A software toolkit for statistical data analysis

G.A.P. Cirrone, S. Donadio, S. Guatelli, L. Lista, A. Mantero, B. Mascialino, S. Parlati, M.G. Pia
INFN
A. Pfeiffer, A. Ribon
CERN
P. Viarengo
IST, Genova, Italy

We present a project in progress to develop a software toolkit for statistical data analysis. The toolkit is based on advanced software technologies, integrating generic programming techniques with object oriented methods, and adopts a rigorous software process, to ensure a high quality of the product. Thanks to the component-based architecture and the usage of the standard **AIDA** interfaces, this tool can be easily used by other data analysis systems or integrated in experimental frameworks. The initial component of the system addresses goodness of fit tests; its applications include the comparisons of data distributions in a variety of use cases typical of HEP experiments: regression testing (in various phases of the software life-cycle), validation of simulation through comparison to experimental data, comparison of expected versus reconstructed distributions, comparison of different experimental distributions - or of experimental with respect to theoretical ones - in physics analysis, monitoring detector behavior with respect to a reference in online DAQ. The system will provide the user the option to choose among a wide set of goodness-of-fit tests (chi-squared, Kolmogorov-Smirnov, Anderson-Darling, Lilliefors, Kuiper, Cramer-von Mises, etc.), specialised for various types of binned and unbinned distributions. Its flexible design makes it open to further extension to implement other tests. This system would represent a significant improvement with respect to the current availability of comparison tests in HEP libraries, limited to the chi-squared and Kolmogorov-Smirnov algorithms. We present the architecture of the toolkit, the detailed design of the basic statistical testing component and preliminary results of its application, in particular concerning the physics validation of the Geant4 Simulation Toolkit. We discuss the openness of the project, welcoming contributions from experts and user requirements from experiments.

1. Introduction

Statistical methods play a significant role throughout the life-cycle of HEP 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 HEP. Nowadays the situation is unchanged even among the libraries of the new generation. The aim of this project is to build an open-source, up-to-date and sophisticated object-oriented statistical toolkit for HEP 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 in order to test whether the distributions of two variables are compatible.

2. The Goodness of Fit Statistical Toolkit

The applications of statistical comparisons of distributions in HEP are manyfold: regression testing (in various phases of the software life-cycle), validation of simulation through comparison to experimental data, comparison of different experimental distributions - or of experimental with respect to theoretical ones - in physics analysis, monitoring detector behavior with respect to a reference in online DAQ. From a mere statistical point of view, the problem consists in testing the non-parametric null hypothesis

$$\mathbf{H_0} : \mathbf{F} = \mathbf{G}$$

against an alternative one

$$\mathbf{H_1} : \mathbf{F} \neq \mathbf{G} \quad \mathbf{or} \quad \mathbf{F} < \mathbf{G} \quad \mathbf{or} \quad \mathbf{F} > \mathbf{G}.$$

Of course, 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.

2.1. GoF statistical features

With the purpose of quantifying the measure of the deviation between the two distributions, many software toolkits for HEP data analysis solve the problem by means of the well known and wide-spread chi-squared test. This test is studied to describe discrete distributions, but it can be useful also in case of unbinned distributions. In this case the researcher is compelled to 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 HEP data analysis should supplement the chi-squared test with other statistical tests, involving individual sample values.

In order to compare unbinned distributions, the **GoF** toolkit includes a wide set of tests dealing with Kolmogorov's empirical distribution function (EDF). Using this toolkit the user is able to compare two EDFs selecting tests based on the supremum statistics:

- Kolmogorov-Smirnov test [2],
- Goodman approximation of Kolmogorov-Smirnov test [3],
- Kuiper test [4],

and together with tests based on the measure of integrated deviations of the two EDFs, multiplied by a weighting function:

- Cramer-von Mises test [5] [6],
- Anderson-Darling test [7].

Due to its mathematical formulation the Anderson-Darling test is favourable in case of fat-tailed distributions. A recent paper by Aksenov and Savageau [8] states that this last test statistic is suitable in case of any kind of distribution, independently on its particular skewness.
For these features, the **GoF** toolkit contains the generalization of these tests containing a weighting function to the case of binned distributions:

- Fisz-Cramer-von Mises test [9],
- k-sample Anderson-Darling test [10].

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 semplicity, 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 undoubtedly 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].

2.2. GoF toolkit architecture

The system 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 Traceability.
The project adopts a solid architectural approach in order 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.
Both object-oriented techniques and generic programming allow a component-based design of the toolkit. This feature is very important as it facilitates the reuse of the toolkit as well as its integration in other data analysis frameworks.
Figure 1 represents the core components of the **GoF** toolkit. Its main features are summarised in two points:

- 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 templated 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).
Figure 2 details all the algorithm implemented up to now: every algorithm 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 deals with.
The component-based design allow for an easy extension of the **GoF** toolkit to new algorithms without interfering with the existing code, employing the *Factory method* [12].
From the user's point of view, the object-oriented techniques adopted together with the standard **AIDA**(*Abstract Interfaces for Data Analysis*) [13] interfaces are able to shield the user from the complexity of both the architecture of the core components and the computational aspects of the mathematical algorithms implemented. 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 also does not have to know the exact mathematical formulation of the distance nor of the asymptotic probability distribution he is computing. Therefore the user can concentrate on the choice of the algorithm relevant for his 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.

3. Examples of practical applications of the GoF toolkit

Thanks to the great variety of its sophisticated and powerful statistical tests, the **GoF** toolkit has been adopted by some projects, having as a crucial point the comparison of distributions of specific physical quantities. The three examples that follow have as a common denominator the essential need for an accurate validation of the simulations versus experimental data-sets. The field of applications are the following:

1. **Physics validation:** GEANT4 [14] decided to adopt the **GoF** toolkit for the microscopic validation of its physics (both Standard and Low Energies processes are involved) with a powerful statistical tool.
2. **Astrophysics:** ESA Bepi Colombo mission [15] decided to use it with the aim of comparing Bessy test beam experimental data with Geant4 simulations of X-ray fluorescence emission.
3. **Medical physics:** CATANA INFN [16], the unique Italian group performing hadrontherapy and treating patients affected by uveal melanoma, use the **GoF** toolkit in order to make comparison of physical quantities of interest (as Bragg peak, isodose distributions).

4. 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 such a variety of sophisticated and powerful algorithms in HEP.
By employing a rigorous software process, using object-oriented techniques as well as generic programming, the toolkit features a component-based design. This facilitates the re-use of the toolkit in other environments. The adoption of AIDA interfaces simplifies the use of the toolkit further.
The code is downloadable from the web [1] together with all the documentation concerning the User Requirements Document and the Traceability Matrix.
Finally, for all the features described, the **GoF** toolkit constitutes a step forward in HEP data analysis quality and could be easily used by other experimental software frameworks.

References

[1] http://www.ge.infn.it/geant4/analysis/HEPstatistics/
[2] A.N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione", Giorn. Ist. Ital. Attuari, 4, 1933: 1-11.
[3] L.A. Goodman, "Kolmogorov-Smirnov tests for psychological research", Psychol. Bull., 51, 1954: 160-168.
[4] N.H. Kuiper, "Tests concerning random points on a circle", Proc. Koninkl. Neder. Akad. van. Wetenschappen A, 63, 1960: 38-47.
[5] H. Cramèr, "On the composition of elementary errors. Second paper: statistical applications", Skand. Aktuarietidskrift, 11, 1928: 171-180.
[6] R. von Mises, "Wahrscheinliehkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik", Leipzig, F. Deuticke, 1931.
[7] T.W. Anderson, D.A. Darling "Asymptotic theory of certain goodness of fit criteria based on stochastic processes" Ann. Math. Statist., 23, 1952: 193-212.
[8] S.V. Aksenov, M.A. Savageau, "Mathematica and C programs for minimum distance estimation of the S distribution and for calculation of goodness-of-fit by bootstrap", 2001, in press.
[9] M. Fisz, "On a result by M. Rosenblatt concerning the von Mises-Smirnov test", Ann. Math. Statist., 31, 1960: 427-429.
[10] J.M. Dufour, A. Farhat, "Exact nonparametric two-sample homogeneity tests for possibly discrete distributions", Cahier 23-2001, Universitè de Montrèal.
[11] M.A. Stephens, "Introduction to: Kolmogorov (1933) on the empirical determination of a distribution", in S. Kotz, N.L. Johnson, "Breakthrougth in statistics", vol II, Springer Verlag, New York, 1992.
[12] E. Gamma, R. Helm, R. Johnson, J. Vlissides, "Design Patterns", Addison Wesley Professional Computing Series, 1994.
[13] http://AIDA.freehep.org
[14] http://geant4.web.cern.ch/geant4/
[15] http://sci.esa.int
[16] http://www.lns.infn.it/catanaweb/

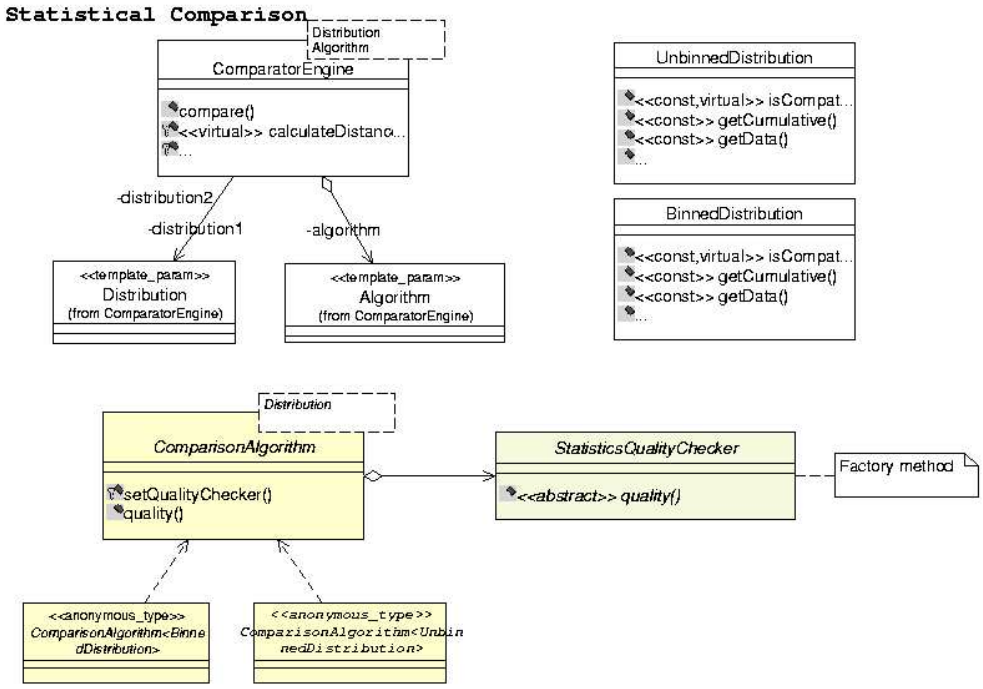

Figure 1: Statistical toolkit core design: one object (*Comparator Engine*) is responsible of the whole statistical comparison process.

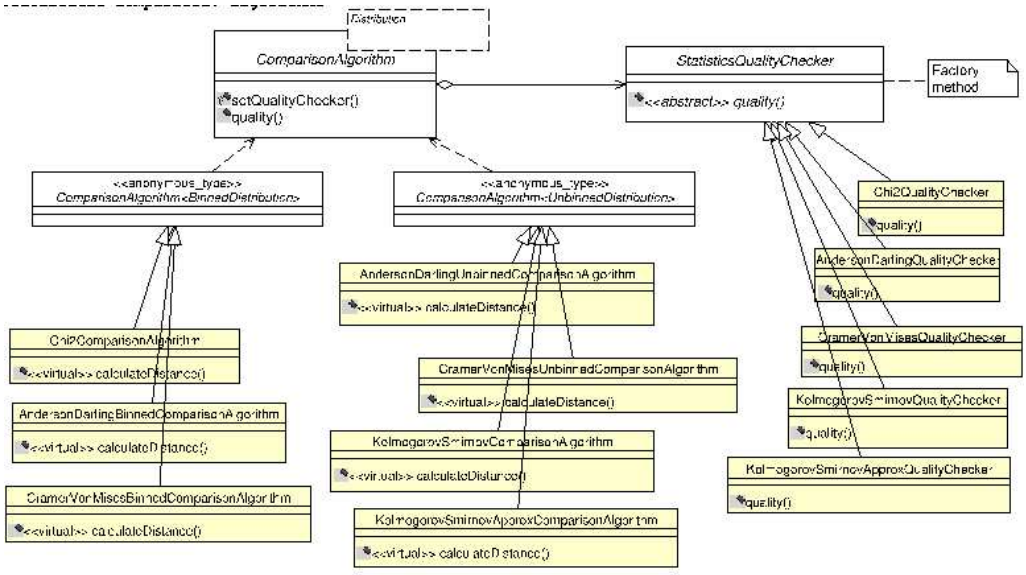

Figure 2: Detail of the statistical toolkit design: algorithms implemented for binned (Chi-squared, Fisz-Cramer-von Mises and k-sample Anderson-Darling tests) and unbinned (Kolmogorov-Smirnov, Goodman-Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests) distributions.