is based of Monte Carlo trials: a large number of pseudo-experiments is performed, each consisting in drawing randomly two samples from the same parent distribution. The aim is to compare the p-values returned by the statistical test with the ones calculated directly from the distribution of distances, using the definition of p-value (*i.e* the probability to get a distance greater than or equal to the one observed) when two samples are drawn from the same parent distribution.
The statistical and mathematical consistency of the algorithms included in the **GoF** Toolkit has been evaluated reproducing examples from some reference statistics books ([13, 14, 15] among the others). This validation is intended to demonstrate that the code is consistent with the mathematics of the algorithms comparing the numerical results obtained by means of the **GoF** Toolkit with the ones published by the authors. It must not be considered as an intrinsic comparison among the specific algorithms.
In any test the **GoF** Toolkit reproduces exactly the numerical result of the test statistics computed by the authors.

CONCLUSIONS

The **GoF** Toolkit is an easy, up-to-date, and powerful tool for data comparison in physics analysis. It is the first statistical Toolkit providing a variety of sophisticated and powerful algorithms in physics analysis.
The Toolkit employes a rigorous software process, it uses object-oriented techniques as well as generic programming and features a component-based design. The component architecture and the adoption of AIDA interfaces facilitates the re-use of the Toolkit in other environments.
The code is downloadable from the web [1] together with ample documentation.
For all the features described, the **GoF** Toolkit constitutes a step forward in physics data analysis quality.

REFERENCES

[1] http://www.ge.infn.it/geant4/analysis/HEPstatistics/

[2] A.N. Kolmogorov, Giorn. Ist. Ital. Attuari 4 (1933) 1.

[3] N.V. Smirnov, Ann. Math. Stat. 19 (1948) 279.

[4] L.A. Goodman, Psychol. Bull. 51 (1954) 160.

[5] N.H. Kuiper, Proc. Koninkl. Neder. Akad. van. Wetenschappen A 63 (1960) 38.

[6] T.W. Anderson and D.A. Darling, Ann. Math. Stat. 23 (1952) 193.

[7] H. Cramèr, Skand. Aktuarietidskrift 11 (1928) 171.

[8] R. von Mises, Wahrscheinliehkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik, Leipzig, F. Deuticke, 1931.

[9] M. Fisz, Ann. Math. Statist. 31 (1960) 427.

[10] M.L. Tiku, Biometrika 52 (1965) 630.

[11] S. Kotz and N.L. Johnson (eds), Breakthrougth in statistics, vol II, Springer Verlag, New York, 1992.

[12] `http://AIDA.freehep.org`

[13] D. Piccolo, Statistica, 1^{st} ed., Ed. Il Mulino, Bologna Italy, 1998.

[14] G. Landenna and D. Marasini, Metodi statistici non parametrici, 1^{st} ed., Ed. Il Mulino, Bologna Italy, 1990.

[15] M.G. Kendall and A. Stuart, The advanced theory of statistics, 2^{nd} ed., C. Griffin & Company Limited, London, 1968.